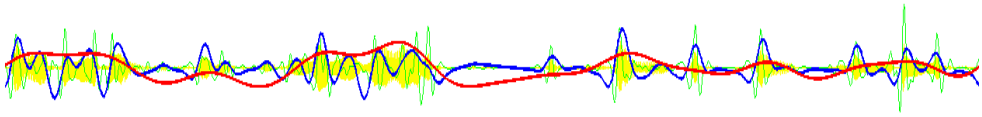
PROSODIC RHYTHM IN THE SPEECH AMPLITUDE ENVELOPE :

AMPLITUDE MODULATION PHASE HIERARCHIES (AMPHS) AND AMPH MODELS



VICTORIA LEONG (CHEAH) VIK EE

HOMERTON COLLEGE

UNIVERSITY OF CAMBRIDGE

THE DISSERTATION IS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

SEPTEMBER 2012

PREFACE

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the word limit of 80,000 words (obtained by special permission). The actual word count of the dissertation (excluding figures, tables, bibliography and appendices) is 76,969 words.

ACKNOWLEDGEMENTS

The work in this thesis was made possible by funding from the Harold Hyam Wingate Foundation, and the Funds for Women Graduates Main Grant. I am grateful for their financial support.

This thesis would not have been possible without the clear vision, sage advice and unwavering support of my supervisor, Professor Usha Goswami. I am grateful to have had the privilege of being her student. She has been an inspiration. I am also fortunate to have benefited from the technical expertise of Dr Michael Stone and Dr Richard Turner. Thank you for teaching me about speech modulation.

There were times during this PhD when the going was hard. I could not have finished this thesis without the constant love and support of my family. Thank you, Calvin, for always believing in me. This thesis is dedicated to all of you, especially Dad.

One could not have asked for better companions on this journey than the excellent folk at the CNE. Special thanks to Nichola Daily who always had a listening ear, wise counsel, hot tea and practical help for beleaguered students. Thanks also to the many staff and students of the lab (past and present), who have made the lab a special and vibrant place.

And finally, the only lasting joy and purpose in this human endeavor comes from understanding the work of the divine God. Soli deo gloria!

THESIS SUMMARY

"Prosodic Rhythm in the Speech Amplitude Envelope : Amplitude Modulation Phase Hierarchies (AMPHs) and AMPH Models"

Victoria Leong (Cheah) Vik Ee

Speech contains prosodic rhythm patterns. As even newborn infants can perceive speech rhythm, such patterns must arise from temporal regularities in the acoustic signal. Previous accounts of speech rhythm have typically been duration-based. Here, a novel *amplitude*-based account of prosodic rhythm is presented, based on patterns of amplitude modulation (AM) in the speech envelope. In symmetry to nested neuronal oscillations, AMs in the speech envelope are conceptualised as forming a nested hierarchy, with tiers capturing prosodic units such as stress feet (~ 2 Hz) and syllables (~ 4 Hz). This AM hierarchy captures different metrical patterns in children's nursery rhymes (e.g. trochees or iambs) as different phase-locked patterns between Stress AMs and Syllable AMs in the hierarchy.

In this thesis, two Amplitude Modulation Phase Hierarchy (AMPH) models are described. The first AMPH model uses a theoretically-derived 5-tier AM hierarchy. In a tone-vocoder experiment, the assumptions of the first AMPH model are tested with human listeners. The AMPH model is found to correctly predict listener's perception of metrical stress patterns on the basis of the Stress-Syllable phase relationship. The second S-AMPH model improves on the first model by using a new 5 x 3-tier spectro-temporal AM hierarchy, derived 'ground-up' from the modulation statistics of the envelope. Both AMPH and S-AMPH models are functionally evaluated in terms of syllable detection and prosodic stress assignment, using samples of metronome-timed and freely-produced nursery rhyme speech.

Finally, the S-AMPH model is used as an analysis tool to characterise rhythmic differences between (1) child-directed speech vs adult-directed speech; and to investigate (2) speech rhythm perception and production in adults with and without dyslexia. The S-AMPH analysis provides unique insights into the spectro-temporal structure of child-directed speech, and into the nature of the dyslexic rhythm deficit.

In conclusion, the AMPH models provide a novel amplitude-based account of speech rhythm perception. They also represent an advancement in methodology for speech rhythm research.

CONTENTS PAGE

PART I :

GENERAL INTRODUCTION

Chapter 1 : Introduction & Literature Review

1 INTRODUCTION & LITERATURE REVIEW

1.1 EARLY LANGUAGE ACQUISITION

Infants around the world spontaneously acquire spoken language. They acquire spoken language because this is the medium that members of their social group and culture have chosen for the purposes of constructing shared meaning. Therefore, in order to participate as social agents in the world, capable of interacting with and exerting control over their environment, infants must acquire spoken language. Spoken language is a system of symbols, where words (the units of meaning) are acoustic symbols that stand for real-life objects or concepts. However, unlike hieroglyphics where the symbols resemble the objects that they stand for, spoken words (with the exception of onomatopoetic words) typically do not sound like the objects they represent. For example, we call a cat *"cat"* rather than "*meow"*. Therefore there is no self-evident mapping between the acoustic properties of the sound symbol and its meaning, and the infant must learn the meaning of a sound symbol from experience.

Moreover, human adults speak in sentences that are concatenations of many words, so that a target word like *"cat"* will often be embedded alongside other words in a sentence such as *"the cat is on the chair"*. In the acoustic signal, these seven words will not occur as seven discrete islands of sound (like Morse code). Rather, what the infant will hear is a continuous babble of sound that is richly patterned in the spectral and temporal domains, more similar to music than to Morse code. Since words are not clearly delineated by pauses or discontinuities (Cole & Jakimik, 1980), the infant must discover these word boundaries for themselves (i.e. the speech segmentation problem). Infants must capture the critical sound pattern for each word without any irrelevant adjoining material so that their mental representation includes all the sound constituents of *"cat"*, but is not over-specified as *"the-cat"* or *"cat-is"*. Moreover, infants must be able to recognise the same sound symbol in the face of acoustic variation from different speakers, or different contexts. Therefore, to acquire spoken language, infants must carry out a Herculean task of acoustic pattern recognition. Unlike adult listeners, infants do this without the aid of an alphabet of graphemes and without recourse to a mental lexicon of words to constrain their search.

Typically, infants acquire spoken language with phenomenal success and speed. By the end of the first year of life, a typical British infant will already have produced his or her first word and be able to comprehend around 50 spoken words (Oxford Communicative Development Inventory Database; Hamilton et al, 2000). Across other cultures and languages, infants show a similar or even faster rate of word acquisition. For example, Table 1.1 shows the typical number of words and gestures comprehended and produced by infants between 8 to 16 months for six different languages. Over eight months, infants on average see a 10-fold increase in the number of words and gestures they understand. The number of words they can produce increases even more - a staggering 40-times on average.

Table 1.1. Cross-linguistic norms for infant word production and comprehension. Norms were measured using adaptations of the MacArthur Communicative Development Inventories (Fenson et al, 2006), and retrieved online from the CLEX database (Nørgaard Jørgensen et al, 2010).

MEAN PRODUCED WORDS & GESTURES						
Age (months)	**American English**	**Danish**	**Swedish**	**Norwegian**	**Mexican Spanish**	**Croatian**
8	2	0.7	0.3	6.9	3	1.1
10	4.2	1.3	1.9	5.1	5	5.5
12	10.2	4.8	5.6	8.4	13.1	8.3
14	26.8	9.5	12.4	17.2	20.5	23.5
16	60.5	24.2	32.8	32.5	28.9	73.9
MEAN COMPREHENDED WORDS & GESTURES						
Age (months)	**American English**	**Danish**	**Swedish**	**Norwegian**	**Mexican Spanish**	**Croatian**
8	42.3	16.8	7.2	24	61.6	12.3
10	55.7	30.6	29.2	39.7	62.3	49.3
12	84.8	56.5	58.5	79.8	104.2	92.8
14	153.2	85.9	116.1	113.6	138.9	158.9
16	193.1	135.3	163.5	158	184.2	212.9

In many situations, infants will receive a rich supply of linguistic input to help them achieve this remarkable feat of language learning. However, other infants will have to make do with less. For example, Shneidman & Goldwin-Meadow (2012) observed that 12-month-old American infants typically heard almost 900 utterances per hour, of which 70% were directed specifically at them. On the other end of the scale however, Mayan infants of the same age only heard around half that number of utterances per hour (~450). Out of these, only 20% were directed at them while the remaining 80% of utterances were overheard. Moreover, for Mayan infants, a much larger proportion of the spoken input came from other child speakers, rather than from adult speakers. Shneidman & Goldwin-Meadow found that the amount of directed speech input received by Mayan children was an important predictor for their later vocabulary development, while overheard speech was not (supporting the view that child-directed speech is adaptive for language learning). Yet despite these vast differences in the quantity and quality of the initial linguistic input, both Mayan and American children (and indeed most children around the world) will eventually go on to master their native language. What strategies do infants and children use to accomplish this task?

1.1.1 STRATEGIES FOR LANGUAGE ACQUISITION

Very soon after birth, infants already appear to show a preference for encoding 'syllable-like' units of speech. For example, Bertoncini & Mehler (1981) found that infants less than 2 months of age could potentially represent syllables as sequences of alternating consonant (C) and vowel (V) segments. In their study, they habituated infants to sequences that had either a 'legal' (CVC, e.g. *"tap"*) or 'illegal' syllable structure (CCC, e.g. *"tsp"*), but contained the same initial and final consonant. They then tested the habituated infants with variants of these sequences in which the first and last consonant were switched (i.e. *"tap"* became *"pat"* or *"tsp"* became *"pst"*). If infants were indeed sensitive to syllable units, and represented these syllable units as alternating consonant-vowel sequences, then they should dishabituate only when the consonant switch was presented in the context of a 'legal' syllable structure (i.e. when *"tap"* became *"pat"*). As predicted, more infants dishabituated to the *"tap"-"pat"* switch (12/15 infants) than to the *"tsp"-"pst"* switch (6/15 infants).

In this study, the authors created syllable-legal and illegal stimuli based on their own prior knowledge of phonemes and syllable structure. They then assumed that if infants

responded differently to the two types of stimuli, it must have been because the infants also had implicit knowledge of phonemes and syllable structure. However, it is highly unlikely that the infants represented the sequences in the way that the experimenters did (i.e. consisting of three phonemes, of which the middle was either a consonant or a vowel). Rather, it is more likely that the acoustic properties of the central segments /a/ and /s/ differed in a way that allowed infants to bind *"tap"* as one temporally continuous unit, but not *"tsp"*. In other words, the fundamental definition of a syllable (for the infant) was probably not the class of its segmental constituents, but the global ***spectral-temporal 'coherence'*** of the sequence as a whole (which in turn was probably affected by factors like the sonority of its constituents). Nonetheless, this study suggests that even 2-month old infants appear to be sensitive to the 'syllable-ness' of a sound sequence, and that this factor apparently moderates their encoding of an auditory stimulus.

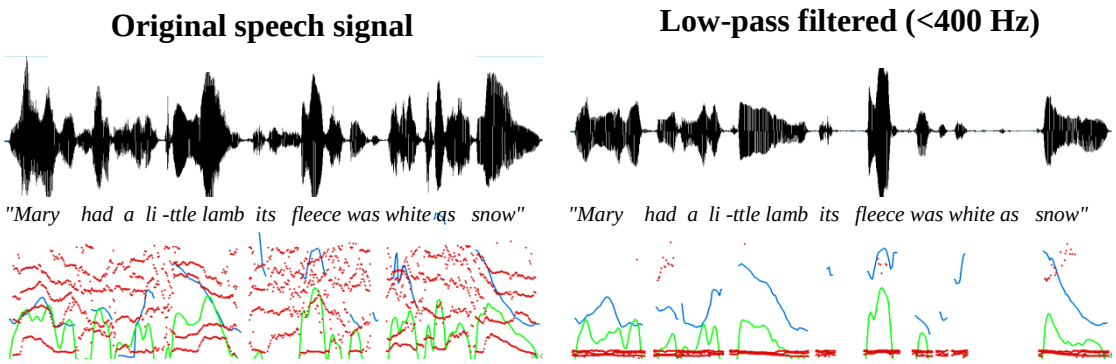
Infants also demonstrate a very early sensitivity to prosodic rhythm patterns in speech. Human speech heard from inside the womb is effectively low-pass filtered by the uterine wall, isolating low frequency information and accordingly foregrounding prosodic and rhythmic structure (Armitage et al, 1980). Therefore, while in the womb, fetuses are already being to be exposed to the global prosodic patterns of their native language. In the last trimester of pregnancy, fetuses may even be able to extract and remember certain temporal and prosodic features of speech, such as those pertaining to speaker identity. For example, newborn infants can already distinguish their mother's voice from that of another female (DeCasper & Fifer, 1980). More remarkably, newborn infants even appear to remember a story that was read to them daily for 6 weeks while in utero (DeCasper & Spence, 1986), adjusting their sucking rates to hear the familiar story rather than an unfamiliar one. This effect remained even when infants were hearing a strange female voice read the story rather than their own mother, indicating that they had formed a preference for the spoken material itself, not just the speaker.

In studies more specific to prosodic rhythm, when newborn infants are presented with low-pass filtered sentences from languages of different rhythm classes[1] (e.g. British English, Dutch and Japanese sentences presented to French babies, see Nazzi et al, 1998), they discriminate successfully between languages with different rhythm classes (English, 'stress-timed' and Japanese, 'mora-timed'). However, they do *not* discriminate between languages with the same rhythm class (English and Dutch, both 'stress-timed'). Similar results are found

[1] Language rhythm classes are discussed in Section 1.3.1.

when a more radical phoneme replacement method (e.g. 'saltanaj') is used to remove phonemic and phonotactic differences between sentences (e.g. Ramus et al, 2000). If neonates can still access and use rhythm cues in the speech signal even when its segmental content has been replaced, this suggests that the supra-segmental (rhythmic) and segmental (phonemic) features of speech are specified by separate sets of acoustic features and temporal statistics. In particular, low-pass filtering the speech signal (e.g. under 400 Hz as used by Nazzi et al, 1998) apparently retains the rhythmic structure of speech while degrading its phonemic structure.

Figure 1.1. Effect of low-pass filtering on the sentence "Mary had a little lamb...". The original speech signal is shown on the left and the 400 Hz low-pass-filtered signal is shown on the right. The top panels show the acoustic waveform, the bottom panels show the formant frequencies (red dotted line), fundamental frequency (blue line) and intensity contour (green line). The scale of the y-axis for the bottom panels is 0-7000 Hz for the formant frequencies, 75-500 Hz for the fundamental frequency, and 50dB-90dB for the intensity contour.



As shown in Figure 1.1, the effect of such low-pass filtering is to remove the vast majority of formant structure (shown in the figure as red dotted lines), whilst retaining the fundamental frequency or pitch contour (blue line) and selected portions of the intensity contour (green line). Furthermore, the top panel of the figure illustrates that low-pass filtering also distorts the waveform disproportionately so that certain types of speech sounds (e.g. high-frequency fricative sounds like /s/) are more likely to be filtered out than others (e.g. lower-frequency vowel sounds). Infants are still able to extract rhythm patterns from this much-reduced version of the speech signal (Nazzi et al, 1998). This suggests that prosodic rhythms are carried primarily by global, relatively slow-varying patterns of spectral and amplitude modulation, as exemplified by the preserved pitch contour and intensity contour

respectively. Furthermore, if infants are already sensitive to such slow-varying speech features at birth (e.g. from pre-sensitisation in the womb), they may be able to harness the syllabic- and prosodic-level information specified in these slow-varying acoustic features to 'boot-up' language learning.

This suggestion is not new. Prosodic information has long been proposed to play an important role in 'bootstrapping' early language acquisition (Gleitman and Wanner, 1982). To solve the problem of speech segmentation, it has been suggested that infants tune in to the common prosodic stress patterns of their native language, and use these patterns to parse the speech signal into candidate words via a 'metrical segmentation strategy' (Cutler & Norris, 1988). For example, in the English language, it is estimated that 90% of content words begin with a strong initial stressed syllable, such as *"DA-ddy"* or *"BA-by"* (Cutler & Carter, 1987). By around 9 months of age, English-learning infants show sensitivity to this prosodic statistic, preferring words with a 'Strong-weak' (S-w) syllable pattern over those with a 'weak-Strong' (w-S) syllable pattern (Jusczyk et al, 1993; Echols et al, 1997). For example, when presented with the sentence *"her gui-TAR is too fancy"*, 7.5-month-old infants preferentially segment *"TAR-is"* (S-w) as a word instead of *"gui-TAR"* (w-S), following the 'S-w' heuristic (Jusczyk et al, 1999). By 10.5 months, infants no longer make this error, possibly due to sensitivity to other cues such as allophonic differences or learning the transitional probabilities between segments and syllables (i.e. 'statistical learning', Saffran et al, 1996). However, it is not the case that these older infants are no longer sensitive to prosodic stress patterns, or no longer use stress in speech segmentation. In fact, in the presence of conflicting 'statistical' and 'prosodic' cues, 11-month-old infants still preferentially use prosodic cues over statistical cues as word boundaries (Johnson & Seidl, 2008). Hence, rather than losing their sensitivity to stress, older infants instead appear to be simply integrating a broader array of segmentation cues. Consistent with this interpretation, computational models of speech segmentation perform better when *both* statistical cues and prosodic cues are combined (Christiansen et al, 1998).

1.1.2 CROSS-LINGUISTIC DIFFERENCES IN EARLY LANGUAGE ACQUISITION

Furthermore, the prosodic rhythm class of a language may also determine the linguistic level or 'grain-size' at which infants initially begin to segment speech. A specific formulation of this view is the 'rhythm activation hypothesis' put forward by Nazzi et al (2006). This hypothesis proposes that the dominant rhythmic unit of a language should also form the main initial unit of prosodic segmentation. Accordingly, infants learning a 'stress-timed' language such as English, Dutch or German should develop a strategy to segment trochaic (S-w) ***stress*** units, while infants learning 'syllable-timed' languages like French, Spanish or Italian should initially segment ***syllable*** units instead.

The findings by Jusczyk et al (1999) are consistent with the view that English-learning infants initially develop a trochaic stress-based segmentation strategy. Similarly, infants learning Dutch (a stress-timed language) also begin segmenting Strong-weak words between 7.5 and 9 months of age (Houston et al, 2000; Kuijpers et al, 1998). Infants learning stress-timed German appear to acquire the trochaic bias at an even earlier age, listening longer to a trochaic pattern (e.g. *"GA-ba"*) than an iambic pattern (*"ga-BA"*) for the same word from as early as 6 months of age (Hohle et al, 2009). Unlike infants learning stress-timed English, Dutch or German, 6-month-old infants learning syllable-timed French do *not* show a preference for either trochaic or iambic stress patterns, although they can discriminate between them (Hohle et al, 2009). Furthermore, French infants do not begin segmenting whole bi-syllable words from continuous speech until 16 months of age, although they do segment the initial and final syllables of such bi-syllable words from 12 months of age (Nazzi et al, 2006).

These language differences are also apparent at the neural level. For example, Friederici et al (2007) found that even 4-month-old German and French infants showed different event-related potential (ERP) responses to trochaic and iambic sound patterns. In an oddball paradigm, the infants were presented with a string of repetitions of the word *"baba"*. There were two version of the stimulus string. In the first version, the standards were stressed in a trochaic manner and occasional deviants were iambically-stressed (e.g. BAba BAba BAba baBA BAba...). In the second version, the standards were iambically-stressed and occasional deviants were trochaically-stressed (e.g. baBA baBA baBA BAba baBA...). German infants showed an ERP 'mismatch' response when the deviant stimulus was an iambic bisyllable amidst a train of trochees (i.e. version 1). However, the infants did *not*

show a mismatch response when the deviant stimulus was a trochee amidst a train of iambs (i.e. version 2). Therefore, even at 4 months, German infants already showed a neural processing bias toward trochaic stress patterns. Conversely, French infants showed a mismatch response when the deviant was a trochee (i.e. version 2), but not when it was an iamb (i.e. version 1). This result for the French infants was different to that predicted by the 'rhythm activation hypothesis'. According to this hypothesis, French infants should show an *equal* mismatch response to both trochaic and iambic deviants since they are preferentially segmenting speech at the syllable level. However, although French is a syllable-timed language, phrase-final syllables commonly carry prosodic stress (di Cristo, 1998), which could explain infants' preference for the iambic stress-final (w-S) pattern.

Friederici et al's (2007) result is convergent with a study by Mampe et al (2009), in which they demonstrated that cross-linguistic prosodic differences between German and French were even evident in the temporal pattern of newborn infants' cries. In their study, Mampe et al (2009) compared the temporal shape of spontaneous cries produced by 30 French and 30 German newborn infants. After normalising for cry duration, they found that French infants produced cries that took relatively longer to reach peak intensity (i.e. longer rise times) and longer to reach maximum pitch, while German infants produced cries that reached peak intensity and maximum pitch significantly earlier, and then slowly tapered off. If the cry is taken to represent a prosodic stress foot, then German babies were placing prosodic stress (i.e. high pitch and intensity) at an earlier portion of the foot than French babies. Therefore, newborn German infants' cries were more 'trochaic' and French infants' cries were more 'iambic' in temporal pattern, consistent with the neural mismatch responses measured in 4-month-old infants by Friederici et al (2007). That is, if German infants were producing more 'trochaic' cries themselves, then they would be more likely to recognise a train of trochees as being standard (like their own cries), and an iambic pattern as being deviant. Conversely, French infants would be more likely to recognise an iambic train as being standard, and a trochaic pattern as being deviant. Moreover, Mampe et al's study with *newborns* indicates that exposure to low-pass-filtered speech in-utero may be sufficient to impart an implicit knowledge of native prosodic patterns to infants.

However, while French infants appear to preferentially discriminate the iambic pattern by 4 months of age, and may even produce a similar pattern themselves when crying, they do not appear to use this iambic motif as a metrical segmentation strategy (unlike English-learning infants who do use the trochee). This may be because final syllable-

lengthening in French is not so much a cue for word boundaries, as it is for phrase boundaries. Therefore, contrary to the 'rhythm activation hypothesis', French-learning infants may instead initially segment speech into larger phrasal units, rather than smaller syllable units. In this case, the basic unit of prosodic segmentation would not correspond to the perceived rhythm class, but to the linguistic level at which prosodic stress provides the most reliable and pronounced boundary cues. This is not to say that French (or indeed English) infants are not sensitive to syllable units in speech. On the contrary, the 2-month old French infants in Bertoncini & Mehler's (1981) study showed an exquisite sensitivity to syllable structure. Rather, perhaps all infants are born equipped as 'syllable-detectors'. What the prosodic patterns of their native language specify are the ways in which these discrete syllables should be ***bound*** or ***grouped*** into higher-level units of meaning, such as words or phrases. By this view, infants develop a metrical *binding* strategy that is dependent on the rhythmic characteristics of their native language. Accordingly, while English infants initially attempt to bind syllables into binary trochaic proto-words, French infants initially attempt to bind syllables into stress-final proto-phrases containing a variable number of syllables (and only later deconstruct these into constituent words, using other distributional cues). Consequently, while English infants become committed to a binary word representation early on, French infants may maintain a more flexible word length representation until later. This explanation could account for the disparate behavioural and neural results observed in French-learning infants.

Whether early or late in the first year of life, across languages, infants eventually begin to use the characteristic statistical distributions and rhythmic patterns of their native language to actively constrain speech processing (e.g. in finding candidate words within the speech stream). In English- and Dutch-learning infants, this occurs around 7.5 to 9 months, while in German-learning infants, this could occur earlier at around 6 months of age. French-learning infants (as discussed previously), may maintain a more flexible representation until as late as 16 months. Therefore within the first year of life, speech perception becomes an active, constructive process, with infants generating 'hypotheses' about what constitutes a meaningful sound pattern or 'word' (eg. word = trochee-patterned syllable sequence). Neurally, during this same period (the first year of life), thalamo-cortical afferent connections are rapidly being formed in the infant auditory cortex (Moore, 2002), allowing the neocortex to become increasingly engaged in interpreting the incoming auditory input. These new connections in the auditory cortex could possibly underlie infants' 'tuning' to the prosodic

patterns of their native language during the later half of their first year. For a review of the stages of maturation in auditory cortical development, please see Appendix 1.1.

In summary, acquiring knowledge about native prosodic rhythm patterns is an important first step in language development. Without knowledge of these prosodic patterns, infants could not gain a foothold on language so quickly. The acoustic basis of these prosodic patterns are the focus of this thesis. In later chapters, two novel models will be proposed to demonstrate how key prosodic information (e.g. syllable location and prosodic stress patterns) may be derived solely from low-level acoustic information (e.g. amplitude modulation patterns in the speech envelope), without recourse to higher-order lexical knowledge. Therefore these models operate with the same constraints that newborn infants face. If rhythm can be inferred solely from the speech signal, then speech rhythm may be a truly emergent, or self-evident, property of the speech signal. Put another way, speech rhythm may be our perceptual experience of what is the fundamental temporal structure of speech. By this view, all later cognitive activity upon the speech input (including interpretations of meaning) must necessarily be initially constrained by this fundamental temporal structure. If speech rhythm is indeed emergent, and an important constraint upon speech processing, it is no wonder that infants show such acute sensitivity to prosodic information early in language learning.

1.2 LINGUISTIC RHYTHM

Rhythm commonly refers to an alternating pattern of 'Strong' and 'weak' elements (Schane, 1979; Lerdahl & Jackendoff, 1983). In the linguistic context, this rhythmic alternation is most clearly illustrated at the level of syllables, which can be stressed (Strong, 'S') or unstressed (weak, 'w'). Moreover, this Strong-weak alternation also occurs at higher levels of rhythmic organisation. For example, the word *"MI-ssi-SSI-ppi"* contains four syllables that alternate in prosodic stress to give a 'S-w-S-w' pattern of syllable stress. However, the four syllables may also be grouped at a higher level of organisation into two pairs of prosodic 'stress feet'[2], where each stress foot has a 'S-w' motif. At this higher level of

[2] The prosodic 'foot' (a term originally borrowed from poetic meter, Selkirk, 1980) is a basic metrical unit of rhythm that refers to a group of syllables with one stressed syllable. For example, binary feet consist of two syllables, and may have either a strong-weak (trochaic) or weak-strong (iambic) stress pattern. In poetry, foot patterns are used to describe poetic meters such as the iambic pentameter, defined as a basic line of five iambic feet (Hanson, 2006).

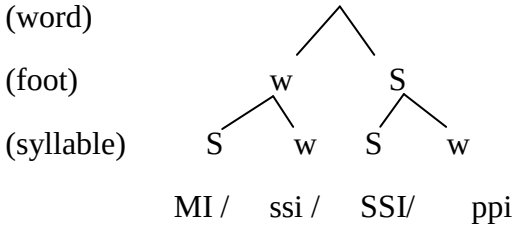
organisation (the stress foot), rhythmic Strong-weak alternation is also present, since the second stress foot (corresponding to *"ssi-ppi"*) is stronger in prominence than the first stress foot (corresponding to *"mi-ssi"*). In metrical phonology, this hierarchical prosodic pattern can be represented either as a grid or as a tree that captures the relative prominence of each element (Selkirk, 1980, 1984, 1986; Liberman & Prince, 1977; Hayes, 1995). Both representations are depicted in Figure 1.2.

Figure 1.2. Grid and tree representations of hierarchical prosodic structure

(a) Grid Representation (x = prominence)

(Level 3)	(		x	)
(Level 2)	(x		x	)
(Level 1)	(x	x)	(x	x)
	MI /	ssi /	SSI/	ppi

(b) Tree Representation (w = weaker than; S = stronger than)



In both representations, the hierarchical tiers of the grid or tree represent prosodic levels such as syllables, stress groupings, primary lexical stress (for each word), and phrasal stress accents. In the grid representation, prosodic prominence at each level is marked with an 'x'. Therefore, at Level 1 (syllable), the four syllables are each marked with an 'x'. At Level 2 (stressed syllable), the locations of the two stressed syllables (*"mi"* and *"ssi"*) are each marked with an 'x'. At Level 3 (primary lexical stress), only the primary stressed syllable for the word (third syllable *"ssi"*) is marked with an 'x'. In the tree representation, the focus is on the *relative* prominence between adjacent elements (nodes). At each level, stronger (S) and weaker (w) nodes are identified. Moreover, nodes at each hierarchical level also represent prosodic units that encompasses one or more 'daughter' nodes at lower levels. Related 'parent' and 'daughter' nodes are indicated as 'branches' in the tree. So in the tree example of Figure

1.2, the whole word (highest level of the hierarchy) is seen as comprising two prosodic feet (second level), one weak and one strong. Each of these feet is in turn made up of two syllables (lowest level), and both prosodic feet have the same metrical pattern (S-w).

Note that both tree and grid representations enable an analogy with metrical structure in music. In music, the term 'meter' refers to the number of beats per bar. This may be duple (e.g. 2/4, two quarter-notes per bar), triple (e.g. 3/4, three quarter-notes per bar) or a compound (e.g. 6/8, two sets of three eighth-notes each). If musical beats are analogous to syllables, and 'bars' are analogous to prosodic feet, then the meter will equate to the number of syllables per prosodic foot. Hence, the hierarchical rhythmic structure of speech has strong analogies to the hierarchical rhythmic structure of music. Of course, this idea is not novel. Lerdahl and Jackendoff (1983) originally proposed that the metrical structure of a musical piece may also be represented using a hierarchical grid structure capturing the alternation of strong and weak beats. They argued that prosodic stress patterns in speech and beat patterns in music may share core principles of rhythmic organisation, especially those of alternation and hierarchy. As will be shown, the two hierarchical AM models proposed in this thesis explicitly support these intuitions about rhythm. Specifically, the linguistic rhythmic hierarchy may have its basis in hierarchical amplitude modulation patterns in the speech envelope. Accordingly, Strong-weak rhythmic alternation may arise from cyclical oscillation patterns in these amplitude modulations. These ideas are developed further in Chapter 2, Section 2.1. Here, however, prior approaches to describing linguistic rhythm will be reviewed.

1.3 PRIOR APPROACHES TO DESCRIBING SPEECH RHYTHM

1.3.1 LANGUAGE RHYTHM CLASSES & RHYTHM-METRICS

Linguists have long suggested that languages can be classed as having different rhythmic typologies (e.g. Abercrombie, 1967; Pike, 1945). For example, it was suggested that languages like English and Dutch could be grouped together as they shared a 'Morse-code'-like rhythm of long and short pulses, associated with alternating stressed and unstressed syllables (stress timing). Other languages, like Spanish and Italian, could be characterised by a 'machine-gun'-like rhythm arising from evenly-spaced syllables (syllable timing). To explain these typologies, Abercrombie (1967) proposed that durational isochrony (regular timing) at different linguistic levels underscored the rhythmic differences between languages. By this account, languages were 'stress-timed' if the intervals between successive stressed syllables were constant, but 'syllable-timed' if the intervals between the syllables themselves were constant.

Empirical support for distinct language rhythm classes came from infant studies, which showed that even neonates appeared to classify languages by rhythm type. For example, as discussed earlier in Section 1.1.1, Nazzi et al, 1998 found that French neonates could distinguish between languages from *different* rhythm classes (such as English vs Japanese), but not between languages from the *same* rhythm class (such as English vs Dutch). Moreover, infants could still perform the rhythm classification even when the phonetic differences between languages were removed by re-synthesis (Ramus et al, 2000) or low-pass filtering (Nazzi et al, 1998), preserving only speech prosody. These striking results indicated that the 'prototypical' temporal statistics used by infants for rhythm classification arose directly from the speech signal and were computed without recourse to lexical knowledge. However, while there is support for the concept of rhythm classes from infant studies, other research has not supported Abercrombie's proposed mechanism of durational isochrony (eg. Dauer, 1983; Roach, 1982). For example, Dauer (1983) found that stress intervals in English grew in length with increasing numbers of syllables, rather than maintaining isochrony.

An alternative segmental durational account of rhythm classes proposes that rhythm differences arise from *phonological* differences between languages. Dauer (1983) proposed that the durational variability of syllables in stress-timed vs syllable-timed languages could be caused by differences in syllable structure, vowel reduction and stress influence on vowel

duration. Several different 'rhythm-metrics' were subsequently proposed to capture these segmental differences. These metrics were based on the statistical properties of segmental duration variation, rather than on isochrony per se. Measures like %V, ΔV, ΔC (Ramus et al, 1999) quantified the relative proportions of vocalic intervals and the standard deviation of vocalic and consonantal durations in speech. Pairwise variability indices (PVIs, Grabe & Low, 2002) and rate-normalized measures like VarcoV and VarcoC (Dellwo & Wagner, 2003) focused on the relative variability in the length of successive consonantal and vocalic intervals. These various measures were successful in classifying prototypical languages like English, Spanish and Japanese, indicating that segmental durational statistics were indeed associated with perceived rhythm class. However, only limited success was achieved in describing non-prototypical languages such as Greek, Malay, or Welsh (Grabe and Low, 2002).

Therefore, while successful under some conditions, segmental durational statistics have also failed to provide a universal account of speech rhythm, suggesting that a new broader conceptualisation of speech rhythm may be required. More recently, Arvaniti (2009) has proposed an emphasis on the role of perception, arguing that perceptual timing factors, rather than acoustic durations, may account for rhythm in speech. Patel (2008) further argues that the apparent rhythmic qualities of speech may not be due to periodic factors at all, but may be better attributed to 'non-periodic' factors such as higher-order temporal, accentual and phrasal patterns (e.g. grouping structure and accentual clash avoidance). In contrast to music, which has a strong periodic framework by intentional design, Patel suggests that rhythm in language is a by-product of its phonology (syllable structure, vowel reduction, etc). While the arguments raised by Arvaniti and Patel (perceptual timing effects, higher-order grouping cues) are certainly relevant, it may also be the case that previous segmental durational statistics simply have not captured *all* the rhythm information that is present in the speech signal. For example, if only durational effects are considered, then any rhythm cues from intensity or pitch changes will be ignored. If only short phonetic segments are measured, then rhythm patterns that arise at longer and slower timescales (e.g. syllables, words, phrases) will not be well represented. Therefore, a more complete account of speech rhythm should include these factors that have previously been ignored by segmental durational statistics.

1.3.2 PERCEPTUAL-CENTRES

Central to the rhythmic timing field is the perceptual-centre or 'p-centre' literature, which was intended to model the perceptual ‘moment of occurrence’ of events in any sensory modality (Morton et al, 1976; Marcus, 1981). For example, if we wish to dance in time with music, then theoretically we need to time the p-centres of our movements with the p-centres of musical notes. In music, the p-centre of a note will depend on the instrument producing it: a bowed instrument (e.g. violin) will produce a note with a later p-centre due to the slower attack time than an instrument like a trumpet (Gordon, 1987).

P-centres in speech are commonly associated with the onsets of syllable vowel nuclei, and are thought to be cued primarily by changes in loudness or signal amplitude (Allen, 1972; Scott, 1993; 1998; Villing, 2010). Accordingly, attempts to identify and model the acoustic correlates of p-centres in speech have focused on the speech amplitude envelope (the speech envelope is described further in Section 1.6). For example, Howell (1984, 1988a, 1988b) proposed a syllabic 'center of gravity model', in which the distribution of energy within the amplitude envelope was the key determinant of p-center location. Other models by Pompino-Marschall (1989) and Harsin (1997) make use of loudness functions or the rate-of-change of modulation in the envelopes of different spectral bands. To date, there is no consensus on the exact acoustic correlate of 'p-centres' in speech (see for example, Patel et al, 1999). However, even though the precise perceptual-acoustic relationship between p-centres and the speech envelope is not known, it is generally accepted that patterns of amplitude modulation in the speech envelope contribute toward the perception of syllable p-centres.

1.3.3 RHYTHMIC CONSTRAINTS IN SPEECH PRODUCTION

Rhythm-metric approaches investigate the acoustic cues to rhythm, while the p-centre approach describes its psychological correlates. A third avenue of enquiry pertains to the *production* of rhythmically-timed speech. For example, Cummins and Port (1998) argued on the basis of speech production data that speech rhythm depends on the hierarchical organization of temporally-coordinated prosodic units of production. In a 'speech cycling' experiment, they accordingly investigated the rhythmic constraints on speech production. Participants were required to repeat a short phrase (e.g. *“beg for a dime”*) in time with a series of alternating high and low tones. The syllables *"BEG"* and *"DIME"* were stressed,

creating two metrical feet (*"BEG-for-a"* and *"DIME"*) within the whole phrase. The high tone was the synchronising signal for phrase onset (when to produce *"BEG"*), and the low tone cued production of the final stressed syllable (*"DIME"*).

The time between high and low tones was kept constant at 700 ms, however, the time between low and high tones was varied in different conditions between 300 ms and 1633 ms. This manipulation was meant to vary the 'onset phase' (relative position) of the low tone within the overall high-high cycle. For example, if the time between low and high tones was 300 ms, this created a total high-high cycle duration of 700 ms + 300 ms = 1000 ms. In this case, the authors defined the 'onset phase' of the low tone within the high-high cycle as 700 ms / 1000 ms = 0.7. When the time between low and high tones was 1633 ms, this created a total cycle duration of 700 ms + 1633 ms = 2333 ms, within which the low tone had an onset phase of 700 ms / 2333 ms = 0.3.

Participants were presented with stimuli where the onset phase of the low tone was evenly distributed over all possible phase values between 0.3-0.7. However, an envelope-based analysis of the actual speech patterns produced by participants showed that they tended to 'over-regularise' the production of the target stressed syllable ("*DIME"*). Rather than faithfully following the cue of the low tone over all onset phase values, participants' actual productions of the target word were heavily biased toward certain specific phase positions within the overarching phrase repetition cycle - 0.33, 0.5 and 0.67. In essence, if the overall phrase is regarded as the temporal unit or cycle that repeats, then the production of key prosodic sub-units was *constrained* to occur only at certain time points in the cycle, corresponding to a sub-division of the cycle into two or three parts (i.e. a duple or triple meter). This result showed that participants' perceptuo-motor behaviour was sensitive to the hierarchical organisation of the pacing beats, with *relative phase* being the key organising or constraining factor (i.e. the relative time point at which *"DIME"* was positioned within the overarching phrase).

Port (2003) subsequently proposed the existence of neural oscillations that produce pulses on every rhythmic cycle, where the pulses act as attractors for the timing of syllable beats (i.e. for aligning the 'p-centres' of the syllables). Port also extended the Cummins and Port (1998) model to propose that rhythmic meter arises from the integer-ratio phase-locking of several oscillators in time. This has the effect of creating a complex, hierarchically-nested structure whose purpose is to generate these pulse attractors.

In a typical motor entrainment task, rhythm and meter are externally imposed by way of a pacing metronome, necessary to give experimenters an objective 'reference phase' for computational purposes (Repp, 2005; McAuley et al., 2006; Corriveau & Goswami, 2009). Cummins and Port elegantly demonstrated that human rhythm timing and production mechanisms are dependent on the *phase* of such an external pacing metronome. However, since we are also able to synchronise our speech to that of other speakers without an external pacing metronome (Cummins, 2003), the speech signal itself may also contain hierarchical phase cues for synchronisation. In this thesis, the possibility that these hierarchical phase cues to rhythm reside in the amplitude envelope of speech will be explored.

1.3.4 COMPUTATIONAL MODELS OF SPEECH RHYTHM

Finally, two computational models of speech rhythm will be reviewed, as exemplars of two general classes of computational models. The first model (the auditory primal sketch) attempts to replicate a physiological response to rhythmic input. As such, it is a model of the auditory representation of rhythm. The second class of models is purely theoretical and not based on biological mechanisms. Rather, the aim of these models is to closely approximate the observed experimental data, and to explain differences between sets of experimental data in mechanistic terms. The two Amplitude Modulation Phase Hierarchy (AMPH) models proposed in this thesis are closer to models in the first category. Like the auditory primal sketch theory, the AMPH models also deal with the actual acoustic data (rather than durational statistics), transforming this raw acoustic information into indices for rhythm using 'neural-plausible' mechanisms. However, while the auditory primal sketch theory focuses primarily on early processing (i.e. in the auditory periphery), the AMPH models are based on central cortical mechanisms. It is assumed that the role of the auditory cortex is to interpret, rather than merely to faithfully represent, the auditory input. As such, the mechanisms proposed in the AMPH models describe how the raw auditory input is transformed into prosodic patterns that are *perceived* by the human listener.

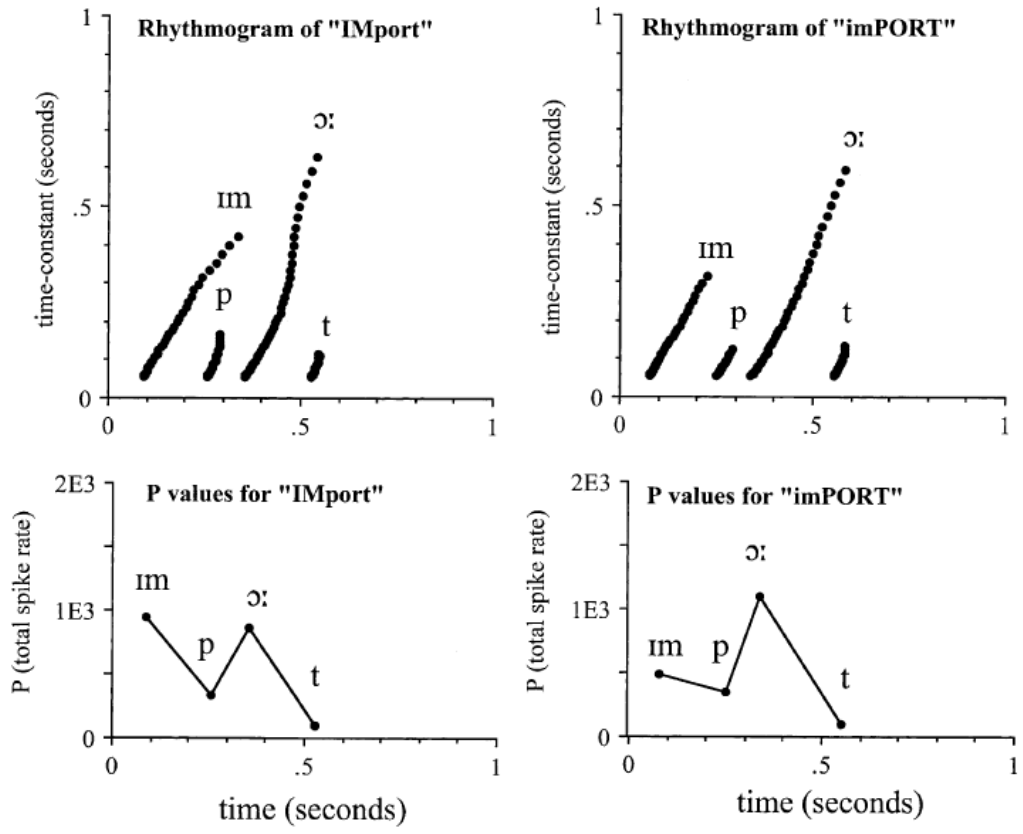
1.3.4.1 Auditory Primal Sketch Theory

Todd and colleagues (Todd, 1994; Todd & Brown, 1996; Todd et al, 1999; Todd et al, 2002; Lee & Todd, 2004) proposed an auditory model of speech rhythm in which rhythmic organisation was determined by the relative auditory prominence of phonetic events. Their

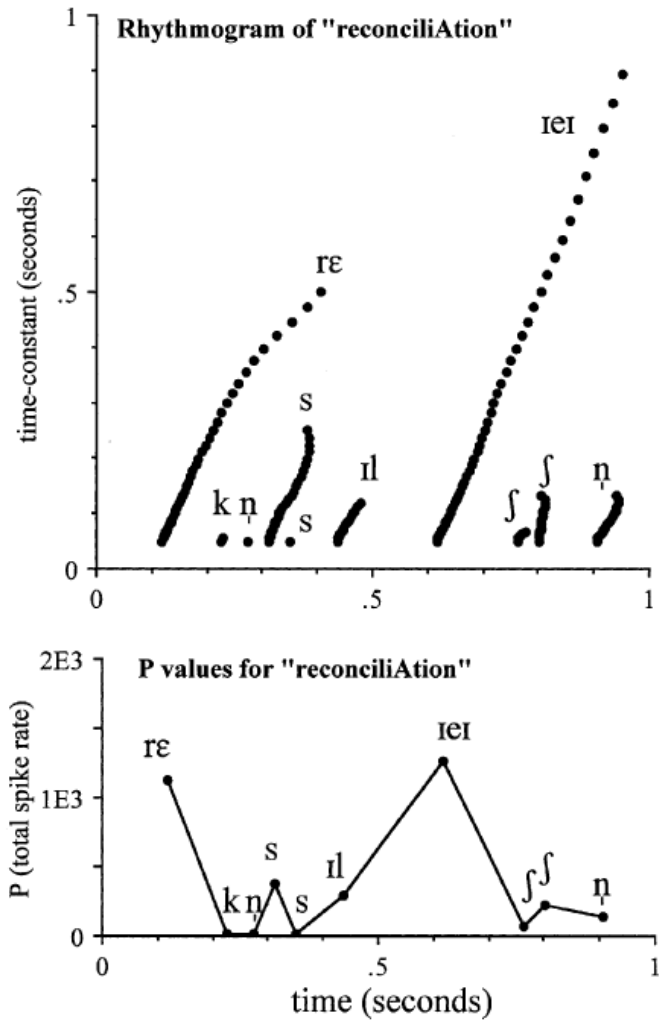
model was based on the idea of an 'auditory primal sketch', derived by smoothing a simulation of the auditory nerve response using a bank of low-pass filters with different time constants. Peaks were then identified in the smoothed signal for each output filter channel, and plotted as a graph, or 'rhythmogram'. For example, Figure 1.3a below shows the rhythmograms for the words *"IMport"* and *"imPORT"*, which differ in syllable stress pattern. In the rhythmogram, each vertical line corresponds to a different phoneme segment or cluster of phonemes, and the height of the vertical line is determined by responses from auditory filters with increasing time constants. Therefore, a segment with very high prominence will elicit a response from even very slow auditory filters, giving a longer vertical line, whereas a segment with low prominence will only elicit responses from the faster filters, giving a shorter vertical line. The authors then compute a 'P' (prominence) value for each segment by integrating the peak output values of the filter channels over time. This prominence value is influenced by the intensity, duration and frequency of the event.

Figure 1.3. Reproduced from Lee & Todd (2004).

(a) Rhythmogram (top) and computed prominence values (bottom) for the words "IMport" vs "imPORT".



(b) Rhythmogram (top) and computed prominence values (bottom) for the word "reconciliation"



Although the two rhythmograms appear similar, the two utterances actually differ in the relative prominence (P value) of the first and third vowel segments, shown in the corresponding prominence plots below. For "IMport", the first segment is slightly more prominent than the third segment, while for "imPORT", the first segment is much less prominent than the third segment. Therefore, the model appears to effectively capture phonetic segments, and is able to compute a prominence value for each of these segments.

However, there is no explicit representation of syllables or words in the model. For example, Figure 1.3b shows the rhythmogram and prominence values for the word *"reconciliation"*. Phonetic segments are again well-represented here (although there is not a one-to-one mapping between phonemes and rhythmogram segments), and the 'stressed' portions of the word are correctly assigned greater prominence values. However, it is hard to infer the syllable pattern from either the rhythmogram, or from the assigned P values. For example, if vowels were assumed to be more prominent than consonants (as was the case for the word *"import"*), then one would expect to see five or six vocalic segments with greater prominence than their surrounding consonantal segments. Instead, the prominence graph shows just two major vocalic peaks and two additional smaller peaks corresponding to the fricative consonants. Therefore, the prominence information computed in the rhythmogram would appear to be insufficient for the auditory system to infer the presence of six syllables in the word.

In fact, a similar problem with specifying syllabicity is present in the previous example with *"IMport"* and *"imPORT"*. If it is assumed that each peak in prominence (above a certain threshold) indicates a new syllable, then *"IMport"* could be correctly inferred to have two syllables. However, the prominence values of the first two segments in *"imPORT"* are very close together, and very low. This would have to be interpreted as evidence for either three syllable in the word, or just one. The issue of syllabicity is important because perceived rhythm patterns in speech depend on the alternation of strong and weak *syllables*, not strong and weak *segments*. Therefore, if a model does not accurately specify syllables, it is hard to see how it would accurately specify the rhythmic patterns that depend on these syllables.

In summary, the auditory primal sketch model is an impressive and physiologically credible model of *segmental* prominence. However, simply knowing the location of prominent segments is insufficient to infer prosodic patterning if one does not also know how many syllables lie between these prominences. This limitation is acknowledged by the authors (Lee & Todd, 2004), who claim that *"it would be a simple matter, and one requiring no controversial psychological assumptions, to incorporate into the model a routine for detecting the presence or absence of periodic acoustic energy and thereby enable it to identify likely syllable nuclei"*. Until this syllable feature is incorporated, the auditory primal sketch model remains primarily a segmental model. It is therefore likely to share the successes and short-comings of other segmental rhythm-metric measures.

1.3.4.2 Coupled Oscillator Models

The Coupled-Oscillator model (COM) of O'Dell and Nieminen (1999) uses the relative timing or phase relationships between multiple oscillators to model differences in speech rhythm across language classes (see also Barbosa, 2002). The COM is a mechanistic model of the relationship between the interstress interval (ISI), or duration of the prosodic foot, and the number of syllables contained in the foot. O'Dell and Nieminen's original model was motivated by the observation by Eriksson (1991) that the ISI between prosodic feet was a simple linear function of the number of syllables '*n*' in each foot, such that ISI = b*n* + a (where a and b are constants). Eriksson further suggested that languages with putatively different rhythm classes could differ in the value of the constant term 'a'. Following this description, O'Dell & Nieminen produced a coupled-oscillator model that would behave in a manner described by Eriksson's formula. Central to the model were two oscillators, termed a

'stress group oscillator' and a 'syllable oscillator', and the general coupling function between these two oscillators (computed using their averaged phase difference). Durational changes (e.g. in ISI) were then explained in terms of the behaviour of these coupled oscillators.

O'Dell and Nieminen (1999) demonstrated that their COM was able to distinguish between languages with different rhythm patterns such as English and Spanish (via oscillator coupling strength). More recently, O'Dell et al. (2007, 2008) used the temporal structural of conversational Finnish speech to differentiate the effect of different hierarchical levels of rhythm (modelled as different coupled oscillators) via Bayesian inference. They reported that stress and mora were rhythmic factors in Finnish, and that there was also a possible role for foot-based timing.

Therefore, coupled oscillator models seek to explain the observed rhythmic parameters in speech (e.g. ISI durations, number of syllables per foot) in terms of mechanical interactions between hypothetical oscillators. This approach is powerful and can produce novel insights into the origins of rhythmic structure. However, the oscillators used in these models are purely hypothetical, and are meant only as abstract representations of linguistic entities such as stress feet and syllables. Unlike 'physiological' models like the auditory primal sketch model, the activity of these COM oscillators is often not motivated by psychological or neural mechanisms, or even constrained by actual acoustic data. A second major difference is that COM models (like O'Dell & Nieminen, 1999) are typically models of *duration* rather than of prominence. Finally, COM models have typically been used to explain *global* differences between language classes, rather than to provide *local* interpretations of actual rhythm patterns (i.e. in prosodic stress transcription).

1.4 THE NEED FOR A COMPLEMENTARY AMPLITUDE-BASED ACCOUNT OF SPEECH RHYTHM

Having provided an overview of the very different previous approaches to measuring speech rhythm, the 'amplitude modulation' approach used in this thesis is now introduced. In terms of the acoustic cues to prosody and stress, it is known that stressed syllables tend to be higher in amplitude, longer in duration and have a distinctive fundamental frequency pattern (Hirst, 2006). Therefore, the alternating 'Strong-weak' syllable patterns that generate the percept of speech rhythm would also associated with patterns of change in all three acoustic dimensions (amplitude, duration and frequency). Traditionally, fundamental frequency was thought to play a primary role in prosodic stress perception (Fry, 1954). However, more recent studies using natural speech have found that amplitude and duration cues play a stronger role than fundamental frequency in prosodic prominence, and by extension in speech rhythm (Greenberg, 1999; Kochanski et al, 2005). Accordingly, methods of describing and measuring speech rhythm can be broadly classified as being either *duration*-based in approach (e.g. 'rhythm-metrics'; O'Dell and Nieminen's Coupled Oscillator Model, 1999) or *amplitude*-based in approach (e.g. p-centres).

In natural speech, duration and amplitude cues to prosodic stress typically co-vary (Kochanski et al, 2005), therefore both duration-based and amplitude-based accounts of speech rhythm would be expected to yield complementary results. However, in the speech rhythm community, much more attention has been paid to duration-based accounts of speech rhythm than to amplitude-based accounts of speech rhythm. Amplitude-based measurements are commonly used in the p-centre literature (e.g. for detecting the 'beats' in single syllables), but have not been used as a more general metric for rhythm and prosodic patterning in speech (although see Silipo & Greenberg, 1999 and Tilsen & Johnson, 2008). Therefore, although amplitude cues are *expected* to contribute toward speech prosody and rhythm patterning, it is not known exactly *how* amplitude variations in the acoustic signal contribute toward the percept of speech rhythm. In this thesis, this gap in knowledge is addressed. Two amplitude-based explanatory accounts of speech rhythm are developed and assessed - the AMPH models. It is not intended that these AMPH models should replace or supersede the existing duration-based accounts of speech rhythm. Rather, the work here provides a complementary account of speech rhythm, using the dynamic amplitude cues that are present in the speech amplitude envelope as amplitude modulation (AM) patterns.

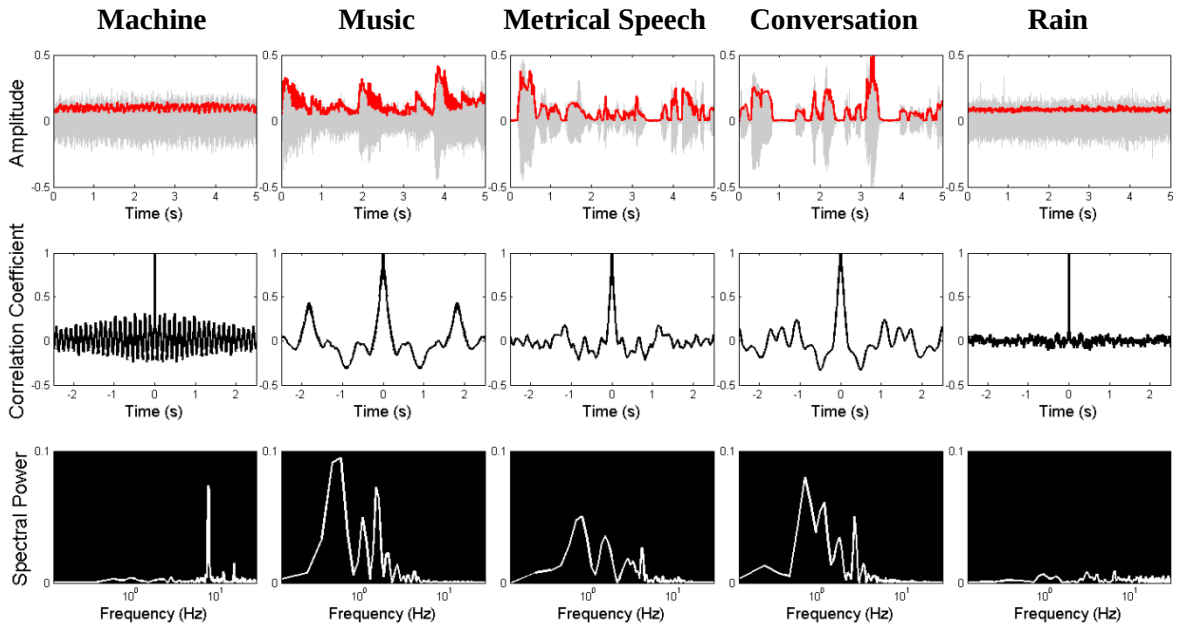
1.5 THE AMPLITUDE MODULATION STATISTICS OF SOUND

Natural sounds are complex and richly-structured in the time and frequency domains. This rich structure in the acoustic signal can be described in terms of two-dimensional patterns of temporal and spectral amplitude modulation (e.g. Chi et al, 1999; Elliott & Theunissen, 2009). The statistics of these amplitude modulation patterns can provide valuable cues to the temporal structure of the sound. For example, Turner (2010) and McDermott & Simoncelli (2011) demonstrated that natural sounds such as rain, fire, birdsong and speech displayed different 'auditory textures' which could be summarised through statistics such as amplitude modulation depth, modulation time-scale and cross-frequency-channel modulation dependency. The general importance for human psychological performance of learning the statistical structure of natural sounds is well-recognised (Winkler et al, 2009). For example, Bayesian approaches are attractive for characterizing the statistical regularities underpinning speech rhythm because they 'learn' statistical structure from the input. For natural sounds like rain, wind and fire, Turner (2010) demonstrated that Probabilistic Amplitude Demodulation (PAD) provided an effective Bayesian learning approach for extracting amplitude modulation structure. Consequently, he argued that natural sounds are characterised by amplitude (local sound intensity) modulation patterns which are correlated over long time scales and across multiple frequency bands.

If an important component of natural sounds is their modulation content, then amplitude modulation (AM) may be central to the perception of speech rhythm. Consistent with this proposal, the speech amplitude envelope contains strong periodicity at slow rates consistent with our experience of prosodic rhythm. For example in Figure 1.4, the autocorrelation function and power spectrum of the amplitude envelope of speech (metrical nursery rhymes and spontaneous conversation) is contrasted with the same measures for a mechanical sound, music and a natural sound (rain). Visual inspection shows that both types of speech contain strong periodicity, particularly at slow rates ~1-2 Hz and ~5 Hz. This is in contrast to rain, which shows little periodicity overall, and to the machine sound, which shows a single spike in periodicity at a higher frequency of ~10 Hz. In fact, the periodic profile of speech is most similar to that of Western music, which is also dominated by slow periodicity ~1-2 Hz. This is not surprising given that the music of a culture is thought to reflect the rhythm patterns of its language (e.g. Patel et al, 2006). This simple analysis

highlights the fact that speech has a strong temporal structure, as evidenced by periodic patterns in its amplitude envelope.

Figure 1.4. The periodicity profile of five different sounds (5s segments). L-R columns: hand mixer, Western music, metrically-spoken nursery rhyme, spontaneous conversation, rain. The top row shows the sound-pressure waveform with its amplitude envelope overlaid in bold. The amplitude envelope was extracted using the Hilbert transform and low-pass filtered under 40 Hz. Prior to autocorrelation, the envelope was down-sampled to 500 Hz, and major trends in the envelope were removed by sequential polynomial curve fitting (1st-4th order), taking the residual after each fit. The de-trended envelope was then autocorrelated with itself over time-lags -2.5s to 2.5s, yielding the autocorrelation function. Finally, the power spectrum of the autocorrelation function was computed using the fast Fourier transform. The middle and bottom rows show the resulting envelope autocorrelation function (+/- 2.5s) and power spectra for each sound respectively.



1.6 THE AMPLITUDE ENVELOPE & THE MODULATION SPECTRUM

Speech is perhaps the most complex acoustic signal that the brain decodes, with important temporal structure at different timescales, as exemplified by formants (concentrations of energy in narrow frequency bands, timescale tens of ms), syllables (timescale hundreds of ms) and sentences (timescale seconds). Consequently, natural speech contains modulations in amplitude and frequency over a variety of timescales. These amplitude and frequency modulations are produced by articulators that dynamically regulate airflow through the vocal passage by changing its shape and length. Articulators also move over different time scales. For example, relatively slow movements such as the rise and fall of the jaw produce slower modulations associated with rhythm and prosody, while the more rapidly moving articulators, such as the lips and tongue, produce quickly-changing modulations that yield phonetic patterns (Nittrouer, 2006). However, the motion of slow and fast articulators is strongly co-ordinated in time. For example, Kelso et al (1986) demonstrated that there are stable oscillatory phase relations (cyclical relative temporal alignments) between jaw movement cycles and upper lip movement onsets across variations in speaking rate and stress patterns. Vocal tract movements also produce systematic facial and head movements that convey the syllable structure of speech (Yehia et al, 2002; Munhall et al., 2004).

The speech signal is commonly analysed using either the speech spectrogram, which emphasises frequency changes over time, or as a sound-pressure waveform, which emphasises amplitude changes over time (e.g. the amplitude envelope). These complementary ways of representing the speech signal are shown in Figure 1.5. Note that these different representations have led to different theoretical frameworks for linking acoustic statistical structure to linguistic/phonological experience. The spectrogram foregrounds phonetic changes, and is commonly analysed by associating changes in formant patterns, or 'formant transitions' (the rapidly-changing concentrations of energy in narrow frequency bands) to articulated consonants and vowels. Conversely, the sound-pressure waveform foregrounds syllabic structure and prosodic changes associated with slowly varying amplitude changes in the signal. This can be illustrated by referring to Figure 1.5. In the uttered phrase "MA-ry MA-ry", the [i] in syllable [ɹi] is associated with a rapid rise in the frequency of the second formant due to the tongue being positioned high at the front of the mouth (highlighted in black boxes in Figure 1.5a). Conversely, the large rise in amplitude at

the beginning of the phrase (Figure 1.5b), and again with the repetition of the stressed syllable [mɛː] in "MA-ry" are related to a larger mouth aperture (lowered jaw) producing greater airflow.

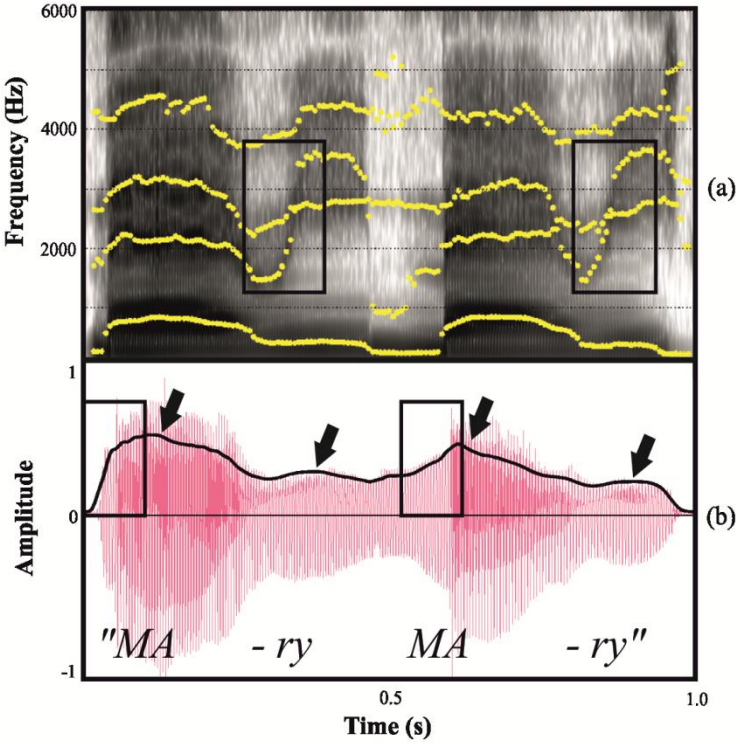


Figure 1.5. Two complementary representations of the acoustic signal for the spoken words "MA-ry MA-ry" (a) Frequency spectrogram and first four formants (dotted lines) (b) Sound pressure waveform (pink) and amplitude envelope (bold line). Boxes in (a) indicate the rapid frequency rise in the second formant corresponding to the vowel [i] in /ry/. Boxes in (b) indicate the rapid increase in amplitude (intensity) associated with the onset of the stressed syllable 'MA'.

In signal processing terms, the speech signal can be modelled as the product of a quickly-varying ***carrier*** (fine structure) and a slower-varying ***amplitude envelope*** (bold outline in Figure 1.5b) that dynamically modulates the amplitude of the carrier. The envelope itself contains multiple rates of amplitude modulation (AM). For example, Figure 1.5b shows that the envelope is dominated by four slow peaks (highlighted with arrows), each associated with an articulated syllable (~4 Hz). However, the envelope also contains smaller, faster fluctuations up to 50 Hz that contain linguistic cues to phonetic manner of articulation, voicing, and vowel identity (Rosen, 1992). The range of modulation rates in the envelope can be expressed as a 'modulation spectrum' which plots the relative power at each modulation rate (Plomp, 1983a; Greenberg, 2006). An example of the typical modulation spectrum of speech (for the 1-2 kHz spectral band) is shown in Figure 1.6 (reproduced from Greenberg et al, 2003).

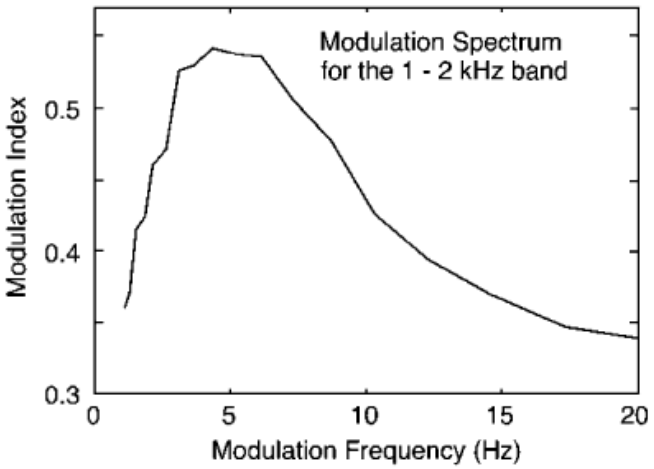


Figure 1.6. Modulation spectrum for the 1-2 kHz frequency band, computed from 2 minutes of material from the SWITCHBOARD speech corpus. Reproduced from Greenberg et al, 2003.

The modulation spectrum typically shows the highest power between 2–12 Hz, peaking at around 3-5 Hz irrespective of differences in language or speech rate (Shannon et al., 1995; Houtgast & Steeneken, 1985; Greenberg et al, 2003; Greenberg, 2006). As the average duration of a syllable is 200 ms, modulations around 5 Hz are likely to relate to syllable-pattern information in speech (Greenberg et al, 2003; Ahissar et al, 2001; Luo & Poeppel, 2007). The peak in the modulation spectrum at the syllable rate thus indicates that the dominant modulation rate in the amplitude envelope is related to the syllable structure of speech. Modulations slower than the syllable rate relate to prosodic stress patterns (Plomp, 1983b; Greenberg et al, 2003; Ghitza & Greenberg, 2009). If different rates of amplitude modulation (AM) in the envelope can be associated with prosodic units such as syllables and stress patterns, then the statistics of these slow AMs may provide the temporal regularities that underlie our perceptual experience of speech rhythm.

1.7 EXTRACTING RHYTHM COMPONENTS FROM THE AMPLITUDE ENVELOPE

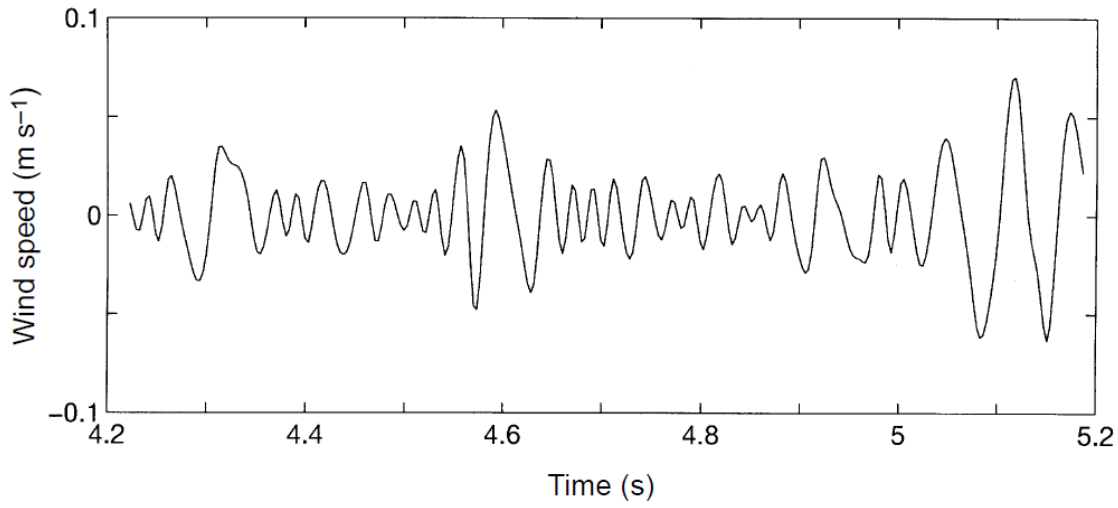
Recently, there has been growing interest in using amplitude envelope-based methods to investigate the temporal structure of the speech signal (e.g. Greenberg, 2003; Tilsen & Johnson, 2008). For example, Tilsen & Johnson (2008) extracted the amplitude envelope (under 10 Hz) from the 700-1300 Hz frequency band of speech, where the speech material was a corpus of conversational English. They divided the speech samples into 2-3 s chunks and computed the power spectrum (modulation spectrum) for each chunk. They then looked for peaks in the power spectrum of each chunk (indicative of strong periodicity), and computed how often these peaks occurred for different modulation rates across all the chunks. They found that overall, around 27% of the chunks contained clear peaks in their modulation spectrum between 1-5 Hz, out of which 18% came from modulations between 1-3 Hz. This analysis demonstrated that even conversational speech contained a significant proportion (~a quarter) of strongly rhythmic stretches (chunks), where the strongest rate of modulation in these rhythmic stretches of speech was around 1-3 Hz (the prosodic stress rate).

If slow amplitude modulations in the envelope do indeed carry rhythmic information, as Greenberg (2003) and Tilsen & Johnson (2008) suggest, it would be interesting to extract these modulations from the envelope, and look for correlates between the acoustic modulation pattern, and the rhythm that listeners perceive. The conventional method for isolating modulations at different rates from the envelope is to use filtering (e.g. low-pass, high-pass or bandpass). This conventional approach (bandpass filtering) is used as the primary method in this thesis to isolate various 'bands' of modulation in the envelope. However, there have been more recent attempts to 'discover' and extract intrinsic slow modulation components from the amplitude envelope, without the use of filtering. Two such novel methods are described here, and the second method (PAD) is also used in these thesis. In a tone-vocoder experiment (Chapter 3 of this thesis), the envelope components generated via bandpass filtering or PAD are explicitly compared by human listeners.

1.7.1 Empirical Mode Decomposition (EMD)

Empirical mode decomposition (EMD) was pioneered by Huang et al. (1998) as a method for adaptively representing non-stationary signals as sums of zero-mean amplitude modulated and/or frequency modulated components, termed 'intrinsic mode functions' (IMFs). This method can be applied to the amplitude envelope of speech to discover its 'dominant' oscillatory components. By definition, an IMF has the same number of extrema (maxima/minima) and zero crossings, and has a symmetric shape around zero. Unlike bandpass-filtered signals, however, IMFs can contain both amplitude and frequency modulation (see Figure 1.7). Practically, these IMFs are extracted using a 'sifting' process in which large trends (non-stationarity) in the data are iteratively removed. In this process, all the maxima (peaks) in the signal are interpolated (cubic spline method) to form an 'upper envelope', and all the minima (troughs) in the signal are similarly interpolated to form a 'lower envelope'. The mean of the two envelopes at each time point is then calculated, and this mean (effectively a slow trend) is subtracted from the original signal to yield a residual. This trend removal process is then repeated iteratively with the residual until the new residual meets the IMF criteria (i.e. regarding extrema, crossings and symmetry). This residual is then called with first IMF, and typically corresponds to the fastest oscillation rate in the signal.

Figure 1.7. Reproduced from Huang et al (1998). Example of a typical intrinsic mode function (IMF) with the same numbers of zero crossing and extrema, and symmetry about zero.



The first IMF in the example shown from Huang et al, 1998 was obtained after 9 iterations. After the first IMF is obtained, this is then subtracted from the original signal, and the entire sifting process is repeated with the residual to uncover other slower IMFs. The recovered IMFs themselves can then be analysed for their characteristic frequency and pattern. Arvaniti (2012) used this approach to compare the intrinsic mode functions extracted from speech samples of different languages. They found that the first and second IMFs extracted corresponded well to syllabic and supra-syllabic elements in speech respectively. Moreover, across languages, the frequency of the second IMF (supra-syllabic element) clustered around 2 Hz. The authors interpreted this 2 Hz result as evidence that languages tend to have a common stress rate that is close to the 'natural tempo' (Clarke, 1999) .

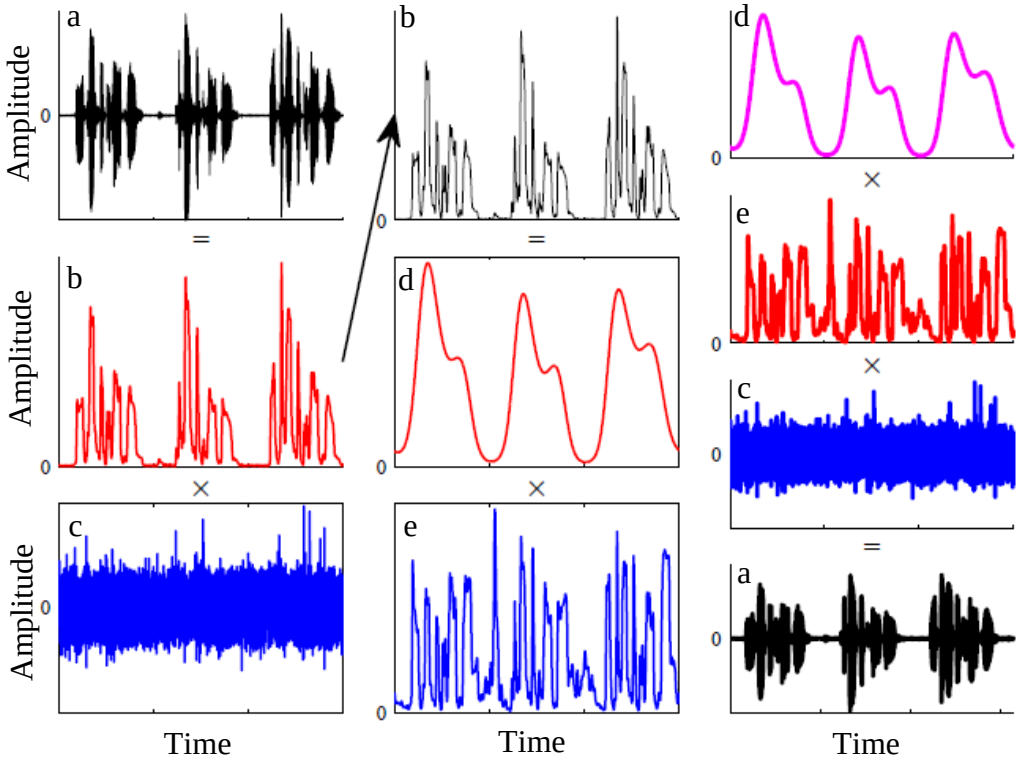
1.7.2 Probabilistic Amplitude Demodulation (PAD)

The Probabilistic Amplitude Demodulation (PAD, Turner, 2010; Turner & Sahani, 2011) method generates a 'model' of the signal (where the model comprises a positive slow envelope, and a fast carrier), and uses Bayesian statistical inference to identify the envelope with the best fit to the data. In the first step, the envelope is modelled by applying an exponential nonlinear function to a stationary Gaussian process. This produces a positive-valued envelope whose mean is constant over time. Importantly, the degree of correlation between points in the envelope is controlled by the parameters of the model (entered manually or 'learned' from the data). This correlation determines the typical time-scale of variation in the envelope, which translates into its dominant modulation rate. The carrier is modelled as a Gaussian process which is uncorrelated in time (like white noise). The envelope and carrier are then be combined into one possible solution for the original data.

The choice of the most appropriate envelope and carrier solution for the data is then cast in Bayesian terms, and solved as a problem of probabilistic inference. For example, the parameters of the model determine the 'prior distributions' describing all possible envelopes and carriers that could be produced. The 'posterior distribution' describes the conditional probabilities associated with all the possible envelopes and carriers, given the original data (i.e. p(env,car\data)). The desired solution are the specific envelope and carrier that have the highest conditional probability, and represent the best fit for the data. Therefore, in the second step, a gradient-based method is used to search for this optimal solution where the probability is maximal, yielding the envelope with the best fit to the data. A useful feature of

PAD is that this process can be run recursively to recover different envelopes with different dominant modulation rates, as shown in Figure 1.8. This is done by changing the demodulation parameters after each iteration so that progressively slower and slower envelopes are returned, forming a 'modulation hierarchy'. Functionally therefore, both EMD and PAD methods can extract 'naturalistic' modulations from the amplitude envelope on different timescales. In the EMD method, these component modulations are termed IMFs, whereas in PAD they form tiers of a modulation hierarchy.

Figure 1.8. Reproduced from Turner (2010). Example of a modulation hierarchy derived by recursive application of PAD. First, the data, 'a', are demodulated using PAD set to a fast timescale (left column). This yields a relatively quickly-varying envelope ('b') and a carrier ('c'). Next, the demodulation process is re-applied to the extracted envelope 'b' (middle column), using a slower timescale than before. This yields a slower daughter envelope ('d') and a faster daughter envelope ('e'). This set of daughter envelopes form the two tiers of the modulation hierarchy. Mathematically, these two tiers ('d' & 'e') can be multiplied back with the very first carrier ('c', bottom left) to yield the original signal, 'a' (shown in the right column).



1.8 METHODS FOR AUTOMATIC SYLLABLE DETECTION

Since the 1970s, loudness or amplitude-based procedures have already been employed in the automatic detection and segmentation of continuous speech into syllables. For example, Sargent et al (1974) used peak-to-peak amplitude measurements in automatic syllable detection. This was based on the concept that sharp increases and decreases in amplitude typically accompany the onset and offset of syllable nuclei. This method achieved an average success rate of 86% for a corpus of 3.5 minutes of speech produced by 9 speakers. In other early studies, Mermelstein (Mermelstein & Kuhn, 1974; Mermelstein, 1975) proposed a 'convex hull algorithm' that used minima in the loudness function of speech in the automatic detection of syllable boundaries. When applied to a corpora of slow, careful speech in quiet (~400 syllables), this 'trough-finding' method yielded an impressively high success rate of around 90%. However, the author noted the need for a moderate amount of post-processing to weed out fricative "syllabic fragments" that were erroneously detected as full syllable units. Pfitzinger et al (1996) also used a loudness-based automatic syllable detector, but they applied this to a more extensive corpora that included both read and spontaneous speech. In their method, the speech signal was first band-limited (e.g. 250-2500 Hz), and then the amplitude envelope was obtained by low-pass filtering with a cutoff of ~10 Hz. Syllable nuclei were located by identifying peaks in the slow-varying envelope. Pfitzinger et al (1996) obtained accuracy levels of 87% and 79% for read and spontaneous speech respectively.

More recent syllable detection methods have, in general, continued to be based on the concept of energy peak detection. However, improvements have been made in terms of the algorithms for boundary/peak detection, the incorporation of information from other acoustic cues (e.g. frequency), and the use of more sophisticated modelling (e.g. hidden markov models, HMMs) or supervised machine learning methods. Since the older studies have tended to use different speech corpora from the more recent studies, it is difficult to make direct comparisons regarding accuracy. Nonetheless, since the 1970s, a broad range of new approaches have been used to tackle the problem of syllable detection. A short selection of these studies is summarised here :

(1) Unsupervised systems.

- Jittiwarangkul et al (1998) proposed a new syllable boundary detection method based on iterative forwards and backwards searching for local maxima and minima in the smoothed energy contour of speech. When applied to the absolute energy contour of

speech, this search method produced a syllable detection accuracy of 93% for a set of 36 utterances produced by 11 speakers.

- Zhang & Glass (2009) developed a novel method of locating syllable vowel nuclei in which amplitude envelope peak-detection was guided by a rhythm-based prediction of the *future* location of peaks. In their method, the 'instantaneous' speech rhythm was estimated from the preceding peaks in the envelope, and this was used to predict intervals where the next syllable nucleus could appear. When tested on the TIMIT corpus, this rhythm-guided method was found to be able to locate ~87% of syllable vowels accurately.
- Xie & Niyogi (2006) developed an unsupervised system for automatic detection of syllabic nuclei that used a combination of 2 acoustic cues : periodicity and energy. Syllable nuclei were identified by locating regions of speech with *both* high periodicity and high energy. This detection system was found to have an accuracy rate of 81.6% when tested on the TIMIT corpus, and continued to perform robustly in the face of noise degradation.
- de Jong & Wempe (2009) used a similar method to Xie & Niyogi (2006) for the detection of syllable nuclei, via the combination of intensity and voicing (periodicity) cues. In their script written for the software programme Praat (Boersma & Weenink, 2007), potential syllable nuclei were first identified by finding local peaks in the intensity contour of the speech sample. These peaks were then evaluated for voicing, and only voiced peaks were retained as bona fide syllable vowel nuclei. The authors then used this syllable information to compute the speech rate of the Dutch utterances. Although the accuracy of syllable vowel detection was not evaluated per se, the authors reported a very high correlation (up to r =.88) between the speech rates computed automatically via their detection algorithm, and the speech rates measured manually by hand.

(2) Supervised systems.

- Howitt (2000) used the speech energy in a fixed frequency band of 300-900 Hz and applied a recursive convex hull algorithm to identify local peaks in energy. Three acoustic features from these peaks (depth, duration and level) were then combined using a multi-layer perceptron in order to identify 'vowel landmarks'. After training, the success rate for Howitt's perceptron vowel landmark detector was 88% for the TIMIT database.

- Shastri et al (1999) developed a sophisticated neural network model to detect and segment syllables from a modulation spectrogram (<16 Hz) representation of continuous speech. Their 'Temporal Flow Model' supported arbitrary link connectivity across multiple layers of nodes, and allowed for both feedforward and recurrent links. This complex system of links enabled the model to smooth and differentiate between signals, measure the duration of features, detect onsets, maintain context and carry out spatio-temporal feature detection and pattern matching. Therefore, the model was able to simulate cognitive functions such as short-term memory and context sensitivity. When trained and tested on a corpus of fluent 'telephone speech', the Temporal Flow Model was found to be able to predict the onset of syllables with an accuracy of ~84%.

Other studies have gone a step further to combine the automatic detection of syllables with the automatic *labelling* of the prosodic prominence and stress of these syllables. A similar approach is used in this thesis. For example, Tamburini (2003) used a modified version of Mermelstein's convex hull algorithm to identify the location of syllable vowel nuclei in connected speech. He then developed a 'prominence function' to automatically determine the prominence of each detected syllable. This function took into account a combination of acoustic parameters including overall RMS energy, 'mid-frequency emphasis' (energy in the 500-2000 Hz band), nucleus duration, and pitch variation. Tamburini (2003) reported that the prominence detector successfully classified 80% of syllables from the TIMIT corpus as bring either prominent or non-prominent. Fujisawa and colleagues (Fujisawa et al, 1998; Minematsu et al, 1999) used a more complex approach where they specifically modelled the structure and position of syllables within words using context-sensitive hidden markov models (HMMs). For example, in Fujisawa et al (1998), the HMMs used a combination of parameters such as LPC mel cepstrum coefficients, power (amplitude) and fundamental frequency (F0) to determine the local position of the word accent (stressed syllable). After being trained on multisyllabic words from the ATR English database, these HMMs were found to correctly detect around 90% of the stressed syllables in utterances by native English speakers.

Most recently, Kalinli (Kalinli & Narayanan, 2007; Kalinli, 2011) has begun to develop a new generation of biologically-inspired attention-based models of syllable and word prominence. These unsupervised models first extract salient auditory features from the speech stimulus with reference to processing mechanisms in the central auditory system.

These auditory features include intensity, temporal and frequency contrast, orientation and pitch. These features are then assembled into a single master saliency map, which is used to detect prominent syllables and words. In their initial study (Kalinli & Narayanan, 2007), the accuracy of the model was reported to be 75% for syllable prominence and 78% for word prominence. In a later study where a similar model was used specifically for syllable nucleus detection (Kanlinli, 2011), the method achieved an impressive 92% accuracy rate on the TIMIT speech corpus. These new psychologically-inspired models of syllable prominence detection show much promise for the future. In this thesis, a new method for syllable detection and prosodic prominence assignment is likewise inspired by neural cortical mechanisms of speech processing, namely the entrainment of neuronal oscillations to amplitude modulation patterns in speech (Giraud & Poeppel, 2012).

1.9 Cortical Oscillations and Modulation Hierarchies in the Auditory System

Empirical studies of animals and humans support the view that the auditory system possesses 'modulation channels' whose function is to extract patterns of amplitude modulation at different rates. Amplitude modulation channels have been demonstrated in neurophysiological studies with animals (Schreiner & Urbas, 1986; Langner & Schreiner, 1988). For example, Schreiner & Urbas (1986) demonstrated that neurons in the auditory cortex of cats were able to follow amplitude modulations of pure tones. These neurons showed ‘best’ modulation frequencies from 3 Hz to 100 Hz, with a median value of about 20 Hz. Psychophysiological masking experiments have also indicated the existence of similar modulation channels in the human auditory system (Houtgast, 1989; Bacon & Grantham, 1989). Dau and colleagues (Dau et al 1996; Dau et al 1997a; Dau et al 1997b) described a model for the operation of a 'modulation filterbank' in the human auditory system. In essence, they proposed a system that filters different rates of amplitude modulation in the acoustic signal into logarithmically-scaled 'modulation bands'. For very low modulation rates (<10 Hz), their 'modulation filterbank' specifically preserves modulator phase at the output of the modulation filter. This feature was included following observations from their psychophysical experiments (Dau, 1996) where, for modulation rates of up to about 10 Hz, subjects could discriminate changes in the starting phase of sinusoidal amplitude modulation

of a 5-kHz carrier. This concept of a 'modulation filterbank' is also implemented in the two proposed AM Phase Hierarchy models in this thesis. These filterbanks use a series of bandpass filters to separate slow modulations in the envelope into a hierarchical series of 'AM tiers'. Like Dau and colleagues, the modulation filterbank also preserves phase information from these slow modulations, which are then used to compute prosodic rhythm patterns in speech.

The neural location of putative human 'modulation channels' that could mediate rhythm detection is currently unclear. However, animal studies suggest that the temporal resolution of neurons in the auditory system tends to decrease as one ascends from the periphery to the cortex (Schreiner & Urbas, 1986). For example, while neurons in the cochlear nucleus have a maximum response at modulation rates between 100-300 Hz (Fernald and Gerstein, 1972; Moller, 1974), AM responses in the inferior colliculus are tuned to modulation rates below 40 Hz (Rees & Moller, 1983). The median best modulation frequency measured in the auditory cortex is even lower at 20 Hz (Schreiner & Urbas, 1986). Consistent with animal studies, the temporal resolution of the AM response in humans also appears to degrade from the brainstem to the auditory cortex, with most cortical regions tuned to low AM frequencies around 4 - 8 Hz (Giraud et al, 2000). As such, it seems that the central, cortical neural mechanisms that might be best ‘tuned’ for rhythm detection at slow AM rates (i.e. below 8 Hz, 'stress' and 'syllable' rates).

In line with this, there is converging evidence from MEG and EEG studies that cortical oscillatory mechanisms play a role in tracking the slow modulation structure of speech. The brain generates a neuroelectric steady-state response (SSR) to amplitude modulations in speech (Ahissar et al, 2001; Aiken & Picton, 2008; Luo & Poeppel, 2007) and also to AM noise (e.g. Liegeois-Chauvel et al, 2004). This SSR tracks the stimulus modulation pattern and is itself oscillatory. In principle, therefore, if the speech envelope contains modulations at different timescales (rates), cortical oscillations should track these components at equivalent temporal rates. Indeed, the convergence between the characteristic timescales in speech and the dominant frequency bands in neuronal oscillations has been used to argue that oscillatory alignment (‘phase locking’) may be an important neural mechanism for speech parsing and hierarchical integration (Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012; Schroeder et al, 2008; Zion Golumbic et al, 2012). For example, the neural delta band (1-3 Hz) entrains to events that occur at ~300-1000 ms intervals, and is well-positioned to capture prosodic stress occurring on average every 493 ms (Dauer, 1983).

The neural theta band (3-7 Hz) tracks events occurring at ~150-300 ms intervals. This coincides with the average timing of syllable onsets, which have a typical duration of ~200 ms (Greenberg, 2006). Finally, neural beta (12-25 Hz) and gamma (25-80 Hz) bands may track fast acoustic events that require temporal precision on the order of tens of milliseconds, such as formant transitions and differences in voice onset time. Therefore, oscillatory phase-locking to AM patterns at different rates may support speech processing.

In humans, low frequency oscillations in the theta band have been shown to phase-lock to amplitude modulation patterns in speech, and the strength of such phase-locking is associated with speech intelligibility (Ahissar et al, 2001; Luo & Poeppel, 2007; Luo et al, 2010). Moreover, neural phase-locking also reflects attentional-tracking in situations that involve speaker separation (Ding & Simon, 2012). This neural pattern of phase-locking is so specific that temporal modulation patterns in speech can even be reconstructed from intracranial EEG patterns (Pasley et al, 2012) as well as from MEG activation patterns (Ding & Simon, 2012). Note however, that Peelle, Gross & Davis (in press) have recently demonstrated phase-locking in the MEG theta range to *unintelligible* speech. Although the participants were adults, this shows in principle that phase-locking could be present in infants even before speech is understood. Peelle et al further showed that phase-locking was enhanced for intelligible speech, suggesting that theta phase-locking is both stimulus-driven and reflective of successful speech processing.

Neuronal oscillations are global fluctuations in the excitability of neuronal populations. These electrical fluctuations are measurable in vivo as the local field potential (LFP) of groups of neurons, which give rise to scalp-measured oscillatory activity. Neurophysiological studies have indicated that the timing of neuronal spiking activity is locked to the phase of the LFP (Kayser et al, 2009; Montemurro et al, 2008). These studies suggest that the oscillatory LFP controls the timing of neuronal excitability and gates spiking activity (Schroeder & Lakatos, 2009; Elhilali et al, 2004). By this account, the oscillatory phase of the LFP defines alternating periods of optimal excitability (during which a stimulus would elicit strong spiking activity), and low excitability (during which few or no spikes would be elicited). Hence, sensory information may be integrated during high excitability 'windows' and then transmitted onwards to other brain areas during periods of low excitability. If periods of optimal excitability are timed to coincide with significant (high amplitude) events in the acoustic stream, this would align optimal neural processing periods with the arrival of speech information chunks, and low excitability with intervening silences

and noise (Zion Golumbic et al, 2012; Giraud & Poeppel, 2012). Hence, neuronal oscillations may provide a 'temporal windowing' function that allows speech encoding at lowered neural cost (spikes).

Such sparse sampling of the speech signal may be adaptive because human speech is itself a sparse, stochastic signal characterised by strong bursts of energy interspersed with periods of low or no activity. These gaps or pauses occur between words and phrases, but also between syllables and phonemic segments. If the brain were to encode the entire speech signal with equal fidelity, precious cognitive resources could be wasted on encoding periods of low or no speech activity with the same level of detail as periods of high speech activity. While the presence of longer gaps or pauses in speaking can sometimes be meaningful in and of themselves (e.g. in signalling prosodic boundaries), a 'low resolution' encoding should suffice to indicate that such pauses have occurred. Moreover, in real-world listening conditions, the actual acoustic content of gaps and pauses is not pure silence, but background noise that is irrelevant to the spoken content, and could even be distracting for the listener. Therefore judicious temporal windowing or sampling could allow the brain to selectively capture only the most salient information in the signal (i.e. sections with high activity) with a high level of detail, and neuronal oscillations could operate as such a temporal windowing mechanism. Furthermore, Poeppel (2003) suggests that such temporal sampling could occur on multiple time scales (eg. syllable and phoneme), by multiple oscillatory rates (eg. theta and gamma) to capture critical information at different linguistic levels.

Finally, neuronal oscillations can exhibit hierarchical cross-frequency coupling or nesting. This means that the activity of faster frequencies is dynamically modulated by the activity of slower frequencies, resulting in 'nested' relationships. A well-established example of such nesting is hippocampal theta-gamma nesting, in which the phase of slow theta oscillations modulates the power of fast gamma oscillations (Canolty et al, 2006). Hierarchically-nested activity of this nature has also been demonstrated in the mammalian auditory cortex between delta, theta and gamma oscillations (Lakatos et al, 2005). Such hierarchical nesting may provide a mechanism for speech information that is sampled at slow (syllable, theta) and fast (phoneme, gamma) timescales to be aligned and integrated (Poeppel, 2003; Giraud & Poeppel, 2012). If the neuronal oscillations that entrain to speech are hierarchical in nature, this raises the possibility that the speech envelope itself may also contain hierarchically-nested modulations on different timescales. This concept of hierarchical-nesting in the speech envelope is investigated in this thesis.

1.10 AMPLITUDE MODULATIONS AND SPEECH INTELLIGIBILITY

Amplitude modulations in the speech envelope have most commonly been investigated with regard to speech intelligibility. For example, in two seminal studies, Drullman and colleagues (Drullman et al 1994a, Drullman et al 1994b) examined the range of modulation frequencies in the envelope that were the most important for speech intelligibility. They systematically low-pass filtered or high-pass filtered the speech envelope at different cutoff frequencies, combined this filtered envelope back with the original fine structure, and tested the effect on speech intelligibility in each case. In the low-pass filtering exercise (Drullman et al, 1994a), the filter cutoffs increased logarithmically from 0 to 64 Hz (i.e. 0 Hz, 0.5 Hz, 1 Hz, 2 Hz, 4Hz, etc). Drullman et al (1994a) found that speech intelligibility increased as the low-pass filter cutoff increased from 0 up to 16 Hz. However, increasing the filter cutoff beyond 16 Hz (to 32 Hz or 64 Hz) did *not* significantly improve participants' performance any further. This result suggested that modulation frequencies *below* 16 Hz were the most important for speech intelligibility, while modulation frequencies *above* 16 Hz made only a marginal contribution (when all the lower frequencies were intact).

In their companion study (Drullman et al, 1994b), the speech envelope was *high-pass* filtered with logarithmically-increasing cutoff frequencies from 0 Hz up to 128 Hz. Therefore, an increasing proportion of low-frequency modulations was removed at each filter cutoff. As before, the filtered envelope was re-combined with the original fine structure. This time, Drullman et al (1994b) witnessed *no reduction* in speech intelligibility when only modulations below 4 Hz were removed. However, as the filter cutoff increased above 4 Hz, removing more and more modulations *above* 4 Hz, speech intelligibility began to decline significantly. This result suggested that modulation frequencies *above 4Hz* were the most important for speech intelligibility, while modulation frequencies *below* 4 Hz made only a marginal contribution (when all the higher frequencies were intact).

When the results of both studies are taken together, this suggests that modulation frequencies between 4-16 Hz in the speech envelope are the most important for speech intelligibility. Speech intelligibility suffers when modulations in this range are removed from the envelope. It is interesting to note that Drullman's range of 4-16 Hz *includes* the neural theta (3-7 Hz) and alpha (7-12 Hz) bands of oscillation. If neuronal oscillations track the activity of the speech envelope in a rate-dependent manner (i.e. theta oscillations track amplitude modulations between 3-7 Hz, and alpha oscillations track amplitude modulations

between 7-12 Hz), as suggested by Giraud & Poeppel (2012), this could explain why neural activity in the theta band is so strongly associated with speech intelligibility (e.g. Luo & Poeppel, 2007). However, by this argument, *alpha* oscillations should also be implicated in speech intelligibility. This is indeed the case, but unlike theta involvement, alpha involvement appears to be *negatively* related to speech intelligibility. For example, Obleser & Weisz (in press) reported that *suppressed* alpha activity was associated with *better* speech intelligibility.

This apparent trade-off between theta and alpha activity in the neural response to speech suggests that neural activation is not entirely stimulus-dependent. Rather, there may be *selective* enhancement and suppression of different modulation frequencies in speech, leading to different patterns of neural activity in different oscillatory bands. One reason that this may occur is if slower modulations around the theta range (3-7 Hz) correspond to longer *stressed* syllables, whereas faster modulations in the alpha range (7-12 Hz) correspond to shorter *unstressed* syllables (e.g. Greenberg et al, 2003). The processing of theta-rate stressed syllables could be enhanced because these often relate to important 'content' words in speech, as compared to 'function' words, which tend to be unstressed. Also, since the vast majority of words in the English language start with an initial stressed syllable (Cutler & Carter, 1987), these stressed syllables often signal important word boundaries. However, empirical research is required to investigate if, and why, such selective enhancement or suppression of modulation frequencies in the envelope would occur.

1.11 NURSERY RHYMES, SPEECH RHYTHM AND PHONOLOGICAL DEVELOPMENT

In this thesis, nursery rhymes are used as the basis for the two new amplitude envelope-based models of speech rhythm. Nursery rhymes are simple poems that possess a basic regular metrical rhythmic structure. These familiar and popular children's poems (also known as 'Mother Goose rhymes') are commonly recited or sung to young English-learning children at a preschool or nursery age. Simpler rhymes (e.g. 'Rock-a-bye baby' and 'Twinkle twinkle little star') are also commonly sung to infants. Therefore, nursery rhymes are particularly suited for the aims of the current investigation, since they are a type of naturally-

occurring rhythmic speech that infants and children will commonly encounter during language development.

Gueron (1974) studied the metrical structure of 130 Mother Goose nursery rhymes. She concluded that all but one of the nursery rhymes had a simple 'Strong (S) - weak (w)' alternating metrical pattern of : (w) S w S (w) S w S (w), with the weak elements in parenthesis omitted in some rhymes. In Gueron's analysis, 'S' elements were usually realized by a single stressed syllable while 'w' elements were usually realised by between one to three unstressed syllables. The characteristic metrical patterning of nursery rhymes suggests that these poems capture the basic prosodic rhythms used in English. Consequently, nursery rhymes are good materials to use as the basis of a model of speech rhythm as they clearly represent the major metrical rhythm patterns present in spoken English. Learning nursery rhymes also affects phonological development, with empirical evidence that children's early knowledge of nursery rhymes predicts later individual differences in phonological awareness and success in learning to read (Maclean, Bryant & Bradley, 1987; Bryant et al, 1989). Consequently, children's nursery rhymes may be socio-cultural devices that have evolved to support early language learning.

What makes a nursery rhyme effective for language learning? It has been suggested that the rhyming words in nursery rhymes serve to highlight the phonological 'rime' units in words, thereby boosting children's phonological awareness (Maclean, Bryant & Bradley, 1987). However, while rhyming words feature in many nursery rhymes, they are not unique to nursery rhymes. Indeed, rhyming words are a necessary ingredient of any literary poem or song, child- or adult-focused alike. Moreover, not all nursery rhymes contain an abundance of rhyming words. For example, 'London Bridge is Falling Down' arguably contains no rhyming word pairs at all. Many nursery rhymes use rhyming word pairs only parsimoniously, as 'bookends' to phrases, sentences or stanzas. For example, in the rhyme 'Jack and Jill' (below), rhyming word pairs occur in two locations.

"Jack and Jill went up the hill
*to fetch a pail of **wa-ter**,*
Jack fell down and broke his crown
*and Jill came tumbling **af-te**r"*

Rhyming word pairs occur within sentences ('Jill'/'hill', 'down'/'crown') where they occupy the same relative syllable position (3&7). They also occur at the end of a sentence

('**water**'/'**after**'), where again they occupy the same relative syllable position. These positional constraints on rhyming words suggests that they are used to highlight the overall structural symmetry of the poem. In this case, Jack and Jill consists of a binary hierarchical structure of two phrases per line, two lines per sentence, and two sentences per stanza, the boundaries of which are marked by rhyming word pairs. Moreover, the entire rhyme is prosodically patterned throughout with a binary alternating 'Strong(S)-weak(w)' stress pattern, which gives it a bouncing beat as in "JACK and JILL went UP the HILL to...". This alternating 'S-w' motif also occurs at longer timescales in the poem, forming a prosodic hierarchy that can also be represented as nested oscillation patterns, as shown in Figure 1.9. In Chapter 2, the idea that the linguistic prosodic hierarchy can be represented as nested amplitude modulation patterns is explored further.

Figure 1.9. Hypothetical example of a modulation hierarchy for the rhyme 'Jack and Jill' showing binary nested modulations at different timescales, each corresponding to a different prosodic structure. Unit segments at each timescale are highlighted as separate blocks. Each segment's prosodic strength is indicated with arrows marked 'S' (strong) or 'w' (weak), as determined by its phase of occurrence with respect to the next higher tier.

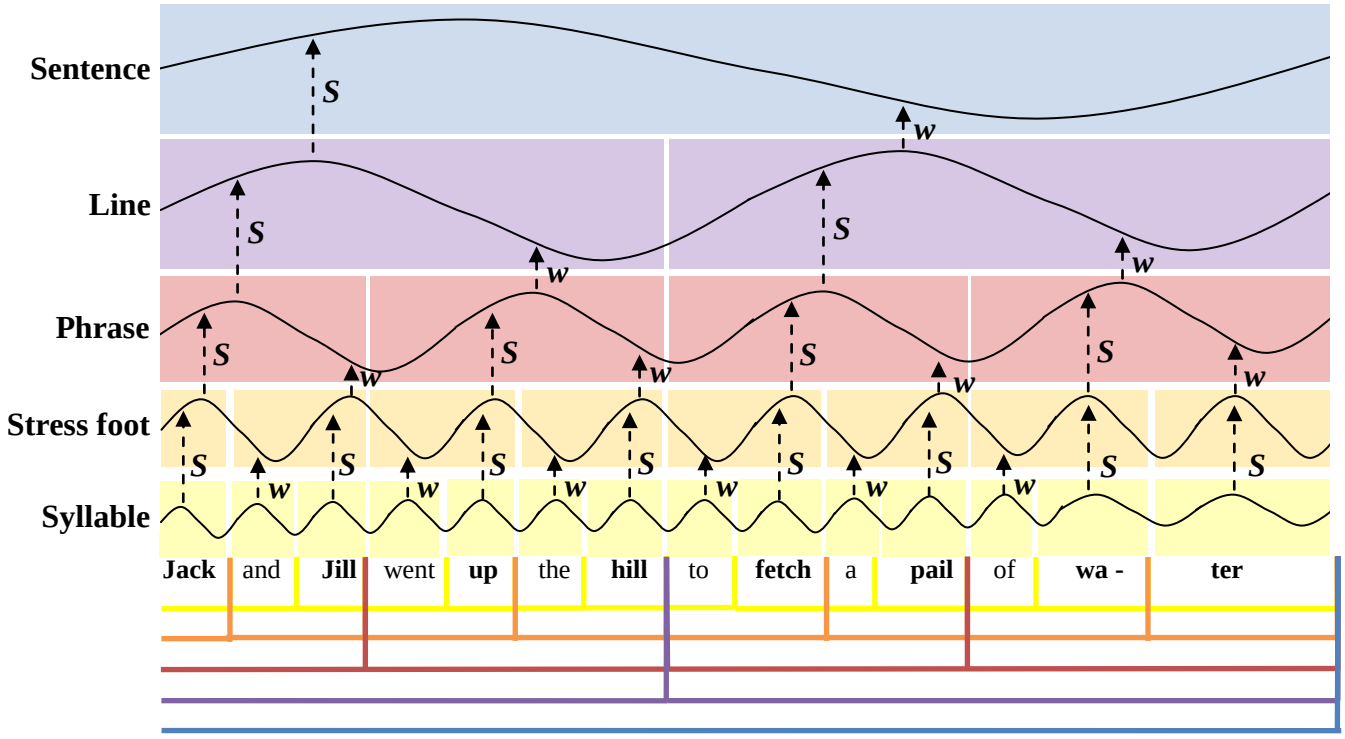


Figure 1.9 illustrates that the words in the nursery rhyme 'Jack and Jill' naturally produce a strong hierarchical prosodic and rhythmic structure. This rhythmic structure is further accentuated by the strategic positioning of rhyming word pairs. As indicated from Gueron's (1974) analysis, such regular hierarchical rhythmic patterning may be a ubiquitous feature of nursery rhymes in general. Therefore, the *rhythm* structure of nursery rhymes may help children to develop an awareness of prosodic structure, inasmuch as the *rhyming words* help children to develop phonological awareness of onset and rime units. If the rhythms in nursery rhymes encourage language development, then speaking to children in a rhythmic manner (irrespective of the spoken material itself) may also generally be adaptive for language learning. The idea that an exaggerated prosodic pattern is beneficial for language learning is not new. For example, adults spontaneously switch to a prosodically-exaggerated register when speaking to infants and children. This speaking style has variously been referred to as 'motherese', 'infant-directed speech' or 'child-directed speech' (Fernald, 1989; Fernald & Simon, 1984). In contrast to the way we speak to other adults, 'motherese' is higher pitched, contains smoother and wider pitch excursions, is slower, and contains more pauses and repetitions (Broen, 1972; Fernald & Simon, 1984; Fernald, 1989; Albin & Echols, 1996). This prosodic exaggeration may increase auditory salience for the infant, allowing auditory information to be processed and remembered more efficiently (Divenyi & Hirsh, 1978). However, the prosodic exaggeration in motherese could also change the *temporal structure* of speech, for example by increasing the hierarchical patterning or temporal regularity of amplitude modulation patterns. Such temporal structural enhancement could benefit speech processing by making word and phrase boundaries more prominent, thereby making fluent speech easier to segment. In Chapter 7, this research question is examined directly when the modulation structure of child-directed speech is compared with that of adult-directed speech. On the other hand, if children have a developmental deficit which makes them less sensitive to the prosodic exaggeration in motherese and in nursery rhymes, they would also receive less benefit from these natural 'language learning devices'. For children with developmental dyslexia, this is indeed the case.

1.12 PROSODIC SENSITIVITY IN DEVELOPMENTAL DYSLEXIA

Developmental dyslexia is a neurodevelopmental condition found across languages, for which the cognitive hallmark is impaired phonological processing (Snowling, 2000; Ziegler & Goswami, 2005). The auditory parameter most consistently found to be impaired has been perception of onsets (rise times) in the amplitude envelope (Goswami et al., 2002; Goswami, 2008; Goswami, 2010; Goswami, 2011; Hämäläinen et al., 2005; Hämäläinen et al., 2009). The onset rise time refers to the time taken for a sound to reach its peak amplitude after its initial onset. This is illustrated in Figure 1.10 where the tone on the left has a fast onset rise time (15ms), while the tone on the right has a much slower onset rise time (300ms). When played to listeners, the tone on the left is perceived as having a sharp, strong onset (like a trumpet note), whereas the tone on the right is perceived as having a more gentle and gradual onset (like a bowed violin note). In the amplitude envelope of speech, the most prominent onsets typically correspond to the onsets of syllables. This is illustrated in the Figure 1.11, where the individual syllable onset rise times are shown as red arrows, overlaid on the green amplitude envelope.

As discussed previously in Section 1.3.2, the perceptual 'moment of occurrence' or p-centre of a syllable is typically associated with the onset of its vowel nucleus (Morton et al, 1976; Allen, 1972; Scott, 1993; 1998). These vowel onsets are in turn acoustically-associated with amplitude modulation patterns in the speech envelope. Therefore, problems with detecting amplitude changes in the envelope (e.g. onset rise times) should also affect the detection of p-centres in speech. Poor p-centre detection in turn should affect the perception of rhythm and prosodic patterns in speech. This predicts that individuals with dyslexia (who are impaired on rise time perception) should also have difficulties with the perception of rhythm and stress in speech. Several previous studies have suggested that this is indeed the case for English-speaking children (Kitzen, 2001; Goswami et al, 2010). Furthermore, young Dutch children at risk of dyslexia have also been shown to have difficulties in producing imitations of non-words with irregular stress patterns (de Bree et al, 2006). Developmentally, this deficit may take the dyslexic child on a different, less optimal trajectory of language acquisition. In the first year, an infant with a reduced sensitivity to speech rhythm may be less accurate in using prosodic cues to segment words from the speech stream. Through early childhood, as the child is developing his or her phonological representations, rhythmic stress patterns in speech may be underspecified, leading to impoverished representations of speech

sounds. For typically developing children, nursery rhymes could be powerful phonological learning devices helping them to quickly acquire the normal rhythmic patterns of spoken English. By contrast, dyslexic children would be less accurate in perceiving the rhythms of nursery rhymes, slower to acquire these long-scale rhythmic templates, and less able to use them for speech processing.

Figure 1.10. Acoustic waveform of two pure tones with a fast (left) and slow (right) onset rise time. Rise times are indicated with arrows. Reproduced from Richardson et al (2004).

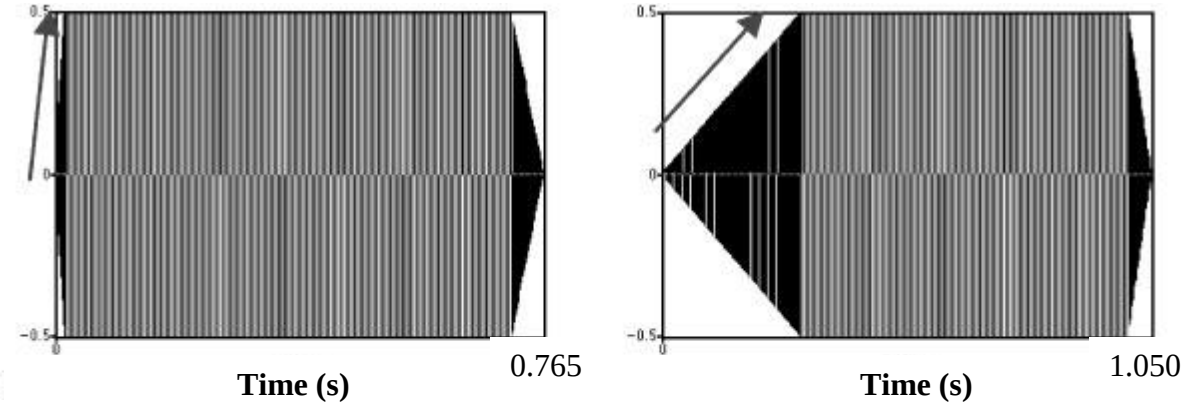
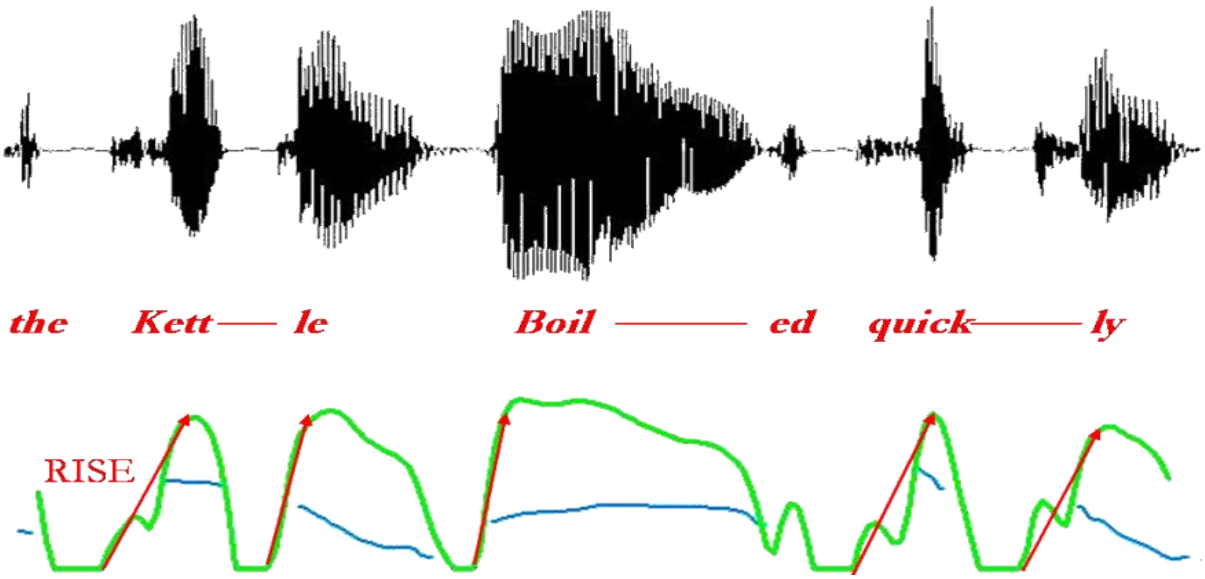


Figure 1.11. Example of onset rise times in the speech amplitude envelope and their relationship to syllable onsets. The acoustic waveform of the sentence is shown on top in black. The amplitude envelope is shown at the bottom in green. Large onset rises in the amplitude envelope are marked with red arrows, and correspond well to the onsets of syllables in the sentence.



In the initial phase of this PhD project (prior to the development of the speech rhythm models), a study was carried out with adult dyslexics to test their perception of prosodic stress patterns in multisyllabic words. The experimental stimuli comprised two sets of 4-syllable words that differed in their lexical stress pattern, imparting a different characteristic rhythm to each set of words. 20 words followed a 'S-w-w-w' lexical stress pattern, such as '**DI**-ffi-cul-ty', while the other 20 words followed a 'w-S-w-w' stress pattern, such as 'ma-**TER**-ni-ty'. Two tokens were then produced for each word. One token represented its correct stress pattern (eg. '**DI**-ffi-cul-ty', S-w-w-w), while the other token represented the opposite, incorrect stress pattern (eg. 'di-**FFI**-culty', w-S-w-w). Using these tokens, pairs of words were created where the stress patterning varied but the phonological content was kept constant (eg. '**DI**-ffi-cul-ty' vs 'di-**FFI**-cul-ty'), or where both stress patterning and phonological content varied (eg. '**DI**-ffi-cul-ty' vs 'ma-**TER**-ni-ty'). These two different types of word pairs were presented in two separate stress discrimination experiments.

In the first experiment, participants heard pairs of words that were phonologically identical, but had either the same (e.g.'**DI**-ffi-cul-ty' vs '**DI**-ffi-cul-ty') or a different (e.g. '**DI**-ffi-cul-ty' vs 'di-**FFI**-cul-ty') stress pattern. The two spoken word tokens were presented one after the other via headphones. Participants then indicated via a button press whether they thought the tokens had the same or a different stress pattern. In the second experiment, participants heard pairs of words that were phonologically different, but had either the same e.g. (eg. '**DI**-ffi-cul-ty' vs '**MI**-li-ta-ry') or a different e.g. (eg. '**DI**-ffi-cul-ty' vs 'ma-**TER**-ni-ty') stress pattern. Participants again indicated whether they had heard the same or a different stress pattern across the word pair. As predicted, dyslexics performed significantly more poorly than non-dyslexic participants in *both* experiments. Since dyslexics performed poorly even when the phonological information of the word pair was identical, this indicated that the source of the dyslexic deficit was in the perception of *acoustic* cues to prosodic stress, such as differences in amplitude, duration or frequency between stressed and unstressed syllables. Consistent with this interpretation, participant's performance on both tasks was strongly related to their auditory psychoacoustic threshold for onset rise time detection, though not for frequency or (static) intensity change detection. The published details of this study are included as Appendix 1.2 : Leong, V., Hamalainen, J., Soltesz, F., & Goswami, U. (2011). Rise time perception and detection of syllable stress in adults with dyslexia. *Journal of Memory and Language*, 64, 59-73. In Chapter 8 of this thesis, this research thread on dyslexia is continued with an AM-based investigation into the perception and production of rhythmic speech in adults with dyslexia.

1.13 THESIS OVERVIEW

In this thesis, two novel models of speech rhythm are proposed, based on amplitude modulation patterns in the speech envelope. Both models use an hierarchically-nested representation of amplitude modulation patterns in the speech envelope (the 'AM hierarchy'), in symmetry to hierarchically-nested neuronal oscillations. Like the newborn infant, both of these models are capable of detecting prosodic rhythm patterns solely from the acoustic information in the speech signal, without the need for any prior manual speech labelling or phonetic segmentation. This 'tabula rasa' approach allows for 'naive' speech segmentation schemes to emerge (e.g. via a metrical segmentation strategy) without recourse to lexical knowledge about words or phonemes.

In **Part II** of this thesis, the first Amplitude Modulation Phase Hierarchy (AMPH) model for speech rhythm is introduced (Chapter 2). In this original AMPH model, the AM hierarchy is derived theoretically, on the basis of previous findings and literature. The Stress Phase Code is also introduced. This is a computational scheme for identifying prosodic 'Strong-weak' stress patterns using the phase relationships between two key tiers in the AM hierarchy (Stress AM & Syllable AM). Finally, the core assumptions of the AMPH model are tested in a tone-vocoding experiment with human listeners (Chapter 3).

In **Part III** of this thesis, a new Spectral Amplitude Modulation Phase Hierarchy (S-AMPH) model is introduced. The S-AMPH model addresses several short-comings of the original AMPH model. First, a spectral sub-band representation of the speech envelope is used, rather than the wholeband speech envelope. This imparts the model with greater dexterity in identifying the rhythm-bearing syllable vowel patterns in speech. Second, a new 'emergent' AM hierarchy is derived, based on the modulation statistics of the acoustic signal. These two improvements collectively result in a new spectro-temporal representation of the speech envelope, which is described in Chapter 4. In line with this new spectro-temporal representation, new prosodic indices for computing 'Strong-weak' stress patterns are developed in Chapter 5. Finally, the original AMPH and new S-AMPH models are functionally evaluated in terms of their success in automatic syllable detection and prosodic stress transcription (Chapter 6).

In **Part IV** of this thesis, the S-AMPH model is used as an analytical tool to compare the underlying temporal structure of different types of speech. Two different experimental case studies are presented. In Chapter 7, the S-AMPH model is used to determine how the

spectro-temporal structure of child-directed speech (CDS) differs from that of adult-directed speech (ADS). In Chapter 8, the perception and production of rhythmic speech is investigated in adults with and without developmental dyslexia. The novel AM-based analysis method is used alongside more traditional linguistic analysis methods.

Finally, **Part V** provides an overall discussion and synthesis of the key themes arising from this thesis, and possible future directions for this line of research.

DEFINITIONS OF COMMON TERMINOLOGY USED

- In this thesis, the terms '**speech rhythm**', '**speech prosody**' and '**speech rhythmic structure**' all refer to the alternating pattern of 'Strong' (stressed) and 'weak' (unstressed) syllables in speech.

- The term '**prosodic foot**' or '**stress foot**' refers to the organisation of groups of Strong and weak syllables into motifs. Common foot motifs in English are the two-syllable trochee ('S-w') and iamb ('w-S'). However, longer feet with more syllables also occur, such as the three-syllable dactyl ('S-w-w') or amphibrach ('w-S-w').

- If a sentence consists of a *regularly-repeating* foot motif (e.g. nursery rhyme sentences), these sentences are said to possess a '**metrical**' rhythm.

- In this thesis, the '**meter**' of a sentence refers to the number of syllables in its repeating foot motif, and is analogous to the number of beats per bar in music. For example, a sentence that consists of repeating *two-syllable* trochees is described as having a '**duple**' meter. A sentence that consists of repeating *three-syllable* dactyls is described as having a '**triple**' meter.

- Where attention is drawn to the specific *sequence* of Strong and weak syllables within the foot (i.e. trochee or iamb) rather than simply the *number* of syllables, the term '**metrical pattern**' is used.

- The **amplitude envelope** is also alternately referred to as the **speech envelope.**

PART II :

THE AMPLITUDE MODULATION PHASE HIERARCHY MODEL (AMPH)

AIMS OF THE MODEL

Previous rhythm-metric approaches to describing speech rhythm have focused on *durational* changes between phonetic segments in speech. The limited success achieved by these methods (e.g. see Arvaniti, 2009) has motivated an on-going search for new and better ways to represent the rhythm information in speech (e.g. Todd, 1994; O'Dell and Nieminen, 1999; Tilsen & Johnson, 2008). In this spirit of improvement and discovery, a new model of speech rhythm is proposed here, adopting an *amplitude*-based approach to speech rhythm. The aims of the model are to make explicit (tease out) any cues to speech rhythm that are present in the amplitude envelope of speech, and to provide a computational scheme for how these amplitude cues relate to the rhythm patterns perceived by the listener.

In line with these aims, two major research questions are posed :
(1) ***Where*** *is speech rhythm information located in the modulation spectrum of the envelope?*
(2) ***How*** *is speech rhythm information 'coded' within the amplitude modulation patterns in the envelope?*

To answer the first question, the modulation spectrum of the envelope is sub-divided into 5 modulation rate bands. Rather than using an arithmetic division of the spectrum (i.e. using linear or logarithmic spacing between bands), the boundaries of the modulation rate bands are *theoretically*-determined with reference to the linguistic prosodic hierarchy. That is, each modulation band is designed to capture a different linguistic tier from the prosodic hierarchy, such as stress feet, syllables or phonemes. The resulting 5-tier (band) **AM hierarchy** is therefore a *concrete* representation of the *abstract* linguistic prosodic hierarchy. In linguistic terms, speech rhythm arises from the organisation of alternating Strong and weak *syllables* into prosodic *stress feet*. That is, several syllables at one tier of the hierarchy are grouped together to form a single rhythmic motif at a higher tier of the hierarchy - the prosodic Stress foot. By analogy, in the AM hierarchy, speech rhythm should arise from the nesting of several *Syllable* AM cycles within a higher level *Stress* AM cycle, collectively forming the rhythm pattern of a prosodic Stress foot. Therefore, in the model proposed here, speech rhythm information is primarily associated with Syllable AM and Stress AM tiers of the AM hierarchy.

If nested Stress AM and Syllable AM patterns form a prosodic stress foot, how is the Strong-weak patterning of syllables within the foot represented? In other words, how does one convert the *continuously*-varying AM patterns into a *discrete* representation of Strong and weak syllables, as perceived by the listener? To answer this second research question, a Stress Phase Code is proposed. This is a computational scheme to convert AM patterns into Strong-weak syllable patterns. In this Stress Phase Code, the key statistic used is the local instantaneous *phase* relationship between the Stress and Syllable AM tiers (hence '*phase* code'). This Stress-Syllable phase relationship determines the perceived prominence of individual syllables, and by extension, the rhythm pattern of a sentence. Therefore, since this model uses the *phase* information between tiers of an *AM hierarchy* to provide a description of speech rhythm, it is called the **AM Phase Hierarchy** model, or **AMPH** model in short.

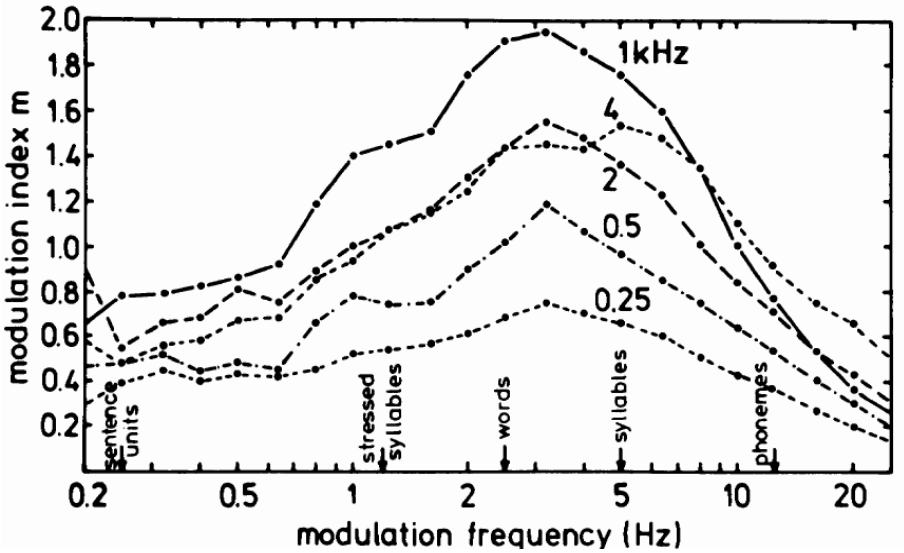
Chapter 2 explains the derivation of the AMPH model as a signal-based method for computing prosodic rhythm from the amplitude modulation structure of speech. Chapter 3 tests the psychological validity of the assumptions of the AMPH model in a tone-vocoding experiment. The aim of this experiment is to see if human listeners perceive prosodic rhythm using the same amplitude modulation cues used by the AMPH model to compute rhythm.

2 THE AMPLITUDE MODULATION PHASE HIERARCHY (AMPH) MODEL OF METRICAL SPEECH RHYTHM

2.1 THE MODULATION SPECTRUM AND THE PROSODIC HIERARCHY

Recall from the Introduction (Section 1.6) that the modulation spectrum is the power profile of the various amplitude modulation rates present in the speech envelope. Although different researchers use slightly different upper rate limits to define the speech envelope, Rosen (1992) considers amplitude modulations rates of up to 50 Hz to be part of the speech envelope. Modulations of up to 50 Hz are thought to contain prosodic cues, as well as segmental cues to manner of articulation, voicing and vowel identity. Accordingly, this relatively high upper rate limit for the envelope is used in the AMPH model (although the focus of the model is on the slower modulation rates).

Figure 2.1. Reproduced from Plomp (1983b). Modulation spectrum of running speech for 5 octave-spaced speech frequency bands (centre frequencies at 0.25 kHz, 0.5 kHz, 1 kHz, 2 kHz and 4 kHz as marked). The y-axis shows the modulation index, 'm', which is the ratio between the (peak) amplitude of the envelope and the amplitude of the un-modulated carrier. The x-axis shows the range of modulation frequencies in the envelope from 0.1-40 Hz. Labels marked by the original author (e.g. 'sentence units' or 'stressed syllables') indicate intuitions regarding different linguistic units and their key associated modulation frequencies (arrows).



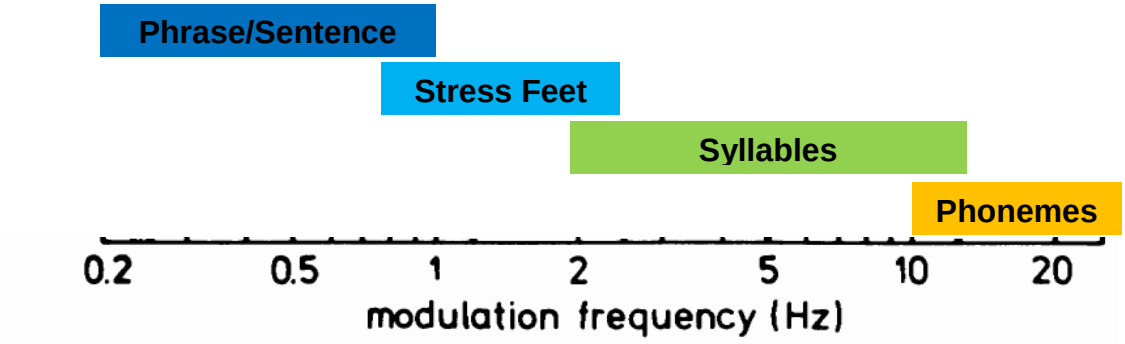
When considering the modulation spectrum of the speech envelope, it is commonly thought that speech units of different lengths give rise to amplitude modulations at different rates. For example, Figure 2.1 shows a plot of the modulation spectrum reproduced from Plomp (1983b), in which the correspondence between different speech units (words, syllables, etc) and different modulation rates has been annotated. In Figure 2.1, Plomp (1983b) associates the main peak in the modulation spectrum (~3-5 Hz across all frequency bands) with words and syllables. Slower modulations to the left of this peak (~1 Hz) are associated with longer stressed syllables, whereas faster modulations to the right of this peak (>12 Hz) are associated with shorter phonemes.

Later studies by Greenberg and colleagues (e.g. Greenberg et al, 2003; Greenberg, 2006; Ghitza & Greenberg, 2009) have supported Plomp's approximate division of the modulation spectrum (for the most part). For example, Greenberg et al (2003) confirmed that the long-term modulation spectrum of speech does indeed consistently peak around 3-5 Hz. Moreover, since the typical measured duration of syllables in speech is ~200 ms (i.e. ~5 Hz), this modulation peak at ~3-5 Hz is indeed likely to correspond to syllable patterns in speech. Like Plomp, Greenberg (Greenberg et al, 2003; Ghitza & Greenberg, 2009) also proposes that modulations slower than the 3-5 Hz peak correspond to stressed syllables, which are longer in duration than unstressed syllables, and therefore are associated with a slower modulation rate.

However, in Greenberg et al's (2003) data, virtually no syllables (even stressed syllables) had a longer duration than ~500 ms (2 Hz). This suggests that modulation rates under 2 Hz must correspond to larger linguistic units, such as multi-syllable words or stress feet, rather than to individual stressed syllables (as annotated by Plomp in Figure 2.1). For example, the trochaic stress foot describes a 'Strong-weak' bi-syllable pattern, and is found in words like *"DOC-tor"* and *"MU-mmy"*. If each syllable is taken to have a length of 200 ms, the total length of the trochee foot would be 400 ms, which corresponds to a modulation rate of 2.5 Hz. Therefore, slow modulations around 2.5 Hz or under should reflect the modulation patterns of prosodic stress feet. Consistent with this view, Dauer (1983) found that the average duration of inter-stress intervals in English was 493 ms, corresponding to a stress rate of around 2 Hz. Therefore, Plomp's annotated division of the modulation spectrum should be revised to incorporate prosodic stress feet, as illustrated in Figure 2.2. Notice that the linguistic units in Figure 2.2 are also components of the *linguistic prosodic hierarchy* (e.g. Selkirk, 1980, 1984, 1986). Therefore, as argued in the introduction to Part II, the

modulation spectrum may be divided (on the basis of theoretical assumptions) into modulation bands or tiers that capture different tiers of the linguistic prosodic hierarchy. By analogy to the linguistic hierarchy, these modulation bands or tiers would also form an *AM hierarchy* of modulation patterns, at prosodically-important timescales.

Figure 2.2. Illustration of the proposed correspondence between linguistic units in the prosodic hierarchy and associated modulation rates in the modulation spectrum, forming an AM hierarchy[3]*.The x-axis and labels are replicated from Figure 2.1, with the inclusion of prosodic stress feet as a new unit. Rather than associating linguistic units with certain key modulation rates (as in Figure 2.1), they are associated with partially overlapping modulation rate bands.*

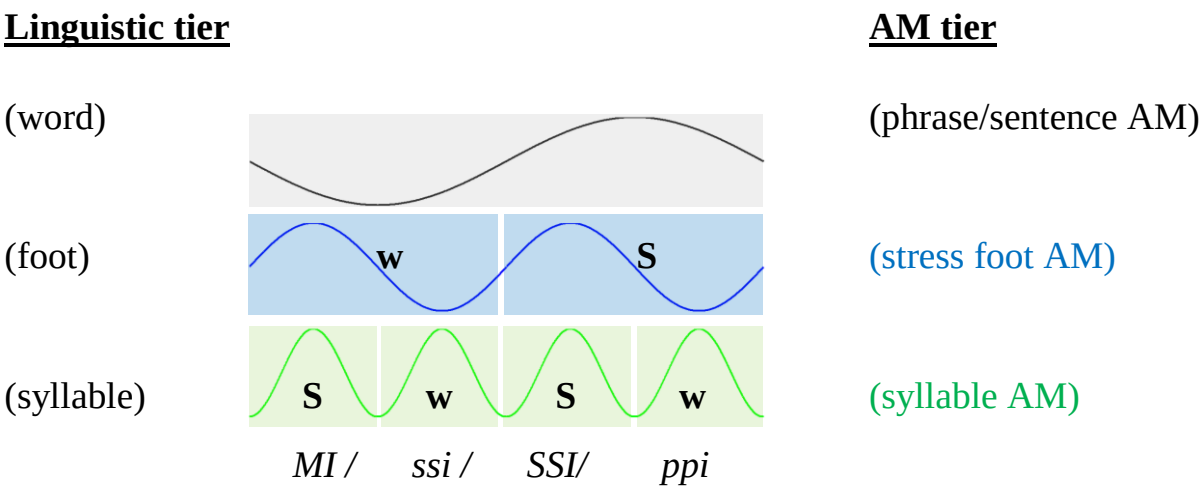


Recall from Section 1.2 of the Introduction that the linguistic prosodic hierarchy is a way to represent the abstract prosodic structure of speech. This prosodic structure is commonly visualised in grid or tree form, where prosodic patterns emerge from the hierarchical nesting of elements in lower tiers within higher tiers (Selkirk, 1980, 1984, 1986; Liberman & Prince, 1977; Hayes, 1995). By extension, the *AM hierarchy* should also be able to capture *acoustic* prosodic patterns in speech. These prosodic patterns could be formed from the hierarchical nesting of faster AM patterns (e.g. Syllable tier) within slower AM patterns (e.g. Stress tier). Accordingly, the AMPH model proposed here represents speech rhythm patterns as Syllable tier AM cycles that are hierarchically-nested within Stress tier AM cycles. Crucially, each Stress tier cycle is equivalent to a prosodic stress foot, and the Syllable AM cycles nested *within* this Stress cycle represent syllables within the prosodic

[3] The AM hierarchy illustrated in Figure 2.2 consists of 4 different modulation bands, or tiers. However, in the AMPH model there are 5 tiers in the AM hierarchy. This is because the AMPH model is based on strongly rhythmic nursery rhyme speech. Therefore the 'Syllable' band is further sub-divided into 2 separate tiers. These 2 tiers correspond to syllables that occupy a single whole rhythmic beat, and syllables that only occupy part of a rhythmic beat ('sub-beat'). The derivation of the 5 AM tiers in the AMPH model is further described in Section 2.4.

stress foot. This functional equivalence between linguistic units (stress foot and syllables) and their respective AM tier cycles is shown in Figure 2.3. In the linguistic prosodic hierarchy, the relative prosodic strength of a unit (e.g. syllable) is denoted with different letters, either 'S' for Strong or 'w' for weak (see Figure 2.3). In the AM hierarchy, the prosodic strength of each AM cycle (representing a single linguistic unit) is determined by its phase relationship to the next highest tier (e.g. the stress tier). This 'phase coding' of prosodic strength is described further in Section 2.5.2.

Figure 2.3. Hypothetical example of hierarchical nesting within the linguistic prosodic hierarchy, and equivalent nesting of AM cycles in the AM hierarchy. In both hierarchies, each tier represents a different prosodic level. Units at a lower level are nested within units at a higher level. This nesting is shown using coloured blocks of different size. In the linguistic hierarchy, each single linguistic unit (e.g. syllable or foot) is denoted with a single letter. This letter corresponds to the unit's prosodic strength, as either 'S' (strong) or 'w' (weak). In the AM hierarchy, each single linguistic unit corresponds to a single AM cycle. For example, in the bottom syllable AM tier, the four syllables are represented as four AM cycles, each boxed in green.

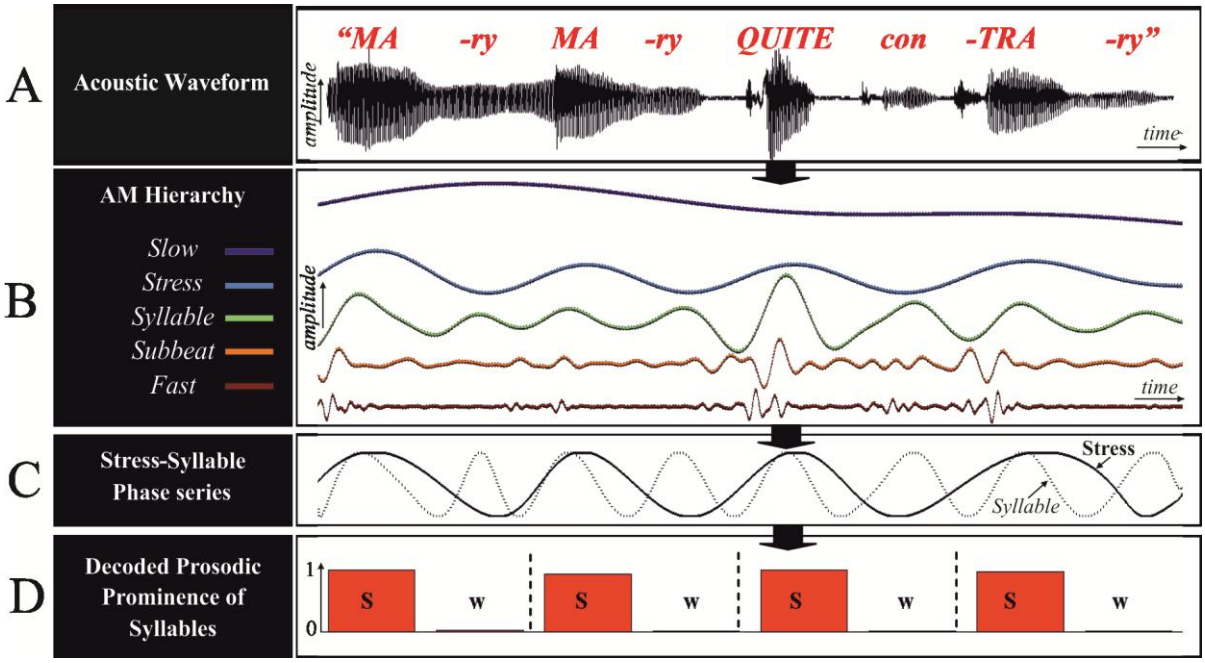


Therefore the AM hierarchy (made up of modulations from the speech envelope), represents prosodic stress patterns as hierarchically-nested AM cycles. This means that if the prosodic stress pattern of an utterance is *unknown*, its stress pattern can be inferred from the acoustic signal by looking at nesting patterns within the AM hierarchy of the speech envelope. The AMPH model is a systematic way to extract these rhythm patterns from the speech envelope. The key processes in the AMPH model are outlined next.

2.2 Overview of the AMPH Model

The essential premise of the AM Phase Hierarchy (AMPH) model is that prosodic rhythm patterns in speech may be inferred from the phase relationships between hierarchically-nested slow amplitude modulations (AMs) in the acoustic signal. When speech is temporally-regular (e.g. metronome-timed speech), the various AMs hierarchical tiers are stably phase-locked, providing consistent phase information that can be used to infer rhythmic meter and prosodic pattern. Therefore, metronome-timed nursery rhyme speech is used as the basis for the AMPH model[4].

Figure 2.4. Schematic overview of the processing stages in the AM Phase Hierarchy Model using an example of the 8-syllable trochaic nursery rhyme sentence "MA-ry MA-ry QUITE con-TRA-ry", where stressed syllables are indicated in capital letters. (A) Original sound pressure waveform of the speech signal, showing amplitude changes over time. (B) Extracted AM hierarchy consisting of 5 AM tiers, each at a different modulation rate. Each AM tier is shown using a different colour. (C) Oscillatory phase of the Stress AM (solid line) and Syllable AM (dotted line), projected onto a cosine function to show the oscillatory shape. (D) Decoded prominence value for each syllable using the Stress Phase Code. Strong (stressed) syllables with prominence values >0.5 are indicated as 'S', weak (unstressed) syllables with prominence values <0.5 are indicated as 'w'.



[4] In Part III of the thesis, the same 'phase coding' principles are applied in a revised version of the AMPH model. This new model is then tested on freely-produced (non-metronome timed) speech.

Figure 2.4 shows a summary flowchart of the key processing stages in the AMPH model. First, an hierarchy of AMs is extracted from the wideband amplitude envelope, selecting theoretically-driven temporal rates ranging from slow (<1 Hz) to phonetic (here 20 – 50 Hz). Each tier of the AM hierarchy relates to linguistic units of different length, such as syllables or prosodic feet. Next, the angular oscillatory phase is computed for 'Stress' and 'Syllable' tiers of the hierarchy using the Hilbert transform. By taking only the angular phase of the AMs, transient fluctuations in power are discarded, treating each AM as if it were a pure sinusoid. For illustration, Figure 2.4 projects the Stress and Syllable phase series onto cosine functions in order to show their equivalent oscillatory shape (third panel). However, the phase series themselves (used for computation) vary between $-\pi$ and π radians in an approximately linear fashion.

Two types of rhythmic information are extracted from the Stress-Syllable phase series. First, rhythmic meter (e.g. duple or triple meter) is inferred from the long-term ratio of angular phase-locking between Stress and Syllable-phase series (described further in Section 2.5.1). Second, the prosodic prominence pattern of strong and weak syllables is assessed based on local (momentary) phase relationships within the Stress-Syllable phase series using a 'Stress Phase Code' (Section 2.5.2). This Phase Code assigns a numerical prominence value (0 to 1) to each syllable, as shown in the fourth panel of Figure 2.4. Finally, meter and stress pattern information are combined in a rhythm-based segmentation scheme for the sentence (Section 2.6).

2.3 MATERIALS

A set of 6 familiar English nursery rhyme sentences with contrasting metrical rhythm patterns formed the basis for deriving the AM Phase Hierarchy model. These are listed in Table 2.1, taking the first line of the nursery rhyme in each case. The nursery rhymes were spoken by a female native speaker of British English who was articulating in time to a 4 Hz (syllable rate) metronome beat. The speaker was instructed to produce the metrical pattern of each nursery rhyme as clearly as possible. Utterances were digitally recorded using a TASCAM digital recorder (44.1 kHz, 24-bit), and the metronome was not audible in the final recording.

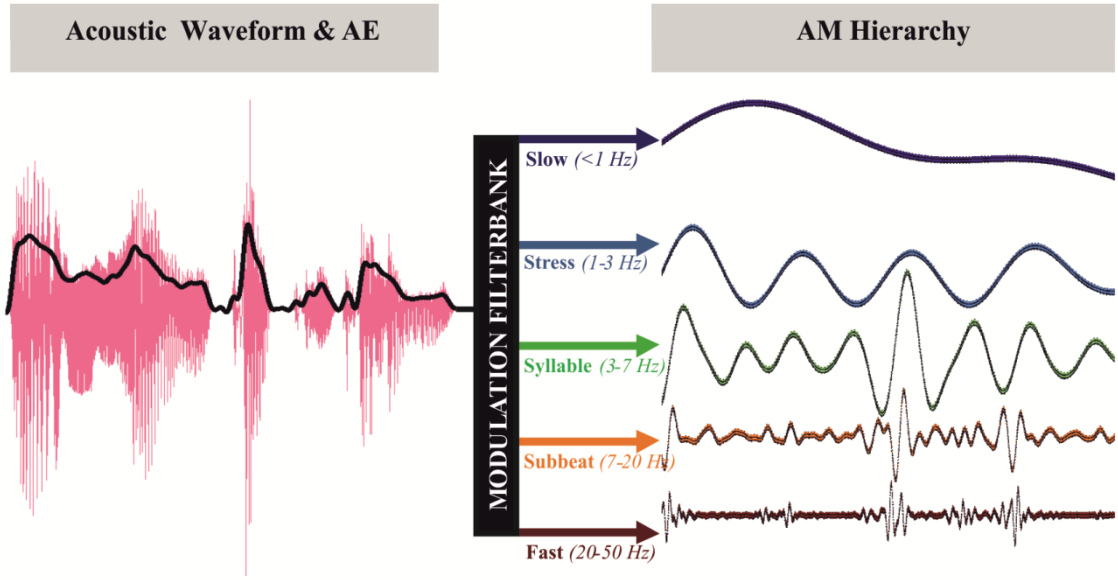
Table 2.1. List of nursery rhyme sentences and their metrical rhythm pattern

METRICAL RHYTHM PATTERN *(**S** = Strong, w = weak)*		**NURSERY RHYME SENTENCE** *(CAPS = Strong syllable)*
Duple meter	**S** w **S** w **S** w **S** w *(trochaic)*	"MA-ry MA-ry QUITE con-TRA-ry"
		"SIM-ple SI-mon MET a PIE-man"
	w **S** w **S** w **S** w **S** *(iambic)*	"as I was GO-ing TO st IVES"
		"the QUEEN of HEARTS she MADE some TARTS"
Triple meter	**S** w w **S** w w **S** w w **S** *(dactyl)*	"PU-ssy-cat PU-ssy-cat WHERE have you BEEN"
	w **S** w w **S** w w **S** w w **S** *(amphibrach)*	"to MAR-ket to MAR-ket to BUY a fat PIG"

2.4 EXTRACTING THE AM HIERARCHY

First, the amplitude envelope was extracted from the wideband speech signal via a demodulation procedure - the Hilbert transform. The wideband envelope of speech contains a dominant contribution from low frequency components such as voiced sounds, as these carry the most acoustic energy. This low-frequency bias is similar to the experience of the foetus in-utero (Armitage et al., 1980). Hence, the AMPH uses the most energetically-dominant spectral components for rhythm detection, consistent with the acoustic environment of a fetus in-utero. Since different rates of modulation in the amplitude envelope pertain to different types of linguistic information, the Hilbert envelope was passed through a series of band-pass filters in order to isolate these different modulation rates. This 'modulation filterbank' approach is illustrated in Figure 2.5. The 'modulation filterbank' (MFB) consisted of a series of adjacent finite impulse response (FIR) filters. Appendix 2.1 provides details of the filtering parameters used in the AMPH. For a detailed description of the filterbank design and features, see Stone & Moore (2003, p.3). The current MFB was adapted from this spectral filterbank for use as a modulation filterbank.

Figure 2.5. Extraction of the AM hierarchy from the acoustic signal. This involves passing the wholeband amplitude envelope (AE, black line) through a modulation filterbank (MFB). Filter frequencies shown here are for illustration only, these were tuned to the individual speaking rate.



Inspection of Figure 2.5 reveals that 5 AM tiers were generated by the modulation filterbank: (1) **Fast**; (2) **Sub-beat**; (3) **Syllable**; (4) **Stres**s; (5) **Slow**. Rather than being fixed, the filter parameters of each tier in the filterbank were adjusted according to the speaking rate for each speech sample. This was accomplished by identifying the speaker's dominant '**Syllable**' rate of speaking, and centering the modulation filterbank around this '**Syllable**' rate. The Syllable rate for the sample was identified by taking the highest peak in the modulation spectrum for that sample, within the 3-7 Hz range.

The '**Stress**' filter captured modulations that occurred at half to one-third the rate of the 'Syllable' rate. In normal conversation, longer prosodic feet than the bi-syllable trochee and iamb occur. Hence, if the average normal syllable rate is 5 Hz, the average stress rate would be less than half this rate (<2.5 Hz). This is consistent with the average duration of inter-stress intervals in English of 493 ms (2.03 Hz), as noted by Dauer (1983). The centering procedure ensured that the most behaviourally-relevant modulations were isolated by the filterbank. For example, if the speaker's syllable rate dropped to an abnormally slow rate of 2 Hz, modulations at this rate would be correctly identified as 'Syllable' rather than 'Stress', and the parameters of the 'Stress' filter would be lowered accordingly.

The '**Sub-beat**' filter captured modulations that occurred between 2 to 3 times faster than the Syllable rate, corresponding to occasions when more than one syllable was uttered per beat[5]. For example, in the English nursery rhyme 'Humpty Dumpty', the first 3 syllables of the prosodic phrase "*sat on the wall*" are typically uttered more quickly to fit the beat of one regular syllable (i.e. their relative timing is the same as for the single syllable "*wall*"). Since these syllables are 1/3 the duration of "*wall*", modulations produced by their utterance would be three times faster than the 'Syllable' rate and would appear in the Sub-beat tier.

Finally, '**Slow**' and '**Fast**' filters captured remaining modulations that fell below the 'Stress' rate and above the 'Sub-beat' rate respectively. Although not explicitly used by the AMPH model for rhythm detection, these 'Fast' and 'Slow' modulation rates nevertheless contained important information for speech intelligibility. For example, fast modulations in the speech envelope up to 50 Hz are thought to contain linguistic cues to phonetic manner of articulation, voicing, and vowel identity (Rosen, 1992). At the slowest end of the modulation spectrum, Fullgrabe et al (2009) demonstrated that even modulations as slow as <1 Hz could

[5] Therefore both the 'Syllable' and 'Sub-beat' tiers capture syllable sounds. However, Syllable cycles correspond to whole beats whereas Sub-beat cycles correspond to divisions of the main beat. This splitting of syllable-rate modulations into two separate tiers explains why there are 5, rather than 4 tiers in the AM hierarchy, as originally depicted in Figure 2.2.

contribute to speech intelligibility in adverse listening conditions, possibly by conveying information about sentence phrasing.

These 5 tiers formed the AM hierarchy which was meant to represent the linguistic prosodic hierarchy. Only 'Stress' and 'Syllable' AM tiers were given names that corresponded to specific linguistic counterparts (the stress foot and the syllable), because these correspondences were strongly indicated by the previous literature. Also, since the syllable rate was explicitly computed for each sample, there was reasonable confidence that the Syllable tier would contain actual syllable-related modulations. Similarly, there was good reason to expect that modulation patterns from prosodic stress feet would fall into the Stress tier, since these modulations would be at integer dividends of the Syllable rate. By contrast, the other AM tiers were given more generic rate-related names (e.g. 'Fast' or 'Slow') because there was no guarantee that they would actually contain, for example, phoneme-related or phrase/sentence-related modulations. Since only Stress and Syllable tiers were used for rhythm computation in the AMPH model, the linguistic content in these other AM tiers was not explored further in this thesis.

2.5 EXTRACTING RHYTHM INFORMATION FROM PHASE RELATIONSHIPS

Three features emerge from the AM hierarchy that would appear to be useful for rhythm perception. First, recall that each cycle of modulation in the Stress and Syllable tiers can be associated with a distinct linguistic unit. For example, one modulation cycle in the 'Syllable' AM tier in Figure 2.3 typically corresponds to a single articulated syllable. Similarly, one modulation cycle in the 'Stress' AM tier should correspond to a prosodic foot. By analogy to music, syllables in a metrical foot are like musical beats in a bar. AM cycles may provide the perceptual cues that induce a percept of rhythmic beats and bars.

Secondly, as illustrated in Figure 2.3 earlier, the AM tiers form nested sets, or a 'nested hierarchy'. That is, one modulation cycle at a higher level of the hierarchy encompasses a set of several cycles at the next lower level of the hierarchy. This type of nesting is also a feature of recent neural oscillatory models of speech perception (e.g. Ghitza, 2011). Conceptually, nesting is exemplified by Russian Matryoshka dolls, where each inner doll is enclosed by a slightly larger doll on the outside, which is itself enclosed by an even larger doll on its outside, all the way to the outermost doll, forming a nested physical hierarchy of layers. By analogy, AM tiers form a nested temporal hierarchy. Slower cycles that are temporally higher up in the AM hierarchy span sets of faster AM cycles at a lower level of the hierarchy. Of course, the analogy is not perfect, since in Matryoshka dolls the set at each level or layer only contains one object (doll). In contrast, in the AM hierarchy, each set (= slower AM cycle) may contain multiple faster modulation cycles. For example, one 'Stress' cycle may span two, three or even four 'Syllable' cycles. In metrical poems where 'Stress' and 'Syllable' AMs are temporally phase-locked (i.e. their phase relationship is fixed by poetic meter), Syllable modulation cycles form stable ordered sets within each Stress cycle. These ordered modulation sets may underlie human perception of a regular poetic meter.

Finally, phase relationships within the AM hierarchy provide 'new' information for encoding temporal patterns. For example, if the Syllable AM was extracted from the hierarchy and used to modulate a sine tone carrier, and this modulation pattern was presented to a listener either immediately or after a delay of 100ms, the relative timing of the pattern of beats heard by the listener would be exactly the same on both occasions. However, if the Syllable AM was instead presented together with the Stress AM, and the Syllable AM was

now played either simultaneously or with a 100ms delay with respect to the Stress AM, the listener would notice a difference when the onset of the Syllable AM was delayed. Even though the component Stress and Syllable AM tiers would be identical in each case, the change in phase alignment between the two AM tiers would yield new temporal information, in the form of a different interference or summation pattern. Consequently, when syllable patterns are presented as part of a temporal hierarchy rather than in isolation, the auditory system is provided with additional information about the phase relationships *between* tiers in the hierarchy, over and above the information contained within each tier. This new phase information may be exploited to provide cues about rhythmic patterning, temporal order or sequence. For example, common metrical foot motifs differ in their cyclical ordering of strong and weak syllables. The amphibrach motif ('w-S-w') is obtained by shifting the initial strong syllable in the dactyl ('S-w-w') down by one position. Accordingly, cyclical phase relationships between Stress and Syllable tiers of the AM hierarchy could provide information about underlying metrical foot patterns.

In the following two sections, these proposed relationships between features of the AM hierarchy and metrical rhythm information are instantiated as computational schemes. In Section (2.5.1), a description is provided of how rhythmic **meter** may be inferred from the *n:m* mode of phase-locking between Stress and Syllable phase series. In Section 2.5.2, a 'Stress Phase Code' for computing **syllable prominence**, and by extension, **metrical rhythm pattern** is described.

2.5.1 Inferring Rhythmic Meter from n:m Phase-Locking Ratio

If an oscillatory wave (such as a filtered AM) is thought of as alternately cycling between peak and trough states, its state at any given instant may be described in terms of the phase of a sinusoidal cycle, varying from $-\pi$ to π radians. In this thesis, an angular phase value of 0π radians marks the peak of the wave, while a phase of $-\pi$ radians or π radians marks a trough (i.e. 'cosine' phase). The phase values of $-\pi$ and π radians are equivalent because, by analogy to a circle, the starting and ending points after traversing a full cycle are the same. If one considers a wave that is oscillating in time, the phase of the wave at any point in time is a simple function of its starting phase, its frequency, and the time elapsed. If one now considers two such oscillating waves with the same frequency, the phase difference between the two waves at any point in time is constant, given by the difference in their initial

starting phases (since frequency and time elapsed are the same for both waves). For example, if they started at the same phase, they will always have the same phase at every future point in time - they are in phase. If one started with a peak (0 radians) while the other started with a trough (-π radians), they will always be π radians out of phase at every future point in time.

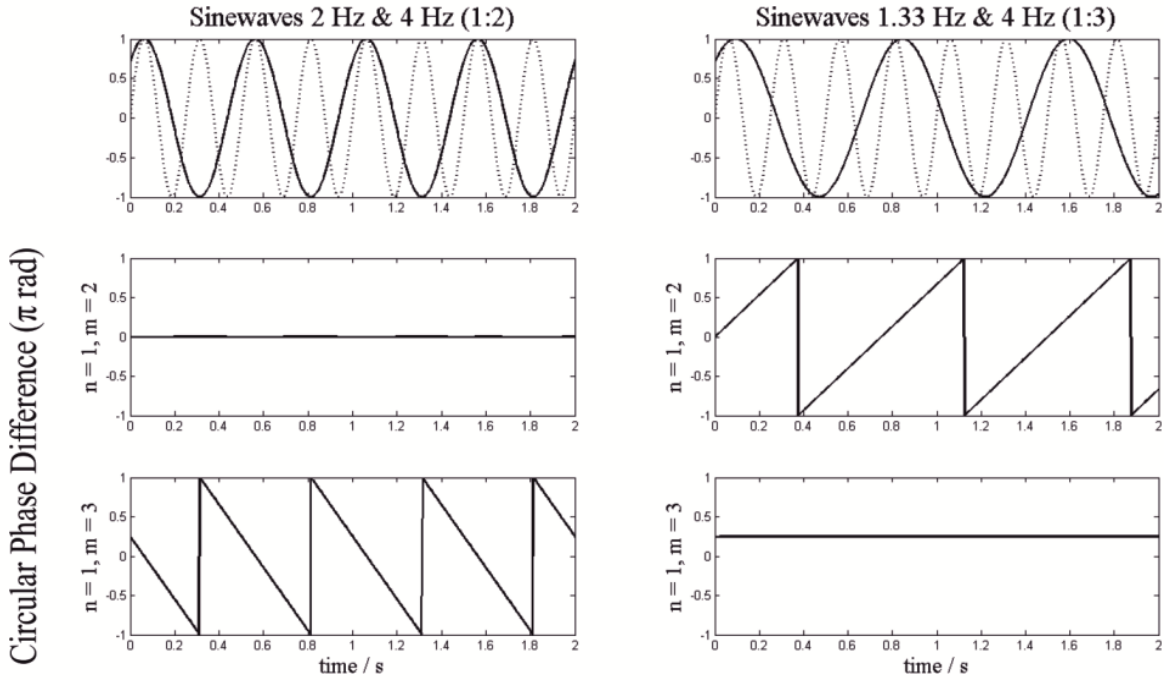
However, the situation becomes more complex when one considers the phase difference, or phase relationship between two waves that do *not* have the same frequency, as is the case for the Stress AM and Syllable AM. Although one can conceptualize phase-coupling or phase-locking between the two waves, this would have to be described in terms of angular ratios. In this case, one could no longer describe the two waves as being either in- or out-of phase. Rather, their instantaneous phase relationship would itself change over time, but in a predictable fashion. Mathematically, the phase relationship between two coupled oscillators at different frequencies can be expressed in a general formula for *n:m* phase-locking (Tass et al, 1998; Rosemblum et al, 2001). If Φ1(t) and Φ2(t) represent the phase series of each oscillator respectively, for a unique value of *n* and *m* (where both are integers) the circular phase difference between the two series is the sum of a constant value a (where $-\pi < a < \pi$) and a noise component, δ. This is summarised in equation (1) :

$$n\,\Phi1(t) - m\,\Phi2\,(t) = a + \delta \qquad \text{(Eq. 1)}$$

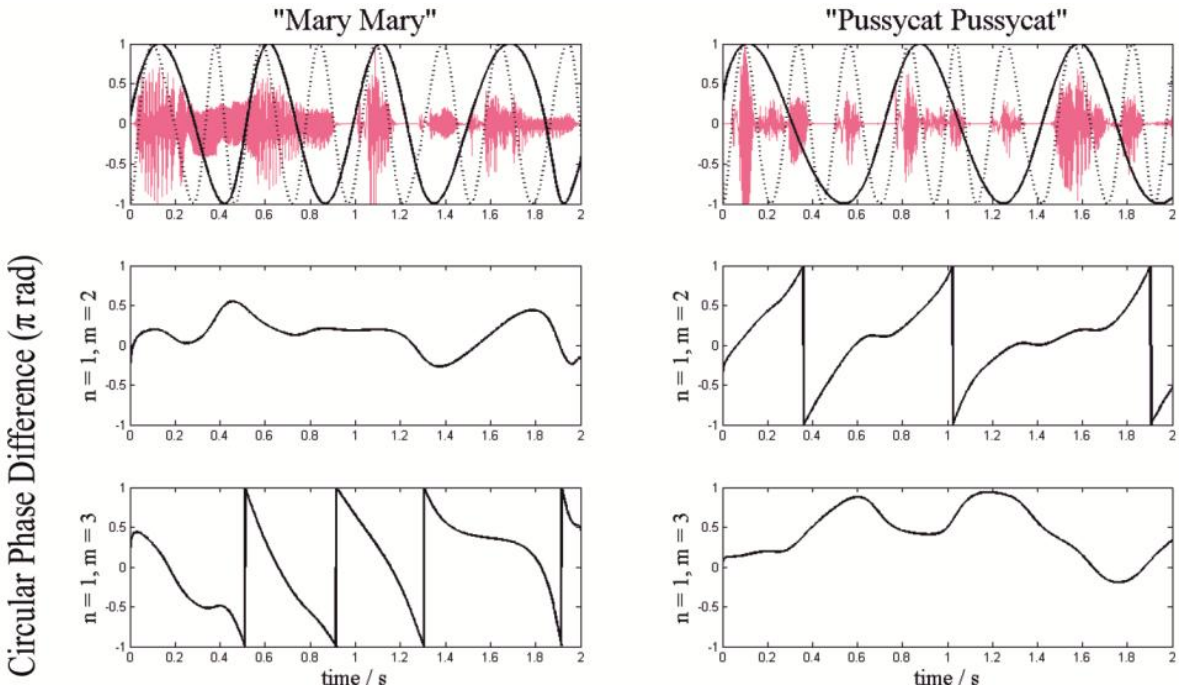
For example, Figure 2.6a shows two pure sine waves, where the slower wave is either half (left-hand column) or a third (right-hand column) the frequency of the other. In the middle and bottom row of the figure, the circular phase difference between the two waves is displayed where '*n* = 1, *m* = 2' (middle row) or '*n* = 1, *m* = 3' (bottom row). When the value of '*m*' correctly reflects the frequency ratio between the two waves, their circular phase difference is approximately constant. When '*m*' does not reflect the frequency ratio between the waves, their circular phase difference is not constant, and itself performs a periodic rotation. Hence, the constant-value *n:m* phase-locking solution reflects the frequency relationship between the two waves. Since this ratio may also be interpreted as the number of cycles at the faster rate that are required to form perfectly nested sets within the slower rate, the *n:m* ratio also denotes the rhythmic **meter** of the sequence.

Figure 2.6. Examples of n:m phase locking with (a) pure sine waves and (b) metrical nursery rhyme sentences.

(a) Pure sine waves with a frequency ratio of 1:2 (top, left) or 1:3 (top, right). The circular phase difference between the two waves is plotted when m = 2 (middle) and m = 3 (bottom).



(b) Duple and triple-metered nursery rhymes "Mary Mary" (left, duple) and "Pussycat Pussycat" (right, triple). In the top row, the waveform for each sentence is shown with Stress AM phase (bold) and Syllable AM phase (dotted) overlaid (phase plotted as cosine function). The circular phase difference between Stress and Syllable AMs is shown when m = 2 (middle) and m = 3 (bottom).



Having illustrated the principle of *n:m* phase-locking with periodic sine waves, this principle is now applied to the Stress-Syllable AM phase hierarchy to see if speech produced metrically does indeed contain such phase-locked 'nested sets'. In metrical poetry such as nursery rhymes, the pattern of stresses and syllables is highly regular. Consequently, it is expected that the 'Stress' and 'Syllable' AM phase series extracted from the spoken nursery rhymes should be phase-locked in the same way as the sine waves in the previous example. For example, in the children's nursery rhyme *"MA-ry MA-ry QUITE con-TRA-ry"* (shown in Figure 2.6b, left panels), every alternate syllable is stressed regularly. Since there are four stresses across eight syllables, the 'Stress' AM (bold line) is half the rate of the 'Syllable' AM (dotted line). These stresses occur at regular intervals with respect to the syllables, hence 'Stress' and 'Syllable' AMs maintain a predictable phase relationship throughout the sequence. Accordingly, their phase difference should also form a perfectly predictable pattern.

One would therefore expect the Stress-Syllable phase series to display *n:m* phase-locking of an order that reflects the rhythmic meter of the poem (i.e. 1:2 or duple). In contrast, for "PU-ssy-cat PU-ssy-cat WHERE have you BEEN" (Figure 2.6b, right panels), every third syllable is stressed. One would therefore expect the *n:m* phase-locking ratio for this sentence to be 1:3, reflecting a triple meter. The Stress-Syllable phase series for the two sentences are shown for *n:m* modes of 1:2 (middle panels) and 1:3 (bottom panels). For the sentence "Mary Mary" (left), Figure 2.6b shows that when *n:m* is 1:2 (left middle panel), the phase difference between Stress and Syllable AMs varies about a constant value. When *n:m* is 1:3, the phase difference becomes a periodic rotation. In contrast, the sentence "Pussycat Pussycat" (right middle panel) shows the opposite pattern. Now an *n:m* of 1:2 is periodic but 1:3 approximates a steady value (with noise). As such, one may conclude that the appropriate *n:m* phase-locking mode for "Mary Mary" is 1:2 but for "Pussycat Pussycat" it is 1:3. In fact, this phase-locking mode correctly reflects the actual rhythmic meter of each sentence. Hence, the *n:m* mode of AM phase-locking could potentially be used to compute rhythmic meter in hierarchically-phase-locked systems such as metrical poems and classical music.

2.5.2 COMPUTING SYLLABLE PROMINENCE USING THE STRESS PHASE CODE

So far, it has been demonstrated that poetic meter can be inferred from integer ratio descriptors of the long-term phase-locked relationship between Stress AMs and Syllable AMs within the AM hierarchy. Now, it is proposed that local (momentary) phase relationships between Stress and Syllable AMs constitute a 'Phase Code' for *relative* syllable prominence. By using the Phase Code to compute the prominence of each individual syllable within a sequence, the 'Strong-weak' metrical patterning of the entire sequence can be revealed.

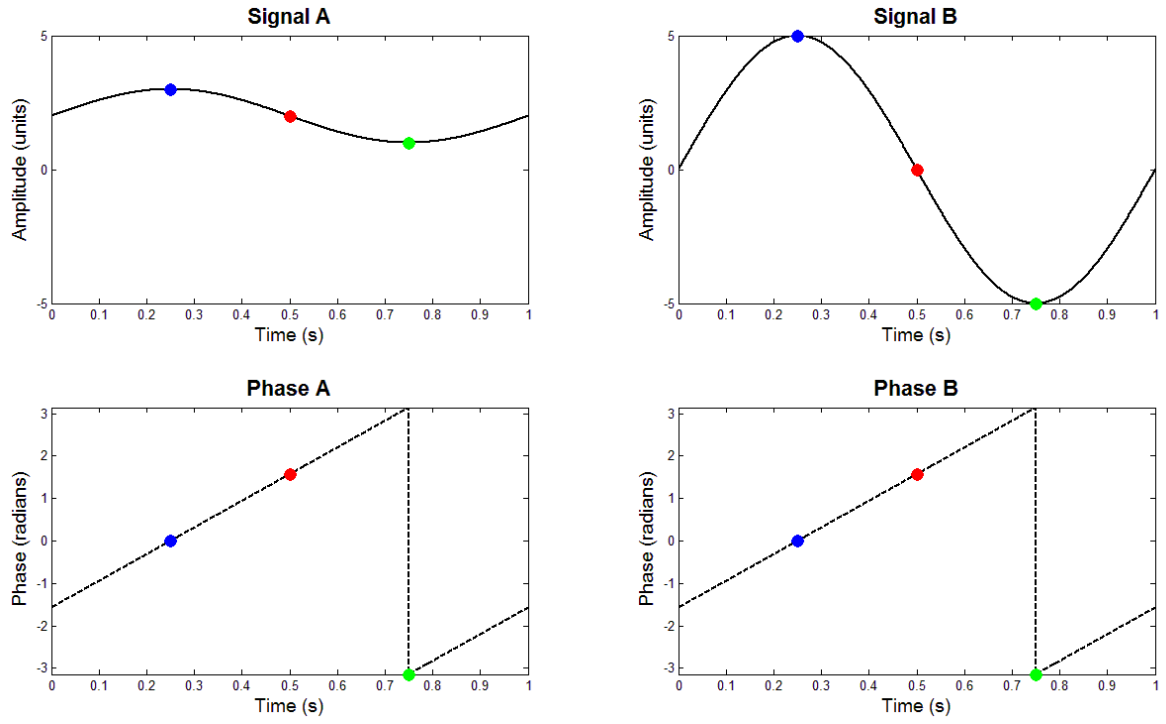
2.5.2.1 Why Use Phase to Compute Prominence?

The local phase at any given point of an AM describes the energy at that point, *relative* to the points before and after it. Unlike amplitude or loudness, which are absolute values, phase is an entirely relative property. For example, Figure 2.7 shows an example of two AM signals, Signal A and Signal B. As shown in the top panel, the two signals differ in terms of their absolute power or amplitude (B is louder than A), and in their modulation depth (B is more deeply modulated and A). Imagine that for each signal, we want to describe the amplitude of the red point, *relative to* the amplitude of the blue point before it and the green point after it. From visual observation of the two signals, this is appears to be easy - the red point is lower in amplitude as compared to the blue point, but higher in amplitude as compared to the green point. In fact, for both Signals A and B, the red point occupies the same *relative* amplitude, being vertically exactly halfway between the blue and green points. Therefore, if the three points corresponded to the loudness (amplitude) of 3 different syllables, one would infer that the relative loudness of the middle syllable was the same in both Signals A and B.

While this conclusion is apparent from visual inspection, it does not emerge naturally when one considers the actual amplitude values of the points in question. For Signal A, the 3 points have amplitudes of 3,2 and 1 respectively. However, for Signal B, the 3 points have amplitudes of 5, 0 and -5 respectively. Therefore the *absolute* amplitudes of the two middle (red) points for Signals A and B are *not* equivalent (2 vs 0). How then does one show that the two red points in Signals A and B actually have the same *relative* amplitude? This is achieved by comparing the amplitude difference between the red and blue points, *relative to*

the amplitude difference between the red and green points, or in other words, performing a local 'normalisation' of the two amplitude differences before and after the point in question.

Figure 2.7. Illustration of how phase captures relative amplitude for given points in a signal. Signals A (left) and B (right) are shown in the top panel. Their respective oscillatory phase values are shown in the bottom panel.



In essence, this is what computing the oscillatory phase achieves. As shown in the bottom panel of Figure 2.7, the oscillatory phase angle of the two red points in Signals A and B (shown in corresponding red points on the phase plot) is exactly the same, as are the phase angles of the two blue and green points. In fact, the phase analysis reveals that the relative pattern of Signals A and B are exactly equivalent - after normalising for their differences in absolute amplitude. This ability to determine relative amplitude is very useful when determining the prosodic prominence of spoken syllables. Speakers vary greatly in the way that they produce stressed and unstressed syllables. Sometimes, the unstressed syllables in one portion of speech may have an even larger (absolute) amplitude than stressed syllables in another portion of speech. If one were to use an absolute amplitude threshold to determine syllable prominence (stress), this would lead to many incorrect assignments of syllable prominence. Rather, syllable prominence is a relative property, as noted by Liberman & Prince (1977). This makes 'phase', a descriptor of relative energy, a particularly suitable

index for syllable prominence. Consequently, in the Stress Phase Code, oscillatory phase (rather than the absolute amplitude) is used to compute syllable prominence.

2.5.2.2 The Stress Phase Code

The Stress Phase Code is a computational scheme that converts continuous Stress and Syllable AM patterns into discrete Strong-weak syllable patterns. This is done by (1) 'sampling' the Syllable AM at key points (peaks) to locate individual syllables; and (2) using the concurrent Stress phase at these Syllable peaks to determine syllable prominence.

(1) 'Sampling' the Syllable AM at Peaks

In the AM hierarchy, cycles of the Syllable AM represent individual syllables. To locate where the individual syllables are, only a single point is needed from each AM cycle. For convenience, this point was chosen as the *peak* of the AM cycle. Therefore, the continuous Syllable AM was 'sampled' at its peaks to determine the location of individual syllables. These peaks were identified by finding the 0 radian upward-crossing points within the Syllable phase series.

(2) Using Stress Phase to Determine Syllable Prominence

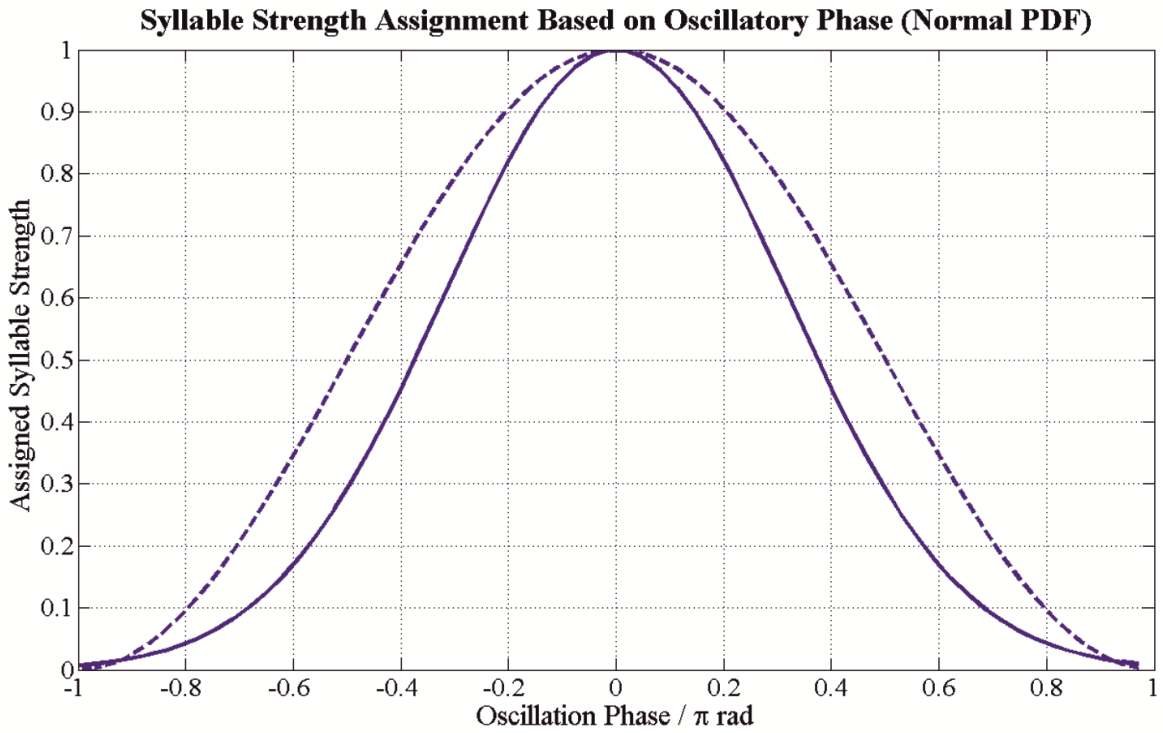
In the AM hierarchy, tiers are hierarchically-nested such that lower tiers are 'modulated' in amplitude by higher tiers. Consequently, the amplitude (or loudness) of a Syllable AM cycle will depend on the modulation that is imposed by the higher-order Stress AM cycle that it is nested under. If the imposed Stress modulation is greater, it follows that the Syllable receiving this modulation will also be louder and more prominent. Conversely, if the imposed Stress modulation is smaller, it follows that the Syllable receiving this modulation will also be softer and less prominent.

As discussed in the previous section, *phase* is a better relative measure for prominence than absolute power or amplitude. Therefore, instead of using the absolute value/amplitude of the Stress modulator to determine the prosodic prominence of the syllable receiving this modulation, the phase of the Stress modulator is used instead. According to this scheme, when the Stress AM is at a peak (0π radians phase), syllables occurring at this peak phase will receive more modulation and therefore be more prominent. Conversely when the

Stress AM is at a trough (-π/π radians phase), syllables occurring at this trough phase will receive less modulation and be less prominent.

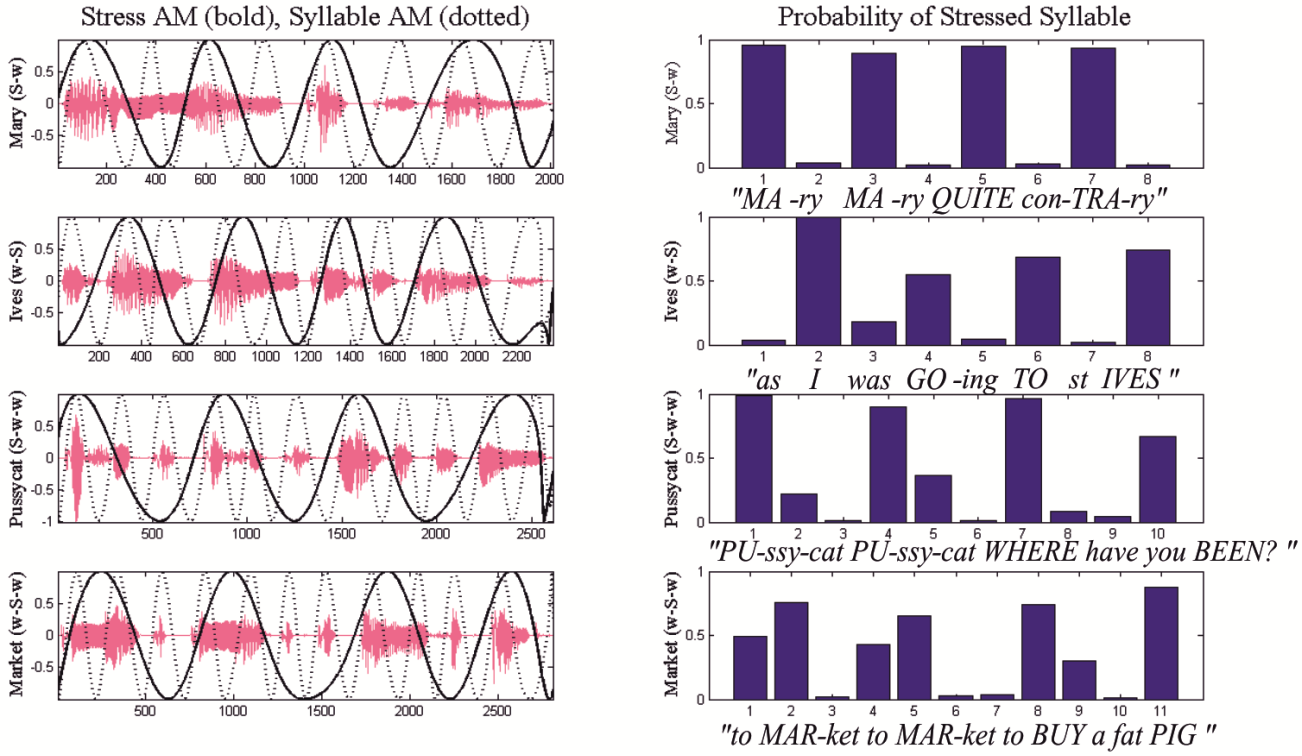
Therefore, the relative prominence of syllables can be expressed in terms of the concurrent oscillatory *phase* of the Stress AM, as shown in Figure 2.8. For each syllable peak that is identified, its concurrent Stress AM phase is noted. This Stress phase value is then converted into an index of prosodic strength via a normal probability density function (PDF) curve, as shown in Figure 2.8. The normal PDF was intentionally chosen because its shape is similar to that of the actual oscillatory shape over the same phase values (shown in the dotted line in Figure 2.8). The shape of the normal PDF assumes that syllables which occur at the peak of the Stress AM would be perceived as strong (stressed). Accordingly, these syllables are assigned the maximum prominence value of 1. By contrast, syllables which occur at the trough of the Stress AM are likely to be perceived as weak (unstressed). Accordingly, these syllables are assigned the minimum prominence value of 0. Therefore, using this phase-to-prominence conversion scheme, Stress phase 'codes for' syllable prominence.

Figure 2.8. Illustration of the oscillation phase convention used in this paper (dotted line). Solid line indicates the assigned syllable strength (y-axis) under the Stress Phase Code.



The final output of this computational scheme is a discrete series of syllables, each assigned with a numerical strength or prominence value. Examples of this discrete output are shown in Figure 2.9, where 4 nursery rhyme sentences with different metrical patterns were processed using the Stress Phase Code computational scheme. These were the English nursery rhymes 'Mary Mary quite contrary', 'As I was going to St Ives', 'Pussy cat pussy cat' and 'To market to market'. Their strong-weak patterning would be *"MA-ry MA-ry QUITE con-TRA-ry", "as I was GO-ing TO st IVES", "PU-ssy-cat PU-ssy-cat WHERE have you BEEN"* and *"to MAR-ket to MAR-ket to BUY a fat PIG"*, where capitalised syllables representing stressed syllables. The series of panels on the left of Figure 2.9 depict the original Stress and Syllable AM tiers of the four nursery rhymes, overlaid on the sound pressure waveform for each sentence. The series of panels on the right of Figure 2.9 show the syllable strengths that were assigned to each syllable based on the Stress Phase Code. Inspection of the figure confirms that the assigned syllable strengths conform closely to the four original prosodic patterns of each nursery rhyme.

Figure 2.9. Illustration of how the Stress Phase Code may be used to compute syllable prominence patterns. The phase of Stress (bold) and Syllable (dotted) AMs for each nursery rhyme sentence are shown on the left (plotted as cosine function for visualisation). Resulting detected syllable beats and their assigned prosodic strength is shown on the right.



It is worth emphasising that the methods of syllable detection and prosodic strength assignment used by the AMPH model are developmentally plausible, because no prior manual segmentation or labelling of the waveform is required. Syllable peaks are detected as '0' radian upward-crossing points within the Syllable phase series, and these are labelled with prosodic strength according to the corresponding Stress phase. Furthermore, using Stress phase as a marker for syllable prominence (rather than sound intensity or duration) means that the method is robust to local fluctuations in loudness and rate of speaking.

2.6 Combining Meter and Prosodic Pattern into Segmentation Schemes

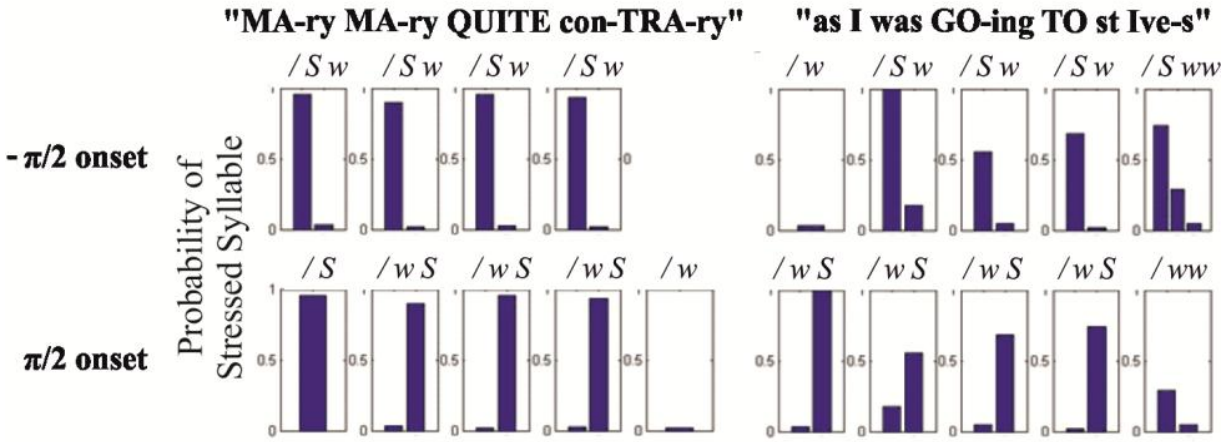
The final step of the AM Phase Hierarchy model is to group or segment the sequence of strong and weak syllables (resulting from the Stress Phase Code computational scheme) into prosodic feet according to the rhythmic meter that was inferred earlier from the *n:m* phase-locking mode. This is not a trivial problem. Without knowledge of the lexical content of the sentence, the eight syllable beats in "MA-ry MA-ry QUITE con-TRA-ry" may be segmented into pairs beginning at the first syllable as "MAry / MAry / QUITEcon / TRAry", or at the second syllable as "MA / ryMA / ryQUITE / conTRA / ry". This word segmentation problem is reminiscent of that faced by infants (see Jusczyk et al, 1999). Jusczyk et al. showed that 7.5-month-old infants made segmentation errors when they were presented with English sentences containing words that did not follow the characteristic trochaic pattern (e.g. *"Her guitar is too fancy"*). Infants mis-segmented the nonword *"TA-ris"* on the basis of strong-weak syllable stress (rather than hearing *"gui-TAR"*, which is weak-strong). Older infants no longer made this mis-segmentation. Infant data such as these imply that the temporal modulation structure of speech affords several possible segmentation schemes.

The AMPH model reaches the same dilemma when oscillatory cycles of the prosodic Stress AM are used to provide segmentation markers. Oscillatory cycles may begin at any phase value, so long as they traverse a full cycle. Therefore, Stress AM cycles may be defined in multiple ways on the basis of different starting phase values. To illustrate this, the AMPH model generated two alternative segmentation solutions for the nursery rhymes 'Mary Mary quite contrary' and 'As I was going to St Ives'. These segmentation solutions were defined based on a starting Stress phase of either $-\pi/2$ or $\pi/2$ radians. These values were

chosen because they are maximally spaced (π radians apart) on the oscillatory phase cycle, and occur just before an oscillatory peak (0 radian) or trough (π radians). Since Stress oscillatory peaks and troughs often coincided with the vowel nuclei of syllables, by beginning the oscillatory segmentation cycle just before the peak/trough, syllable onsets could be included within the same segmentation cycle as their vowel nuclei.

As can be seen in Figure 2.10, using these two starting Stress phase values resulted in segmentation solutions that corresponded closely to trochaic and iambic rhythm patterns respectively. In both cases considered in Figure 2.10, an onset of -π/2 radians resulted in a trochaic (S-w) pattern, whereas an onset of π/2 radians resulted in an iambic (w-S) pattern. If one considers this finding in the light of the empirical data from Jusczyk et al. (1999), this may imply that English-learning infants initially default to a segmentation scheme with a single onset phase (e.g. -π/2 radians), thereby segmenting trochaic words from the speech stream (see Cutler & Carter, 1987). Other oscillatory phase onsets may be used by English-learning infants only later in development, following longer experience with their native language (e.g. using π/2 radians to perform iambic segmentation).

Figure 2.10. Alternative segmentation solutions for "Mary Mary" and "St Ives" using a Stress AM starting phase of +pi/2 or -pi/2. Note that there appear to be extra syllables at the end of the sentence "St Ives". These correspond to very small modulations associated with the final "s" that nevertheless contain phase information. These non-syllable modulations are addressed in the new S-AMPH model (Part III).



2.7 CHAPTER SUMMARY

The AM Phase Hierarchy model offers a signal-based method for calculating the prosodic and rhythmic information conveyed by any speaker from the amplitude modulation structure of their speech. According to the AMPH model, metrical rhythm patterns are detected by breaking down (demodulating) the speech signal into theoretically-defined component modulation rates, based on the linguistic prosodic hierarchy. This generates a 5-tier AM hierarchy where each AM tier corresponds to a different prosodic level. Within this AM hierarchy, the Stress and Syllable AM tiers are especially rich in rhythm information. For example, the long-term *n:m* phase-locking ratio between these two tiers can be used to infer rhythmic meter. Furthermore, the discrete sequence of Strong and weak syllables in an utterance can be computed using a Stress Phase Code. This uses the instantaneous Stress-Syllable phase relationship (the Stress AM phase concurrent with Syllable AM peaks) as an index of syllable prominence. Finally, these two types of information (meter and Strong-weak syllable sequence) can be combined into rhythmic segmentation schemes.

Having described the theory and mechanisms underlying the AM Phase Hierarchy model, the next step is to conduct an empirical test, to see whether the AM tiers and phase relations specified in the model do indeed form the basis of listeners' perception of metrical rhythm patterns in speech. In the next Chapter, a tone-vocoding experiment is conducted to test the two main tenets of the AMPH model - the importance of Stress and Syllable AM rates for rhythm, and the importance of the relative *phase* between Stress and Syllable AMs for determining syllable prominence.

3 TESTING THE ASSUMPTIONS OF THE AMPH MODEL : A TONE-VOCODER EXPERIMENT

Recall from the Introduction to Part II that two major research questions were initially posed. These questions were :

(1) ***Where*** *is speech rhythm information located in the modulation spectrum of the envelope?*
(2) ***How*** *is speech rhythm information 'coded' within amplitude modulation patterns in the envelope?*

The answers to these questions formed two basic tenets of the AMPH model, as described in the previous Chapter. Here, the psychological validity of these tenets (with respect to human listeners) is tested in a tone-vocoder experiment.

(1) Speech rhythm information is primarily carried by Stress- and Syllable-rate AMs within the speech envelope

The AM Phase Hierarchy model assumes that Stress and Syllable AMs together contain sufficient information to specify metrical rhythm patterns in speech. To test this assumption, 5-tier AM hierarchies were extracted from nursery rhyme sentences with different metrical patterns. These AM patterns were made audible through tone-vocoding, and played back to listeners either as single AM tiers or as pairs of AM tiers. In each case, participants were asked to identify the original nursery rhyme based on the rhythm pattern that they heard. It was predicted that Stress+Syllable AMs would contain more metrical rhythm information than the other AM tiers or their combinations, and hence result in the *best* metrical rhythm identification when presented together.

(2) Speech rhythm patterns are 'coded' via the phase relationship between Stress and Syllable AMs

Another central tenet of the AMPH model is that syllable prominence is specified by the local phase relationship between the Stress AM and the Syllable AM. Since the Stress Phase Code assigns syllable prominence *circularly* by phase, this predicts that incremental phase displacements of the Stress-Syllable AM relationship should cause *circular*

perturbations in the perceived syllable prominence[6]. Accordingly, phase-shifts in the Stress-Syllable AM relationship of up to 1π radians (half a cycle) should move participants' perception of a given syllable toward the opposite prominence (e.g. from strong to weak), but larger shifts of up to 2 π radians (a full cycle) should bring perception back to the original value (e.g. strong).

By extension, if a series of syllables contains a regularly alternating strong (S) and weak (w) pattern such as 'S-w-S-w-S-w-S-w', a 1π radians phase-shift applied to every syllable in the series should result in the opposite pattern of 'w-S-w-S-w-S-w-S'. Conversely, a 2π radians phase-shift should elicit no net change in the perceived rhythm pattern. To test this prediction, the phase relationship between Stress and Syllable AMs was manipulated by parametrically phase-shifting the Stress AM by either 1π radians (turning peaks to troughs) or 2π radians (keeping peaks as peaks) while holding the syllable AM constant. The 1π phase shift should result in incorrect assignment of the prosodic pattern, since peaks and troughs in the Stress modulator would be inverted. By contrast, a 2π phase shift would maintain the phase-relationship between Stress AM and Syllable AM, allowing correct assignment of the prosodic pattern. However, if participants were not using phase, then metrical rhythm judgments should be better in the 1π-shift condition than the 2π-shift condition (since larger phase-shifts would also introduce more acoustic artefacts).

[6]Consider the analogy of a wheel. Imagine placing the index finger of each hand at the same point on a wheel. Now, keeping one finger in its original location, use the other to rotate the wheel clockwise. As the wheel rotates, your two fingers become more widely spaced apart until they are diametrically opposite. This point is half a rotation cycle. After this point though, any further rotation now brings your two fingers closer together until they eventually meet again at the original point.

3.1 EXPERIMENTAL METHOD

3.1.1 PARTICIPANTS

Twenty-three adults (7 male; mean age 26.0 yrs, range 22.0 years - 37.5 years) participated in the study. All participants had no diagnosed auditory, language or learning difficulties and spoke English as a first language. Twelve participants had had more than 5 years of musical training while the remaining eleven had less than 5 years or no musical training. In an initial analyses of the results, the factor of musical training did not affect participants' performance on the tone-vocoder task (i.e. musical training status was not significant as a between-groups factor in a repeated measures ANOVA). Therefore, musical training is not considered further here.

3.1.2 MATERIALS

Four duple meter nursery rhyme sentences were used in this vocoder experiment (see Table 3.1). Each sentence contained 8 syllables and had a duple metrical rhythm of alternating strong (S) and weak (w) syllable beats. Two nursery rhymes started with a strong stressed syllable and continued in a 'strong-weak' or trochaic pattern (e.g. "MA-ry MA-ry..." and "SIM-ple SI-mon..."). The other two nursery rhymes began with a weak unstressed syllable and continued with a 'weak-strong' or iambic pattern (e.g. "as I was GO-ing..." and "the QUEEN of HEARTS..."). These two distinctive Rhythm Patterns (RPs) are shown in Table 3.1.

Each sentence was approximately 2 seconds in length. The nursery rhymes were spoken by a female native speaker of British English who was articulating in time to a 4 Hz (syllable rate) metronome beat. The speaker was instructed to produce the metrical pattern of each nursery rhyme as clearly as possible. Utterances were digitally recorded using a TASCAM digital recorder (44.1 kHz, 24-bit), and the metronome was not audible in the final recording.

Table 3.1. List of nursery rhyme sentences used in the tone vocoder experiment and their metrical Rhythm Pattern (RP).

METRICAL RHYTHM PATTERN *(**S** = Strong, w = weak)*		**NURSERY RHYME SENTENCE** *(CAPS = Strong syllable)*
Duple meter	**S** w **S** w **S** w **S** w *(RP 1, trochaic)*	"MA-ry MA-ry QUITE con-TRA-ry"
		"SIM-ple SI-mon MET a PIE-man"
	w **S** w **S** w **S** w **S** *(RP 2, iambic)*	"as I was GO-ing TO st IVES"
		"the QUEEN of HEARTS she MADE some TARTS"

3.1.3 TASK

Participants heard four single-channel tone-vocoded nursery rhyme sentences, presented one at a time. They were asked to indicate which of the four possible target rhymes they thought that they had heard via a button press. Participants were told to base their judgment on the rhythm pattern of the stimulus. All participants were first given 20 practice trials during which they heard the four nursery rhymes as originally spoken, without vocoding. This enabled participants to learn the metrical rhythm pattern of each rhyme, and to become familiar with the response button mapping. Subsequently, participants performed the task with tone-vocoded stimuli only. The tone-vocoded stimuli retained the temporal pattern of each nursery rhyme sentence, but were completely unintelligible. Cartoon icons representing the four response options were displayed on the computer screen throughout the experiment to help to reduce the memory load of the task. These icons are shown in Figure 3.1. Auditory stimuli were presented binaurally using Sennheiser HD580 headphones at 70dB SPL. The experimental task was programmed in Presentation (Neurobehavioural Systems) and delivered using a Lenovo ThinkPad Edge laptop.

Figure 3.1. Cartoon icons displayed on-screen throughout the experiment to remind participants of the four nursery rhyme response options and their respective response buttons. The corresponding rhymes are (L to R) : St Ives, Mary Mary, Queen of Hearts and Simple Simon.

3.1.4 SIGNAL PROCESSING METHODS

All signal-processing steps were carried out using MATLAB (R2009a, Version 7.8.0, The Mathworks Inc).

3.1.4.1 AM Hierarchy Extraction

Two different demodulation methods were used to extract the AM hierarchies, creating two different but complete sets of vocoded stimuli. The reason for using two methods was to ensure that the experimental results obtained were not due to methodological artifacts introduced by the demodulation or filtering procedures. For example, artificial modulations could be introduced into the stimuli by filter 'ringing'. These spurious modulations would occur at the same frequency as the filter, and be indiscernible from the true signal. Consequently, it was important to have a methodological control for any such artifacts. The standard method involved using the Hilbert transform to extract the amplitude envelope, which was then passed through the modulation filterbank (MFB) to extract the AM hierarchy. For this method, the edge frequencies for the 5 tiers of the filterbank were 0.5-0.8

Hz (Slow), 0.8-2.3 Hz (Stress), 2.3-7 Hz (Syllable)[7], 7-20 Hz (Subbeat) and 20-50 Hz (Fast). The choice of these filtering parameters is explained further in Appendix 2.1.

The second, control method of AM hierarchy extraction was Probabilistic Amplitude Demodulation (PAD; Turner, 2010), and did not involve the Hilbert transform or filtering. As described in Chapter 1, Section 1.7.2, PAD extracted the AM hierarchy directly from the speech signal using a Bayesian inference-based approach. A description of the PAD 'demodulation cascade' process which was used to produce the PAD AM hierarchy, and a comparison of MFB-derived and PAD-derived modulators is provided in Appendix 3.1.

All participants heard both sets of stimuli. It was reasoned that if participants produced the same pattern of results with two methods of AM extraction that operate using very different sets of principles, these results were likely to be due to real features in speech rather than artifacts.

3.1.4.2 Tone Vocoding

To make the AMs in the hierarchy audible, each AM tier was used to modulate a 500 Hz sine-tone carrier (i.e. single channel tone-vocoding). Note that the phonetic fine structure of the signal was intentionally discarded, and only AMs derived from the amplitude envelope were used to modulate the sine tone carrier. A multi-channel vocoder was *not* used to ensure that the sentences would be completely unintelligible. Since the dependent variable was how well participants could identify each sentence from its rhythm pattern alone, all other cues to sentence identity were removed. Multi-channel vocoders would have increased the intelligibility of the sentences, providing listeners with extra non-rhythm-related cues.

To create single-tier AM stimuli (e.g. Stress only), the appropriate AM tier was extracted from the hierarchy and combined with the 500 Hz sine-tone carrier. Since PAD AMs were entirely positive-valued, these were multiplied directly with the carrier. For MFB AMs which had negative-valued portions, a 30ms-ramped pedestal at RMS power was added prior to combining with the carrier. To create double-tier AM stimuli (e.g. Stress+Syllable), the two AM tiers were first combined via addition (MFB) or multiplication (PAD) before combining with the carrier. All stimuli were equalised to 70dB. The resulting tone-vocoded

[7] The syllable rate measured for all nursery rhymes was 4.04 Hz, which confirmed that the speaker conformed closely to the metronome rate

sentences had clear temporal patterns ranging from Morse-code to flutter, but were otherwise completely unintelligible.

3.1.4.3 Phase-Shifting

The aim of phase-shifting was to parametrically change the phase-relationship between AM tiers to measure whether this also systematically changed the rhythm pattern perceived by the listener. Since the sentences were either trochaic or iambic in pattern, the aim of the procedure was to make trochaic sentences sound iambic, and vice versa. In a modulation hierarchy, the slower AM should impose perceptual constraints on faster AMs (e.g. Stress AM phase determines the prosodic prominence of Syllable AM beats). Hence, for pairs of AM tiers, phase-shifting involved shifting the slower AM with respect to the faster AM, which was held constant. For single AM tiers, phase-shifting was also performed as a control, but this was not expected to produce a significant change in participants' judgements.

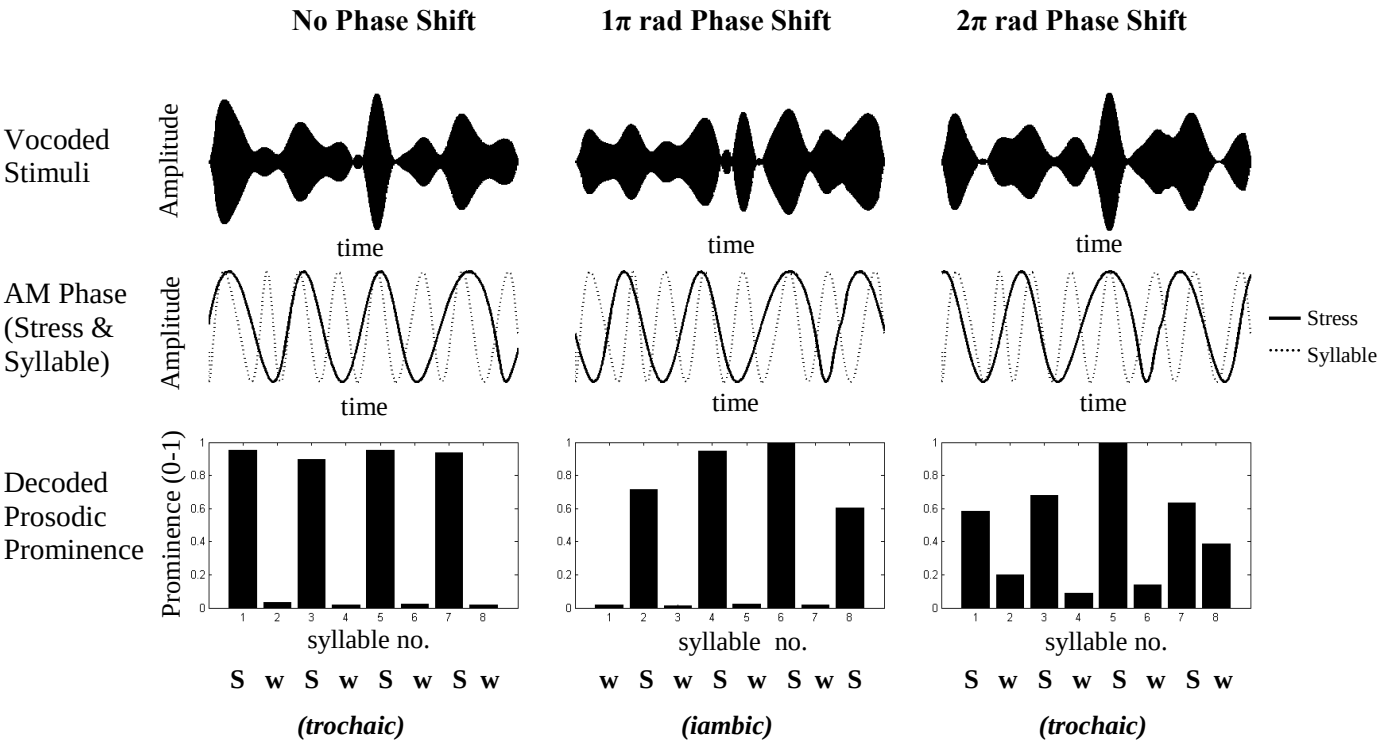
Due to the use of the metronome, the nursery rhyme sentences were perfectly regular in metrical structure, therefore their AMs were also highly regular in their temporal pattern of peaks (P) and troughs (t), resembling a pure sinusoid. Thus, phase-shifting was implemented by shuffling sections of the signal from the start to the end. For example, for a nursery rhyme with an AM pattern of 'P-t-P-t-P-t-P-t', shuffling the first peak (P) from start to end would result in a pattern of 't-P-t-P-t-P-t-P', the same result as if each element was individually phase-shifted by 1π radians.

Phase-shifting via shuffling was a superior method to deletion or silence insertion (delay) because it allowed all the original information within each sentence to be retained, changing only its temporal order. The length of phase-shifted stimuli could also be kept the same as non-phase-shifted stimuli by this method. For the 2π radians shift, the sample length shuffled was a full period cycle corresponding to a representative single frequency within the bandwidth of the AM tier in question. The representative Syllable frequency was determined by finding the peak RMS power in the 3-7 Hz range of the modulation spectrum for each sample. This turned out to be 4.04 Hz, which was very close to the metronome pacing beat of 4.0 Hz. The representative Sub-beat frequency was determined by taking the mean of 2 and 3 times the Syllable frequency (10.1 Hz), to allow for both duple and triple patterns in this tier. For consistency, the representative Stress frequency was also determined by taking the mean of half and a third of the Syllable frequency (1.68 Hz). This resulted in cycle lengths of 595

ms for the Stress tier (1.68 Hz), 248 ms for the Syllable tier (4.04 Hz), and 99 ms for the Sub-beat tier (10.1 Hz) that were used for shuffling. For a 1π radians shift, the length shuffled was half of that used for the 2π radians shift.

For all stimuli (phase-shifted and non-phase-shifted), a 50ms ramp was applied to the start and end of the AMs to make the phase-shift boundary less abrupt. Stimuli were manually checked to verify that the shuffling process produced the desired phase changes for all stimuli. Even though some minor sound artifacts were introduced as a result of the shuffling process (e.g. at phase-shift boundaries), the resulting metrical patterns emerged as predicted, e.g. trochaic (no shift) ==> iambic (1π shift) ==> trochaic (2π shift). This is illustrated in Figure 3.2.

Figure 3.2. Illustration of the effect of phase-shifting on the metrical pattern of 'Mary Mary'. (Top row) Tone-vocoded MFB stimuli used in the experiment. (Middle row): Corresponding Stress (bold) and Syllable (dotted) AM phase patterns. Phase values are projected onto a cosine function for visualisation purposes. Only Stress AMs were phase-shifted while Syllable AMs were held constant. (Bottom row) Decoded prosodic pattern of syllables. Strong syllables ('S') have a prominence value of >0.5, weak syllables ('w') have a prominence value of <0.5.



It is important to note that the overall modulation shapes of the non-shifted and 2π-shifted stimuli were substantially different. However, both stimuli had the same metrical pattern according to the Stress Phase Code, as shown in the bottom panel of Figure 3.2. Hence, if listeners judged both stimuli as having the same metrical pattern, this would not be due to perceptual similarity or familiarity, but because the key metrical statistics (phase relationships) were similar.

3.1.5 DESIGN

The experiment followed an AM tier (5) x Phase Shift (3) x Demodulation Method (2) design. Five different AM tiers or tier combinations were used for vocoding[8]. These were: (1) Stress only; (2) Syllable only; (3) Sub-beat only; (4) Stress+Syllable; (5) Syllable+Sub-beat. Each of these AM combinations was presented in three phase shift conditions : (1) No phase shift; (2) 1π radians phase-shift; and (3) 2π radians phase-shift.

Fewer phase-shifted stimuli (1π radians or 2π radians) were presented than non-phase-shifted versions to allow participants to maintain a strong representation of the correct metrical pattern for each nursery rhyme. Thus, participants heard the normative (non-shifted) version five times for each nursery rhyme, but they only heard each of the phase-shifted variants (1π radians or 2π radians) twice.

Phase-shifted and non-shifted stimuli were presented within the same experimental block in a randomised fashion. Stimuli that were vocoded using MFB-produced AMs and PAD-produced AMs were presented in separate experimental blocks, giving a total of 360 trials for the entire experiment (5 AM tier combinations x 9 phase variants [5 x 0π radians, 2 x 1π radians, 2 x 2π radians] x 4 nursery rhymes per block x 2 demodulation methods).

[8] The full range of AM rates was used in a pilot study, including Slow and Fast AMs. It was found that participants were at chance for the Slow AM tier, and performed equally well for both Sub-beat and Fast AM tiers. Based on these results, the current subset of Stress, Syllable and Sub-beat AM tiers was chosen to reduce the number of conditions needed in the experiment.

3.2 RESULTS

Performance was scored in terms of whether participants identified each nursery rhyme correctly (the accuracy score) and also whether they identified the correct rhythm pattern (RP score, trochaic or iambic; recall that 2 nursery rhymes exemplified each RP). AM combinations that provide strong metrical rhythm information should boost both accuracy and RP identification. Furthermore, if the AM combination was providing metrical rhythm pattern information, participants should be more likely to confuse sentences with the same RP than to confuse sentences with different RPs. Hence, the hallmarks of an AM tier combination providing strong metrical rhythm information would be a high RP score and a high ratio of same:different RP confusions. Table 3.2 provides a summary of mean scores for all conditions. For the Accuracy scores, the level of chance performance was 25% (1 out of 4 nursery rhyme choices). For RP scores, the level of chance performance was 50% (1 out of 2 RPs).

Table 3.2. Accuracy scores and RP scores for AM combinations and phase shift conditions. Means shown are averages across PAD and MFB methods.

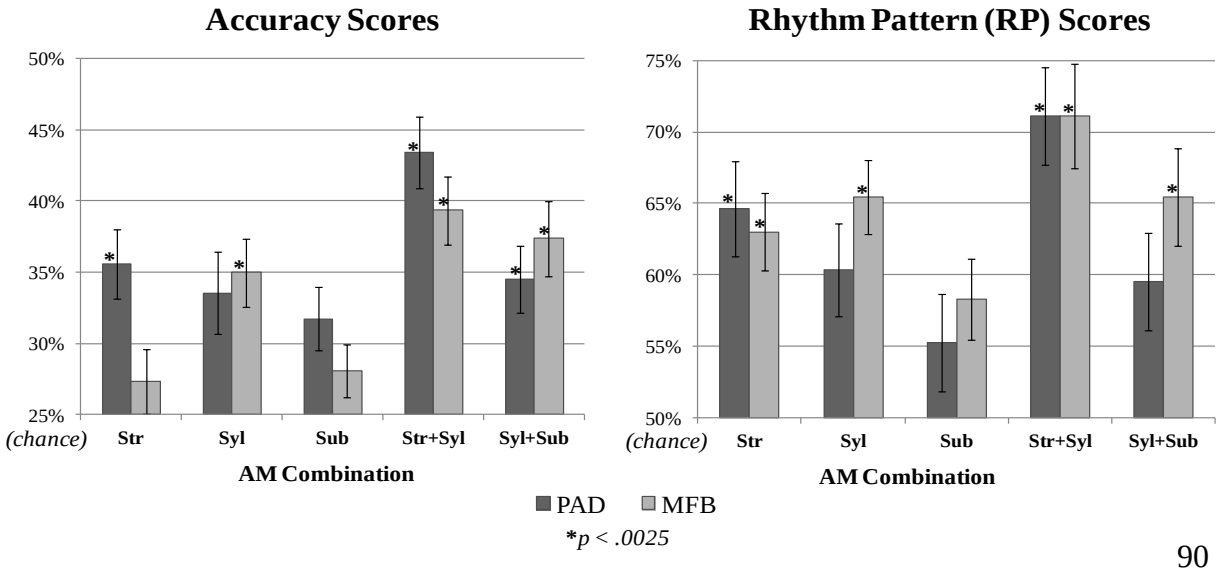
		Accuracy Scores (%)			**RP Scores (%)**		
	Phase Shift	*0 rad*	*1π rad*	*2π rad*	*0 rad*	*1π rad*	*2π rad*
AM tier combination	**Stress**	31.4	20.7	27.9	63.8	43.2	54.5
	(SE)	(1.8)	(2.3)	(2.9)	(2.4)	(3.2)	(3.0)
	Syllable	34.3	29.6	26.6	62.9	54.3	50.7
	(SE)	(1.9)	(2.4)	(2.9)	(2.6)	(2.6)	(3.4)
	Sub-beat	29.9	29.6	22.0	56.7	58.0	46.5
	(SE)	(1.4)	(2.9)	(2.0)	(2.3)	(2.8)	(2.5)
	Stress + Syllable	41.4	22.0	38.9	71.1	42.4	67.1
	(SE)	(1.8)	(2.5)	(2.9)	(3.0)	(2.9)	(2.8)
	Syllable + Sub-beat	35.9	29.3	29.1	62.5	53.3	54.1
	(SE)	(2.1)	(2.8)	(2.5)	(2.7)	(3.2)	(2.3)

3.2.1 THE NO PHASE SHIFT STIMULI

3.2.1.1 Accuracy and RP Scores

Accuracy and RP scores for the five AM tier combinations that were *not* phase-shifted (0π radians) are shown in Figure 3.3, broken down by demodulation method (PAD or MFB). Participants' accuracy in identifying the correct sentence ranged between 28%-42%, while their RP identification ranged between 55%-70%. These low scores were not surprising given that the sentences were completely unintelligible. To test performance against chance, one-sample t-tests were conducted against the test values of 0.25 (accuracy) or 0.5 (RP) for each AM combination and method. Since a total of twenty t-tests were conducted, a Bonferroni-corrected significance value of $p<.0025$ (.05/20) was used. Inspection of Figure 3.3 shows that across conditions, participants always performed above chance when hearing Stress AMs and Syllable AMs together, but always performed at chance when hearing Sub-beat AMs only. Combinations containing either Stress AMs or Syllable AMs sometimes elicited performance above chance. This indicates that Stress AMs and Syllable AMs individually contained some rhythm information, but that participants performed best when both these AM tiers were provided together. In contrast, Sub-beat AMs alone did not appear to contain sufficient rhythm information for participants to make successful metrical rhythm judgments.

Figure 3.3. Accuracy and RP scores for the five AM combinations, for each Method (PAD and MFB). Error bars indicate the standard error of the mean, and are not suitable for inferring statistically-significant differences between repeated measures data. (*) Indicates performance above chance.



To compare performance between AM combinations more directly, two 2 x 5 Repeated Measures ANOVAs were conducted, taking RP or Accuracy scores respectively as the dependent variable, and Demodulation Method (2) and AM tier or tier combination (5) as within-subjects factors. Scores in all conditions were normally distributed ($p > .05$ in Kolmogorov-Smirnov test of normality). For both RP and Accuracy scores, there was a significant main effect of AM combination (RP: $F(4,88) = 9.15$, $p < .0001$; Accuracy: $F(4,88) = 10.53$, $p < .0001$), but no difference between demodulation methods (RP : $F(1,22) = 2.27$, $p = .15$; Accuracy : $F(1,22) = 2.82$, $p = .11$) and no interaction between AM tier x Method (RP : $F(4,88) = 0.75$, $p = .56$; Accuracy : $F(4,88) = 2.03$, $p = .10$). This confirmed that both PAD and MFB demodulation methods were producing similar patterns of listening performance.

The AM tier main effect was analysed further by performing a Tukey HSD post-hoc analysis. For both RP and Accuracy scores, performance with Stress+Syllable AMs was significantly superior to all four other AM tiers (RP : $p < 0.025$ and Accuracy : $p < 0.05$ for all four comparisons). This confirmed that metrical pattern identification for the Stress+Syllable AM tier combination was reliably better than for any other AM combination tested. Since listeners hearing Stress+Syllable AMs outperformed listeners hearing Syllable+Sub-beat AMs, the superior performance with Stress+Syllable AMs could not simply be due to a greater modulation bandwidth being presented to listeners (which was actually greater for Syllable+Sub-beat AMs). It must have been due to the quality of the rhythm information provided by combining this particular pair of AM tiers.

Secondly, performance with Stress+Syllable AMs was better than performance with either Stress AMs or Syllable AMs alone. This indicated that participants were able to combine syllable-rate information with stress-rate information productively, and that the two forms of rhythm information were not redundant. In contrast, performance with Syllable+Sub-beat AMs was not significantly better than with Syllable AMs alone ($p = 0.91$ for Accuracy) indicating that Sub-beat modulations were not providing additional rhythm cues over and above those already present in the Syllable AM.

Hence, Accuracy and RP scores both indicated that the combination of the Stress+Syllable AM tiers provided the most metrical pattern information, consistent with the first tenet of the AMPH model.

3.2.1.2 Confusion Errors

According to the AMPH model, the metrical pattern information provided by the combination of Stress+Syllable AMs relates directly to the perceptual impression of strong-weak (trochaic) and weak-strong (iambic) prosodic patterns. Hence the two RP groups of nursery rhymes should be distinguished on the basis of the pattern of local phase relationships between the Stress AM and Syllable AM. However, it is also possible that participants were relying on other temporal cues such as subtle differences in utterance speed or syllable spacing to make their judgments. In this case, performance for Stress+Syllable AMs should be unrelated to whether or not participants actually heard metrical rhythm patterns.

To distinguish between these two competing explanations, the pattern of confusion errors produced by participants was analysed. If participants were using RP cues to make their judgment, then they should make more confusions between rhythmically-similar rather than unrelated nursery rhyme sentences. For example, participants should be more likely to confuse 'Mary Mary' with 'Simple Simon'. Conversely, if participants were relying on other temporal cues, their errors should be evenly distributed across the different nursery rhymes.

To quantify the pattern of confusion errors, a normalised scoring system was used. Each nursery rhyme could be confused with one of three other possible nursery rhymes. One of these rhymes had the 'Same' rhythm pattern as the target, and the other 2 rhymes had a 'Different' rhythm pattern. Consequently, participants should be twice as likely to make a Different confusion as a Same confusion if they were responding randomly. The observed percentage of Different confusions was therefore normalised by a factor of half, and this was compared to the full percentage of Same confusions. Since there was no difference between demodulation method in terms of accuracy, the errors made for MFB and PAD stimuli were pooled and entered into a single 2 x 5 repeated measures ANOVA with Confusion type (Same or Different) and AM tier or tier combination (5 levels) as within-subjects factors.

The results of the ANOVA indicated that there was a significant main effect of Confusion type ($F(1,22) = 27.0$, $p<.0001$). Participants made significantly more confusions between nursery rhymes with the same rhythm pattern (28.7%) than between nursery rhymes with different rhythm patterns (18.3%). There was also a significant main effect of AM tier ($F(4,88) = 4.02$, $p <.01$). As expected, participants made the fewest errors in the Stress+Syllable AM condition. Finally, there was a significant interaction between Confusion

type x AM tier ($F(2.84,62.5) = 3.71$, $p < .05$, Greenhouse-Geisser epsilon = 0.71). Tukey HSD post hoc tests revealed that there were significantly more same than different confusions made in all AM tiers or tier combinations *except for* the Sub-beat band ($p = .34$). These results indicate that participants were indeed performing the task on the basis of RP (strong-weak/weak-strong) prosodic information rather than on the basis of speed or spacing information. As such, one may infer that the superior performance previously observed for the Stress+Syllable AM combination was indeed due to the provision of superior metrical rhythm pattern information.

3.2.1.3 Summary of the No Phase Shift Data

Participants' performance in the no-phase shift condition clearly indicated that the combination of Stress+Syllable AMs elicited the best metrical rhythm judgment. This was true whether performance was measured in terms of Accuracy, or in terms of identifying the correct Rhythm Pattern. Furthermore, when the pattern of confusion errors was analysed, participants were found to confuse sentences with the *same* rhythm pattern more often than they confused sentences with a *different* rhythm pattern. This confusion pattern was true for all AM tier combinations except for the Sub-beat band. This confirmed that participants were indeed basing their judgments of sentence identity on the rhythm pattern that they heard. Therefore, the results of the no phase shift data are consistent with the first assumption of the AMPH model, that Stress and Syllable AMs are the primary carriers of metrical rhythm information in the speech envelope.

3.2.2 PHASE SHIFT EFFECTS

As an empirical test of the 'Stress Phase Code' used in the AM Phase Hierarchy model, participants also attempted to recognise the tone-vocoded nursery rhymes following phase-shifts of 1π radians or 2π radians. Phase-shifts perturbed the phase relationship between AM tiers in the stimuli. For AM combinations carrying rhythm information, it was predicted that phase-shifts of 1π radians would cause participants to perceive sentences as having the opposite rhythm pattern (e.g. trochaic to iambic), while phase-shifts of 2π radians would maintain the original rhythm pattern (see Figure 3.2). This would lead to a ***'V-shaped'*** pattern of drop and recovery across the three phase-shift conditions (no shift to 1π radians shift to 2π radians shift). Meanwhile, for AM tiers that are *not* carrying metrical information (e.g. the Sub-beat tier), phase-shifting should either impair metrical rhythm judgments equally for both types of phase-shift, or decrements in judgment should be greater with the 2π-shift compared to the 1π-shift, as the 2π-shift involves greater signal distortion.

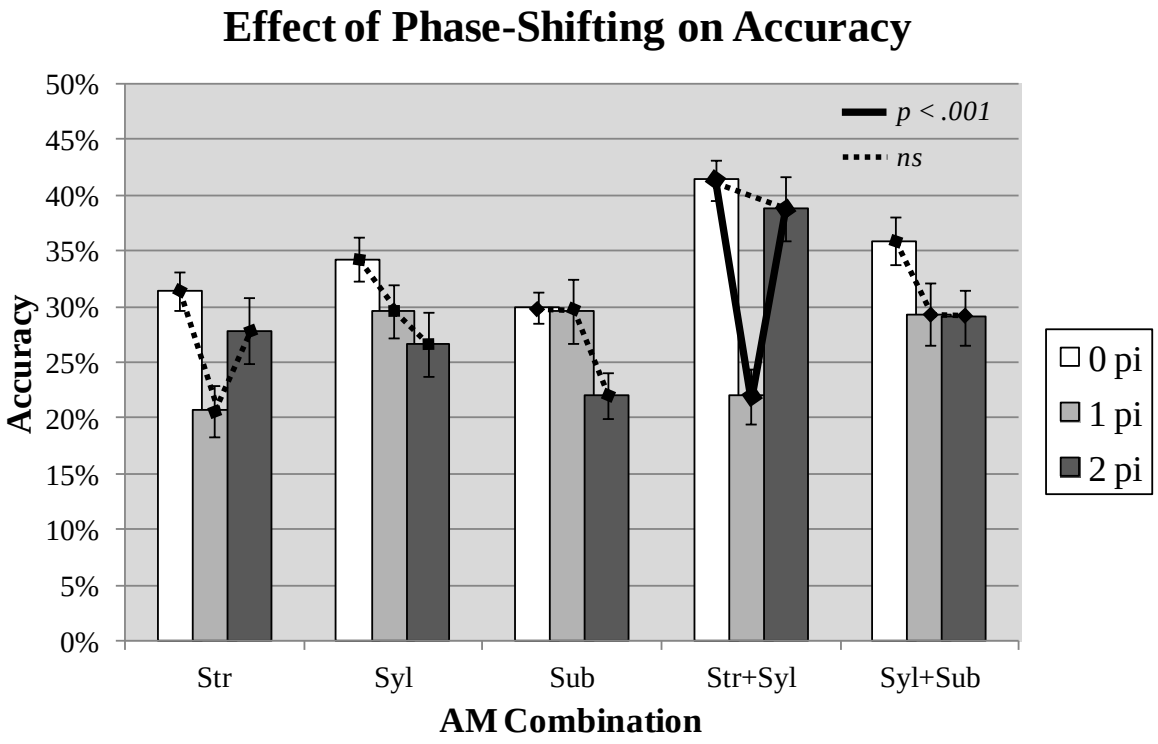
3.2.2.1 Accuracy Scores

The effects of phase-shifting on performance accuracy in each AM tier or tier combination are shown in Figure 3.4. Inspection of the Figure suggests that the predicted 'V-shaped' drop in accuracy for a 1π-radians shift and recovery for a 2π-radians shift indeed occurred in the Stress+Syllable AM tier combination, and possibly also in the Stress-AM tier. To investigate this effect, a 5 (AM tier or tier combination) x 3 (Phase shift, 0π, 1π, 2π radians) x 2 (Demodulation Method: PAD, MFB) repeated measures ANOVA was carried out, taking Accuracy of identifying the target nursery rhyme as the dependent variable. An interaction between AM tier and phase shift would indicate that a phase-shift effect occurred in some AM combinations, but not in others.

The ANOVA showed a significant main effect of AM tier ($F(4,88) = 5.98$, $p < .0001$), and a significant main effect of phase shift ($F(2,44) = 11.3$, $p < .0001$), but as previously, no significant effect of AM extraction method ($F(1,22) = 3.71$, $p = .067$), although this came close to significance. The predicted interaction between AM tier and phase shift was significant, $F(4.93,108.35) = 4.78$, $p < .0001$, Greenhouse-Geisser epsilon = 0.62. A Tukey-HSD post hoc analysis was used to compare differences between 0 and 1π shifts, and 1π and 2π shifts respectively for each AM combination. Tukey post-hoc tests showed that significant effects were limited to the Stress+Syllable AM tier combination. The phase shift effects in

the Stress+Syllable AM tier combination occurred exactly in the direction predicted by the AM Phase Hierarchy model, namely a significant drop in accuracy with a 1π-shift, but a significant recovery of accuracy with a 2π shift.

Figure 3.4. Effect of phase-shifting on accuracy, scores averaged across PAD and MFB methods.



Indeed, there was no significant difference in performance between 0 and 2π-shifted Stress+Syllable AM stimuli. This showed that the rhythm information in 2π radians phase-shifted stimuli was equivalent to that in non-phase-shifted nursery rhymes, despite the acoustic distortions introduced by phase-shifting. Recall from Figure 3.2 that the non-shifted and 2π-radians shifted stimuli were actually completely different in modulation pattern. Therefore, for participants to achieve a statistically-equivalent performance in both conditions, they must have been relying on the relative Stress-Syllable phase information - which was the only factor that remained unchanged after the phase-shifting.

Therefore, the results of the phase-shift conditions indicate that, as predicted, listeners were relying on the phase relationship between Stress and Syllable AMs to make rhythm pattern judgments. No other AM tier or tier combination showed the predicted phase-dependent 'V-shaped' response, indicating that rhythm judgment for other AM tier

combinations was *not* phase-dependent. Although the Stress tier by itself also showed the predicted 'V-shaped' pattern, this drop and recovery was not statistically-significant, suggesting that additional Syllable-rate information was required for phase-coding of rhythm to operate in full.

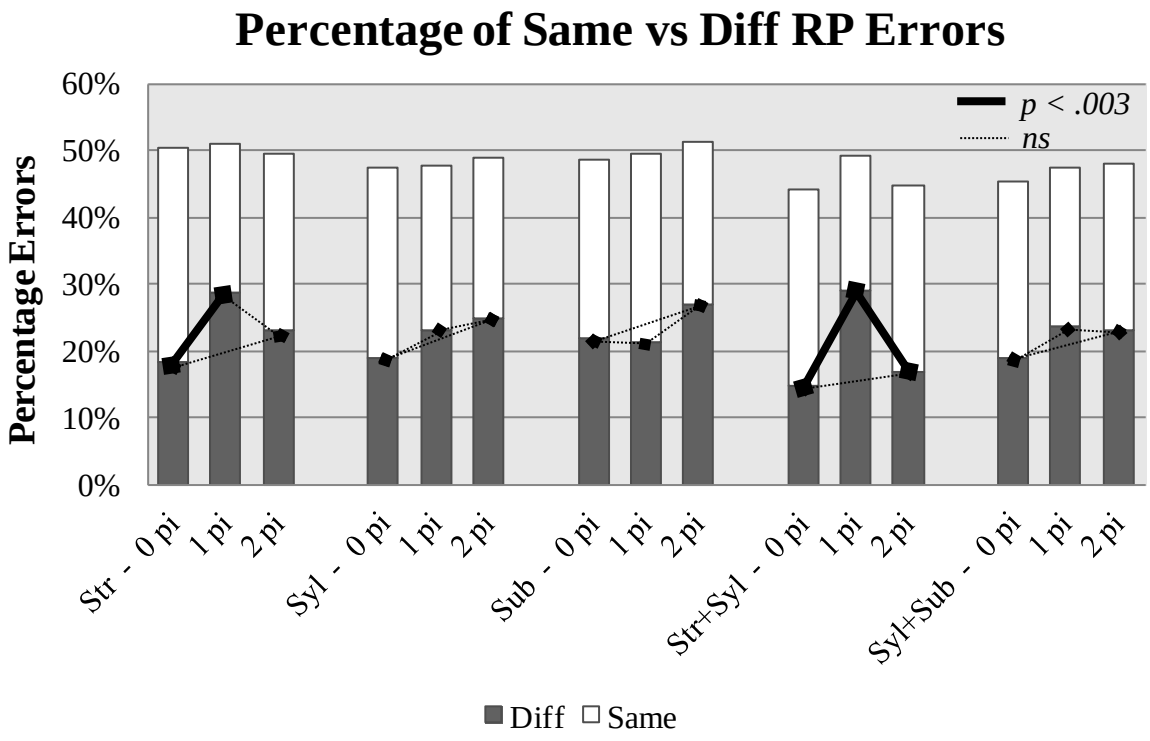
3.2.2.2 Confusion Errors

It is also important to examine the pattern of confusion errors produced by phase-shifting. A different pattern of errors should now be observed, relative to the non-phase-shifted stimuli, as the perceived rhythm patterns (RPs) should change systematically with the degree of phase shift. In particular, if participants were misled by the 1π radians phase-shift into perceiving the opposite RP to the actual rhythmic pattern of a given nursery rhyme, they should now show more Different RP confusions. For example, "MA-ry MA-ry" originally had the same rhythm pattern as "SIM-ple SI-mon". However, the 1π radians stress phase-shifted version ("ma-RY ma-RY") would now be rhythmically similar to "the QUEEN of HEARTS" and "as I was GO-ing...". Hence, if participants were using phase-information for rhythm pattern perception, they should now make more Different confusions for the 1π radians shifted stimuli, but continue to make more Same confusions for the 2π radians shifted stimuli (as these nursery rhymes should sound equivalent to non-phase-shifted nursery rhymes).

The percentage of Same and Different confusions produced by phase-shifting each AM tier is shown in Figure 3.5. Paired t-tests were used to compare the percentage of Different confusions made for zero phase shift vs 1π shift, 1π shift vs 2π shift, and zero phase shift vs 2π shift for each AM tier. This necessitated 15 comparisons, hence a Bonferroni-corrected p-value of 0.003 (0.05/15) was used to determine significance. For the 0π-1π shift, the only significant differences occurred for the Stress AM tier and the Stress+Syllable AM tiers, where significantly more different confusions were made under a 1π radians phase shift than under no phase shift (see Figure 3.5). For the 1π-2π shift, there were now significantly less different confusions for the Stress+Syllable AM tier, but not for the Stress AM only tier. For the 0π-2π shift, there were no significant differences in any AM tier or tier combination. Hence, as predicted by the AMPH model, phase-shifting the Stress AM (either alone or in combination with the Syllable AM) by 1π radians resulted in participants confusing nursery rhymes that had opposite metrical structure (trochaic versus iambic). The proportion of different confusions fully returned to baseline with a 2π phase-shift for the Stress+Syllable

AM tier, but showed only intermediate recovery for the Stress only AM tier. In contrast, phase-shifting Syllable AMs and Sub-beat AMs had no systematic effect on the nature of rhythmic confusions.

Figure 3.5. Effect of phase-shifting on percentage of same versus different Rhythm Pattern confusion errors, scores averaged across PAD and MFB methods.



3.2.2.3 Multi-Dimensional Scaling (MDS)

In a final analysis step, participant's response patterns in the 0π, 1π and 2π radians phase shift conditions were used as the basis for multi-dimensional scaling (MDS). The aim was to represent participants' perception of the rhythmic similarity between nursery rhymes as a map in 'perceptual space'. The shape of these maps would indicate whether rhythmic perception changed systematically as a function of phase shift. To construct these maps, 4 x 4 confusion matrices were computed from participants' responses. These confusion matrices capture information about how often one sentence is confused for another, providing a rich source of information about the structure of participants' psychological representations of the stimuli (Shepard, 1972).

Table 3.3 shows an example of a 4 x 4 confusion matrix produced for the Stress+Syllable AM tier combination, in the no phase-shift condition. In the table, PAD and MFB responses were averaged for simplicity of inspection, but these were computed separately in the actual analysis. Table 3.3 shows that participants made more confusions *within* sentences with the same rhythm pattern than *across* sentences with different rhythm patterns. For example, when participants heard the nursery rhyme 'St Ives', they responded that they had heard 'St Ives' 36% of the time (correct response). They chose an incorrect response with the same rhythm pattern ('Queen of Hearts') a further 32% of the time, but they only chose responses with a different rhythm pattern ('Mary Mary') 16% of the time each. Confusion matrices were computed for each of the 5 AM tier combinations, 3 phase-shift conditions, and 2 AM extraction methods.

Table 3.3. Example of a confusion matrix for Stress+Syllable AMs, in the no phase-shift condition. Grand averages over 23 participants are shown

(Values shown in the table are response percentages)			**RESPONSE**			
			RP 1		**RP 2**	
			Mary Mary	Simple Simon	St Ives	Queen of Hearts
STIMULUS (Sentence presented)	**RP 1**	Mary Mary	**62.7%**	<u>15.7%</u>	12.7%	8.7%
		Simple Simon	<u>42.7%</u>	**29.1%**	13.1%	14.9%
	RP 2	St Ives	16.2%	15.7%	**36.2%**	<u>32.2%</u>
		Queen of Hearts	14.0%	20.5%	<u>28.3%</u>	**37.5%**

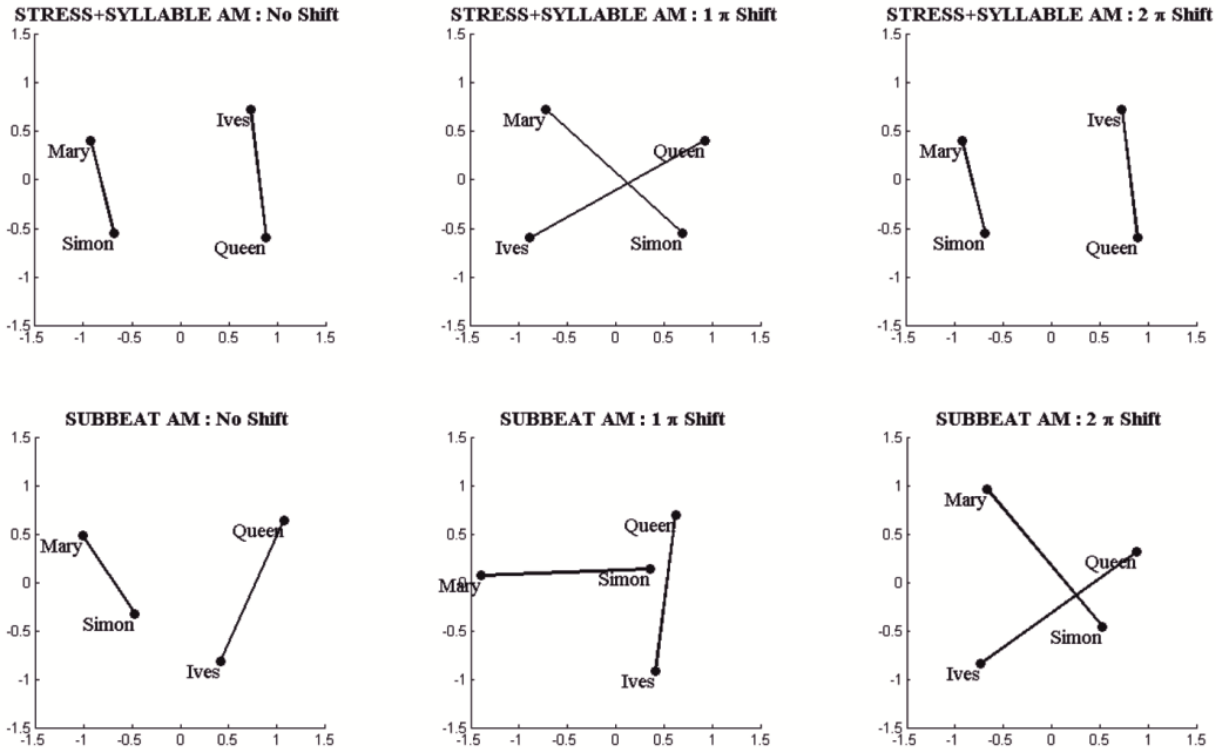
bold = correct response

<u>underline</u> = confusion within same Rhythm Pattern group

These confusion matrices were then converted into 'psychological' similarity maps using multidimensional scaling (MDS). In these maps, the psychological proximity (similarity or confusability) between the various nursery rhyme sentences was represented in

terms of the spatial proximity (distance) between points in 2-dimensional[9] space, where each point represents one nursery rhyme sentence. Sentences that are more similar are mapped closer together, whilst sentences that are more dissimilar are mapped further apart. The MDS maps obtained for the Stress+Syllable AM tier combination and the Sub-beat AM tier across the three possible phase-shifts (0, 1π radians, 2π radians) are shown in Figure 3.6. These two AM tiers were selected for display because the data so far indicate phase-shift effects for the Stress+Syllable AM tier combinations but not for the Sub-beat AM tier alone. In Figure 3.6, sentences with the same rhythm pattern are joined with a black line.

Figure 3.6. MDS maps for the three phase-shift conditions (columns), for Stress+Syllable AMs (top row) and Subbeat AMs (bottom row). Since only the distance between points is meaningful and not their absolute position, MDS solutions were reflected about the x- or y-axis before overlay to allow for easy visual comparison between conditions. Lines in the plot join sentences with the same Rhythm Pattern. Note that the MDS solutions obtained for the Stress+Syllable 0 shift and 2π shift conditions were identical even through their respective response matrices were different.



[9] A 2-dimensional representations for the MDS solution was chosen because a 1-dimensional representation provided a poor fit for some matrices (goodness-of-fit 'stress' values >0.1). On the other hand, 2-dimensional representations provided a good fit for all matrices (goodness-of-fit 'stress' values <0.001).

In the no phase shift condition, sentences with the same rhythm pattern should be located closer to each other, than to sentences with the opposite rhythm pattern. Therefore, in the MDS maps, the distance between *connected* points (i.e. the length of the black line) should be shorter than the distance between *unconnected* points. For the Stress+Syllable AM combination in the no-phase-shift condition (top left subplot), this was indeed the case. For example, participants mapped the nursery rhymes 'Mary Mary' and 'Simple Simon' closer to each other than to the other two nursery rhymes ('St Ives' and 'Queen of Hearts'). However, for the Subbeat AM combination in the no-phase-shift condition (bottom left subplot), this systematic grouping of sentences by rhythm pattern was *not* observed. For example, 'St Ives' was mapped very far from 'Queen of Hearts' (the same RP), but very near to 'Simple Simon' (the opposite RP).

With a 1π radians shift, each nursery rhyme should now move systematically closer to nursery rhymes in the *opposite* rhythm group, so that the distance between *unconnected* points is now shorter than the distance between *connected* points. This indeed occurred for the Stress+Syllable AM combination (top middle subplot), as shown by the 'crossed' map that was produced. For example, the sentence 'Simple Simon' was now closer in space to 'St Ives' and 'Queen of Hearts' (from the opposite rhythm group) than to 'Mary Mary' (from the same rhythm group). Recall that the maps were based on confusion matrices, therefore this pattern indicates that 'Simple Simon' was now more often confused for 'St Ives' and 'Queen of Hearts' than for 'Mary Mary'. Importantly, this map indicates that participants *systematically* (rather than randomly) misperceived the rhythm pattern of each sentence so that their responses were now completely opposite to the no phase shift condition. Therefore the significant drop in accuracy observed in the 1π radians phase shift for the Stress+Syllable AM combination (see Section 3.2.2.1) was not due to random error, but due to a genuine and systematic shift in participants' perception of rhythm. By contrast, the MDS map for the Subbeat AM in the 1π-phase-shift condition (bottom middle subplot) showed *random* re-alignment of the sentences in psychological space, which was not accompanied by any significant change in participants' accuracy performance (see Section 3.2.2.1).

Finally, with a 2π radians shift, the maps should now be restored to look like that of the original no phase shift condition. For the Stress+Syllable AM combination (top right subplot), this restoration occurred exactly as predicted, consistent with the recovery in accuracy performance observed in Section 3.2.2.1. In contrast, the MDS map for the Subbeat

AM (bottom right plot) was *not* restored to look like the no-phase-shift condition, but took on a completely different configuration.

Therefore, for the Stress+Syllable AM combination, participants' psychological maps occurred exactly as predicted for each phase-shift condition. Crucially, their maps in the 1π shift condition showed a *systematic* re-alignment consistent with an orderly shift in rhythm perception, rather than an *unsystematic* re-alignment produced by random error (e.g. if participants did not know what they were hearing and responded at chance). In contrast, the MDS maps for the Sub-beat AM did not show the predicted grouping patterns, or phase-shift effects. Hence for the Sub-beat AM stimuli, participants were not using metrical rhythm pattern to group nursery rhymes. Indeed, they appeared to be grouping the nursery rhymes at random when listening to Sub-beat AMs only.

3.2.2.4 Summary of the Phase Shift Data

When phase shifts were used to test the 'Stress Phase Code' proposed in the AM Phase Hierarchy model, it was clear that listeners' metrical rhythm perception *did* depend on the relative phase relationship between Stress and Syllable AM tiers. Performance for the Stress+Syllable AM tier combination demonstrated *all* the predicted pattern of changes. No other AM tier or tier combination demonstrated these predicted changes. When the accuracy of identifying nursery rhymes was the outcome measure, only the Stress+Syllable AM tier combination produced the predicted 'V-shaped' pattern of a sharp decline in performance (1π radians shift), followed by full recovery (2π radians shift).

When the proportion of same:different RP confusion errors was considered, it was shown that participants were more likely to confuse nursery rhymes with the *opposite* metrical pattern when Stress+Syllable AMs were phase-shifted by 1π radians. When MDS was used to map the similarity of nursery rhymes in perceptual space, the similarity maps for the Stress+Syllable AM combination were systematically altered according to the degree of phase-shift. Hence, listeners' perceptual experience of metrical rhythm was contingent on the phase relationship between the Stress AM and the Syllable AM, in a manner predicted by the Stress Phase Code.

3.3 CHAPTER SUMMARY

For both non-phase shifted and phase-shifted stimuli, participants performed exactly as expected by the AMPH model. For the non-phase-shifted stimuli, participants produced the best metrical rhythm judgments when hearing Stress+Syllable AMs, as compared to all the other AM tiers and tier combinations, suggesting that these two tiers contained the most metrical rhythm information. For the phase-shifted stimuli, participants showed the expected drop-and-recovery pattern *only* when the Stress+Syllable AMs were shifted by 1π- or 2π-radians respectively. No other AM tier or tier combination produced this characteristic cyclical pattern of responding. Moreover, the analysis of participants' pattern of confusion errors by MDS mapping indicated that the drop in performance for the 1π-radians Stress+Syllable AM shift was not a result of random error (e.g. uncertainty and guessing). Rather, the performance drop was generated by a *systematic* shift in participants' perceptual mapping of rhythm patterns. Therefore, only for Stress+Syllable AM stimuli, participants' rhythm perception was being systematically altered by the degree of phase shift between the two tiers. This would only occur if, like the AMPH model, listeners were using the Stress-Syllable phase relationship to infer the metrical rhythm pattern of the sentences.

Therefore, these empirical listening data provide strong support for the two main tenets of the AMPH model. These are that (1) Speech rhythm information is primarily carried by Stress- and Syllable-rate AMs within the speech envelope; and (2) Speech rhythm patterns are 'coded' via the phase relationship between Stress and Syllable AMs. In these two ways, the AMPH model replicates the behaviour of human listeners when computing the metrical rhythm pattern of a sentence.

PART II SUMMARY & DISCUSSION

In Chapter 2, the AMPH model was presented as a signal-based method for computing the prosodic rhythm information conveyed by a speaker from the amplitude modulation structure of their speech. As demonstrated by the results of the tone-vocoding experiment in Chapter 3, the AMPH model made psychologically-valid assumptions, since human listeners also perceived prosodic rhythm using the same amplitude modulation cues as the AMPH model. For example, listeners were most accurate in making metrical rhythm judgments when hearing the combination of Stress+Syllable AMs, indicating that speech rhythm information resided primarily at these two modulations rates in the envelope (Research Question 1). Listeners were also found to base their judgments of metrical rhythm pattern on the *phase* relationship between Stress and Syllable AM tiers, since their judgments were systematically altered by phase-shifts of this relationship (Research Question 2). Therefore, the AMPH model is a psychologically-relevant way to represent speech rhythm using amplitude modulation patterns in the speech envelope.

The Amplitude-Based AMPH Model and Durational Measures of Rhythm

The AMPH model represents an advance in both methodology and theory for speech rhythm research. The hierarchical AM approach is a new method for isolating the rhythmic properties of speech from its phonetic content, and may be useful as an alternative research tool to traditional 'rhythm-metrics' or speech re-synthesis. Furthermore, the data support the view that rhythm patterns in speech are not found in durational isochronies. Rather, rhythms lie in the statistics of temporal structure, such as phase relationships between modulation rates.

One significant advantage of coding rhythm via AM phase patterns is that such coding would be robust to random durational variations when the speaker speeds up or slows down - a major contributor to anisochrony in speech. Phase relationships between Stress and Syllable AMs within a particular stress foot can remain constant even if the overall duration of the foot varies from foot to foot. This is because once phase-locked, both Stress and Syllable oscillatory cycles will compress or stretch in synchrony. Consequently, if listeners use phase relationships to detect rhythm patterns, they will still be able to perceive rhythm patterns even when the inter-stress duration (the length of prosodic feet) is not isochronous.

This feature of phase-coding may help to reconcile conflicting views concerning the quasi-rhythmic structure of speech (Pike, 1945; Abercrombie, 1967) despite its anisochrony (Dauer, 1983).

However, it is important to note that the amplitude-based rhythm information in the speech envelope is *complimentary to* the other acoustic sources of rhythm information - such as durational variation, or pitch variation. Therefore, the success of the AMPH model does not invalidate previous attempts at describing rhythm through durational variation (i.e. rhythm-metrics). Rather, it serves to illustrate that *multiple* sources of rhythm information are present in the acoustic signal of speech.

For example, durational contrasts within the prosodic foot are expected to work co-operatively with amplitude changes in the phase coding of rhythm when both co-vary (as is usually the case in natural speech, e.g. Kochanski et al, 2005). Although not explicitly manipulated in this experiment, increasing the length of the stressed syllable compared to the unstressed syllable within the same prosodic foot would also increase the temporal peak-to-peak separation of both syllables. This would support their encoding as two separate syllables rather than one long syllable by reducing their energy overlap in time. Therefore duration and intensity cues have complementary roles in rhythm perception and syllable prominence.

The AMPH Model and Speech Segmentation

An additional novel feature of the AMPH model is that syllable beats are detected automatically, enabling a range of possible segmentation schemes to emerge from computed prosodic patterns 'for free'. Hence prior lexical knowledge about the semantic content of the sentence is not required for efficient parsing, although clearly as lexical knowledge is acquired it will support parsing via 'top-down' processing. If human speech perception is indeed tuned to AM hierarchies, then the AMPH approach would enable naïve listeners (e.g. infants) to use prosodic patterns for speech segmentation in the absence of lexical knowledge. Furthermore, the hierarchical nature of the AMPH model resonates strongly with classical models of poetic (Liberman & Prince, 1977) and musical (Lerdahl & Jackendoff, 1983) rhythm and meter. This synergy may suggest that intuitions about the hierarchical nature of metrical structure arise from the fundamentally hierarchical nature of the speech signal itself.

The AMPH Model and Neural & Articulatory Mechanisms

The AMPH model was also motivated by the proposal that speech-to-brain temporal structure mapping is important for speech perception (e.g. Poeppel, 2003; Giraud & Poeppel, 2012). By this view, the brain detects and represents temporal structure in the acoustic environment, drawing from this temporal structure key regularities and statistics that define the perceptual experience of sounds. Focusing on the amplitude envelope as a primary source of this temporal structure, the AMPH model captures highly-ordered patterns of amplitude modulation that affect how listeners experience the *rhythmic* structure of speech sounds. In this sense, the AMPH model reveals latent acoustic temporal structure that could be mapped onto (entrain) neuronal oscillatory patterns in order to generate a percept of rhythm.

Furthermore, since speech is produced by motor articulators like the jaw, lips and tongue, the oscillatory AM tiers and patterns described could well correspond to these motor articulators and their actions (e.g. Tilsen, 2009). Therefore, the phase statistic investigated here could also be an important index for motor synchronisation and timing (e.g. Cummins & Port,1998; Port, 2003), since synchronised motor actions are producing these phase-locked AM patterns in the acoustic signal. Consequently, the AM hierarchical representation of speech rhythm also fits well with the *articulatory mechanisms* that could be generating and synchronising these acoustic AM patterns.

Limitations of the AMPH Model

The AMPH model is conceptually simple, theory-driven, and is psychologically-relevant to how listeners perceive rhythm patterns in metrically-produced speech. Nevertheless, the current AMPH model is also limited in several ways. As noted previously, it does not take into account the possible contribution of non-intensity-related acoustic cues for rhythm perception such as pitch or duration. Second, the AMPH model uses AM hierarchies that are derived from demodulation of the wholeband speech signal. To more accurately reflect physiological processes in the cochlear, where the speech signal is effectively split into multiple frequency channels, a more complete approach may use AMs from the envelopes of *each* frequency channel, using a weighted procedure to combine these 'sub-band' AMs for rhythm calculation. Finally, the AMPH model was developed and tested exclusively using metronome-timed and metrically-regular (nursery rhyme) speech. It would be important to test the model with freely-produced speech, to see if the model is able to

'scale-up' to the challenges presented by such speech. These short-comings of the AMPH model are addressed in Part III of the thesis.

PART III :

THE NEW SPECTRAL AMPH MODEL (S-AMPH)

Chapter 4 : A New Spectro-Temporal Representation of the Amplitude Envelope

Chapter 5 : New Prosodic Indices

Chapter 6 : Functional Evaluation of the S-AMPH & AMPH Models

MOTIVATIONS FOR A NEW MODEL

The original AMPH model used a simple 5-tier hierarchical representation of the wholeband speech envelope. This model demonstrated that metrical rhythm patterns arise from the phase relationship between 'Stress' and 'Syllable' rates of amplitude modulation (AM). While this simple model yielded a relatively good description of metrical structure in a sample of regularly-timed speech, it relied on various simplifications concerning the complex spectro-temporal structure of speech. While this was a necessary expediency to constrain the problem space addressed during a first attempt at such a model, the AMPH model relied on a set of theoretical assumptions which can be questioned.

Therefore, a new Spectral AMPH (S-AMPH) model was derived. This new model makes two major improvements to the original AMPH model. First, syllable (vowel nucleus) detection is improved by using a more complex spectral sub-band representation of the amplitude envelope, instead of using the wholeband envelope. Second, the tier-structure of the AM hierarchy is derived in a 'bottom-up' fashion from the modulation statistics of speech, rather than being decided on the basis of theoretical assumptions.

1. USING SPECTRAL (SUB-BAND) ENVELOPES INSTEAD OF THE WHOLEBAND ENVELOPE TO DETECT SYLLABLE VOWEL NUCLEI

Different types of speech sounds contain energy (modulation) at different spectral frequencies. For example, a vowel sound such as /a/ will typically contain the most energy around 1000 Hz (corresponding to the first two formants), while a fricative sound such as /s/ will contain energy at much higher frequencies around 5000 Hz. In the wholeband envelope of speech, it is not possible to determine whether a particular modulation pattern was produced by a low-frequency sound (like /a/) or a high-frequency sound (like /s/), because the wholeband envelope represents the sum of all the energy across all speech frequencies at each point in time.

This is a problem because not all of the speech sounds that elicit strong amplitude modulation are equally important for determining speech rhythm. For example, rhythmic beats ('p-centres') in speech are most strongly associated with the onsets of syllable vowel

nuclei (Morton et al, 1976; Allen, 1972; Scott, 1993). Therefore, a listener hearing the word *"cats"* would perceive it as having just one rhythmic beat (p-centre), located around the onset of the vowel /a/. However, the wholeband envelope of the utterance *"cats"* contains not one, but two major peaks - the first occurring at the vowel /a/, and the second at the consonant /s/. In the original AMPH model (which used the wholeband envelope), this presented a problem, because there was no way of separating out two such peaks, and no way to evaluate which peak actually corresponded to the rhythm-bearing vowel nucleus. In order to separate out the two modulation peaks associated with /a/ and /s/, one would have to obtain two different 'sub-band' envelopes corresponding to low (~1000 Hz) and high (~5000 Hz) frequency bands in speech respectively. In this case, the rhythm of the word would be given by the modulation pattern of the low frequency sub-band envelope, not the high frequency sub-band envelope.

For this reason, p-centre research has typically focused on a single narrow sub-band of speech frequencies corresponding to the lower (fundamental, first or second) formants of spoken vowels. For example, Cummins & Port (1998) used a spectral band of 700-1300 Hz for their envelope-based p-centre analysis of sentences such as *"big for a duck"*. Patel et al (1999) used a spectral band of 390-1015 Hz for their p-centre analysis of CV syllables containing the vowel /a/ or /i/.

This single sub-band approach is efficient when the stimulus set consists either of single syllables or short sentences containing just a few vowels with similar formant frequencies. However, when the stimulus set consists of long sentences with a diverse range of vowels (as was the case here), this approach is not as satisfactory. First, vowels with formant frequencies that are higher or lower than the selected range could be omitted. For example, the vowel /i/ with a typical first formant of 342 Hz and a second formant of 2322 Hz (for a male speaker, Hillenbrand et al, 1995) would not be represented in the frequency ranges used by Cummins & Port (1998) or Patel et al (1999). Second, speakers may vary the way they produce a particular vowel in different parts of the utterance, as a result of prosodic stress or emphasis. This could result in the vowel formants becoming higher or lower than expected, and again being omitted from the single sub-band representation.

One solution to this problem is to make the bandwidth of the single sub-band very wide in order to accommodate large variations in vowel formant frequency (e.g. 300-3000 Hz rather than 700-1300 Hz). However, in doing this, other non-vowel speech sounds could also be included into the sub-band, masking the actual temporal pattern of the vowels. Another solution is to use dynamic formant-tracking, where the frequency of the formants is tracked

on a moment-by-moment basis, allowing for adjustments to be made dynamically to the bandwidth of the sub-band depending on the speech sounds being uttered. However, this approach relies heavily on the success of the formant-tracking algorithm, and could yield spurious results during sections of speech when the formants are not clearly identifiable.

Here, a different approach is used to identify syllable vowel energy in the frequency spectrum. The frequency spectrum is divided into multiple sub-bands, these sub-bands are of an 'optimal' bandwidth so that each sub-band captures mutually non-redundant modulation patterns. Each sub-band (henceforth simply called 'spectral band') provides a set of 'candidate' syllable vowels, which are peaks in the envelope of that spectral band. These candidate peaks are then evaluated according to set criteria, resulting in the final set of peaks deemed to correspond to actual syllable vowels. Since the spectral bands cover the entire frequency range (100 to 7250 Hz), variations in vowel frequency simply result in the vowel being represented in a different spectral band, rather than being omitted entirely.

Chapter 4, Section 4.3 describes the procedure used to determine the 'optimal' division of the frequency spectrum into spectral bands. This was done via PCA which uses the underlying correlation structure of the speech signal. Therefore, there is no claim or requirement that these spectral bands should correspond to the formant frequencies. Rather, these bands are simply a parsimonious way to represent the spectral variation in speech. Chapter 5, Section 5.2 describes the procedure used for detecting and selecting the candidate peaks from each spectral band, resulting in the final syllable vowel pattern. Chapter 6, Section 6.1 evaluates the effectiveness of this automatic procedure for syllable vowel detection, using two corpora of metronome-timed and untimed speech. Syllable vowel detection accuracy levels of over 80% for untimed speech, and over 97% for metronome-timed speech were achieved.

2. Using a data-driven Instead of a Theory-Driven AM hierarchical structure

The second improvement in the S-AMPH model is a new data-driven derivation of the AM hierarchical structure. In the original AMPH model, five AM rate bands had been proposed based on the rationale that different linguistic units should correspond to different rates of modulation in the speech envelope. It was then demonstrated in the tone-vocoder experiment (Chapter 3, Section 3.2.1) that two of these modulation bands, corresponding to

the prosodic stress foot ('Stress AM') and the syllable ('Syllable AM'), were most important for speech rhythm perception. However, the functional significance of the three remaining modulation bands within the AM hierarchy ('Slow', 'Subbeat' and 'Fast') was unclear since these were not used in the AMPH model's computation of rhythm. It was possible that these three bands reflected other types of linguistic information (i.e. related to phrasing for the Slow band, or to phonemes for the Fast band). However, it was also possible that the original division of modulation bands was not an accurate reflection of the 'true' information content within the modulation spectrum. That is, modulation bands could have been artificially created to comply with theoretical assumptions where there in fact were none.

To address this issue, a new AM hierarchy structure was derived for the S-AMPH model. Rather than deciding the number of AM tiers and the bandwidths of these tiers in a theory-driven 'top-down' fashion, the new AM hierarchy was allowed to emerge in a 'bottom-up' manner from the statistics of the modulation spectrum. Starting with an original set of 24 finely-spaced modulation channels, a PCA procedure was used to dimensionally-reduce these to an appropriate number of non-redundant *modulation rate* bands. Three modulation rate bands emerged from this process, corresponding to 'Stress', 'Syllable' and 'Phoneme' rates of amplitude modulation (discussed in Chapter 4, Section 4.4). These 3 new modulation rate bands, emergent from the structure of the speech data itself, then formed the new S-AMPH model's AM hierarchy.

OVERVIEW OF PART III STRUCTURE

Part III of the thesis is divided into 3 chapters. **Chapter 4** explains the derivation of the new spectral bands and modulation rate bands using PCA, in two dimensionality-reduction exercises. First, the spectral dimensionality reduction process is described, which resulted in 5 spectral bands. Next, the modulation rate dimensionality reduction process is described, which resulted in 3 AM tiers, forming the new AM hierarchy. Together, these processes resulted in the new 5 x 3 spectro-temporal representation of the speech envelope used in the S-AMPH model.

In **Chapter 5**, the new procedures for identifying syllable vowel nuclei and computing prosodic prominence, based on the new 5 x 3 representation, are described. The new Prosodic Strength Index (equivalent to the AMPH Stress Phase Code) is introduced.

In **Chapter 6,** the new S-AMPH and original AMPH models are functionally evaluated in terms of success in (1) automatic syllable vowel identification, and (2) automatic prosodic stress transcription. The original AMPH model had been developed and tested exclusively using metronome-timed speech. While this type of speech represents the rhythmic 'ideal', since the prosodic template is known and produced exactly in an isochronous fashion, it lacks ecological validity. Spontaneous human utterances are neither isochronous nor perfectly metrically-regular. Hence, while metronome speech can be used for demonstrations 'in principle', it is important to see the extent to which any model can 'scale-up' to meet the challenges of real speech. With this in mind, the AMPH and S-AMPH models were tested on both metronome-timed and freely-produced (un-timed) speech.

4 A NEW SPECTRO-TEMPORAL REPRESENTATION OF THE AMPLITUDE ENVELOPE

The original AMPH model had used the wholeband amplitude envelope to derive the modulation hierarchy. In fact, it is known that the human cochlea splits the speech signal into multiple frequency channels. To model the frequency resolution seen in the normal human listener, an estimated 29 frequency channels are needed within the range of 100-7250 Hz (ERB_N-spacing, Glasberg and Moore, 1990; Moore, 2012). For a complex time-varying sound like speech, each of these cochlear channels transmits its own temporal fine structure carrier, modulated by its own amplitude envelope.

The speech envelope patterns in the various cochlear channels are not identical, since at any instant (a) speech sounds do not consist of cross-sectionally equal power at all frequencies, and (b) the essential information in speech involves changes in power over time that differ between frequency channels. In other words, speech sounds selectively activate groups of spectral channels for brief periods of time, and their relative pattern of activation continuously changes over time. This relative pattern of activation across channels is captured in their individual temporal amplitude envelopes. By taking only the wholeband envelope, the original AMPH model in effect used a power-weighted summation of these individual cochlear channel envelopes. Since low-frequency channels tend to have higher power than high-frequency channels, the wholeband envelope is dominated by these low-frequency components, similar to low-pass filtered speech (e.g. speech heard by fetuses in the womb). Therefore, while the wholeband envelope foregrounds prosodic rhythm, the information about relative patterns of channel activation that contributes toward speech intelligibility (including distinguishing between vowel and non-vowel sounds) is consequently discarded and unused.

For example, when only the wholeband envelope is used to vocode speech (i.e. a single-channel vocoder), the resulting vocoded speech is completely unintelligible. As the frequency spectrum is divided into more spectral bands and these sub-band envelopes are used for vocoding, speech intelligibility increases accordingly. A classic study by Shannon et al (1995) demonstrated that vowel and sentence recognition scores of 80% could be achieved

when only three spectral bands were used for noise-vocoding. Furthermore, two bands (i.e. binary information) were already sufficient to convey 90% of the information about the voicing and manner of a medial consonant (i.e. a/C/a), although information about place of articulation required more spectral bands. These results indicate that the information from individual cochlear channels is highly mutually redundant, since ~24 finely-spaced cochlear channels (Shannon used a cut-off of 4 kHz) can effectively be 'collapsed' into just 3 broad spectral bands with relatively modest loss to intelligibility. Therefore, to include spectral patterns of variation into the S-AMPH model, one need not use an exceedingly high-dimensional representation of speech rhythm (e.g. 29 cochlear channels x 5 AM tiers = 145 dimensions). Rather, a smaller number of non-redundant, but broader, spectral bands may be used.

4.1 Using Principal Component Analysis to Identify Non-Redundant Spectral Bands

To achieve this dimensionality reduction in the frequency domain, principal component analysis (PCA) can be applied. The PCA method has long been used for dimensionality reduction in speech, and for identifying spectral differences between speech sounds. For example, Klein et al (1970) used PCA analysis to classify 12 Dutch vowel sounds that had been produced by 50 different male speakers. They found that 4 PCA factors were sufficient to represent 75% of the total variance contained in the 18 original 1/3-octave spectral channels. Pols et al (1973) later used the same dataset to compare the results of PCA-based vowel discrimination with traditional formant-based vowel discrimination. In the formant analysis, they found that the first and second formant frequencies (F1 and F2 respectively) were the most discriminatory acoustic features of the vowels (i.e. the vowels occupied different regions on a log F2 - log F1 plot). Moreover, Pols et al (1973) found that the PCA method (using 3 or more factors) produced similar discrimination performance to the formant analysis. Therefore, principal component analysis is a valid method for speech analysis, and has become widely used in automatic speech recognition algorithms.

The aim of the PCA procedure used in this thesis is to identify boundaries for a new, parsimonious spectral filterbank with only a few channels that have a wide bandwidth. This new spectral filterbank should yield wide spectral bands of speech where the information

within the bands is redundant, but the information *between* the bands is non-redundant. By analogy to vocoding, the new spectral filterbank should contain as few vocoding channels as possible, but still allow for good speech discrimination. Rather than defining the boundaries of the new filterbank by formant frequency (as the formants frequently overlap), a criteria of redundancy is used, based on the PCA procedure.

Therefore, unlike a typical PCA analysis, the key outcome of interest here is <u>not</u> the *component scores* (as in Klein et al, 1970), but the *component loadings*. Component scores are the transformed representations of the data after it has been dimensionally-reduced to its dominant components. However, here the interest is not in the transformed data per se[10], but in *how* that transformation was achieved - that is, what were the underlying patterns of correlation between the channel variables that led to that particular transformation. These underlying patterns of correlation are expressed as the component *loadings*. The component loadings across the channels reflect the similarity or redundancy in the modulation information carried by each cochlear channel. Cochlear channels that carry similar (redundant) patterns of amplitude modulation should also be strongly correlated, and tend to have similar loading (i.e. strongly positive or negative) on the various PCA components that are extracted. Conversely, cochlear channels that carry different (non-redundant) patterns of modulation should be poorly correlated, and should show different loading patterns on the PCA components. Therefore, by analysing the patterns of component loading across the channels, one should be able to identify groups of channels that carry similar (redundant) modulation information, and therefore may be considered as belonging to the same wider spectral 'band' - which is the aim of the exercise.

For example, suppose that cochlear channels 1-3 are all carrying one pattern of modulation, and cochlear channels 4-6 are carrying another different pattern of modulation. When we carry out the PCA analysis, we might see that for PCA component 1, the loading across cochlear channels 1-6 shows a pattern of '- - - + + +'. For PCA component 2, this loading pattern might switch to the opposite as '+ + + - - -'. Therefore, groups of channels carrying similar modulation patterns will show similar loadings for each PCA component. Furthermore, the boundary *between* these two groups of redundant channels can be inferred where the loading pattern changes abruptly, quickly becoming more positive or more

[10] For example, if one were to subsequently filter or perform phase computations on these component scores (as part of the rhythm identification process), it would not be clear how the results would relate back to the original modulation patterns in the acoustic signal.

negative[11]. In this example, this abrupt change occurs between channels 3 & 4. Therefore, spectral regions where the PCA loading pattern *remains stable* should indicate redundancy, whereas spectral regions where the loading pattern *changes* should indicate boundaries between non-redundant spectral bands. These boundary values can then be used to construct a new spectral filterbank to filter speech into non-redundant spectral bands.

4.2 SPEECH MATERIAL USED FOR DERIVING THE NEW SPECTRO-TEMPORAL REPRESENTATION

Again, the speech material used to derive the new model was nursery rhymes. A large multi-speaker corpus of children's nursery rhymes was used to generate the underlying statistics for the revised S-AMPH model. There were 44 different nursery rhymes, each produced by 6 female native British English speakers, giving a total of 264 spoken samples. Over the 6 speakers, the nursery rhyme samples varied between 15.9 and 52.2 seconds in length (average 28.2 seconds). Samples were digitally recorded using a TASCAM digital recorder (44.1 kHz, 24-bit).

The full list of these nursery rhymes, and a brief description of their metrical patterns is provided in Appendix 4.1. The text for the rhymes was compiled from children's books such as 'This Little Puffin' (Puffin Books, 1991) and 'Nursery Treasury' (Miles Kelly Publishing, 2010). The nursery rhymes were freely-produced by the speakers (not metronome timed) in a child-directed style of speaking (CDS). As described in the Introduction (Section 1.11), CDS is an exaggerated prosodic register characterised by higher pitch, smoother and wider pitch excursions, and a slower rate of speaking with more pauses (Fernald, 1989; Fernald & Simon, 1984). It was expected that the rhythm and prosody of the nursery rhymes would be enhanced by the use of this speaking register, making the underlying statistical distributions for these rhythms more distinct.

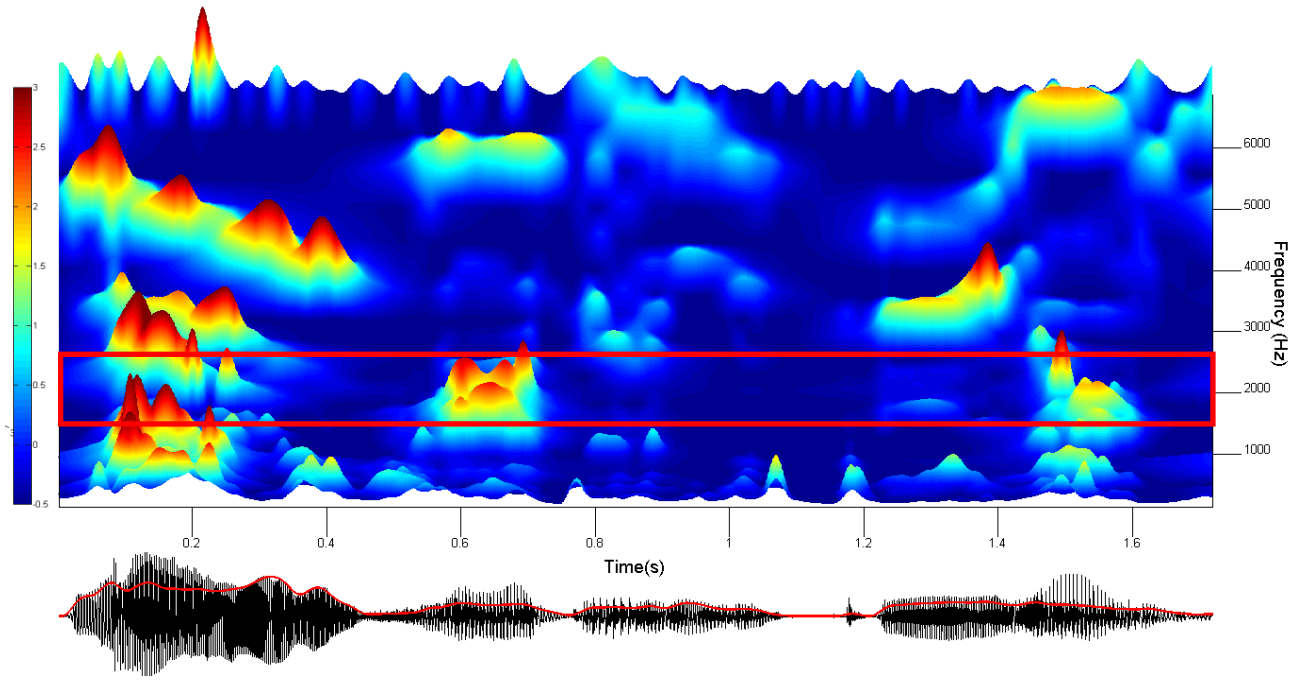
[11] Pols et al (1973) refer to this method of identifying steep slopes in component loading, noting that the slopes in their data corresponded well with F1 and F2 boundaries.

4.3 Spectral Dimensionality Reduction

4.3.1 The Original 'High Dimensional' Cochlear Channel Representation

The aim of the spectral dimensionality procedure was to identify adjacent 'cochlear channels' that had similar modulation patterns so that these could be grouped into larger spectral bands. To generate these cochlear channels, the speech signal was filtered into 29 ERB_N-spaced frequency channels spanning 100-7250 Hz, mirroring the frequency decomposition that occurs at the cochlea of a normal human listener (Glasberg and Moore, 1990; Moore, 2012). The edges and bandwidths of the spectral filterbank are listed in Appendix 4.2. The Hilbert envelope was extracted from each cochlear channel and low-pass-filtered under 40 Hz. This high-dimensional (29 dimension) representation of the speech envelope may be plotted as a modulation 'landscape', which depicts the modulation power in each cochlear channel, as a function of time. An example of this is shown in Figure 4.1, which shows the modulation patterns for the sentence *"Mary had a little lamb"*. In the plot, red indicates high modulation power, and blue indicates low power.

Figure 4.1. (Top) Envelope patterns across 29 ERB_N-spaced cochlear channels. Data has been interpolated across channels to appear continuous. Y-axis indicates channel frequency and z-axis indicates amplitude (also coded in colour). (Bottom) Sound-pressure waveform with 40 Hz low-pass filtered wholeband Hilbert envelope overlaid in red.



From visual inspection, it may be observed that certain clusters of cochlear channels tend to become activated together or become 'co-modulated' in time. For example, a cluster of approximately 8 channels centred around 2000 Hz (red box) collectively show little activity in the first 200 ms, a time when most other channels are active. Conversely, this cluster becomes active at 600 ms (when there is little activity elsewhere) and again at about 1400 ms. Comparing this to the actual spoken phrase, these two clusters of activity correspond to the vowel [æ] in *"had"* and *"lamb"*. Figure 4.1 illustrates that the speech signal does indeed, as proposed earlier, elicit broadly similar temporal patterns of activity within groups of cochlear channels. If these clusters of cochlear channels are co-modulated in time, they also carry mutually redundant information. Therefore, a dimensionality-reduction procedure such as PCA would be appropriate. In this section, the focus is on the spectral PCA analysis. Appendix 4.3 shows other analyses that were also conducted on the speech data, including plotting the RMS power across the frequency spectrum, and examining the intercorrelations between cochlear channels.

4.3.2 SPECTRAL PCA : COMPONENT LOADINGS

Next, PCAs were carried out in the spectral dimension. A separate PCA was conducted for each nursery rhyme sample and each participant, taking the individual timepoints as observations and the 28[12] cochlear channels as variables. Recall that the aim of the analysis is to define underlying patterns of spectral redundancy by noting which groups of channels (variables) load in a similar fashion onto the extracted PCA components. Component loadings reflect the correlation between the variables (channels) and the extracted principal components. Consequently, if two adjacent channels both load strongly (either positively or negatively) onto a given component, it is assumed that they contain similar (mutually redundant) information about that PCA component.

Each separate PCA for each sample yielded 28 principal components. Each of these components had a different loading pattern across the 28 original cochlear channels. After running all 264 separate PCAs (44 nursery rhymes x 6 speakers), the sets of component loadings for each extracted principal component were averaged, resulting in one grand

[12] In the spectral filterbank, the first channel was a low-pass filter rather than a band-pass filter (see Appendix 4.2). This channel captured the 'DC' component of the speech signal and was excluded from the analysis because of the large amplitudes it contained. Therefore, the PCA was conducted on 28 spectral channels, spanning 137 Hz (rather than 100 Hz) to 7250 Hz.

average set of component loadings per principal component. For this exercise, only the *pattern* of loadings across the channels was important and not the sign (positive or negative). Therefore, the *absolute* loading values were taken to compute the average. This 'rectification' of the component loadings was done to avoid mutual cancellation in the averaging process. This cancellation would occur when the loading pattern had an opposite valence across samples. For example, Figure 4.2 shows the loading patterns that could be observed for the first 6 cochlear channels on principal component 1 for two different samples.

Figure 4.2. Hypothetical example of PCA loading patterns for a single component, across 6 cochlear channels, for two nursery rhyme samples. The vertical position of the '+' and '-' markers indicates greater positivity or negativity respectively.

Cochlear channel :	1	2	3	4	5	6
Nursery rhyme 1 :	+	+	+	-	_	-
Nursery rhyme 2 :	-	_	-	+	+	+

In both nursery rhyme samples, there is evidence of clustering between channels 1-3 and 4-6 in terms of similarity in loading pattern. However, if the raw component loadings were averaged across the two samples, the resulting average would be close to zero for all channels. This mutual cancellation was indeed the outcome when the raw component loadings were averaged across all 264 samples. Therefore, absolute component loadings were used to compute the averages instead.

In fact, a switch in the valence of component loadings can be created simply by rotating the eigenvector basis matrix by 180 degrees, which would not change the variance explained (eigenvalues). As such, the +/- signs in component loadings are in effect arbitrary, the relative pattern of loading across channels is more important. The resulting 'rectified' (absolute valued) loading functions allow channel clustering patterns to be observed, but do not allow the computation of principal component *scores* (which were not used in this analysis).

When interpreting these rectified loading patterns, two features are of interest. First, peaks indicate clusters of cochlear channels that all load strongly onto the given principal component. The higher the peak, the stronger the loading. However, these loadings could

have been either positive *or* negative originally for each sample. Second, troughs are also of interest because they indicate *boundary regions* between groups of spectral channels that behave similarly. These troughs could arise from loading sign changes in the original non-rectified loading patterns (e.g. crossings from positive to negative, or negative to positive, that become inflections after rectification), or else local drops in loading strength. Either way, these troughs indicate the abrupt change that one expects when transitioning from a region where spectral channels all carry similar information, to a region where different information is being transmitted. Consequently, these troughs identify potential boundaries of non-redundant spectral bands (which is the aim of this exercise).

The PCA resulted in 28 principal components, each with a different loading pattern, which was too many patterns to consider simultaneously. Therefore, only the top 5 components were considered. These top 5 components contributed the highest amount of variance individually, and cumulatively accounted for 65% of the total variance. The top 5 component loadings for individual speakers are shown in Appendix 4.4, but here we concentrate on the grand average across all samples and speakers.

To identify 'spectral bands' from the rectified component loading patterns, two criteria were used. First, at least 2 of the 5 principal components should show a distinct peak within that spectral band. Second, at least 2 of the 5 principal components should show troughs at the upper and lower boundaries of that spectral band. Figure 4.3 shows the rectified component loadings for principal components 1-5, averaged across all 264 speech samples. In the figure, the components accounting for more variance (lower numbered) are shown in darker, thicker lines. Peaks and troughs in component loading were located by visual inspection, and troughs indicating the boundaries between spectral bands are marked with red dots in the figure. As shown in the figure, 5 spectral bands were identified that fulfilled the stated criteria of peaks and troughs. The boundaries between these 5 spectral bands are indicated as 4 vertical dotted lines. In most cases, the boundary (vertical line) was easy to determine because the troughs of the different principal components were closely aligned. However, for the boundary between spectral bands 3 and 4, the troughs were not perfectly aligned, so the boundary was drawn at the approximate mid-point between the 4 nearest component troughs (red dots).

Figure 4.3 also shows that component 1 (the thickest line) had relatively even loadings across a broad range of spectral frequencies. This was different from the sharp peaks and troughs observed in the rectified loading pattern of the other components, which were

more informative in identifying the spectral bands. The first principal component is unique because it reflects the mean correlation over all the variables, and so is strong when the underlying variables (channels) are highly correlated. Appendix 4.3 shows that the loading pattern of the first principal component (in Figure 4.3) does indeed resemble the grand mean correlation pattern over all the cochlear channels. Moreover, across spectral bands, each band's mean correlation strength was found to be inversely related to its RMS power. This indicates that higher power in a spectral region does not necessarily imply stronger correlation or co-modulation with other spectral regions.

Figure 4.3. Mean rectified loadings for top 5 principal components, averaged across all 264 speech samples. The red dots mark troughs in loading occurring at the boundaries between spectral bands. The boundaries themselves are shown as vertical dotted lines.

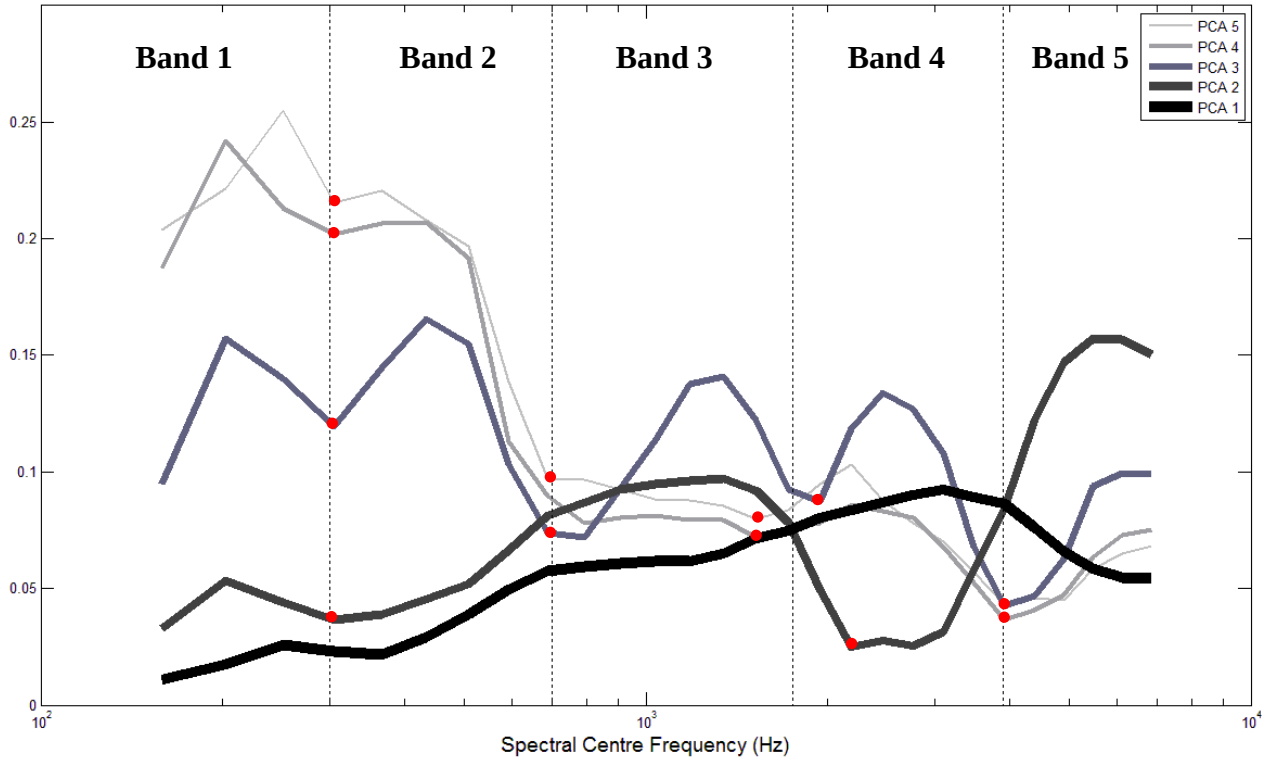


Table 4.1 summarises the 5 spectral bands that were identified in the spectral dimensionality reduction process, as shown in Figure 4.3. Each band groups together roughly the same number (between 5 & 7) of original cochlear channels. Given the strong resemblance between spectral band 1 (100-300 Hz) and the typical range of fundamental frequencies in human speakers (e.g. Baken, 1987 : Adult male 85-155 Hz, Adult female 165-255), it might not be unreasonable to suggest that spectral band 1 corresponds to fundamental

frequency energy. However, there is insufficient evidence to claim the same correspondence between spectral bands 2 to 5 and the first to fourth formants. Nonetheless, these 5 spectral bands achieve the desired aim of a parsimonious, low-dimensional and less-redundant representation of the spectral structure of speech. This 5-band spectral structure will form part of the new S-AMPH model, introducing a manageable amount of spectral complexity into the rhythm detection process.

Table 4.1. Spectral bands indentified from component loading patterns

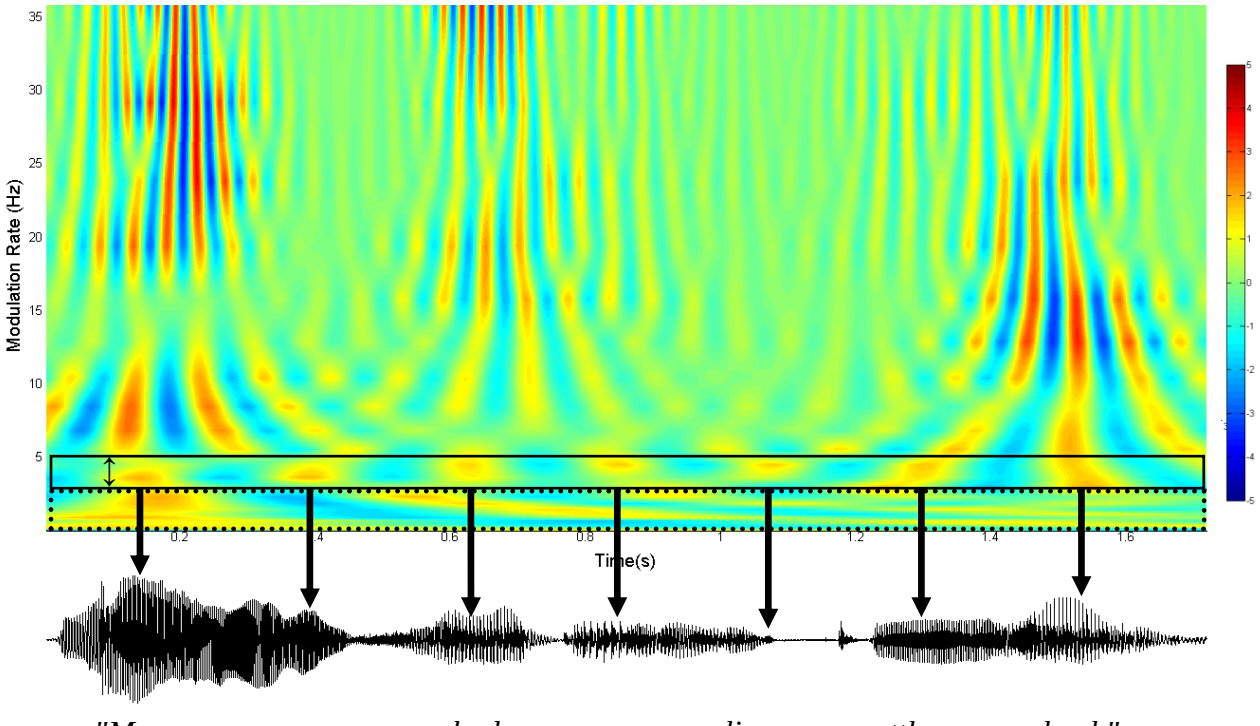
Spectral Band	**No of ERB_N Channels**	**Frequency Range (Hz)**
Band 1	5	100-300
Band 2	5	300-700
Band 3	7	700-1750
Band 4	7	1750-3900
Band 5	5	3900-7250

4.4 Modulation Rate Dimensionality reduction

4.4.1 The Original 'High Dimensional' Modulation Rate Representation

Next, a similar statistical approach was used to identify modulation *rate* bands in the amplitude envelope of speech, which would then constitute tiers in the AM hierarchy of the new S-AMPH model. Recall that the amplitude envelope of speech contains a continuous 'modulation spectrum' of modulation rates. To generate this initial 'high-dimensional' modulation spectrum representation, the Hilbert envelope was passed through a modulation filterbank with a very fine resolution. This filterbank comprised 24 channels logarithmically-spaced between 0.9-40 Hz. The parameters for this fine modulation filterbank are detailed in Appendix 4.2. The resulting modulation rate 'landscape' for the sentence *"Mary had a little lamb"* (for the spectral band of 0.5-2kHz) is shown in Figure 4.4. The red-blue striation pattern reflects oscillatory changes in instantaneous modulation amplitude (red = peak, blue = trough). The width of this striation pattern graduates from wide (bottom) to narrow (top) consistent with the increasing modulation rate and decreasing period, so that modulation rate is doubly represented as the y-axis and as the time-pattern of red/blue alternation.

Figure 4.4. Modulation pattern for 24 logarithmically-spaced modulation channels over time. Data has been interpolated across modulation channels to appear continuous. The y-axis indicates modulation rate and the z-axis indicates z-scored amplitude (also coded using colour).



Strikingly, the pattern of modulation for the rate channel around 3-4 Hz (bold black box) follows the syllable pattern of the sentence quite closely, with peaks in amplitude (black arrows) corresponding approximately to the temporal onsets of vowels. Since vowel onsets are associated with the p-centres in speech (Morton et al, 1976), modulations around this syllable rate may provide information about speech rhythm patterns. Since the sentence in this example was spontaneously-produced and not metronome-timed, there are natural fluctuations in the local syllable rate (number of syllables per second). These rate fluctuations are reflected as small *vertical* shifts (↕) in the sequence of peaks within the modulation band (bold black box). Despite these shifts, peaks and troughs (red and blue coloured spots) at this 'syllable' rate are temporally and spectrally well defined as 'spots' rather than 'smears'.

In comparison, at slower modulation rates, peaks and troughs tend to be smeared horizontally (i.e. in time), while at faster modulation rates they are smeared vertically (i.e. in frequency). The transmission of information by modulations, summarized in this modulation spectrum, requires a trade-off between resolution in time and modulation frequency. Faster modulations above 4 Hz show power changes over time that co-occur with very slow modulations below 4 Hz (dotted box). For example, faster modulations (>4 Hz) collectively show peaks in power at timepoints around 0.2s, 0.7s and 1.5s. These times coincide approximately with moments when peaks in very slow modulation (<4 Hz) occur, and could indicate the prosodic stress or emphasis pattern of the sentence.

This descriptive analysis suggests that the entire modulation spectrum (in this example) may be usefully divided into 3 regions. First, a narrow syllabic rate band at ~4 Hz. This carries strong temporal information related to syllable vowel nuclei (bold black box). Second, a band of slower (<4 Hz, dotted box) modulations that could correspond to the prosodic stress pattern of the utterance. Third, a band of faster (>4 Hz) modulations that are themselves modulated in power by the pattern of slow modulations. To formally assess the presence of 'modulation rate bands', the 24 modulation channels were subjected to a PCA analysis. Moreover, since it had previously been determined that the frequency spectrum could be represented by 5 spectral bands, this PCA analysis was conducted for *each* spectral band. That is, each speech sample was first spectrally-filtered into 5 spectral bands. The Hilbert envelope was then obtained for each spectral band, and this envelope was further filtered into 24 logarithmically-spaced modulation rate channels to give a high-dimensional 5 (spectral band) x 24 (modulation rate) channel representation for each speech sample. The

aim of the PCA procedure was to reduce this large number of 24 modulation channels into a smaller number of non-redundant modulation rate bands.

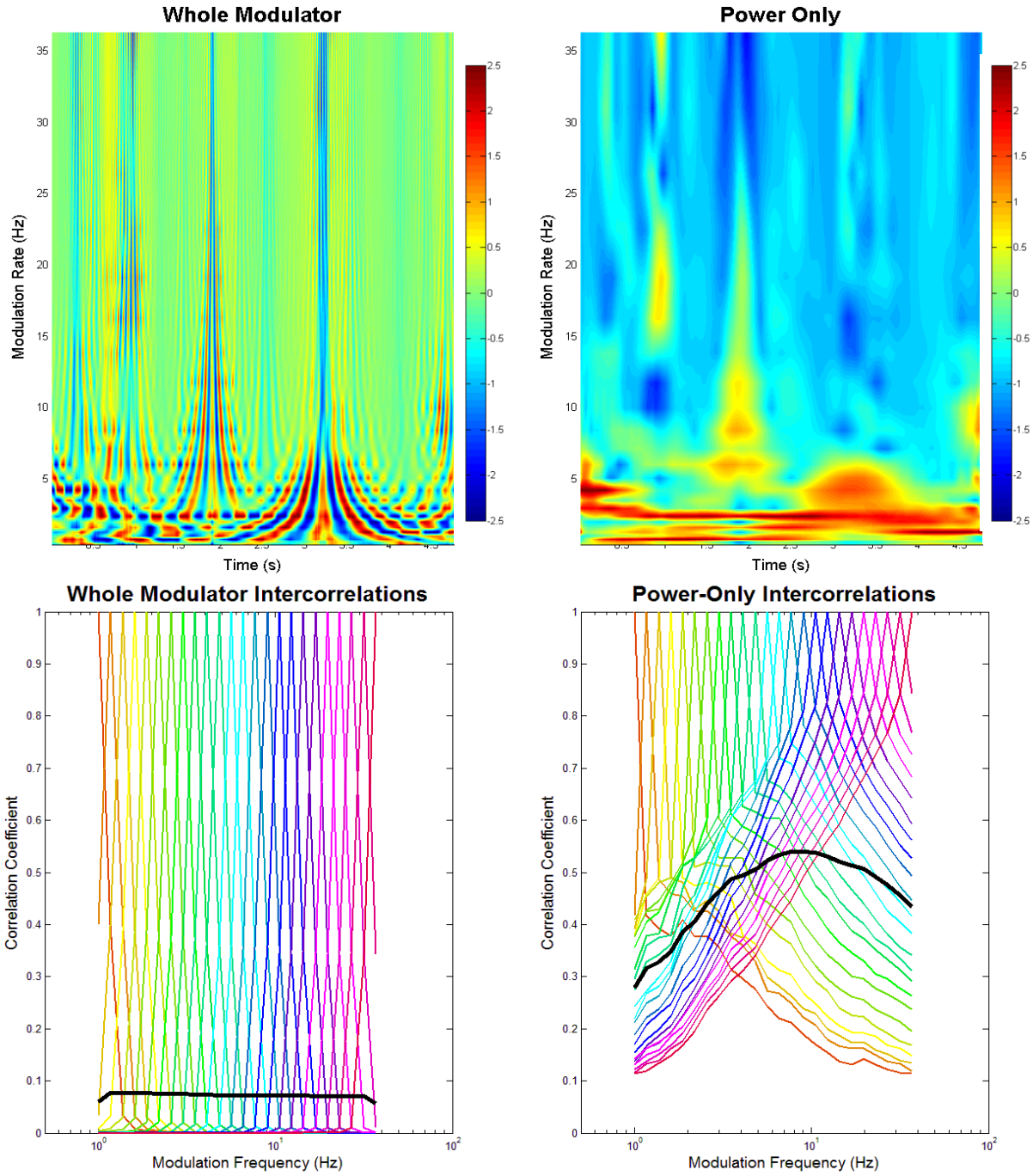
4.4.2 Problems with Low Inter-Correlation Between Channels

The PCA method relies on the underlying correlation structure of the data. Hence, if the variables have a generally low correlation, the method will not produce a strong first principal component, and higher components may be unreliable or not meaningful. In the 5 x 24 representation of the speech envelope, the correlation of each modulation rate channel with the other channels was in fact exceptionally low, with an average correlation coefficient of <0.1 (see bottom left plot of Figure 4.5). This low correlation occurred because the modulation patterns were effectively sinusoids at different rates whose phases were not aligned. Therefore, when a PCA was performed on this uncorrelated data, indistinct results were obtained. For example, the total variance explained by the top 5 PCA components was only around 35%, compared to 65% for the earlier Spectral PCA.

To overcome the issue of low inter-channel correlation stemming from different temporal rates, only the *power* in each modulation channel was used for the PCA analysis[13]. By analogy to the distinction between envelope and fine structure, using the power (envelope) only meant that differences in frequency (fine structure) were discarded. Figure 4.5 shows an example of the original whole modulators (top left) and modulator power only (right top) for a 24-modulation-channel decomposition of the envelope. In the whole modulator plot, rate differences between modulation channels are evident in the red-blue striation patterns which graduate from broad to narrow as the modulation rate increases (up on the y-axis). In the power-only plot, these rate-dependent striations are no longer present, but the gross activation patterns across the channels are maintained. As expected, using the power-only substantially increased the correlation of each modulation channel with other modulation channels (now typically between 0.3-0.5), indicating that a PCA analysis would now be more meaningful. The bottom right-hand plot of Figure 4.5 shows the average inter-channel correlations obtained across all samples and speakers when only the modulator power was used (showing spectral band 3 as an example).

[13] A "rate-normalisation" procedure was also developed as an alternative to using only the power of each modulation channel. Using these rate-normalised modulators for the PCA analysis yielded very similar results to using the power only. The details of the rate-normalisation procedure are reported in Appendix 4.5, but here the focus is on the results of using modulation power only.

Figure 4.5. (Top row) Example of the 24-channel modulation landscape for a single 5s sample of speech. The z-scored whole modulators are shown on the left, and the z-scored modulator power only is shown on the right, for spectral band 3 (700-1750 Hz). (Bottom row) Grand mean channel inter-correlations averaged over all samples and speakers, for spectral band 3 (700-1750 Hz). The individual coloured lines show the correlation of each modulation rate channel with all the other channels. The dark black line shows the grand mean inter-correlation, averaged across all modulation channels. Inter-correlations calculated with the whole modulator are shown on the left, and with the modulator power only on the right.



4.4.3 MODULATION RATE PCA (POWER ONLY): COMPONENT LOADINGS

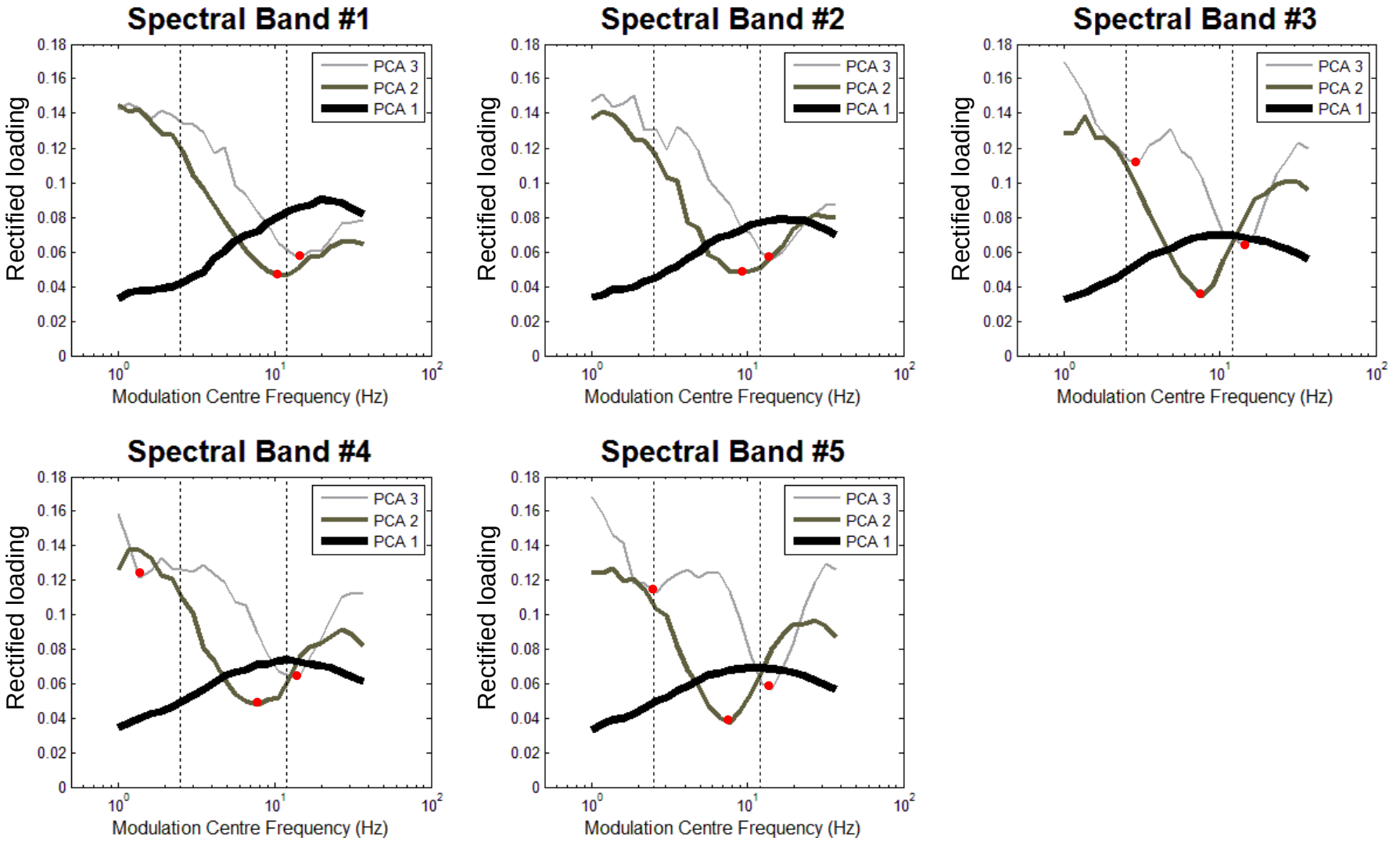
As before, a separate PCA was conducted for each nursery rhyme sample and each speaker, taking the individual timepoints as observations and the *power* of the 24 modulation rate channels as variables[14]. The loading patterns for each of the resulting 24 principal components was then rectified (absolute-valued), and averaged over all 264 samples. This was repeated for each of the 5 spectral bands, and the results from each band were analysed separately. Recall that the aim of the analysis was to define underlying patterns of redundancy by noting which groups of modulation channels (variables) load in a similar fashion onto the extracted PCA components.

In this analysis, only the top 3 principal components were considered, since these already accounted cumulatively for 60-80% of the total variance across the 6 speakers. As before for the spectral PCA, two criteria were used in identifying 'modulation rate bands'. These criteria were modified to reflect that here, only 3 principal components were being considered rather than 5 principal components. First, at least 1 of the 3 principal components should show a distinct peak within that modulation rate band. Second, at least 1 of the 3 principal components should show troughs near the upper and lower boundaries of that modulation rate band.

The mean rectified loading patterns for principal components 1 to 3 are shown in Figure 4.6, where each spectral band is shown in a separate subplot. From visual inspection of the subplots in Figure 4.6, principal component 1 loaded broadly across all the modulation frequencies, and did not show any clear troughs. However, the rectified loading patterns of principal components 2 and 3 were more informative. Across the 5 spectral bands, both these components showed clear troughs around 12 Hz, with component 2 showing slightly earlier troughs ~8-9 Hz, and component 3 showing a slightly later troughs ~15 Hz (note that the x-axis of the plots is logarithmic). In addition, for spectral bands 3 to 5, there was an additional slower trough in component 3 occurring around 2.5 Hz. All the troughs described in this paragraph are marked with red dots in Figure 4.6. This pattern of troughs suggests that there are 3 modulation rate bands, whose boundaries are located ~2.5 Hz and ~12 Hz. These boundaries are marked with vertical dotted lines in the plots.

[14] Note that the signal that was filtered through the 24 modulation rate channels was the *original* waveform (filtered into 5 spectral bands), and *not* the PCA component scores from the Spectral PCA analysis.

Figure 4.6. Mean rectified loading patterns for principal components 1-3, averaged over 44 rhymes and 6 speakers. The results for each spectral band are shown in a separate subplot. The red dots mark troughs in loading occurring at the boundaries between modulation rate bands.



For spectral bands 3, 4 & 5, the criteria regarding the identification of modulation rates bands were met. Each of the 3 modulation rate bands in these spectral bands contained at least 1 peak, and the boundaries between the bands were marked by at least 1 trough. However, the criteria were *not* met for spectral bands 1 & 2, where there was no distinct trough boundary at ~2.5 Hz. Nonetheless, for ease of future computation, a standard representation of 3 modulation rate bands was decided upon for all the 5 spectral bands. This would result in a balanced 5 (spectral band) x 3 (modulation rate band) representation of the envelope, rather than an unbalanced 2 x 2 + 2 x 3 representation.

Table 4.2 summarises the 3 modulation rate bands, and their possible association with 3 types of linguistic units - stress feet, syllables and phonemes. The close correspondence of the three modulation rate bands to neural oscillatory bands in the delta, theta/alpha and beta/gamma range is also noted in the table. Consistent with multi-time resolution models of speech processing (e.g. Giraud & Poeppel, 2012), the speech information on these 3 different timescales (stress, syllable and phoneme) could modulate their corresponding neural oscillatory bands independently, generating 3 separate neural streams of information coding.

Table 4.2. Modulation rate bands indentified from component loading patterns

Linguistic Unit	Main PCA Component Loading	Modulation Rate Band (Hz) and Geometric CF	Corresponding Neural Oscillatory Band (Hz)
Prosodic Stress (Mod band 1)	2 & 3	0.9 - 2.5 (1.4)	Delta : 1-3
Syllable (Mod band 2)	1 & 3	2.5 - 12 (5.5)	Theta : 3-7 Alpha : 7-12
Phoneme (Mod band 3)	2 & 3	12 - 40.0 (21.9)	Beta : 12-25 Gamma : 25-80

There is also a close correspondence between the current PCA-defined modulation rate bands, and the modulation statistics reported by Greenberg et al (2003) based on the SWITCHBOARD speech corpus. According to Greenberg et al (2003), the modulation spectrum shows a positive skew (i.e. sharp fall-off at slower modulation rates, slow fall-off at faster modulation rates) with a peak around 5 Hz representing the dominance of syllable-rate modulations. This pattern is shown in Figure 4.7, which is reproduced from Greenberg et al (2003). In the current analysis, the bandwidths of the modulation rate bands increase logarithmically, mirroring the slow exponential fall-off (positive skew) of Greenberg's modulation spectrum.

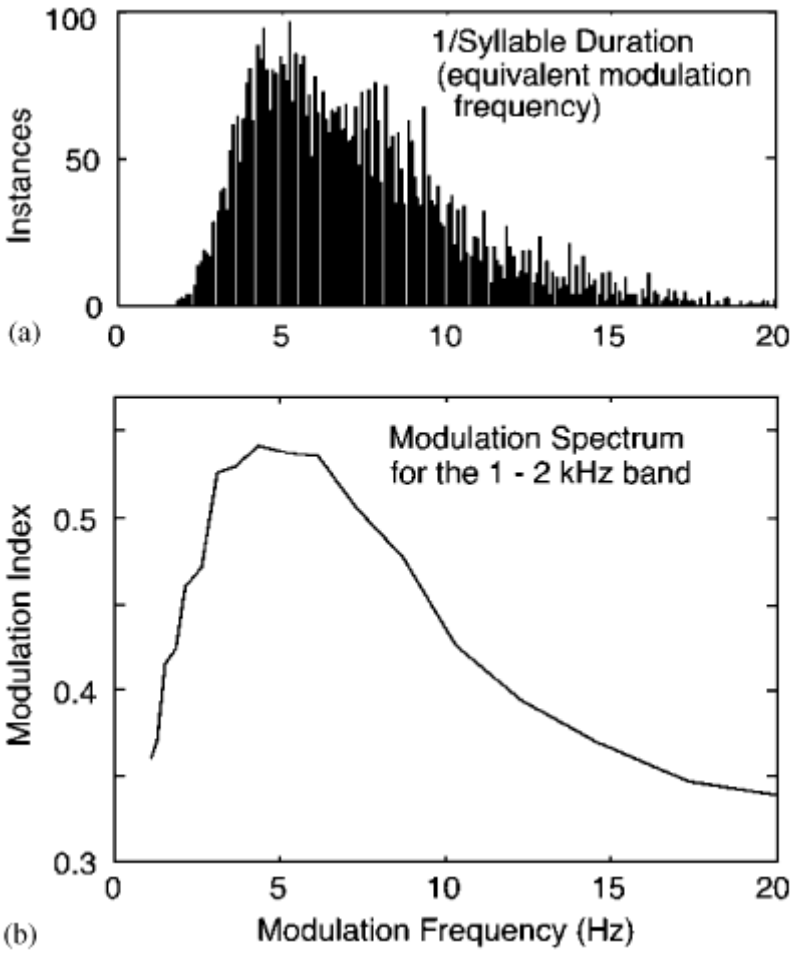


Figure 4.7. Reproduced from Greenberg et al, 2003. Syllable durations (top,(a)) and modulation spectrum (bottom, (b)) for the SWITCHBOARD speech corpus.

The proposed 'Syllable' modulation rate band has a geometric centre frequency of 5.5 Hz, which is close to the peak of Greenberg's modulation spectrum (5 Hz). Moreover, the Syllable rate band also contains the highest RMS modulation power (see Appendix 4.6), which is consistent with Greenberg's view of 'syllable dominance' in the modulation spectrum. According to Greenberg et al (2003), modulation rates under 5 Hz correspond to heavily stressed syllables. Conversely, shorter, unstressed syllables are represented in the spectrum up to around 15 Hz (the end of the positive 'tail'). In their analysis, very few syllables had durations consistent with a modulation rate of under 2 Hz, or over 15 Hz. These

duration limits for 'normal' syllables correspond closely to the proposed boundaries for the Syllable rate band of 2.5-12 Hz. In addition to this, the Syllable band of 2.5-12 Hz is also similar to the 4-16 Hz[15] range of modulation frequencies identified by Drullman et al (1994a, 1994b) as being the most important for speech intelligibility (see Section 1.9 of the Introduction). Hence, the characteristics of the Syllable rate band appear to be well in accordance with previous research.

Pertaining to the proposed 'Stress' rate band (0.9-2.5 Hz), the association of prosodic stress with modulations in this range is supported by the observation that the average duration of stress feet in the English language is ~500 ms or 2 Hz (Dauer, 1983).

Regarding the modulation statistics of linguistic units shorter than the syllable (i.e. >12 Hz in this analysis), Rosen (1992) noted that fluctuations in the speech envelope (defined as between 2-50 Hz) can be associated with segmental cues to manner of articulation and voicing. For example, the noise burst following the release of a stop consonant typically gives rise to a sharp energy peak lasting just a few tens of milliseconds (which would activate the 'Phoneme' rate band) . However, not all phoneme-related activity in speech occurs on such short timescales. Vowels, sonorant consonants like /m/ and /l/, and even fricatives like /s/ can produce fairly long-lasting modulations on the order of hundreds of milliseconds (which could activate the Syllable rate band). Consequently, although the third rate band is ostensibly named 'Phoneme rate', it is most likely to reflect the activity associated with stop consonants, and other similarly brief speech sounds.

Finally, it should also be noted that there is significant similarity between the 3 modulation rate bands identified here, and the original 5 tiers of the AMPH hierarchy. Recall that the AMPH hierarchy consisted of Slow, Stress, Syllable, Subbeat and Fast tiers. The typical modulation rates for these tiers are shown in Table 4.3, compared alongside the current 3 modulation tiers. As may be observed from the table, the boundaries for the AMPH Stress tier and the current Stress modulation rate band are virtually identical, and the boundaries for the AMPH Fast tier and the Phoneme modulation rate band are also quite similar. The major difference lies in the current Syllable modulation rate band, which encompasses both the original AMPH Syllable and Subbeat tiers. This means that modulations in the current Syllable band will be faster on average than the modulations in the

[15] Recall that the low- and high-pass filter cutoffs used by Drullman et al (1994a, 1994b) increased logarithmically (e.g. 2, 4, 8, 16 Hz). Therefore the 4-16 Hz range is only approximate.

original AMPH Syllable tier. However, overall there are strong similarities between the theoretically-proposed AMPH hierarchy, and the statistically-inferred modulation rate bands. Importantly, the two key tiers used for computing rhythm in the AMPH model (Stress and Syllable) are still present as new modulation bands (which was not guaranteed in the PCA analysis). This supports the view that the speech information at these rates is important and dominant in the speech signal, since they emerge spontaneously from a data-driven statistical analysis.

Table 4.3. Comparison between AMPH tiers and current modulation rate bands

<table>
<tr><th>AMPH Hierarchy Tier</th><th>Modulation Rate Band</th></tr>
<tr><td>Slow (0.5-0.8 Hz)</td><td>N.A.</td></tr>
<tr><td>Stress (0.8-2.3 Hz)</td><td>Stress (0.9 - 2.5 Hz)</td></tr>
<tr><td>Syllable (2.3-7 Hz)</td><td rowspan="2">Syllable (2.5 - 12 Hz)</td></tr>
<tr><td>Subbeat (7-20 Hz)</td></tr>
<tr><td>Fast (20 - 50 Hz)</td><td>Phoneme (12 - 40 Hz)</td></tr>
</table>

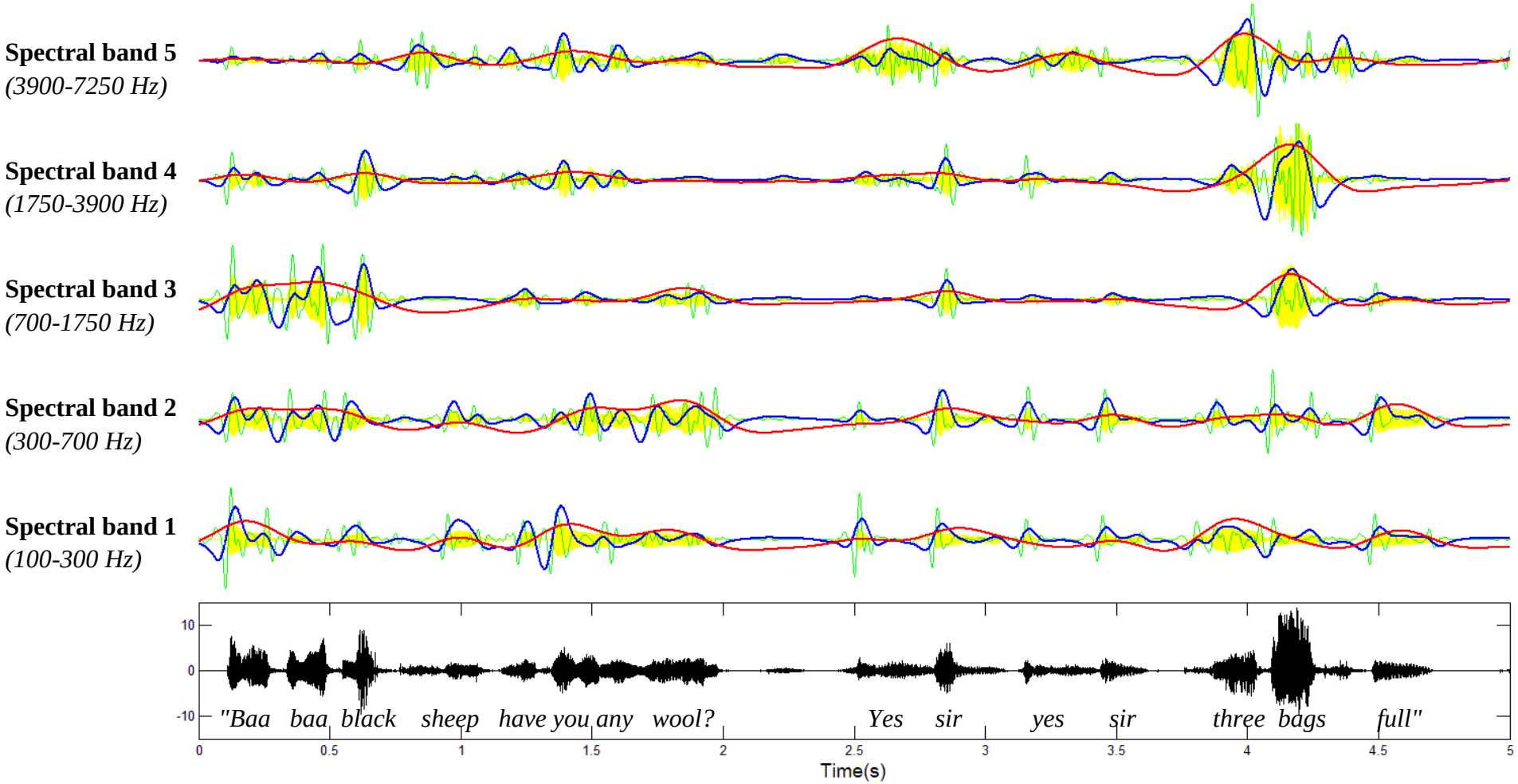
4.5 THE NEW 5 X 3 SPECTRO-TEMPORAL REPRESENTATION OF THE AMPLITUDE ENVELOPE

In summary, the original set of 29 cochlear (spectral) channels has now been reduced to 5 broad spectral bands. The original set of 24 modulation rate channels has been reduced to 3 modulation rate bands, these will now form a new 3-tier AM hierarchy. This 5 x 3 structure is a parsimonious representation of the dominant spectro-temporal modulation patterns in the speech envelope, and forms the basis of the new S-AMPH model.

Figure 4.8 shows an example of the 5 x 3 envelope structure obtained for the nursery rhyme sentence *"Baa baa black sheep have you any wool? Yes sir, yes sir, three bags full"*. In the figure, the 5 spectral bands are shown in rows, and the 3 modulation rate bands are overlaid in colored lines. The red line shows the slowest prosodic 'Stress' band (0.5-2.5 Hz), the blue line shows the 'Syllable' band (2.5-12 Hz) and green line shows the 'Phoneme' band (12-40 Hz). The yellow shaded sections show the sub-band filtered speech signal. The panel beneath plots the original speech waveform in black.

From the figure, it may be observed that the modulation patterns for each spectral band are different. This is to be expected if each spectral band carries non-redundant (unique) information.

Figure 4.8. Example of the new 5x3 AM hierarchy used for the S-AMPH model. Rows show the 5 spectral bands, coloured lines indicate the Stress (red), Syllable (blue) and Phoneme (green) rate modulators. The original waveform is shown in black below, and the filtered signal for each spectral band is shown in yellow.



4.6 CHAPTER SUMMARY & DISCUSSION

In this chapter, an attempt was made to derive a parsimonious, low-dimensional representation of the speech envelope to use as the spectro-temporal representation underlying the new S-AMPH model. To achieve this, PCA analyses were applied in the spectral and modulation rate domains. Component loading patterns were derived and analysed for evidence of channel 'clustering' (i.e. peaks) and boundaries indicating the transition between different spectral/modulation bands (i.e. troughs). Based on these analyses, 5 spectral bands and 3 modulation rate bands were identified.

While standard criteria were used to evaluate the evidence for 'banding' (e.g. number of peaks and troughs), this method still involved subjective interpretation. For example, in the modulation rate analysis, the troughs indicating the Syllable/Phoneme boundary differed in location across principal components, and across spectral bands. Therefore, the 'best fit' for the final boundary had to be determined by eye, in a subjective manner. The presence of such subjectivity necessitates caution when interpreting the 'bands' derived from these analyses. For example, the boundaries between the various bands should not be viewed as being strict cut-offs (i.e. Syllable information is only ever contained within 2.5-12 Hz), but as approximate boundaries with a degree of tolerance for error. Moreover, if a different criteria was set for identifying these bands, a different number of bands may have been identified, or the boundaries may have been slightly different.

Therefore, in this thesis, it is not claimed that these 5 spectral bands and 3 modulation rate bands correspond to any physiological mechanism, either in speech production or perception (e.g. in the way that the cochlear channels are actual physiological mechanisms in auditory perception). Rather, these bands are a convenient, low-dimensional, less-redundant representation of the 'dominant' spectro-temporal patterns in speech. This basic 5 x 3 structure will allow the S-AMPH model to provide a richer representation of speech rhythm (as compared to the original wholeband AMPH model) without creating unreasonable computational demands (e.g. by using a high-dimensional 29 x 24 representation).

5 NEW PROSODIC INDICES

Having derived a new 5 x 3 spectro-temporal representation for the speech envelope in the previous chapter, the next task is to adapt the prosodic indices used in the original wholeband AMPH model to this more complex representation of envelope structure. Recall that the original AMPH model used 'Syllable' Tier peaks to indicate beat location, and computed the prosodic strength of each beat according to the concurrent 'Stress' tier phase, following a Gaussian probability density function. In this revised S-AMPH model, there are now 5 'Syllable' bands and 5 'Stress' bands (from each spectral band). This means that for each particular syllable in the speech signal, there may be up to 5 possible markers delivered by the computational scheme, i.e. 5 correlates of its vowel location, and likewise 5 possible correlates of its prosodic strength. In the following section, the principles for computing key prosodic markers and indices based on the 5 x 3 S-AMPH representation are set out. These markers and indices are : (1) the location of syllable vowel nuclei; (2) the assignment of syllable prominence using the new prosodic strength index (PSI). To develop and fine-tune these new indices, a smaller stress-annotated stimulus set was used.

5.1 SPEECH MATERIAL USED FOR DEVELOPING PROSODIC INDICES

The following two sets of spoken material were used to develop and fine-tune the prosodic indices used in the revised S-AMPH model. These same materials were also used to evaluate the effectiveness of the revised S-AMPH model, the results of which are detailed in Chapter 6. Both sets of spoken material were made up of nursery rhyme sentences. Consequently, both sets of material had a strong underlying metrical structure. One set (Sample Set A) was generated by asking speakers to repeat a single nursery rhyme sentence using different prosodic templates. The second set (Sample Set B) was 20 nursery rhyme sentences that were freely produced (a subset of the 44 nursery rhymes used to derive the 5 x 3 structure in Chapter 4). While the first set of utterances was metronome-timed and perfectly metrically-regular, the second set of utterances contained variations in metrical pattern within the same sentence, as well as across speakers.

5.1.1 METRICALLY-REGULAR (METRONOME-TIMED) SPEECH (SAMPLE SET A)

This stimulus set comprised 9 metrically-controlled, metronome-timed variations of the same nursery rhyme sentence. 3 native English speakers (1 M, 2 F) produced each of the 9 metrical variations, giving 27 sentences in total. The nursery rhyme sentence was *"Jack and Jill went up the hill to fetch a pail of water, Jack fell down and broke his crown, and Jill came tumbling after*". This sentence was spoken in time to a 3 Hz metronome beat, and speakers deliberately changed the syllable stress pattern to fit 9 different prosodic 'foot' patterns. The first two patterns conformed to a Duple meter, or 2 syllables per foot (Trochaic (Sw) and Iambic (wS)). The next three patterns followed a Triple meter (Dactyl (Sww), Amphibrach (wSw) and Anapest (wwS)). The last 4 patterns followed a Quadruple meter (Primus paeon (Swww), Secondus paeon (wSww), Tertius paeon (wwSw) and Quartus paeon (wwwS)). The stress patterns for each metrical variation are shown in Table 5.1 below.

Table 5.1. Different metrical stress patterns for the metronome-timed sentences. Stressed syllables are shown in CAPS.

Syllable Stress Pattern (CAPS = stressed)	**Prosodic Foot**
JACK and JILL went UP the HILL to FETCH a PAIL of WAter...	Trochaic (Sw)
jack AND jill WENT up THE hill TO fetch A pail OF waTER...	Iambic (wS)
JACK and jill WENT up the HILL to fetch A pail of WAter...	Dactyl (Sww)
jack AND jill went UP the hill TO fetch a PAIL of waTER...	Amphibrach (wSw)
jack and JILL went up THE hill to FETCH a pail OF water...	Anapest (wwS)
JACK and jill went UP the hill to FETCH a pail of WAter...	Primus paeon (Swww)
jack AND jill went up THE hill to fetch A pail of waTER...	Secondus paeon (wSww)
jack and JILL went up the HILL to fetch a PAIL of water...	Tertius paeon (wwSw)
jack and jill WENT up the hill TO fetch a pail OF water...	Quartus paeon (wwwS)

The spoken material was kept identical even though not all the stress patterns fit the semantic content of the sentence naturally (for example, function words like 'the' or 'a' are not normally stressed). In view of this semantic-prosodic mismatch, and also because some prosodic patterns were simply more difficult to produce than others, speakers produced each sentence many times until they had 'learned' the pattern and were able to produce it without error.

While the utterances produced by this metrical manipulation were unnatural, they had several important properties. First, any differences detected by the model would be solely attributed to changes in prosody rather than phonological content. Second, the metrical manipulation allowed rare prosodic foot patterns to be included in the investigation, which would otherwise be difficult to find in real speech. Finally, since the exact metrical pattern of the utterances was known a-priori, the results of metrical rhythm detection by the model could be scored against an objective standard. The only additional source of error would be how accurate each speaker was in producing that metrical pattern, and this error was minimised by allowing the speakers to practice.

In freely-produced (untimed) speech, the actual prosodic stress pattern of each utterance is not known a-priori (i.e. speakers may not be using the same prosodic template, or may choose to deviate from a known template), hence the produced prosodic stress pattern must be annotated afterwards by a trained listener. Therefore, not only are the prosodic patterns in untimed speech irregular, but the subjective process of stress assignment by a listener also introduces significant error. For the purposes of evaluating the effectiveness of a model, therefore, un-timed speech is not the ideal material because one cannot be sure what proportion of 'errors' are due to a failure of the model, or to one of the other sources of error mentioned. Hence, the artificial metronome sentences acted as an important 'positive control', while the freely-produced sentences used in Sample Set B acted as a 'reality check'.

5.1.2 FREELY-PRODUCED UNTIMED SPEECH (SAMPLE SET B)

This selected corpus of 20 English children's nursery rhyme sentences was an annotated subset of the larger original set of 44 nursery rhymes that had been used to derive the statistics for the 5 x 3 spectro-temporal representation of the envelope in Chapter 4. 6 native British English speakers (all female) contributed spoken recordings of each of these nursery rhymes. 10 of the nursery rhymes had a dominant duple-beat rhythmic meter, while

the other 10 had a dominant triple-beat rhythmic meter. A decision was made not to assign a more specific prosodic foot type (such as trochee or iamb) to these nursery rhymes, since this would involve a degree of subjectivity, and the sentences themselves often comprised a mixture of different types of feet. Table 5.2 lists the 20 selected nursery rhymes and their dominant rhythmic meter (duple or triple).

Table 5.2. List of 20 nursery rhymes and their rhythmic meter

Nursery Rhyme	**Rhythmic Meter**
Old MacDonald Had a Farm	Duple
Mary Had a Little Lamb	Duple
Polly Put the Kettle On	Duple
Yankee Doodle	Duple
Peter Peter Pumpkin Eater	Duple
Mary Mary Quite Contrary	Duple
Simple Simon Met a Pieman	Duple
Lucy Lockett	Duple
Cobbler Cobbler Mend My Shoe	Duple
Peter Piper	Duple
Little Miss Muffett	Triple
Little Jack Horner	Triple
Little Boy Blue	Triple
Curly Locks	Triple
To Market	Triple
Pussycat Pussycat	Triple
Ladybird Ladybird	Triple
There Was An Old Lady	Triple
Two Cats of Kilkenny	Triple
Lavender's Blue	Triple

Since many of these nursery rhymes are often sung to children in familiar folk tunes, the rhythmic meter was determined with reference to both the musical time signature for the sung rhyme (where available), and by poetic scansion. The meter is the fundamental repeating pattern of beats that corresponds to the poetic foot. In practice, finding the meter involves identifying accented (stressed) beats and counting the number of beats until the next accent (MacPherson, 1930; Scholes, 1977). A 'duple' meter (foot length of 2) was assigned when the time signature of the musical piece indicated 2 or 4 beats per bar (i.e. 2/4 or 4/4), and when the dominant prosodic feet were trochees ('S-w') or iambs ('w-S'). A 'triple' meter (foot length of 3) was assigned when the time signature of the musical piece indicated 3 beats per bar (i.e. 3/4), or when the dominant prosodic feet were dactyls ('S-w-w'), amphibrachs ('w-S-w'), or anapests ('w-w-S').

Since each nursery rhyme was different in length from the others, only the first 24 syllables from the first line of each rhyme were used in the analysis. This was done to standardise the amount of spoken material being compared between rhythmic conditions. The full list of sentences is shown in Appendix 5.1.

As these sentences were spontaneously uttered by each speaker, the produced prosodic patterns varied from sentence to sentence and from speaker to speaker. To ascertain the actual stress patterns that were produced, the nursery rhyme sentences were manually stress-transcribed by a female native English speaker with formal training in Linguistics (not the author). Stress transcription was done by listening to each sentence carefully, and judging whether each syllable sounded stressed or unstressed. The hypotheses of the current study (i.e. duple/triple beat distinction) were not known to the individual doing the transcription so that her judgments would be based on the acoustic patterns of each sentence only.

5.2 LOCATING SYLLABLE VOWEL NUCLEI IN THE ENVELOPE

5.2.1 SYLLABLE VOWEL NUCLEI CORRESPOND TO 'SYLLABLE' MODULATOR PEAKS

In the original AMPH model, peaks in the 'Syllable' modulator tier had simply been used as temporal markers to locate the presence of syllable beats. Arguably, taking slightly earlier or later points in the oscillatory cycle (i.e. before or after the peak) would have sufficed for the same purpose. No attempt was made to link this acoustic landmark to a specific linguistic feature. However, given that peaks generally indicate prominent or significant events, it is possible that these landmarks in the modulation landscape may in fact correspond to perceptually salient segments of speech.

Peaks in the 'Syllable' tier modulator generally occur in the middle of a spoken syllable, and correspond roughly to the point of highest local energy at the vowel nuclei of syllables. Vowels also tend to occur in the middle of syllables (with a simple C-V-C structure), and generally carry the greatest energy in the signal because during vowel phonation the vocal tract is most open. In fact, it may not be unreasonable to propose that listeners' perceptual experience of syllables is in fact derived from these prominent energy peaks in the speech signal. To test if syllable vowel nuclei did indeed correspond to peaks in the 'Syllable' tier of the new S-AMPH hierarchy, the two sets of speech samples were manually annotated to mark the mid-points[16] of their syllable vowel nuclei. The 'Syllable' tier phase value at each of these vowel locations was then determined. It was expected that vowel nuclei should be associated with Syllable modulator phase values of close to 0 pi radians (i.e. the oscillatory peak).

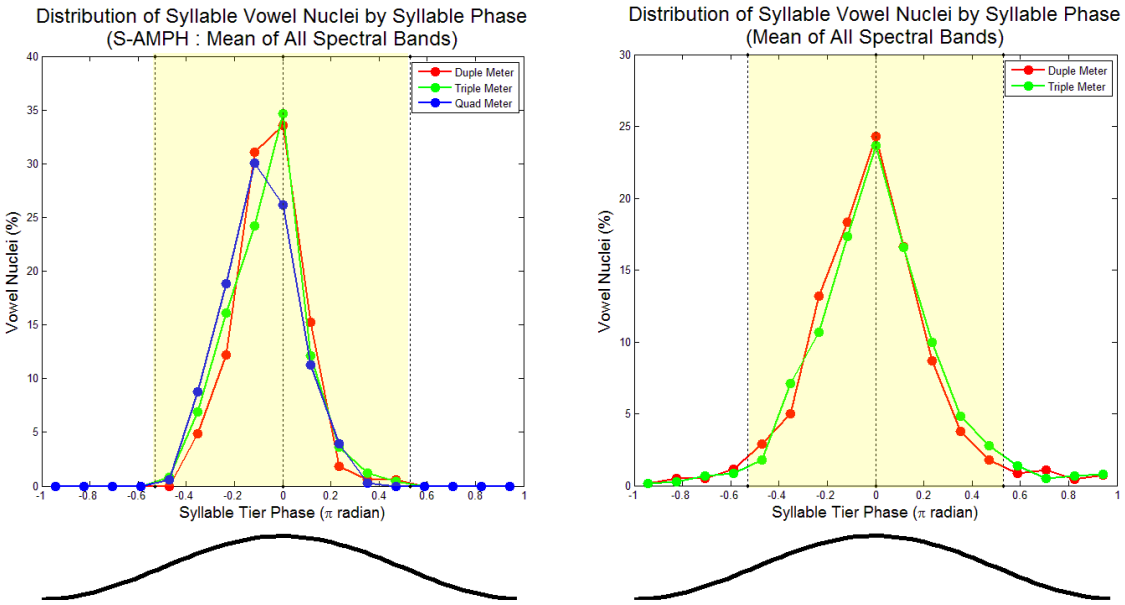
Figure 5.1 shows the distribution of syllable vowel nuclei with respect to the phase of the S-AMPH Syllable modulator, for metronome-timed (left) and untimed (right) speech samples, averaged over all speakers. The Syllable phase distribution was computed for each spectral band, and the average percentages over all 5 bands are shown. The coloured lines indicate sentences with different prosodic meter (i.e. different number of syllables per foot).

[16] Mid-points were used rather than onsets (associated with p-centres) because the mid-points of the vowel nuclei were higher in energy than the onsets, and therefore should correspond better to the peak in energy over the entire syllable.

For the metronome-timed sentences (left), these lines correspond to the averages over sentences with duple, triple or quadruple meter respectively (i.e. for duple meter, this was the average of the trochaic and iambic sentences). For the freely-produced sentences (right), averages were taken over the 10 duple meter and 10 triple meter sentences respectively.

Figure 5.1. Distribution of actual measured vowel nuclei against the phase of the Syllable tier modulator. LEFT : Metronome Speech; RIGHT : Freely-produced speech. The bold curves beneath each plot show the oscillatory shape corresponding to the phase values on the x-axis.

(a) Metronome-timed Speech **(b) Freely-produced speech**



For the metronome speech (left plot), all of the vowel nuclei occurred within a phase zone of -π/2 to π /2 radian of the Syllable modulator (i.e. near the peak of the oscillation, yellow shaded region), and no vowel nuclei occurred outside this phase zone. Although only ~30% of vowel nuclei coincided exactly with the Syllable tier modulator peak maxima (i.e. 0 π radians), all of the remaining vowel nuclei were distributed within ± π/2 radian of this Syllable peak phase value. If the geometric mean centre frequency of the Syllable band is considered (5.5 Hz), this tolerance of ± π/2 radian about the peak corresponds to a time window of ± 45 ms. There was very little difference in the distribution pattern between duple, triple and quadruple meter sentences. This was not surprising given that the phonological content of the metronome sentences was identical.

For the freely-produced speech (right plot), the vast majority of vowel nuclei again fell within +/- π/2 radian of the Syllable peak. However, this distribution pattern had a wider base than that of metronome-timed speech, indicating a slightly larger variance in the distribution of vowel nuclei with respect to Syllable modulation phase. This was expected, given that these sentences were not uttered to a strict timing. Nonetheless, for both speech sample sets, the vast majority of vowel nuclei were located at or in close proximity to the peak of the Syllable tier modulator, and virtually no syllable nuclei occurred at trough regions of the Syllable tier modulator. Therefore, peaks in Syllable modulator do indeed correspond well to the occurrence of syllable vowel nuclei, both in metronome-timed and in un-timed speech.

5.2.2 Syllable Peak Detection & Selection Using 5 Spectral Bands

In the previous section, it was established that peaks in the 'Syllable' tier of the AM hierarchy are good proxy markers for syllable vowel nuclei. Here, the procedure for identifying syllable vowel nuclei using the 5 Syllable modulators in the S-AMPH model is described. There are two parts to the procedure : (1) peak detection, followed by (2) peak selection. In the peak detection step, all possible peaks that could correspond to syllable vowel nuclei across the 5 spectral bands are identified, forming a large pool of 'candidate peaks'. In the peak selection process, this large pool of candidate peak is systematically evaluated to identify the most likely correlates of syllable vowel nuclei, and to remove any possible duplicates of the same vowel nucleus.

5.2.2.1 Syllable Peak Detection Procedure

Prior to the peak-detection procedure, the Syllable modulator in each spectral band was z-score standardised to ensure that the mean of all its values would be 0, and the standard deviation equal to 1. This standardisation would allow a uniform minimum peak height criterion (+0.5 standard deviations) to be applied across all the samples. Next, the speaking rate for each sample was estimated. The estimate of the speaking rate was used to set the minimum peak-to-peak spacing criterion for each sample. Since the speaking rate varied greatly from sample to sample, it was important that this peak-to-peak distance should also

vary according to speaking rate. Otherwise, syllable peaks might be missed if the set distance was too large, or spurious peaks included if the distance was too small.

To estimate the syllable speaking rate for each sample, a Fourier analysis was applied to the Syllable tier to determine the single modulation frequency (rate) with the highest power. A value of 60% of the period of this strongest rate was used as the minimum peak-to-peak distance for each sample. For example, if the estimated syllable speaking rate was 3 Hz, the minimum peak-to-peak distance was 0.6 x (1000/3) = 200 ms. Once these parameters had been established, Matlab's peak-detection algorithm was used to detect all the Syllable peaks in the 5 spectral bands that met the criteria (i.e. min peak height and min peak distance). This resulted in a large pool of candidate peaks across the 5 spectral bands.

5.2.2.2 Syllable Peak Selection Procedure

Next, a 2-step selection process was applied to this pool of candidate peaks, to select those deemed to correspond to syllable vowel nuclei. The two major parameters used in this 2-step selection process were : (1) the relative power of the spectral band that the peak was located in; and (2) the temporal proximity of the peak to other peaks in other spectral bands.

Step 1 : Ranking spectral bands by power

Power was the first parameter to be considered because vowels are voiced, and voiced sections of speech have higher power than unvoiced sections of speech. Therefore, spectral bands with higher power should also be more likely to contain the energy (modulation) peaks associated with voiced vowels. Therefore, in the first step of the selection procedure, the total RMS power of each spectral band was determined, and the 3 spectral bands with the highest RMS power were identified, forming the 'primary band' (highest power), 'secondary band' (2nd highest power) and 'tertiary band' (3rd highest power). Only the peaks arising from these primary, secondary and tertiary spectral bands were retained for further consideration, all other peaks were discarded. Note that this process of power ranking was done individually for each speaker and speech sample, to allow for variations across speakers and conditions.

Since the primary band contained the most vowel energy (i.e. highest power), all the peaks arising from the primary band were automatically selected, and deemed to correspond to unique syllable vowel nuclei. However, it was possible that some syllable vowel nuclei had

a lower or higher frequency than that of the primary band. In this case, they would not be represented in the primary band, but would appear in either the secondary or tertiary band, and would need to be identified in these other bands. However, the majority of peaks in the secondary or tertiary bands would arise from syllables that already had a vowel correlate in the primary band, and would therefore be redundancies (in the sense that they belong to the same syllable, *not* in the sense that they contain identical information, the definition used earlier in the PCA analyses). In order to distinguish between peaks arising from real 'outlying' syllable vowel nuclei, and peaks corresponding to redundancies, the second selection step used a criterion of temporal proximity.

Step 2 : Cross-band peak matching by temporal proximity

For a typical syllable, its vowel nucleus should produce a peak in primary band, while its initial and final consonants may produce peaks in other bands, including the secondary or tertiary bands. These consonant-related peaks, although in different spectral bands, would occur close together in time to the vowel peak in the primary band because they belong to the same syllable. By this reasoning, any peaks in the secondary and tertiary bands that lay in close proximity to a peak in the primary band would be redundancies and should be discarded. On the other hand, an outlying syllable vowel nucleus would appear in either the secondary or tertiary bands (which had the next highest power), but *not* have a temporal correlate in the primary band. Therefore, after all the redundancies in the secondary and tertiary bands had been identified with respect to the primary band, any *remaining* peaks without a primary band correlate should correspond to genuine outlying syllable vowel nuclei.

In the procedure, peaks in the secondary band were compared to the primary band first. Any secondary band peak that lay within 0.5 syllable-lengths of a primary band peak (based on the estimated syllable rate for each sample) was treated as a redundancy and discarded. Any remaining secondary peaks without a primary band correlate were retained and added to the repertoire of selected primary band peaks. Next, peaks in the tertiary band were compared to peaks in the primary band (including added non-redundant peaks from the secondary band). Again, any tertiary band peaks that lay within 0.5 syllable-lengths of a primary/non-redundant secondary band peak were treated as redundancies and discarded. Any remaining tertiary band peaks without a temporal correlate were retained and added to

the already-selected repertoire of primary and secondary band peaks. This final set of selected primary, secondary and tertiary band peaks formed the final set of Syllable peaks deemed to correspond to unique syllable vowel nuclei.

To double-check that the final selection of peaks correctly reflected the actual spoken syllable pattern, these peak sequences for each sample were all validated by ear during the method development process. To do this, the sequence of peak timings was converted into a binary 'temporal mask' (i.e. peak = 1, no peak = 0). These temporal masks were then used to vocode the sentences in a single-channel tone vocoder, effectively converting each sentence into a sequence of tone pulses, where each pulse corresponded to a detected syllable 'beat' (vowel nucleus). The parameters used in peak detection (i.e. 0.5 standard deviation minimum peak height, 60% syllable rate minimum peak-to-peak distance) were also manually fine-tuned using this vocoding process. That is, the entire peak selection and detection process was re-run several times, each time using different peak detection parameters. The parameters that produced the most accurate syllable pattern (determined by ear) were used as the final settings in the S-AMPH model.

5.2.2.3 Example of Syllable Peak Detection & Selection in Operation

An example of peak detection, followed by peak selection is shown in the top and bottom graphs of Figure 5.2 respectively, for a sample of trochaic metronome speech. In the top graph, the coloured dots indicate all the 'candidate peaks' that were detected across all 5 spectral bands. The vertical dotted lines indicate the approximate timing of the metronome beats. From the figure, it may be observed that each syllable typically produces peaks in 3-5 different spectral bands, corresponding to the mixture of vowel and consonant sounds within the syllable. Importantly, some spectral bands seemed to represent the vowel pattern better than other spectral bands.

For example, the peaks in spectral bands 1, 2 & 3 appeared to be reasonably well located near to the vowel nuclei of the syllables in the sentence. In contrast, spectral band 5 showed clear non-vowel peaks corresponding to the affricate [tʃ] at the end of 'fetch', or the [dʒ] at the start of 'Jill' (highlighted in the dark blue boxes). It is also worth noting that no single band represented the vowel nuclei of *all* the syllables in the sentence (i.e. there were always 'missing' syllable vowels). For example, in spectral bands 2, 3 & 4, peaks were not

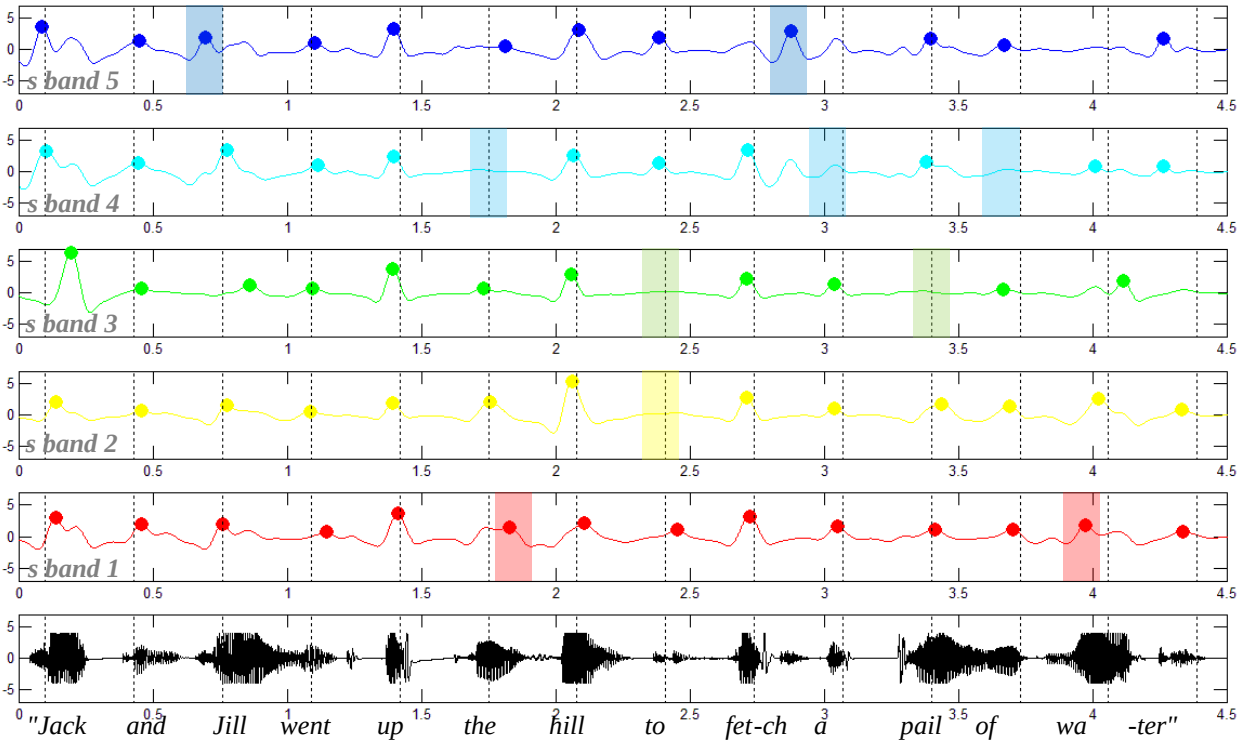
elicited for function words like 'the', 'to', 'a' or 'of' (yellow, green and light blue boxes respectively).

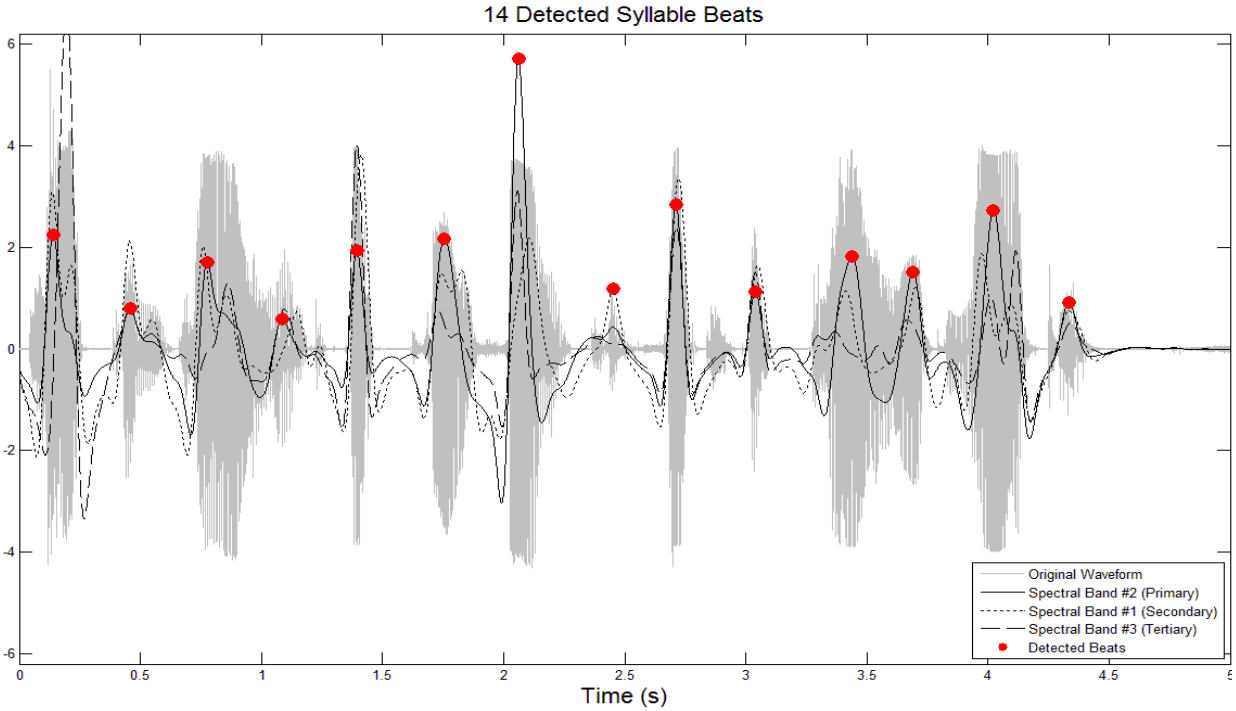
Moreover, although spectral band 1 detected all the 'syllable beats' in the utterance, the timing of these beats was not perfectly accurate. For example, assuming that the speaker was accurately synchonising the uttered syllables to the metronome beats, the band 1 peak for the syllable 'the' was delayed, whereas the peak for the syllable 'wa-' in 'water' was too early (red boxes). This illustrates the problem of using a single sub-band approach (as discussed in the preface to Part III) - not all the syllable vowels will be well-represented in any single frequency band. Therefore, it is important that *more than one* spectral band is used to determine the location of syllable vowel nuclei, as was done here.

Moreover, across different speakers and speaking conditions, the specific combination of spectral bands that best captured the syllable vowels also varied. For example, in this dataset, syllable peaks for the male speaker were commonly well-represented in spectral bands 1 & 2, whereas syllable peaks for female speakers were better represented using higher frequency spectral bands 3 or 4. This underlies the importance of being able to flexibly change *which* spectral bands are used to determine the location of syllable vowel nuclei, according to the nature of the speech sample itself, as was done here.

The bottom panel of Figure 5.2 shows the result of the peak selection process. In this example, spectral band 2 was used as the primary band. The unusually low-frequency vowel [oo] in the word 'to' (at ~2.5s) did not appear in spectral band 2, but was picked up and filled in by secondary spectral band 1. No additional vowels were spuriously added by tertiary spectral band 3, resulting in perfect identification of the 14 syllable vowel nuclei (beats) in the sentence.

Figure 5.2 (Top). Example of the Syllable modulators from each spectral band (rows) and candidate peaks (coloured dots) detected. The sample was a trochaic metronome-timed sentence. Coloured boxes highlight either missing peaks, or extraneous peaks that did not correspond to actual spoken syllables. The original waveform of the utterance is shown in the bottom panel. (Bottom) Results of peak selection from amongst the candidate peaks. The final set of selected peaks are shown as red dots. In this example, the three highest-power bands used for syllable selection were spectral band 2 (primary), band 1 (secondary) and band 3 (tertiary).





5.3 ASSIGNING SYLLABLE PROMINENCE (NEW PROSODIC STRENGTH INDEX)

5.3.1 DISTRIBUTION OF SYLLABLE VOWELS WITH RESPECT TO STRESS PHASE

The most important prosodic statistic used in the Stress Phase Code of the AMPH model was the instantaneous Stress phase corresponding to each Syllable peak. This was used to determine the prosodic prominence of the syllable. In the AMPH model, this Stress phase value was transformed into an indicator of prosodic prominence via a Gaussian probability density function (PDF). This PDF was an approximation, since the exact relationship between Stress phase and perceived prosodic prominence was unknown at the time when the AMPH model was developed. However, it was assumed that prosodically-prominent syllables would occur close to the peak of the Stress oscillatory cycle while prosodically-weak syllables would occur close to the trough of the Stress oscillatory cycle. Based on this assumption, the AMPH model identified 'Strong' and 'weak' syllables using the Stress phase at which they occurred. Here, this assumption is examined by looking at the *actual* distribution of syllables with respect to Stress phase.

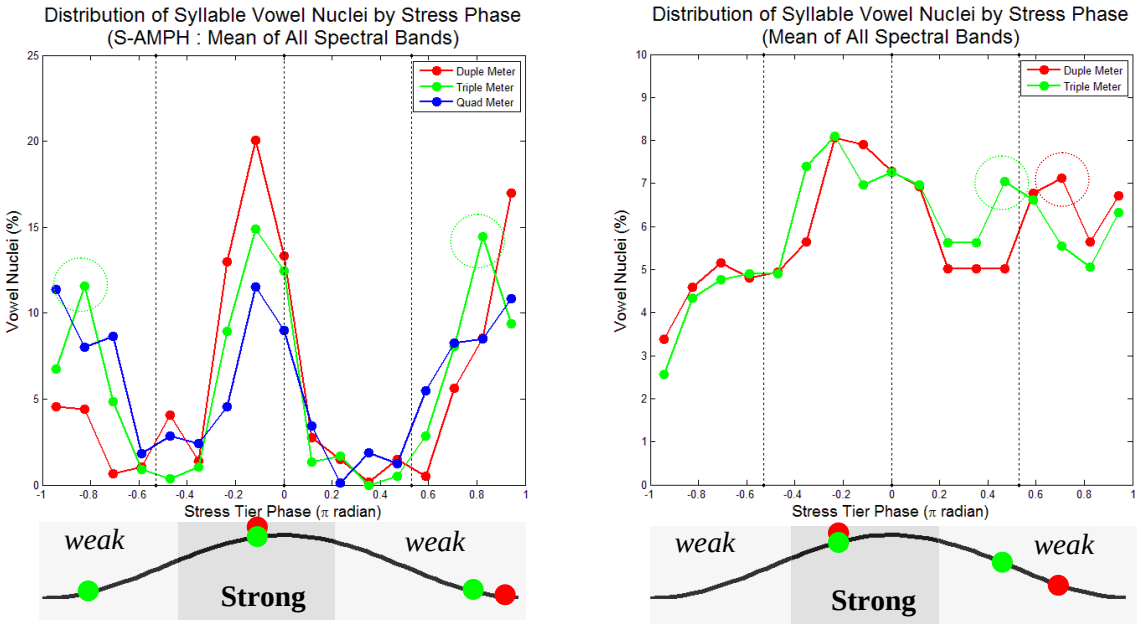
In the AMPH model, the Stress oscillatory cycle is functionally equivalent to a prosodic stress foot (i.e. 'Stress cycle = Prosodic foot'). According to this view, prosodic meter is realised as the functional division of 'Stress phase space' into *n* regions, where *n* is the number of syllables in the prosodic foot. Each syllable in the foot therefore occupies a distinct Stress phase region, but each of these regions need not be equal in width. For example, for a duple meter pattern like the trochee ('S-w'), Stress phase should be divided into two (peak and trough) regions, with strong and weak syllables occurring in each region. For a triple meter pattern like the dactyl ('S-w-w'), Stress phase should now be divided into *three* regions, with the strong syllable occurring near the peak of the oscillation, and the two weak syllables occupying two adjacent but distinct regions near the trough of the oscillation. These predicted distribution patterns are investigated here.

Recall that in a previous analysis (Section 5.2.1), all the syllable vowel nuclei had been manually located in all the sentences. Here, the corresponding Stress modulator phase for each vowel nucleus was recorded. For ease of analysis, Stress phase was divided into 17 equal bins between $-\pi$ and π radian. The percentage of vowel nuclei falling into each phase

bin was then computed, resulting in a vowel-Stress phase distribution pattern. This phase distribution was computed for each spectral band, and then averaged over all 5 spectral bands, for all sentences (by meter) and speakers. The results are shown in Figure 5.3, for metronome-timed speech (left subplot) and untimed speech (right subplot). The coloured lines in each plot show the averages for the different prosodic meters.

Figure 5.3. Phase distribution of syllable vowel nuclei with respect to Stress phase. The black bold line underneath each plot shows the equivalent oscillatory shape of the Stress modulator for each phase value.

(a) Metronome-timed Speech **(b) Freely-produced speech**



5.3.1.1 Metronome-Timed Speech *(left subplot)*

For duple meter sentences (left plot, red line), speakers clearly tended to utter syllables in two major Stress phase regions, producing two spikes in the distribution pattern. These occurred slightly before the Stress peak (-0.1π rad) for stressed vowels and at the Stress trough (pi radian) for unstressed vowels. The two red dots on the oscillation plot below the phase distribution represent these two phase regions, where strong or weak syllables occurred. This bi-modal distribution of vowel nuclei for duple meter sentences was exactly as predicted.

For triple meter sentences (green line), speakers also showed a spike in the syllable distribution just before the Stress peak (-0.1π rad), corresponding to stressed vowels. However, now, there were *two* other distribution spikes occurring in the Stress trough region. These two spikes are circled in green on the plot, and occurred just before and just after the stress trough at 0.8π radian and -0.8π radian. The three green dots on the oscillation plot below the phase distribution represent these three distribution spikes. The drop in the height of the distribution spike at -0.1π rad is consistent with triple meter sentences having proportionately fewer stressed vs unstressed syllables than duple-meter sentences (i.e. 1:2 for triple vs 1:1 for duple). Therefore, the tri-modal distribution pattern of vowel nuclei for triple meter sentences was also exactly as previously predicted.

For the quadruple meter sentences (blue line) however, there was weak evidence that speakers were now dividing Stress phase into four distinct regions. Speakers continued to place vowel nuclei just before the Stress peak (-0.1π rad), corresponding to stressed vowels. The height of this central spike was, as expected, lower than for triple and duple meter sentences. However, at the trough phase regions, the distribution pattern did not show three distinct spikes as expected. Rather, syllable vowels were most concentrated around ± π radian (the Stress phase trough), and there were two symmetric 'shoulders' in the distribution pattern, centred around the Stress phase trough. This distribution pattern suggests that speakers were not timing the unstressed syllables as accurately for these quadruple-meter sentences. This led to a 'smeared' distribution pattern in the trough phase region, rather than the sharp spikes observed for duple and triple meter sentences.

Therefore, for duple and triple meter metronome-timed sentences, the distribution patterns of syllable vowels with respect to Stress phase was exactly as predicted, following the 'Stress cycle = Prosodic foot' hypothesis. However, for quadruple meter sentences, the predicted pattern did not emerge. This could be because speakers are poorer at controlling the timing of unstressed syllables (as compared to stressed syllables). Therefore, when more unstressed syllables occur in a series (i.e. longer prosodic feet), the timing of these syllables becomes more and more variable so that each unstressed syllable no longer occupies a fixed position with respect to Stress phase.

However, for all three types of meters, stressed and unstressed syllables occurred in completely different 'Strong' and 'weak' regions of Stress phase (with very few syllables

occurring at an 'intermediate' phase). This suggests that for metronome-timed speech, Stress phase will be a very useful index for discriminating stressed and unstressed syllables.

5.3.1.2 Freely-Produced Speech *(right subplot)*

For untimed speech, the syllable vowel distribution patterns was much less well-defined. Instead of sharp spikes, the distribution pattern was characterised by gentle rolling 'hills'. That is, syllable vowels occurred at *all* phase values with a baseline percentage of around 5%. However, syllables were slightly more likely to occur at two Stress phase regions, where the percentage of syllable vowel occurrence rose to 7-8%. These two phase regions are indicated with red (duple meter) and green (triple meter) dots on the oscillation plot below the phase distribution graph, and described below.

First, for both duple and triple meter sentences, there was a major 'hill' in the distribution pattern just before the peak of the Stress modulator, at around -0.2π radians. For metronome-timed sentences, a major spike had also occurred in a similar phase region, and corresponded to strong stressed syllables.

The second hill in the distribution pattern for both duple and triple meter sentences occurred along the downward slope of the Stress cycle, around 0.5-0.7π radians. This hill was *earlier* for triple meter sentences (~0.5π radians) than for duple meter sentences (~0.7π radians), which was also the trend observed for metronome-timed speech

However, while only two regions of concentration were expected in the distribution for duple meter sentences, three regions were expected for triple-meter sentences, but only two were observed. Moreover, a significant proportion of syllables occurred *outside* these two phase regions for both duple and triple meter sentences[17]. This suggests that in untimed speech, speakers were *not* constraining the occurrence of syllables into different Stress phase regions as tightly as they did for metronome-timed speech. Rather, syllables were allowed to occur at all Stress phases, with only a slightly stronger tendency to occur in some phase regions than others (e.g. 'weak attractor' regions). Interestingly, there appeared to be the same number of these weak attractor phase regions (two) irrespective of whether the sentence was

[17] Note that this could also be because these sentences were not *perfectly* duple or triple metered (i.e. they included prosodic feet of other lengths as well). The phase regions for these prosodic feet of other lengths could have overlapped with those for the duple/triple meter, resulting in the appearance that syllables occurred at all phase values, rather than within well-defined boundaries.

duple- or triple-timed. Rather, the meter of the sentence was reflected in a slight phase-shift in the attractor regions (slightly earlier for triple meter).

Nonetheless, some distinction between 'Strong' and 'weak' phase regions was observed (the two 'hills'), and these regions were broadly similar to the Strong and weak phase regions observed for metronome-timed speech. This suggests that Stress phase may still be a useful index for whether syllables are strong or weak in untimed speech, although this will not be as sharply discriminatory as for metronome-timed speech.

5.3.1.3 Conclusion

Therefore, the strong version of the 'Stress cycle = Prosodic foot' hypothesis, where each syllable in the foot occupies a distinctly different Stress phase region, only seems to hold for metronome-timed speech (and only for duple and triple meter sentences). In freely-produced speech, the timing constraints on individual syllables do not appear to be strong enough to produce such sharp phase separations.

However, in both metronome-timed and untimed speech, there was evidence for two different 'Strong' and 'weak' phase regions, which could act as attractors for syllable occurrence (e.g. as also proposed by Cummins & Port, 1998). Therefore, the relative prosodic strength of syllables (Strong or weak) could still be inferred from their Stress phase of occurrence. In the next section, the new Prosodic Strength Index is developed on this basis.

5.3.2 THE NEW PROSODIC STRENGTH INDEX (PSI)

In the previous section, it was observed that the prosodic strength of a syllable ('Strong' or 'weak') was related to the Stress phase region at which it occurred. Here, a new Prosodic Strength Index (PSI) is developed which converts a syllable's Stress phase of occurrence into a parametric measure of its prosodic prominence or strength.

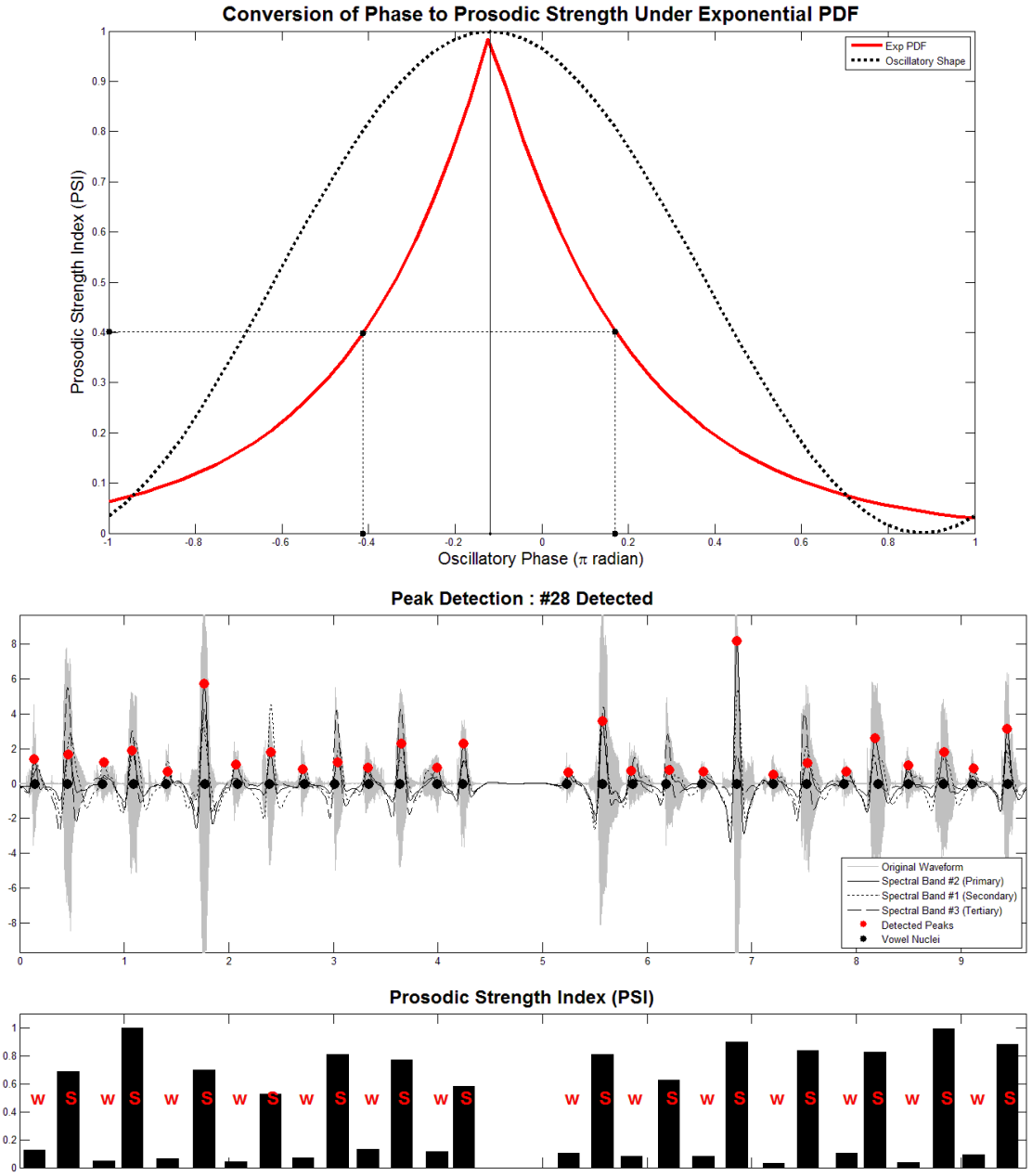
In the metronome-timed sentences, 'Strong' syllables occurred within a Stress phase region of -0.4π to 0.2π radian, with the peak probability of occurrence at -0.1π radian (see Figure 5.3 in the previous section). Conversely, 'weak' syllables were more widely-spaced, but tended to occur most frequently around ±π radian. Therefore, the PSI should have a shape where it is at a maximum around -0.1π radian, and at a minimum around ± π radian.

One mathematical function that possesses this shape is the exponential probability density function (PDF) with a mean of 1, shown in the red line in Figure 5.4. When applied to the *absolute* phase value, this exponential probability density function is unimodal, symmetric, and has a more convex shape than the 'normal' Gaussian PDF used in the original AMPH model (the basis of the Stress Phase Code). Consequently, it is particularly suitable for describing a property (i.e. prosodic strength) that reaches a maximum at one point (i.e. -0.1π radian), then falls off exponentially and symmetrically toward either side of this maximum.

Since the shape of the exponential PDF is more convex than the 'normal' bell-shaped PDF, phase values close to the maximum will increase much more rapidly in their PSI values than phase values far away from the maximum. This results in a more conservative assignment of prosodic strength, since phase values have to be relatively close to the maximum before they can achieve high PSI values. This is consistent with the narrow width of the 'Strong' phase region observed in the distribution of vowel nuclei in metronome-timed speech, where 'Strong' vowel nuclei occurred within a narrow 0.6π radians-wide phase window. This 0.6π radians width is substantially less than the 1π radians width expected if the total phase space of 2π radians was divided equally into 'Strong' and 'weak' regions. Hence, the convex-shaped exponential function appropriately reflects this heightened sensitivity for phase values very close to the maximum (so that small phase changes here translate into large PSI gains).

According to the exponential PDF, PSI values of at least 0.4 (corresponding to phase values between -0.4π and 0.2π radian) would be considered 'Strong', whilst values under 0.4 would be considered 'weak' (as shown in Figure 5.4). For the S-AMPH model, there were 5 Stress tiers that could each contribute a different phase value to the PSI of a given syllable peak. Therefore, the *circular mean* of these 5 different phase values was used to assign the PSI. An example of prosodic strength assignment for an iambic (w-S) metronome-timed sentence is shown in the bottom panel of Figure 5.4, applying a 'Strong' PSI threshold value of 0.4. In this example, all 28 syllables were correctly assigned as being either 'Strong (S)' or 'weak (w)'.

Figure 5.4. (top) Exponential PDF as the basis for the Prosodic Strength Index (PSI). Note that the peak of the function lies at -0.12 pi radian, and phase values between -0.42 pi to 0.18 pi radian achieve PSI values above 0.4. (middle) Example of metronome-timed Iambic (w-S) meter sentence, syllable peaks detected (red dots) and actual location of vowel nuclei (black dots). (bottom) Assignment of syllable prosodic strength using the PSI. Individual bars correspond to syllables, and the height of each bar shows the PSI value. Syllables with a PSI value of ≥0.4 were considered 'Strong (S)', syllables with a PSI value of <0.4 were considered 'weak (w)'.



6 FUNCTIONAL EVALUATION OF THE S-AMPH & AMPH MODELS

In this final chapter of Part III, a comparison of the relative success of the original AMPH model and new S-AMPH model is presented. Two evaluation measures were adopted. These are (1) the identification of syllable vowel nuclei using Syllable modulator 'peaks' ; and (2) the assignment of syllable stress (Strong or weak). For both of these evaluation measures, d' statistics were computed to evaluate the level of success achieved by each model.

The same two sets spoken material that were used to develop the prosodic indices in Chapter 5 were also used for the functional evaluation here. These two sets of spoken material were : (A) the 9 metrical variations of the sentence *"Jack and Jill went up the hill..."*, spoken in time to a metronome beat ; and (B) the 20 nursery rhyme sentences that were freely-produced. The results for identification of syllable vowel nuclei are presented first in Section 6.1, followed by the results for assignment of syllable stress in Section 6.2.

6.1 SYLLABLE VOWEL NUCLEUS DETECTION

6.1.1 EVALUATION PROCEDURE

For both the AMPH and S-AMPH models, an automated peak-detection procedure was used to detect Syllable modulator peaks that corresponded to syllable vowel nuclei. For the S-AMPH model, the parameters used for peak detection and selection were as detailed in the Section 5.2.2. For the AMPH model, syllable modulator peaks were also detected using Matlab's peak detection algorithm, using a minimum peak height of 0.3 standard deviations. No additional peak selection was required for the original AMPH model, as only one set of peaks was produced from the wholeband Syllable modulator (compared to the 5 sets of peaks produced by the 5 spectral bands in the S-AMPH).

In Section 5.2.1, it was observed that the vast majority of syllable vowel nuclei were located in close proximity to the oscillatory peak (0 pi radian) of the Syllable tier modulator. However, while the actual vowel nuclei may be closely associated with Syllable modulator

peaks, the converse may not be true - not all the automatically-detected Syllable modulator peaks may be 'caused' by syllable vowel occurrence. For example, peaks in the Syllable modulator could also correspond to fricative sounds (especially in high frequency bands), or to random background noise. If the Syllable peak detection and selection criteria was too lax, these spurious peaks could be mistakenly identified as real syllable vowels by the model. Conversely, if the Syllable peak detection criteria was too strict, some peaks corresponding to actual vowel nuclei might be missed. This could occur, for example, if the uttered syllable was very soft, or very brief. In signal detection terms, these two possibilities (spurious peaks and missed syllables) would correspond to 'false alarms' and 'misses' respectively, as shown in Table 6.1.

Figure 6.1 shows an example of a freely-produced sentence from the nursery rhyme "Lucy Lockett" where both misses and false alarms occurred. Here, syllable peaks were detected using the S-AMPH model. Note that although the correct total number of syllable peaks was detected (24), these actually included 1 miss (yellow box) and 1 false alarm (red box). Hence, if one were simply to count the total number of syllable peaks that the model detected, without considering whether any of these were spurious, the success of the model would be over-estimated. In view of this, a d' analysis was used to provide a more complete picture of the success of the model in terms of the number of 'hits', 'misses', 'false alarms' and 'correct rejections' produced. The procedures used for computing the responses in each category are detailed next.

Figure 6.1. Example of syllable peak detection for freely-produced nursery rhyme 'Lucy Lockett'. The top panel shows the original waveform (grey), the 3 highest-power Syllable Tier modulators from Spectral Bands 2, 3 & 4 (lines), automatically-detected syllable peaks (red dots) using the revised S-AMPH model, and manually-measured vowel nuclei (black dots).

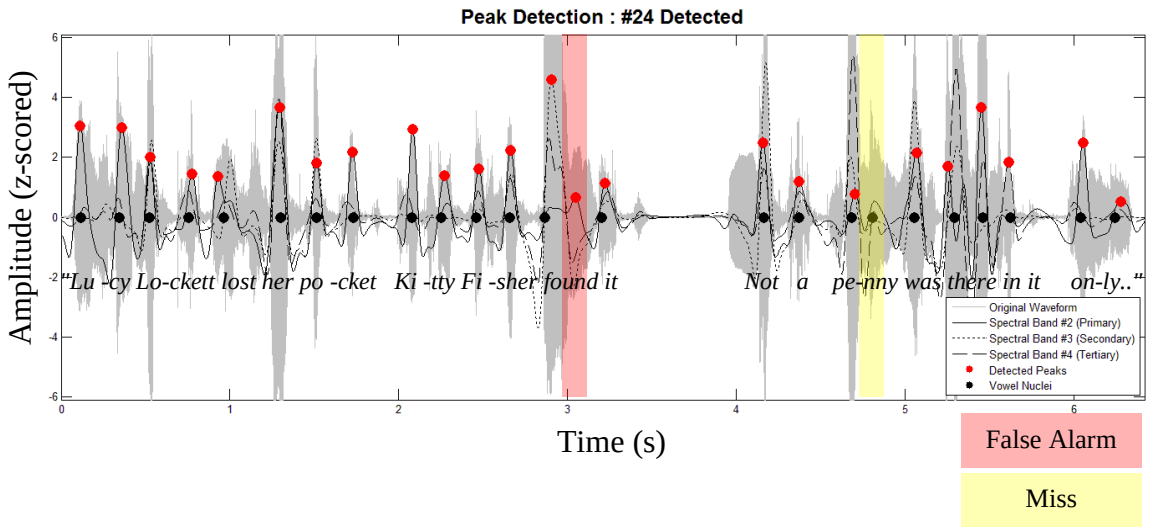


Table 6.1. Summary of signal detection terms and their relevance to syllable vowel nucleus detection

	Syllable Peak Detected	Syllable Peak NOT Detected
Syllable Vowel Nuclei	**HIT**	**MISS**
Not Syllable Vowel Nuclei	**FALSE ALARM**	**CORR REJECTION**

Hits. A Syllable modulator peak was considered as having correctly identified a corresponding syllable vowel nucleus (i.e. a 'hit') if the peak lay within +/- half the mean syllable length (for that sample) of the actual syllable vowel nucleus[18].

Misses. A 'miss' was any syllable vowel nucleus that was *not* detected by the model. This was computed by subtracting the total number of 'hits' from the total number of *actual* syllable vowel nuclei in the sample.

False alarms. A 'false alarm' was a Syllable peak that was identified by the model as corresponding to a real syllable vowel, but did *not* actually correspond to any syllable vowel. The number of false alarms for each sample was computed by subtracting the number of correctly identified syllable vowels ('hits') from the total number of peaks identified by the model as corresponding to syllable vowel nuclei.

Correct rejections. A 'correct rejection' was a spurious peak in the Syllable modulator that did *not* actually correspond to a syllable vowel, and was correctly eliminated by the model. To compute the number of correct rejections, an estimate of the *total* number of 'spurious peaks' that did not correspond to a syllable vowel was required. This estimate was generated by (a) detecting all possible peaks in the sample, setting *no* criteria on the mean peak height or distance (i.e. *any* point that was higher than its surrounding neighbours was included) and (b) subtracting the number of actual syllable vowel nuclei in the sample from the value obtained in (a).

However, since there were 5 spectral bands in the S-AMPH used for Syllable peak detection but only 1 band used in the AMPH, there would be 5 times as many spurious peaks detected by the S-AMPH model as compared to the AMPH model. Therefore, the total number of spurious peaks detected from all 5 S-AMPH bands was divided by 5, so that this would be comparable in number to the number of spurious peaks detected in the AMPH

[18] The mid-point, not the onset of the syllable vowel nucleus.

model. Having obtained a normalised estimate of the number of spurious peaks in the sample, the number of correct rejections was then computed by subtracting the number of false alarms from this (normalised) number of spurious peaks.

6.1.2 RESULTS

6.1.2.1 Metronome-Timed Speech (Sample Set A)

Table 6.2 shows the mean syllable detection percentages (averaged over all 27 samples from 3 speakers) and the resulting mean d' and criterion (response bias) values for the original AMPH and new S-AMPH models[19]. As shown in the table, both models had very high d' scores, with the AMPH model producing a d' of 4.81, and the revised S-AMPH model producing a d' of 5.01. On a non-parametric Wilcoxon matched-pairs test[20], there was no significant difference between the d' of the two models (Z = 0.00, p = 1.00). Therefore, both models performed equally well in terms of syllable detection, correctly detecting 94%-97% of syllable vowel nuclei (hits).

Table 6.2. Syllable detection performance for AMPH and S-AMPH models (metronome speech)

AMPH : d' = 4.81, bias = 0.34

(mean over 27 samples)	Syllable Peak Detected	Syllable Peak NOT Detected
Syllable Vowel Nuclei	94.2% (HIT)	5.8% (MISS)
Not Syllable Vowel Nuclei	5.6% (FALSE ALARM)	94.4% (CORR REJECTION)

S-AMPH : d' = 5.01, bias = 0.14

(mean over 27 samples)	Syllable Peaks Detected	Syllable Peaks NOT Detected
Syllable Vowel Nuclei	97.1% (HIT)	2.9% (MISS)
Not Syllable Vowel Nuclei	1.5% (FALSE ALARM)	98.5% (CORR REJECTION)

[19] Note that the mean d' score shown here is the average d' score taken across all speech samples. This mean d' value is not the same as the single d' score that would be computed from the mean *percentages* shown in the tables. That is, the mean d' is not the same as the d' of the mean percentages. The mean d' is chosen for display because it better reflects the individual sample d' values that were entered into the statistical test.

[20] The distribution of d' scores across the 3 speakers was *not* normal (Kolmogorov-Smirnov test, $p<.05$), therefore a non-parametric test was used.

In terms of response bias, the S-AMPH model had a lower (more neutral) criterion of 0.14, compared to 0.34 for the AMPH model. The more positive criterion value for the AMPH model indicates that it was more conservative than the S-AMPH model in identifying syllable vowel nuclei, since the signal had to be stronger to elicit a detection response (hit or false alarm) from the model.

6.1.2.2 Freely-Produced Speech (Sample Set B)

Although the performance of the two models in syllable vowel detection was equivalent for metronome-timed speech, there was a clear difference in performance for un-timed speech. Table 6.3 shows the mean syllable detection percentages (averaged over all 120 samples from 6 speakers) and the resulting mean d' and criterion (response bias) values for the original AMPH and new S-AMPH models.

For this corpus of un-timed speech, the S-AMPH model produced 16.9% more hits than the AMPH model, and registered less than half the AMPH false alarm rate. As a result, the S-AMPH model had a much higher d' value of 2.29, compared to 0.52 for the original AMPH model. On a paired-samples t-test[21], this difference in d' scores between the two models was highly significant ($t(5) = -6.67$, $p<.01$).

Table 6.3. Syllable detection performance for AMPH and S-AMPH models (un-timed speech)

AMPH : d' = 0.52, bias = -0.20

(mean over 120 samples)	Syllable Peaks Detected	Syllable Peaks NOT Detected
Syllable Vowel Nuclei	66.6% (HIT)	33.4% (MISS)
Not Syllable Vowel Nuclei	47.2% (FALSE ALARM)	52.8% (CORR REJECTION)

S-AMPH : d' = 2.29, bias = -0.01

(mean over 120 samples)	Syllable Peaks Detected	Syllable Peaks NOT Detected
Syllable Vowel Nuclei	83.5% (HIT)	16.5% (MISS)
Not Syllable Vowel Nuclei	19.9% (FALSE ALARM)	80.1% (CORR REJECTION)

[21] The distribution of d' scores across the 6 speakers was normal (Kolmogorov-Smirnov test, $p>.05$), therefore a parametric test was used.

As was the case for metronome-timed speech, the S-AMPH model again had a more neutral criterion value of -0.01, compared to -0.20 for the AMPH model.

6.1.3 Summary & Discussion for Syllable Vowel Nucleus Detection

For metronome-timed speech, both models performed on par (94% vs 97%) for syllable vowel identification. However, for un-timed speech, the performance of the S-AMPH model out-stripped the AMPH model, producing 17% more hits, and half the false alarm rate. Therefore, the performance of the S-AMPH model on syllable vowel identification was superior to the AMPH model, but *only* for un-timed (freely-produced) speech.

In metronome-timed speech, each syllable is clearly separated in time by brief gaps or pauses. These gaps are inserted by the speaker to maintain the isochronous timing between syllables. By contrast, in freely-produced speech, co-articulation occurs so that these brief gaps between the syllables are removed. In the AMPH wholeband envelope, this co-articulation leads to a 'blending' of energy between adjacent syllables so that the peaks from individual syllables merge into one another. This may account for the drastic drop in performance for the AMPH model for syllable vowel identification in freely-produced speech.

However, if the speech signal is divided into several spectral bands, *temporally* co-articulated syllables can still be distinguished if there is sufficient *spectral* separation between their vowel sounds, so that the syllable peaks appear in *different* spectral bands. Hence, the greater spectral resolution of the S-AMPH model pays off for syllable vowel identification in fluent, un-timed speech.

In comparison to the other methods for syllable detection reviewed in Section 1.8 of the Introduction, the S-AMPH model appears to have performed reasonably well. The current amplitude-based method is particularly similar to that used by Pfitzinger et al (1996), who also used peaks in the low-pass filtered (~<10 Hz) envelope as candidates for syllables. Pfitzinger et al (1996) reported accuracy rates of 87% and 79% for read and spontaneous speech respectively. These percentages are not distant from the S-AMPH accuracy rates of >97% and >80% for metronome-timed and un-timed speech respectively (although no

spontaneous speech was used in this evaluation). By comparison, supervised machine-learning methods by Howitt (2000) and Shastri et al (1999) are able to achieve syllable detection accuracy rates of up to 88%, whereas Kanlinli's (2011) biologically-inspired auditory attention model achieves an impressive 92% accuracy on the same TIMIT corpus of read speech. Therefore, in terms of syllable detection, the current S-AMPH method performs as well as other unsupervised amplitude-based methods. From the psychological perspective, this suggests that the amplitude envelope of the raw acoustic signal already provides very strong cues as to the location of most (i.e. ~80%) syllable vowel nuclei in speech, even before the higher-order contextual or lexical knowledge available to the listener is considered. To achieve any further gains in accuracy, more complex methods that take these higher-order factors into account will need to be employed. These include machine learning methods, HMMs, or methods where the search for syllables is guided by attention (e.g. Kalinli, 2011) or speech rhythm (e.g. Zhang & Glass, 2009).

6.2 PROSODIC STRESS ASSIGNMENT

6.2.1 EVALUATION PROCEDURE

The second evaluation procedure involved assessing the performance of the models in syllable prominence or stress assignment (Strong or weak). Recall that for the AMPH model, syllable prominence was computed using the Stress Phase Code (Chapter 2, Section 2.5.2). This employed a Gaussian probability density function (PDF) transformation of the Stress modulator phase value concurrent with each syllable peak. For all speech samples, the threshold value used was 0.5 (as per Chapter 2) so that syllables achieving values greater than or equal to 0.5 were assigned a 'Strong' status. Syllables achieving values less than 0.5 were assigned a 'weak' status.

For the S-AMPH model, syllable prominence was computed using the Prosodic Strength Index (PSI), described in Chapter 5, Section 5.3.2. This employed a different *exponential* PDF transformation. Also, as there were now 5 Stress modulators from the 5 spectral bands (instead of 1 wholeband Stress modulator), the Stress phase value used for computing the PSI was the circular mean of the 5 Stress modulator phase values that were concurrent with the Syllable peak. For the metronome-timed speech sample, the threshold PSI value used was 0.4 (as explained in Chapter 5, Section 5.3.2). Syllables achieving values greater than or equal to 0.4 were assigned a 'Strong' status. Syllables achieving values less than 0.4 were assigned a 'weak' status. However, for the freely-produced speech samples, it was found that the PSI threshold value of 0.4 yielded poor results. Recall that this PSI threshold had been determined using the Stress phase-distribution patterns of the metronome-timed speech corpus, and thus may not be appropriate for the freely-produced speech corpus. Therefore, a lower PSI threshold of 0.22 was used for the un-timed speech corpus. This PSI threshold was selected to match the false alarm rate produced by the AMPH model (~30%) as closely as possible so that the two models could be compared on equal footing. The implications of changing the PSI threshold for the S-AMPH model are discussed later at the end of Chapter 6.

The automatically computed syllable prominence assignments were then compared against the *actual* prosodic status of each syllable in the utterance. As described in Chapter 5, Section 5.1, for the metronome-timed corpus, the actual stress patterns were known and

deliberately produced by the speakers. On the other hand, for the freely-produced speech corpus, speakers were not explicitly instructed to produce the nursery rhyme sentences with any particular stress pattern. Therefore, speakers either used the familiar stress template for each nursery rhyme, or they produced their own stress patterns which differed from the familiar template. To ascertain the exact stress patterns that were produced by each speaker, the nursery rhyme sentences were manually stress-transcribed by a female native English speaker with formal training in Linguistics (not the author).

To evaluate the success of the models in prosodic stress assignment, d' values were computed from the hits, misses, false alarms and correct rejections generated by each model. These 4 types of responses are shown in Table 6.4. Hits were prosodically-stressed syllables that were correctly assigned a 'Strong' status. Misses were prosodically-stressed syllables that were incorrectly assigned a 'weak' status. False alarms were unstressed syllables that were incorrectly assigned a 'Strong' status, and correct rejections were unstressed syllables that were correctly assigned a 'weak' status. For this evaluation process, only correctly identified syllable peaks were included in the analysis (i.e. 'hits' from the previous syllable vowel identification process). This was done so that the effectiveness of prosodic stress assignment could be evaluated independently of the model's success in identifying syllables in the first place.

Table 6.4. Summary of signal detection terms and their relevance to prosodic stress assignment

	Assigned 'Strong'	Assigned 'weak'
Strong syllable	**HIT**	**MISS**
Weak syllable	**FALSE ALARM**	**CORR REJECTION**

6.2.2 RESULTS

6.2.2.1 Metronome-Timed Speech (Sample Set A)

Table 6.5 shows the mean stress assignment percentages (averaged over all 27 samples from 3 speakers) and the resulting mean d' and criterion (response bias) values for the original AMPH and new S-AMPH models. From the table, it may be observed that both the AMPH and S-AMPH models performed very well. A very high proportion (>95%) of stressed syllables were correctly assigned with a Strong status (i.e. hits), and high mean d' values were achieved by both models (4.10 and 4.44 for AMPH and S-AMPH respectively).

The S-AMPH model (using the exponential PDF) achieved slightly (1.6%) more hits than the original model, which had used a Gaussian PDF to calculate the PSI. Moreover, this improvement also came together with a slightly reduced false alarm rate (7.7% vs 9.1%), indicating that the improved performance was not merely an artifact of relaxing the detection criterion. Consistent with this interpretation, the gains for the S-AMPH model were achieved while keeping the response bias the same as for the AMPH model (-0.27 vs -0.26). However, on a non-parametric Wilcoxon matched-pairs test[22], there was *no* significant difference in the d' scores of both models ($Z = 1.60$, $p = 0.11$). Therefore, any improvements produced by the S-AMPH model in prosodic stress assignment were only slight.

Table 6.5. Prosodic stress assignment performance for AMPH and S-AMPH models, for metronome-timed speech

AMPH : d' = 4.10, bias = -0.26

(mean over 27 samples)	Assigned 'Strong'	Assigned 'weak'
Strong syllable	95.0% (HIT)	5.0% (MISS)
Weak syllable	9.1% (FALSE ALARM)	90.9% (CORR REJECTION)

S-AMPH *(PSI threshold of 0.4)***: d' = 4.44, bias = -0.27**

(mean over 27 samples)	Assigned 'Strong'	Assigned 'weak'
Strong syllable	96.6% (HIT)	3.4% (MISS)
Weak syllable	7.7% (FALSE ALARM)	92.3% (CORR REJECTION)

[22] The distribution of d' scores across the 3 speakers was *not* normal (Kolmogorov-Smirnov test, $p<.05$), therefore a non-parametric test was used.

6.2.2.2 Freely-produced speech (Sample Set B)

Finally, the performance of both models in prosodic stress assignment was evaluated for freely-produced speech. Table 6.6 shows the mean stress assignment percentages (averaged over all 120 samples from 6 speakers) and the resulting mean d' and criterion (response bias) values for the original AMPH and new S-AMPH models.

From inspection of Table 6.6, the S-AMPH model appeared to perform better than the AMPH model, producing almost 10% more hits, while keeping the same false alarm rate (recall that the S-AMPH PSI threshold of 0.22 used here ensured that the false alarm rates for the two models would be as similar as possible). However, on a paired-samples t-test[23], the difference in d' scores between models was again *not* significant (t(5) = -1.28, p=.26). Therefore, as was the case for the metronome-timed corpus, although there appeared to be gains in performance for the S-AMPH model, these gains were not robust enough to reach statistical significance.

Table 6.6. Prosodic stress assignment performance for AMPH and S-AMPH models, for un-timed speech

AMPH : d' = 1.19, bias = 0.24

(mean over 120 samples)	Assigned 'Strong'	Assigned 'weak'
Strong syllable	61.7% (HIT)	38.3 % (MISS)
Weak syllable	31.7 % (FALSE ALARM)	68.3% (CORR REJECTION)

S-AMPH *(PSI threshold of 0.22)* **: d' = 1.35, bias = 0.03**

(mean over 120 samples)	Assigned 'Strong'	Assigned 'weak'
Strong syllable	70.2 % (HIT)	29.8 % (MISS)
Weak syllable	30.8 % (FALSE ALARM)	69.2% (CORR REJECTION)

[23] The distribution of d' scores across the 6 speakers was normal (Kolmogorov-Smirnov test, p>.05), therefore a parametric test was used.

6.2.3 SUMMARY & DISCUSSION FOR PROSODIC STRESS ASSIGNMENT

In both metronome-timed and freely-produced speech, there was no significant difference in the d' scores of the AMPH and S-AMPH models, even though the S-AMPH model appeared to show gains over the AMPH model. This suggests that the Stress phase coding of syllable prominence is highly robust, and the shape of the PDF transformation function used (Gaussian or exponential) has no significant effect on the effectiveness of this coding scheme.

However, it should be noted that the performance of both models in stress assignment for freely-produced speech was far from ideal. Neither model succeeded in assigning more than 70% of syllables with the correct prosodic status when speech was more produced in this more 'natural' context. However, while this margin of error (30%) may appear to be large, it is not inconsistent with the performance of other models developed specifically for automatic stress transcription.

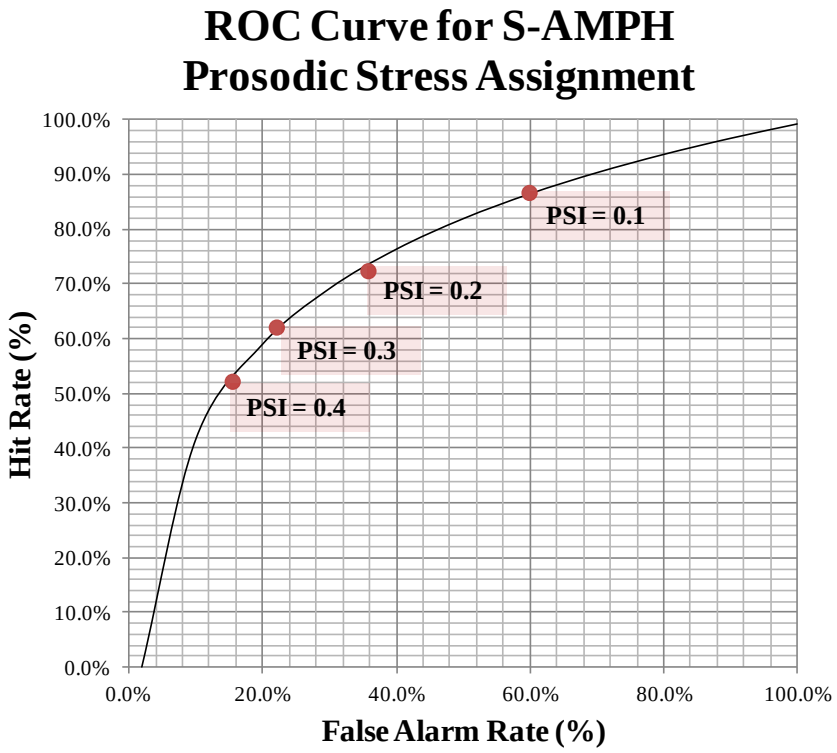
For example, Silipo & Greenberg (1999) developed models for automatic stress transcription of spontaneous speech using amplitude, duration and pitch cues. They tested a variety of models where these cues were either used singly or in paired combination. Silipo & Greenberg reported that the best performance was obtained when *both* duration and amplitude cues were used in combination, yielding correct identification of ~80% of stressed syllables and ~78% of unstressed syllables. Performance using amplitude cues alone was ~64% for stressed syllables and ~65% for unstressed syllables.

For freely-produced speech, the original AMPH model was similar in performance to Silipo & Greenberg's amplitude-only model, correctly identifying 62% of stressed syllables and 68% of unstressed syllables. The performance of the S-AMPH model was statistically equivalent, where ~70% of stressed and unstressed syllables were correctly indentified using just amplitude cues from the speech envelope. Therefore, both AMPH and S-AMPH models performed as well as could be expected, when only amplitude cues were used to infer stress pattern, not taking into account duration or pitch cues.

Finally, the PSI threshold that is used for stress assignment in the S-AMPH model affects the overall performance of the model in terms of trade-offs between hits and false alarms. Figure 6.2 shows this trade-off between hits and false alarm rates for different PSI

threshold values as an ROC (receiver operating characteristic) curve, computed for the un-timed speech corpus. If the S-AMPH model is to be used in future studies for automatic stress transcription, the PSI threshold should be adjusted to reflect the aims and priorities of the study.

Figure 6.2. ROC curve for S-AMPH prosodic stress assignment using different PSI threshold values, for the un-timed speech corpus (sample set B.) Actual computed hit rates (y-axis) and false alarm rates (x-axis) for 4 PSI threshold values are shown as red dots. The solid black curve indicates the logarithmic line of best fit through these points.



For example, if a low false alarm rate is the priority (i.e. no weak syllable should be erroneously labelled as Strong), then a higher PSI threshold should be used. Conversely, if a high hit rate is important (i.e. as many Strong syllables as possible should be detected), then a low PSI threshold should be used. As shown in Figure 6.2, by using a PSI threshold of 0.1, a hit rate of 87% may be achieved, but this creates a very high false alarm rate of 60%. On the other end of the spectrum, false alarms may be reduced to just 15%, if a low hit rate of just 52% is sufficient. The PSI threshold of ~0.2 used in the evaluation of un-timed speech here appears to represent the best compromise. At this threshold value, over 70% of Strong syllables are correctly detected, and around half that percentage (35%) of weak syllables are falsely identified, giving a signal-to-noise ratio of around 2.

PART III SUMMARY

Unlike the theoretically-focussed AMPH model, the S-AMPH was developed as a data-driven model. Here, the intention was to use the underlying temporal statistics of the envelope as the basis for the assumed spectro-temporal hierarchical structure of the model, and to derive all ensuing prosodic indices from actual distribution patterns in the speech envelope. This aim was operationalised by applying principal component analysis (PCA) to reduce a high-dimensional spectro-temporal representation of the speech envelope to a non-redundant, lower-dimensional hierarchical representation.

The result of this process was a new 5 (spectral band) by 3 (modulation rate band) spectro-temporal AM hierarchy. Compared to the original AMPH model, this representation was more complex in the spectral domain (5 spectral bands instead of 1), and less complex in the AM rate domain (3 modulation rate bands instead of 5). Most importantly, these bands were derived statistically from the structure of the envelope itself, rather than determined theoretically.

It is worth noting that the 'Stress' and 'Syllable' tiers of the AM hierarchy emerged spontaneously as part of the S-AMPH modulation band representation (albeit with a wider Syllable bandwidth). These two rates of modulation were fundamental to the speech rhythm computation in the original AMPH model. The new S-AMPH model could not have continued to function on similar principles to the AMPH model (i.e. using Stress-Syllable phase relationships) if these two bands had not emerged in the PCA analysis as *separate* modulation bands, with modulation rates that were similar to the AMPH model. The fact that Stress and Syllable modulation bands *did* emerge from the modulation statistics indicates that the initial theoretically-motivated intuitions about the structure of the speech signal at these two key modulation rates were on the right track. This is also consistent with the results of the tone-vocoder experiment (Chapter 3), which had indicated that listeners relied heavily on modulations in these two rate bands for speech rhythm perception, and were also sensitive to the phase relationship between these two bands.

The hypothesis that the Stress cycle was functionally equivalent to the linguistic prosodic foot was also examined by computing the distributions of syllable vowel nuclei with respect to Stress AM phase in nursery rhyme sentences with different prosodic meters. For

metronome-timed speech, the phase distributions produced by speakers provided support for this hypothesis. First, speakers tended to divided Stress 'phase-space' into 'Strong' and 'weak' regions, placing Strong syllables near the Stress peak, and weak syllables far from the Stress peak. Second, speakers further sub-divided the 'weak' phase region according to the number of weak syllables contained in the prosodic foot.

For un-timed speech, speakers were also found to place syllables more frequently at Strong and weak regions in Stress phase space, but these phase regions acted as weak attractors, rather than imposing the strong constraints observed in metronome-timed speech. These findings indicate that speakers do indeed use Stress phase to constrain the timing of stressed and unstressed syllables. However, a strong view of the Stress cycle = prosodic foot hypothesis is only supported in metronome-timed speech, not in freely-produced speech.

Using the phase regions typically occupied by Strong (stressed) and weak (unstressed) syllables in the data, a new Prosodic Strength Index was developed. This used an exponential probability density function to convert Stress phase values into a prominence index ranging from 0 to 1. By setting a threshold PSI value, syllables could be assigned with a binary stress status ('Strong' or 'weak') depending on whether they achieved PSI values above or below the threshold.

The two models were then functionally compared on two criteria : syllable vowel nucleus detection, and prosodic stress assignment. For syllable vowel nucleus detection, both models performed near ceiling (around 95%) for metronome-timed speech, but the S-AMPH model showed a distinctly superior performance for freely-produced (fluent) speech. For this un-timed speech sample, the S-AMPH model registered a 17% improvement in hit rate as compared to the AMPH model, while more than halving the false alarm rate. Therefore, the multi-band spectral complexity of the S-AMPH model made it better able to handle the challenges of syllable vowel detection in natural speech.

However, for prosodic prominence assignment, both models gave statistically equivalent performances for both metronome-timed and un-timed speech. The accuracy of prosodic stress assignment for un-timed speech was comparable to the performance of amplitude-only models developed specifically for automatic stress-transcription (e.g. Silipo & Greenberg, 1999). Both models used different probability functions to transform the Stress phase values, but performed equally well on the final stress assignment. This suggests that the

fundamental principle of Stress phase indicating syllable prominence is a robust one, not dependent on any particular mathematical transformation.

In Part IV, the new S-AMPH model will be employed to address a variety of experimental questions. Apart from being an analytical tool, the S-AMPH model also represents a novel cognitive and neural framework for understanding speech rhythm, and could provide unique insights into how rhythmic differences arise between individuals, and in different contexts.

PART IV :

USING THE S-AMPH MODEL IN DATA ANALYSIS

TWO EXPERIMENTAL CASE STUDIES

In these chapters, the S-AMPH model was used as an analytical tool to compare the underlying temporal structure of different types of speech. Two different research questions were addressed in two separate experiments[24].

In Chapter 7, the S-AMPH model was used to determine how the spectro-temporal structure of child-directed speech (CDS) differed from that of adult-directed speech (ADS). When adults speak to children, they prosodically-enhance their speech in order to accommodate the needs of the child listener. Here, the spectro-temporal changes accompanying this prosodic enhancement were investigated. In the spectral domain, CDS showed a specific 'boost' at middle frequencies (~1200 Hz), suggesting that vowel sounds were relatively louder and more strongly co-modulated across spectral channels. In the modulation domain, CDS samples were found to be more rhythmically-regular than ADS samples. CDS also showed a more tightly-nested AM hierarchical structure that was indicative of stronger prosodic patterning (e.g. more frequent syllable stress). These global changes in the spectro-temporal structure of CDS are consistent with the enhancement of word and phrase boundaries in the acoustic signal. Such word boundary exaggeration could help the child to segment words from the speech stream more easily, facilitating speech comprehension and new vocabulary acquisition.

In Chapter 8, the perception and production of rhythmic speech was investigated in adults with and without developmental dyslexia. Participants performed three different speech rhythm tasks, testing speech rhythm perception, speech rhythm entrainment (tapping) and speech rhythm production respectively. Performance on these tasks was measured using conventional measures, as well as using indices from the S-AMPH model. In all 3 perception and production tasks, dyslexic individuals consistently showed disruptions to syllable-level timing. Individual differences in syllable-timing (both in perception and production) were strongly related to performance in phonological processing and reading measures. The S-AMPH indices also uncovered differences between dyslexics and controls that were not evident from conventional analysis. Therefore, envelope-based measures could be useful analytical tools to complement more traditional methods of speech analysis.

[24] The experimental design, data collection and analysis were all carried out by the author as part of this thesis.

7 DIFFERENCES IN TEMPORAL STRUCTURE BETWEEN CHILD-DIRECTED AND ADULT-DIRECTED SPEECH

Child-directed speech (CDS) and adult-directed speech (ADS) refer to two different speaking 'registers' or styles. These differences are thought to arise because the speaker is adapting his or her speaking style to the language abilities and needs of his or her audience. In child-directed speech, these adaptations reflect the fact that the child is a novice language-learner, rather than an expert. Much of this adaptation occurs at the lexical and syntactic levels. For example, child-directed speech contains simpler syntactic structures (Sachs et al, 1976), shorter sentences (Barnes et al, 1983), and pertains to topics that are of interest to the child (Ferguson, 1977; Ferguson & Debose, 1977).

However, adaptation also occurs at the perceptual-acoustic level. Child-directed speech is *prosodically-enhanced*, making it more interesting and engaging for the listener, and conveying a positive affect (Fernald, 1989). The perceptual-acoustic properties of CDS have commonly been studied in terms of pitch, duration, speaking rate, pauses, etc (Broen, 1972; Fernald & Simon, 1984; Fernald, 1989; Albin & Echols, 1996). However, child-directed speech could also show different rhythm patterns and a different temporal organisation as compared to adult-directed speech. These adaptations in temporal structure could likewise be helpful for the child listener. This study aims to identify any such differences in temporal rhythmic structure between adult- and child-directed speech, using the S-AMPH as an analytical tool.

7.1 METHODS

7.1.1 PARTICIPANTS

Six female native British English speakers contributed samples of child-directed and adult-directed speech. All of the participants were highly fluent English speakers. Participants were selected on the basis of having had extensive prior experience in working with children. Prior experience with children was important since the participants would have to produce

realistic child-directed speech 'on demand' during the recording session, when no children were actually present. Two of the six participants were Cambridge University lecturers in early years education (having previously been teachers), and a further two participants were currently working as early years teachers. One participant was a speech and language therapist working with children. The last participant was an ex-teacher and doctoral student whose research involved working with children using poetry. All the speakers were familiar with the children's nursery rhymes used in this study.

7.1.2 Speech Recording Procedure

Each participant was recorded individually in a single 2-hour session. The recording session was conducted in a quiet location (either in the participant's home/office, or in a laboratory testing room) to minimise background noise. A TASCAM digital recorder (44.1 kHz, 24-bit) was used for the speech recording, together with an AKG C1000S condenser microphone. The microphone was fixed to a microphone stand, and placed at a comfortable distance and height for the speaker. During the recording session, participants read printed texts out of a folder, or out of a children's book. To assist participants in producing appropriate 'child-directed' or 'adult-directed' speech, picture prompts were used. For 'child-directed speech' (CDS) samples, participants were shown a picture of young children of a nursery age, and told to speak in a lively and engaging manner as if they were reading to the children in the picture. This CDS picture prompt is shown on the left in Figure 7.1.

Figure 7.1. CDS (left) and ADS (right) target picture prompts. The relevant picture was presented to participants as the intended recipient(s) for their utterances.

For 'adult-directed speech' (ADS) samples, participants were shown a picture of a professional-looking adult, and told to speak in a clear and formal way to this individual. The picture prompt for ADS is shown in Figure 7.1 on the right. A more formal version of ADS was requested to ensure that the participants produced clearly-enunciated ADS samples. If the ADS speech was sloppily produced, a difference between CDS and ADS samples could be due to a lack of clarity in the ADS samples, rather than being adult- or child-directed per se.

7.1.3 SPEECH MATERIALS

Each speaker produced four speech corpora, two spoken in CDS and two spoken in ADS. These were (1) Nursery rhymes produced in child-directed speech (CDS Rhyme); (2) Nursery rhymes produced in adult-directed speech (ADS Rhyme); (3) Children's stories produced in child-directed speech (CDS Story); and (4) Spontaneous conversation spoken in an adult-directed manner (ADS Conversation).

(1) CDS Rhyme & (2) ADS Rhyme

In these two corpora, the spoken material (nursery rhymes) was held constant, and speakers were told to change their manner of speaking as if addressing a child (CDS Rhyme) or an adult (ADS Rhyme). A total of 44 familiar children's nursery rhymes were used, and each speaker produced all 44 nursery rhymes first in ADS, then in CDS. These were the same set of nursery rhymes that had previously been used to derive the S-AMPH model in Chapter 3 (where only the CDS recordings were used), and are listed in Appendix 4.1. Since the spoken material was identical, any differences between the two corpora should cleanly reflect perceptual-acoustic variations in speaking style, without being affected by differences in words or syntax.

For the CDS version of the rhymes, participants spoke in a lively and rhythmic fashion, often speaking to the rhythm of tunes that were associated with these nursery rhyme (although the rhymes were not actually sung). For the ADS version of the rhymes, participants' utterances were less metrically-regular, and they attempted to produce the sentences with a 'normal' prosodic pattern, as appropriate to an adult audience. For example, for the nursery rhyme 'Polly Put the Kettle On' shown below, the CDS version followed the

duple beat of the sung version of the rhyme. Therefore stress was placed (CAPS) on every other syllable, and the syllables were grouped evenly into sets of trochaic feet (underlined). In the ADS version, the sentence was instead produced as if a verbal instruction was being given to Polly. Therefore a pause was introduced after the word 'Polly', and the words 'put the kettle on' formed a single long foot instead of being portioned into smaller feet. Therefore, the ADS Rhyme utterances had a less regular rhythm than the CDS Rhyme utterances.

CDS : *"PO-lly PUT the KE-ttle on.."*
ADS : *"PO-lly [pause] , PUT the kettle on.."*

Since the nursery rhymes varied in length, each rhyme was repeated between 1-3 times to produce an adequate amount of spoken material for analysis. The shorter nursery rhymes were repeated more times than the longer nursery rhymes. Appendix 7.1 shows the duration of each nursery rhyme for each speaker, and the number of repetitions produced, for both ADS and CDS modes of speaking. The syllable rate for each speaker and nursery rhyme was computed by dividing the duration of the sound file by the number of syllables in the nursery rhyme text. These syllable rates are also shown in Appendix 7.1. On average across the six speakers, the mean syllable rate was slower for CDS rhymes than for ADS rhymes (3.2 syllables per second versus 3.6 syllables per second). However, a paired t-test indicated that this difference was not significant ($t(5) = 1.36$, $p=.23$).

As nursery rhymes are normally directed to children rather than to adults, speakers could have found it awkward to produce nursery rhymes to adults, making ADS Rhymes unrepresentative of 'natural' ADS speech. Therefore, speech samples (3) CDS Story and (4) ADS Conversation were also collected. These were expected to be more representative of 'naturally-produced' CDS and ADS. Both these speech corpora involved narrative 'story-telling' to the listener, but the spoken material was allowed to differ, in order to be appropriate to the intended child or adult listener.

(3) CDS Story

For the third speech corpus, participants read 5 classic children's stories in a child-directed manner. These stories were taken from the children's book 'The Puffin Baby and Toddler Treasury' (Puffin Books, 1998). The titles of the stories were 'The Gingerbread Man', 'The Three Billy Goats Gruff', 'Goldilocks and the Three Bears', 'The Three Little Pigs' and

'The Ugly Duckling'. The language used in these stories was simple, and appropriate for nursery and preschool aged children. The following is an excerpt from the story 'The Three Little Pigs'.

"Once upon a time there were three little pigs who lived in a very small house with their mother. One day their mother gathered them all together and said, "It is time that you left our little house and built your own homes." The three little pigs said goodbye to their mother and as they set off down the road she called after them, "Beware the wolf doesn't catch you and eat you!"

Depending on the reading speed of the speaker, the recordings of the stories ranged from 4 to 8 minutes in length for each story. To produce speech samples that were comparable in length and quantity to the 44 nursery rhymes, nine continuous sections were extracted from each of the five stories, giving 45 story sections for each speaker. For each speaker, the length of each story segment was the same as the mean length across their 88 CDS + ADS nursery rhyme samples. These mean lengths for each speaker are shown in Table 7.1. Different segment lengths were used for each individual (rather than one standard length for the whole group) because the speakers differed substantially in speaking rate. Therefore, to ensure that the same quantity of spoken material (i.e. syllables and words) was captured for each individual across nursery rhymes and stories, individually-adjusted segment lengths were used. Therefore, if an individual's speaking rate was slower, this would result in a longer mean section length for the nursery rhymes. This longer length would then also be used to segment her read stories so that each story sample would contain on average the same quantity of syllables and words as each nursery rhyme sample.

Table 7.1. Mean lengths for CDS and ADS nursery rhyme samples by speaker, used to determine section length for CDS Story samples.

Speaker	**Mean Section Length (s)**
1	25.6
2	23.9
3	27.5
4	34.3
5	24.5
6	26.1

(4) ADS Conversation

For the fourth speech corpus, spontaneously-produced adult-directed speech was recorded. Here, participants were provided with suggested topics to speak about, and were given a few minutes for mental preparation before they began speaking. These suggested topics were :

- Describe a typical day at work (or home)?
- Describe a book that you've read?
- Describe a film that you've watched?
- What leisure activities and hobbies do you enjoy?
- What are you looking forward to in the summer?

Participants were told to speak about each topic for about 2-3 minutes, before moving on to the next topic. Since no verbal feedback was given and participants spoke continuously, these samples essentially comprised a narrative monologue that lasted around 10-12 minutes. As all the participants were eloquent speakers, the sentences in these conversation samples were typically well-formed and grammatically correct. Similar to the CDS Story recordings, sections were extracted from the continuous ADS recording for analysis. These sections were matched in length to the nursery rhymes produced by each speaker. Since the total amount of spontaneous conversation varied from speaker to speaker, the total number of segmented samples also differed from speaker to speaker, ranging from 23 to 34 samples[25], with a mean of around 27 samples.

Table 7.2 summarises the speaking style and material used for the four speech corpora. In total, each of the 6 speakers contributed 44 CDS Rhyme samples, 44 ADS Rhyme samples, 45 CDS Story samples and ~27 ADS Conversation samples. This gave a total of ~160 samples per speaker (each around 25s-35s in length), and a grand total of ~960 samples across all the speakers. Assuming that each speaker maintained a stable speaking rate, each of her samples (CDS Rhyme, ADS Rhyme, CDS Story or ADS Conversation) contained approximately the same amount of spoken material. This quantity matching was important so that in later analyses, differences between the corpora would not be confounded by differences in the amount of spoken material between samples in the corpora.

[25] The number of ADS conversation samples for the 6speakers were 26, 23, 27, 34, 24 and 26.

Table 7.2. Features of the four speech corpora

	Speaking Style	**Type of Material**
(1) CDS Rhyme	Child-directed	Nursery Rhymes
(2) ADS Rhyme	Adult-directed	Nursery Rhymes
(3) CDS Story	Child-directed	Children's Stories
(4) ADS Conversation	Adult-directed	Free Conversation

7.1.4 ANALYSIS PROTOCOLS

The aim of the analysis was to compare the underlying spectro-temporal structure of the child- and adult-directed speech, using modulation patterns in the speech envelope. Therefore, the dimensionality reduction procedure previously used for deriving the 5 x 3 spectro-temporal representation in Chapter 4 was also applied here. This was done to see if the number of derived bands, or the location of these bands would be different in ADS as compared to CDS. As described in Chapter 4, this procedure involved a separate Spectral PCA analysis (based on 28 spectral channels), and a separate Modulation Rate PCA analysis (based on 24 modulation rate channels).

7.1.4.1 Spectral PCA

Prior to performing the PCA analysis on the 28 cochlear-spaced frequency channels, the RMS power for the spectral channels was computed. This initial step was performed to see if there were any obvious differences in patterns of spectral power across the 4 speaking conditions, before the more complex PCA analysis was applied. A Spectral PCA procedure was then applied to the samples, as described in Chapter 4, Section 4.3. For this PCA analysis, the speech signal was divided into 29 ERB_N-spaced frequency channels, and the Hilbert envelope was taken from each spectral channel. As was done in Chapter 4, the first low-pass channel was discarded, and the remaining 28 spectral envelopes were then entered into a PCA analysis as separate variables. As before, the absolute value of the spectral channel loadings was taken, and these rectified values were averaged across all the samples, for each speaking condition. The mean rectified PCA loading patterns for each speaking condition were then compared.

7.1.4.2 Modulation Rate PCA

For this analysis, the speech signal was first divided into 5 spectral bands (as indicated by the results of the previous Spectral PCA). The Hilbert envelope was obtained for each spectral band and then low-pass filtered under 40 Hz. These 5 Hilbert envelopes from each spectral band were then individually passed through a 24-channel logarithmically-spaced modulation filterbank spanning 0.9-40 Hz. Prior to conducting the PCA, the RMS power of the 24 modulation channels was computed to look for obvious differences in the modulation spectrum. Then, the modulation rate PCA procedure was applied to the 24 modulation-filter outputs from each spectral band, as described in Chapter 4, Section 4.4. This PCA analysis was repeated separately for each of the 5 spectral bands. The absolute value of the modulation rate channel loadings was again taken, and these rectified values were averaged across all the samples for each speaking condition. The mean rectified PCA loading patterns for each speaking condition were then compared.

7.1.4.3 AM Hierarchy Analysis

Finally, 5x3 AM hierarchies were extracted from the speech samples (as indicated by the results of the Spectral PCA and the Modulation Rate PCA). That is, each speech sample was filtered into 5 spectral bands. The Hilbert envelope was extracted from each spectral band, and this envelope was then filtered into 3 modulation rate bands ('Stress', 'Syllable' and 'Phoneme'). The resulting sets of modulation rate bands were then analysed for their rhythmic structure and hierarchical organisation using two indices.

First, the rhythmic regularity at each modulation rate was compared by computing the autocorrelation function (ACF) for the Stress, Syllable and Phoneme AMs. The autocorrelation function computes the correlation of the signal with itself at different time lags. It is a measure of periodicity within the signal, and can be used to detect patterns that repeat over time. To assess the amount of periodic power contained in the ACFs, a Fourier transform was applied to the ACFs, resulting in a periodic power spectrum for each of the Stress, Syllable and Phoneme ACFs.

Second, the modulation hierarchy organisation of the speech samples was analysed by computing the peak-phase distribution between the three modulation tiers. That is, two distributions were computed, the distribution of (1) Syllable peaks with respect to Stress

phase; and (2) Phoneme peaks with respect to Syllable phase. In both cases, binned phase values were used to compute the distribution (17 equally-spaced bins between $-\pi$ and π radians).

To quantify any differences in hierarchical AM organisation, a novel 'conditional entropy' (CE) measure was used. This measure computed the amount of uncertainty (entropy) about events in one modulation tier (e.g. occurrence of Syllable vowel nuclei), as a result of knowing the phase value of the adjacent slower tier (e.g. Stress phase). For example, if the entropy of Syllable vowel occurrence was low as a result of knowing Stress phase (i.e. CE was small), this indicated that there was tight hierarchical phase-nesting between the two modulation tiers, and Stress phase was strongly 'constraining' the occurrence of Syllable vowels. On the other hand, if the entropy of Syllable vowel occurrence was high even after knowing Stress phase, (i.e. CE was large) then Stress phase was unrelated to the occurrence of Syllable vowel nuclei, indicating weak hierarchical phase-nesting between the two modulation tiers.

Appendix 7.2 provides a more detailed explanation of entropy and conditional entropy, and formulae used for computing these values. The appendix also includes worked examples to explain how the peak-phase distribution should be interpreted. Information and entropy measures (including mutual information measures) are increasingly being used in neuroscience to measure properties of the neural signal, in particular the contribution of neuronal oscillatory *phase* to neural coding of sensory stimuli like speech (e.g. Kayser et al, 2009; Cogan & Poeppel, 2011). Here, similar methods are applied to AM patterns within the AM hierarchy, to investigate hierarchical phase-nesting between tiers of the AM hierarchy.

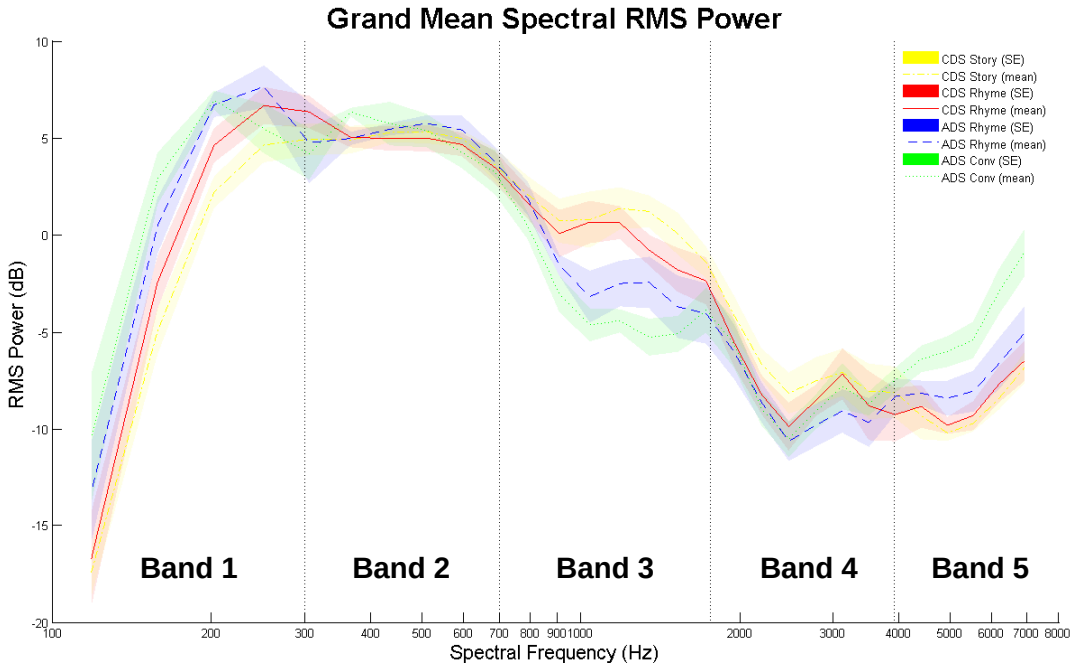
7.2 RESULTS

7.2.1 SPECTRAL PCA ANALYSIS

7.2.1.1 RMS Power Across Cochlear Channels

Figure 7.2 shows the grand mean RMS power[26] over the 28 cochlear channels for each speaking condition. The 4 speech corpora are each plotted in a different colour. From visual inspection, there appear to be clear differences in RMS power between CDS and ADS samples within spectral bands 1, 3 and 5. In the middle spectral band 3 (700 Hz-1750 Hz), the two CDS conditions (yellow and red lines) clearly showed higher RMS power than the two ADS conditions (blue and green lines). The opposite was true for the two extreme spectral bands 1 (100 Hz -300 Hz) & 5 (3900 Hz to 7250 Hz), where the two ADS conditions now showed higher RMS power than the two CDS conditions. Moreover, the order of effects across speaking conditions was maintained even though the direction of differences was reversed.

Figure 7.2. RMS power of the 28 cochlear channels. Vertical dotted lines indicate the boundaries between the 5 S-AMPH spectral bands. The 4 speaking conditions are shown as different coloured lines. Shaded areas indicate the standard error of the mean.



[26] Since the overall RMS power differed across samples and speakers (i.e. some speakers were speaking more loudly than others), for each sample, the average power across all spectral channels was subtracted from each channel, leaving only the difference in power from the average power for each spectral channel. This difference power was then averaged over samples and speakers to give the results shown in Figure 7.2.

In spectral band 3, CDS Story showed the highest RMS power, followed by CDS Rhyme, ADS Rhyme and ADS Conversation. In spectral bands 1 & 5, this order was perfectly reversed, with ADS Conversation now showing the highest RMS power, followed by ADS Rhyme, CDS Rhyme and CDS Story. This orderly pattern of RMS power differences suggests that child- and adult-directed speech differ systematically in their relative spectral composition. For child-directed speech, the power of the middle spectral frequencies was increased relative to the power of very low and very high spectral frequencies. For adult-directed speech, the middle spectral frequencies received less emphasis, relative to very low and very high spectral frequencies.

It is worth noting that the differences between CDS and ADS samples occurred in spectral regions that corresponded fairly well to the S-AMPH spectral band divisions. For example, the increase in power for CDS samples at the middle spectral frequencies occurred between ~800-1750 Hz, which correspond closely with the spectral band 3 region of 700-1750 Hz. This suggests that the 5 (PCA-derived) spectral bands do indeed reflect separate spectral components in speech, since they can be modulated independently (in power) by the speaker. However, recall from Appendix 4.3 that RMS power does not necessarily reflect correlation strength. Therefore, to investigate the underlying spectral correlation structure of the 4 types of speech more closely, a Spectral PCA analysis was conducted.

7.2.1.2 Spectral PCA Analysis

The aim of this analysis was to investigate whether CDS and ADS samples had the same underlying spectral structure (i.e. 5 non-redundant spectral bands). To investigate this, the mean rectified PCA component loading patterns across 28 spectral channels were compared. Recall that spectral channels which show a similar loading for a given PCA component are assumed to carry similar (redundant) information.

a. Individual Principal Component Loadings

The loading patterns for the top 3 PCA components in each speaking condition are shown in Figure 7.3. Only the top 3 components were analysed because cumulatively, these 3 components already accounted for over 50% of the total variance in the samples. The variance accounted for by each principal component is indicated in the titles of Figure 7.3. In Figure 7.3, the x-axis plots the centre frequency for each spectral channel and the y-axis plots

the absolute (rectified) component loading, averaged across all samples and speakers. Each of the 4 speaking conditions is shown in a different colour.

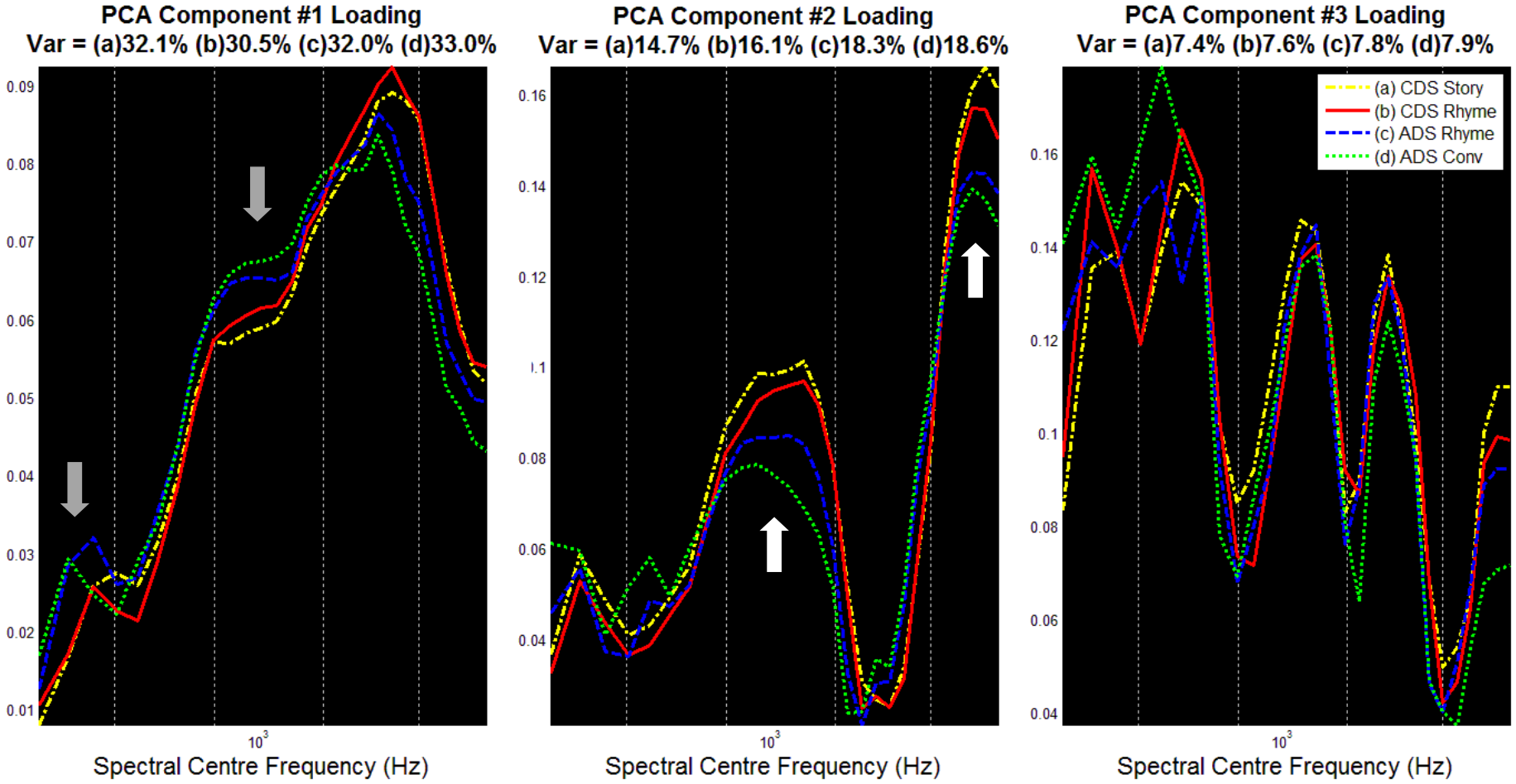
From visual inspection, the overall loading patterns were very similar across the four spoken conditions. The locations of major peaks and troughs (indicating spectral band boundaries) were also very similar across the four conditions, coinciding well with the band boundaries previously identified in Chapter 4. These previously-identified band boundaries are shown as vertical dotted lines. This suggests that the fundamental band structure of the speech samples was similar *across* the four speaking conditions, and agreed well with the spectral band boundaries that had previously been identified.

However, there were also clear and systematic differences between the loading patterns of the four conditions. That is, although the spectral *location* of the peaks and troughs was similar across conditions, the relative *height* of the peaks differed systematically across conditions. For example, in PCA component 2 (middle plot of Figure 7.3), loadings within spectral bands 3 and 5 (indicated by the white arrows) were clearly different across speaking conditions. In both these spectral regions, component loadings were, from highest to lowest : CDS Story > CDS Rhyme > ADS Rhyme > ADS Conversation. This same pattern of component loading was also observed in PCA component 3 (right plot) in the region of spectral band 5 (white arrow). Therefore, CDS samples loaded more strongly than ADS samples in mid-high frequency regions.

The opposite pattern was observed for PCA component 1 in the regions of spectral band 3 and spectral band 1 (grey arrows). Here, the order of component loadings, from highest to lowest were : ADS Conversation/ADS Rhyme > CDS Story/CDS Rhyme. Therefore, ADS samples loaded more strongly than CDS samples in low-mid frequency regions.

Note that in spectral band 3 (the middle frequency band), there was an opposite loading order across conditions for PCA components 1 & 2. In component 1, ADS samples loaded more strongly, but in component 2, CDS samples loaded more strongly. To estimate how these loading order differences would trade-off on average, the loadings for all the conditions were averaged across PCA components 1 to 3. The results of this averaging are shown in Figure 7.4.

Figure 7.3. Results of the Spectral PCA Analysis. Rectified mean loading patterns for PCA components 1, 2 and 3 are shown from left to right. The four speaking conditions are shown in different coloured lines. The amount of variance accounted for by each PCA component, for each speaking condition is shown in the title of each plot. This amount is the average across 6 speakers. Vertical dotted grey lines indicate the boundaries between the five spectral bands at 300 Hz, 700 Hz, 1750 Hz and 3900 Hz. White arrows indicate spectral regions where CDS samples load more strongly than ADS samples. Grey arrows indicate spectral regions where ADS samples load more strongly than CDS samples.



b. Averaged Principal Component Loadings (Components 1-3)

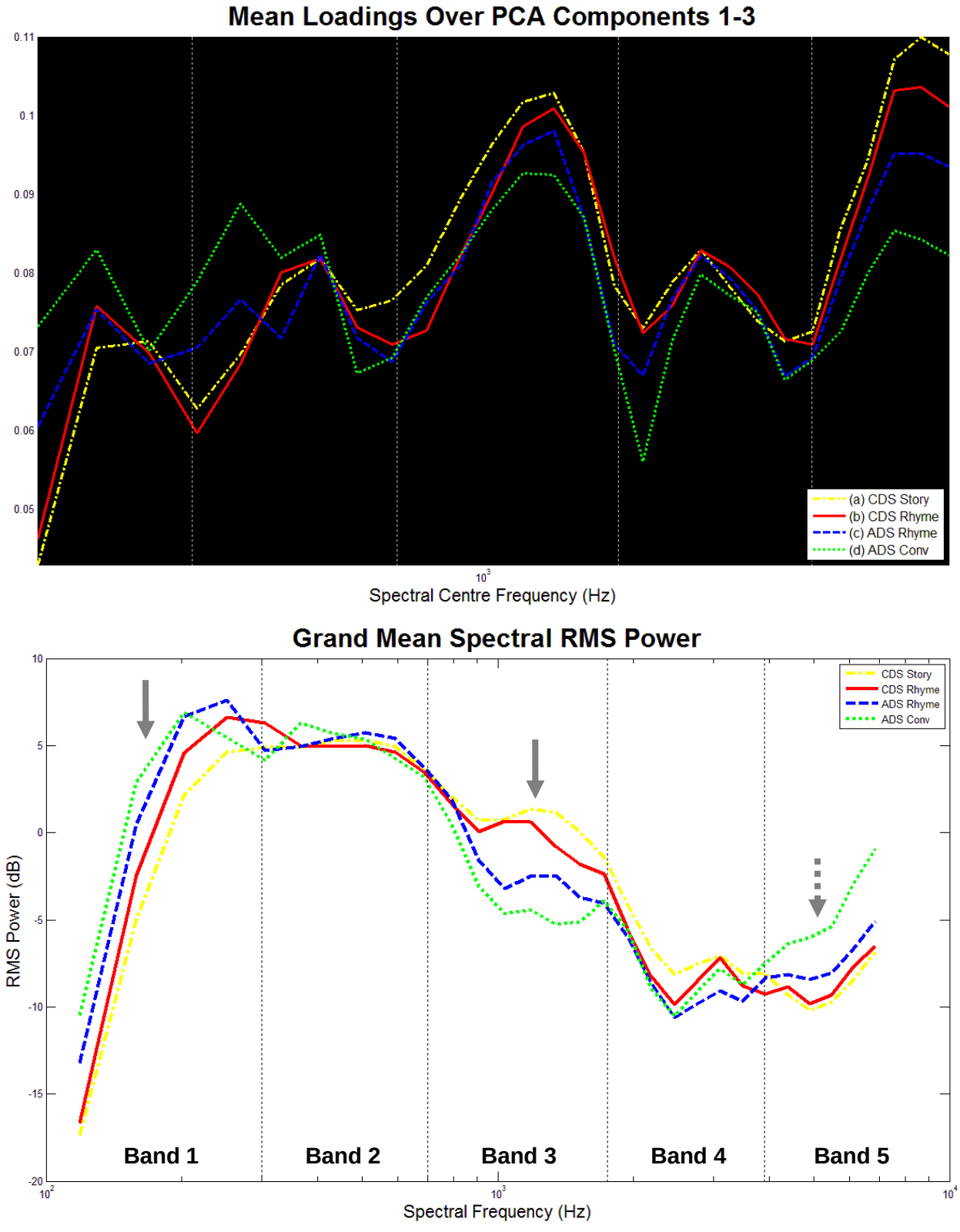
In the averaged loading patterns shown in Figure 7.4, the trends differentiating CDS and ADS conditions are now clear and consistent. CDS conditions (yellow and red lines) loaded *more strongly* than ADS conditions (blue and green lines) at *higher* spectral frequencies (e.g. spectral bands 3 & 5), but loaded *less strongly* at *lower* spectral frequencies (e.g. spectral bands 1 & 2). It is also interesting to note that the overall order of the 4 speaking conditions was systematically preserved at both ends of the frequency spectrum, suggesting a parametric change between conditions. At high spectral frequencies, CDS Story loadings were the strongest, followed by CDS Rhyme, ADS Rhyme and ADS Conversation. At low spectral frequencies, this order was reversed, with ADS Conversation showing the strongest loadings, followed by ADS Rhyme, CDS Rhyme and CDS Story[27].

However, the interpretation of these loading patterns requires careful consideration. Recall that in the previous section, the RMS power over the 28 spectral channels had been computed. This power spectrum is shown again in the bottom half of Figure 7.4, where it can be directly compared with the PCA loading patterns. It may be observed that the higher component loading for CDS samples in spectral band 3 *is* accompanied by a relative boost in RMS power in this spectral band (grey arrow). Similarly, the drop in component loading for CDS samples in spectral band 1 is *also* accompanied by a drop in RMS power in this spectral band (grey arrow). However, the higher component loading for CDS samples in spectral band 5 is *not* accompanied by an increase in RMS power. Rather, in spectral band 5, the RMS power for CDS samples *decreases* relative to ADS samples (grey dotted arrow).

Therefore, in CDS samples (Story & Rhyme), there is a relative 'boost' at middle spectral frequencies (~1200 Hs, spectral band 3) as compared to low spectral frequencies (<300 Hz, spectral band 1). This boost occurs for both RMS power (i.e. middle frequency sounds like vowels are louder), as well as for component loading strength (middle frequency modulation patterns are more similar). However, at very high spectral frequencies (>3900 Hz, spectral band 5), component loadings again increased in CDS, but this is now accompanied by a *drop* in RMS power (i.e. high frequency sounds like fricatives are *softer*).

[27] This order of CDS and ADS conditions was also observed in RMS power differences for the 29 spectral channels in Section 7.2.1.1

Figure 7.4. (top) Mean rectified loadings averaged over PCA Components 1 to 3. The 4 speaking conditions are shown in different coloured lines. Vertical lines indicate the boundaries of the five spectral bands. (bottom) RMS power of the 28 spectral channels, replicated from Figure 7.2. Solid arrows indicate spectral regions where strength of component loading across conditions is positively correlated with RMS power. Dotted arrow indicates a spectral region where strength of component loading is negatively correlated with RMS power.



c. Interim Summary & Discussion of Spectral PCA Results

Both child- and adult-directed speech shared the same fundamental spectral band structure of 5 major spectral bands. The frequencies contained in the 5 spectral bands are listed in Table 7.3, as a reminder.

Table 7.3. 5 spectral bands, as identified in Chapter 4

Spectral Band	**Frequency Range (Hz)**
Band 1	100-300
Band 2	300-700
Band 3	700-1750
Band 4	1750-3900
Band 5	3900-7250

However, CDS speech samples showed a specific *boost* in spectral RMS power and PCA component loading strength at middle spectral frequencies around 1200 Hz, suggesting that vowel sounds may be particularly emphasised in CDS. This result is consistent with the finding that vowel sounds in child-directed speech are hyperarticulated, or more separated in formant space (Ratner, 1984; Burnham et al, 2002). Here, these results suggest that the vowel sounds in child-directed speech are also relatively louder, and more strongly co-modulated (i.e. contain similar patterns of modulation across the channels in each band).

In contrast, CDS showed a relative *reduction* of RMS power for very high frequency (>3900 Hz) and very low frequency sounds (<300 Hz), even though high frequency sounds also showed increased co-modulation. This suggested that while high frequency sounds (like fricatives) were softer in CDS, they still contained strong and consistent modulation patterns.

7.2.2 MODULATION RATE PCA ANALYSIS

The aim of this next part of the analysis was to investigate whether CDS and ADS samples differed in their modulation spectrum, or their modulation rate structure (i.e. the number of non-redundant modulation rate bands).

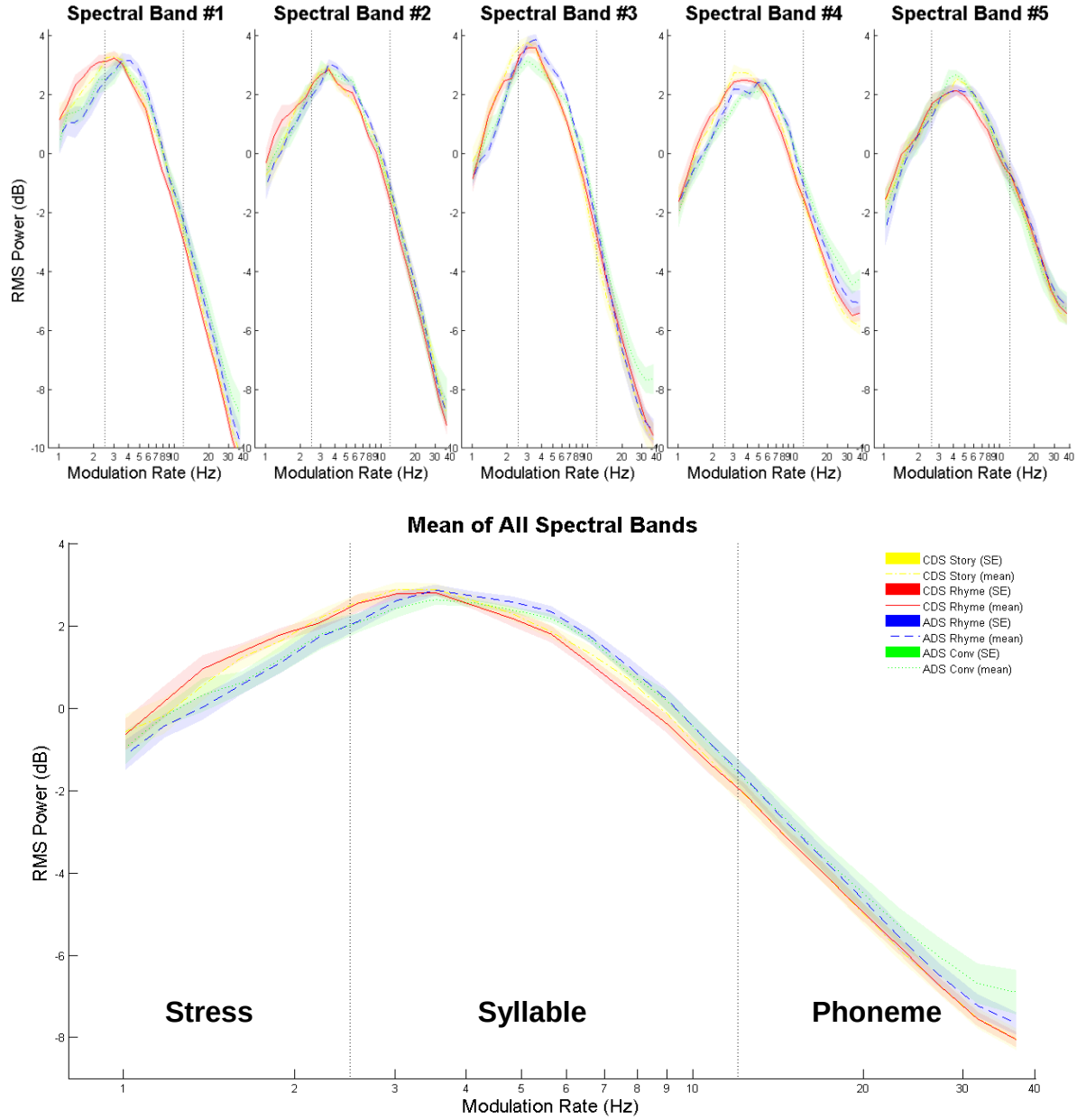
7.2.2.1 Modulation Spectrum (RMS Power Across Modulation Channels)

Since the speech samples across the 4 conditions had the same 5-Band spectral structure, all the samples were filtered into 5 spectral bands, and Hilbert envelopes were extracted from each spectral band. These envelopes for each spectral band were then passed through a 24-channel modulation filterbank. Figure 7.5 shows the mean RMS power[28] at each modulation channel (i.e. the modulation spectrum), for each of the 5 spectral bands (top plots), as well as the grand mean over the 5 spectral bands (bottom plot). Vertical lines indicate the boundaries between 'Stress', 'Syllable' and 'Phoneme' modulation bands.

From visual inspection of the grand mean plot (bottom), CDS samples (red and yellow lines) appeared to have higher power in the Stress modulation band, and slightly lower power in Syllable and Phoneme modulation bands. This suggests that in child-directed speech, speakers tended to place relatively greater emphasis on the slower stress patterns, than on the faster syllable and phoneme-rate patterns. To examine whether there were also differences in the modulation structure of CDS and ADS samples, a PCA analysis was carried out using the 24-modulation channels.

[28] Since the overall RMS power differed across samples and speakers (i.e. some speakers were speaking more loudly than others), for each sample, the average power across all 24 modulation channels was subtracted from each channel, leaving only the difference in power from the average power for each modulation channel. This difference power was then averaged over samples and speakers to give the results shown in Figure 7.5.

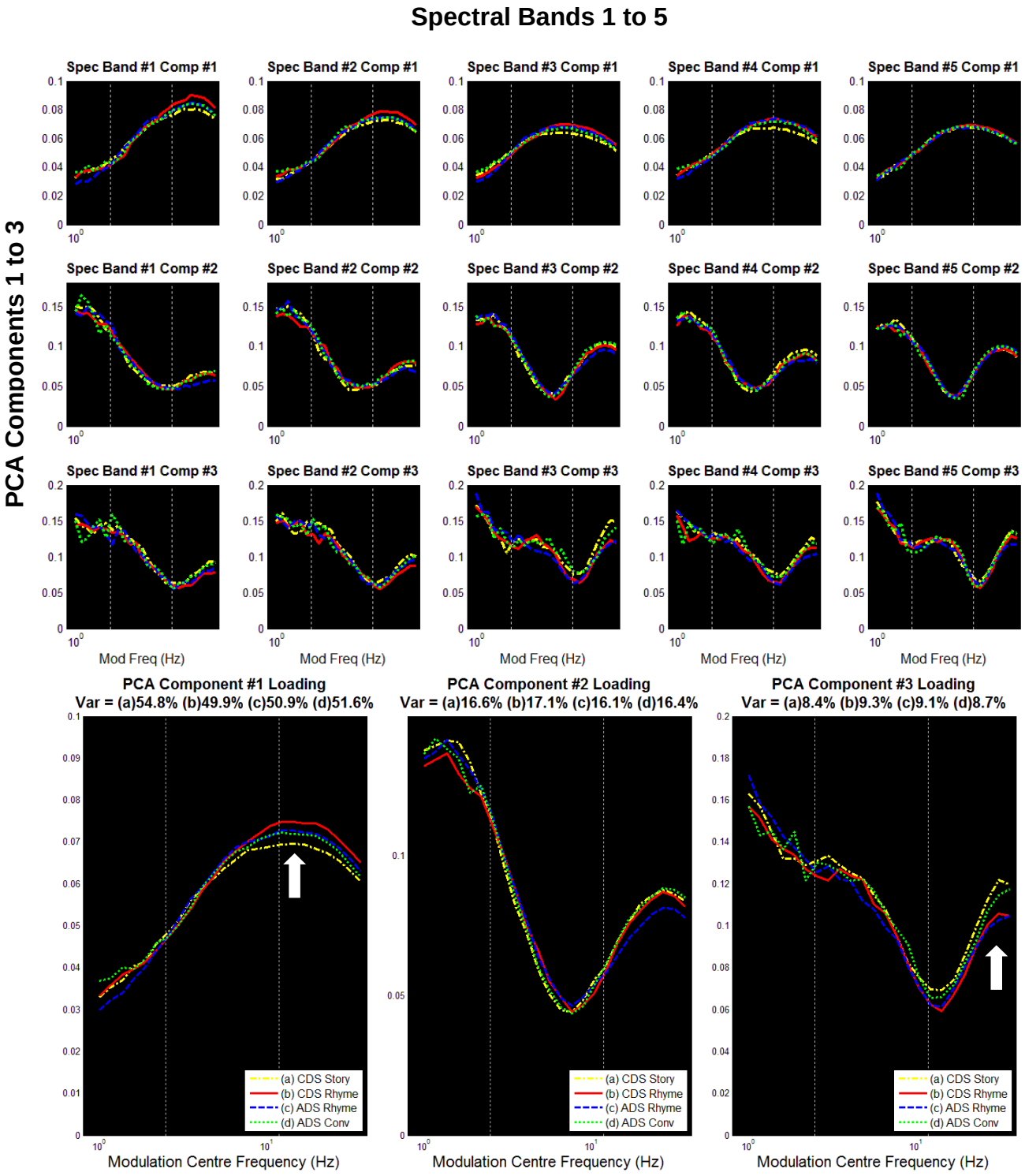
Figure 7.5. Modulation spectrum for each spectral band (top) and averaged over all spectral bands (bottom). Coloured lines indicate the 4 speaking conditions. Shaded areas indicate the standard error of the mean for each condition. Vertical dotted lines indicate the boundaries of the 3 modulation rate bands.



7.2.2.2 Modulation Rate PCA Analysis

The modulation rate PCA analysis was carried out separately for each of the 5 spectral bands. Figure 7.6 shows the mean PCA loading patterns for the top 3 components, for each spectral band (top half of figure), as well as averaged over the 5 spectral bands (bottom half of figure). The top 3 principal components cumulatively accounted for almost 80% of the total variance in the samples. The variance accounted for by each principal component is indicated in the titles of the bottom plot of Figure 7.6.

Figure 7.6. Results of the Modulation Rate PCA Analysis. (top) Mean rectified loading patterns for each spectral band are shown in columns, PCA components 1 to 3 are shown in rows (bottom) Grand mean component loading patterns, averaged across spectral bands 1 to 5. The vertical lines in the plots indicate the boundaries between the 3 modulation rate bands at 2.5 Hz and 12 Hz. White arrows indicate inconsistencies in the order of loading strength across the 4 conditions, observed within the same modulation rate band.



In Figure 7.6, the x-axis plots the centre frequency for each modulation channel and the y-axis plots the absolute (rectified) component loading, averaged across all samples and speakers. Each of the 4 speaking conditions is shown in a different colour.

From visual inspection, the loading patterns were again highly similar across all 4 speaking conditions. The locations of major peaks and troughs (indicating modulation band boundaries) were also very similar across the four conditions, coinciding well with the band boundaries previously identified in Chapter 4 (see Figure 4.5). These previously-identified modulation band boundaries are shown as vertical dotted lines. This suggests that the basic modulation band structure of the speech samples was again similar across the four speaking conditions, and agreed well with the band boundaries that had previously been identified.

Although there were small differences in the loading pattern between conditions at certain modulation frequencies, these differences did not appear to be systematic. For example, in the grand mean plot for PCA component 1 (bottom left subplot), CDS Rhyme loaded the most strongly in the Phoneme Band, and CDS Story loaded the most weakly (indicated with the white arrow). However, for the same Phoneme Band in PCA component 3 (bottom right subplot), this pattern was reversed, with CDS Story now loading the most strongly, while CDS Rhyme loaded weakly (white arrow).

Therefore, since there were no large, consistent differences between the four speaking conditions in terms of their principal component loading pattern, the same band structure of 3 modulation rate bands ('Stress', 'Syllable' and 'Phoneme') was applied to all 4 speaking conditions. The modulation rates contained in the 3 modulation rate bands are listed in Table 7.4, as a reminder.

Table 7.4. 3 modulation rate bands, as identified in Chapter 4

Modulation Rate Band	**Modulation Range (Hz)**
Stress	0.9 - 2.5 Hz
Syllable	2.5 - 12 Hz
Phoneme	12 - 40 Hz

7.2.3 Interim Summary of Spectral & Modulation Rate PCA Results

To summarise the PCA results, all 4 types of speech were well-described by the same spectro-temporal structure of 5 spectral bands and 3 modulation rate bands (forming a 3-tier AM hierarchy).

However, there were also differences between the CDS and ADS samples that occurred *within* these spectral or modulation rate bands. In the spectral domain, CDS samples showed a boost at middle frequencies around 1200 Hz (spectral band 3), consistent with a greater emphasis on vowel sounds. In the modulation rate domain, the modulation spectrum of CDS showed slightly higher RMS power than ADS at Stress rates, but lower power at Syllable and Phoneme rates.

In the next section, the 3 tiers of the AM hierarchy (Stress, Syllable & Phoneme tiers) are analysed in terms of their (1) rhythmic regularity (periodic power); and (2) hierarchical organisation (peak-phase distribution).

7.2.4 AM HIERARCHY ANALYSIS

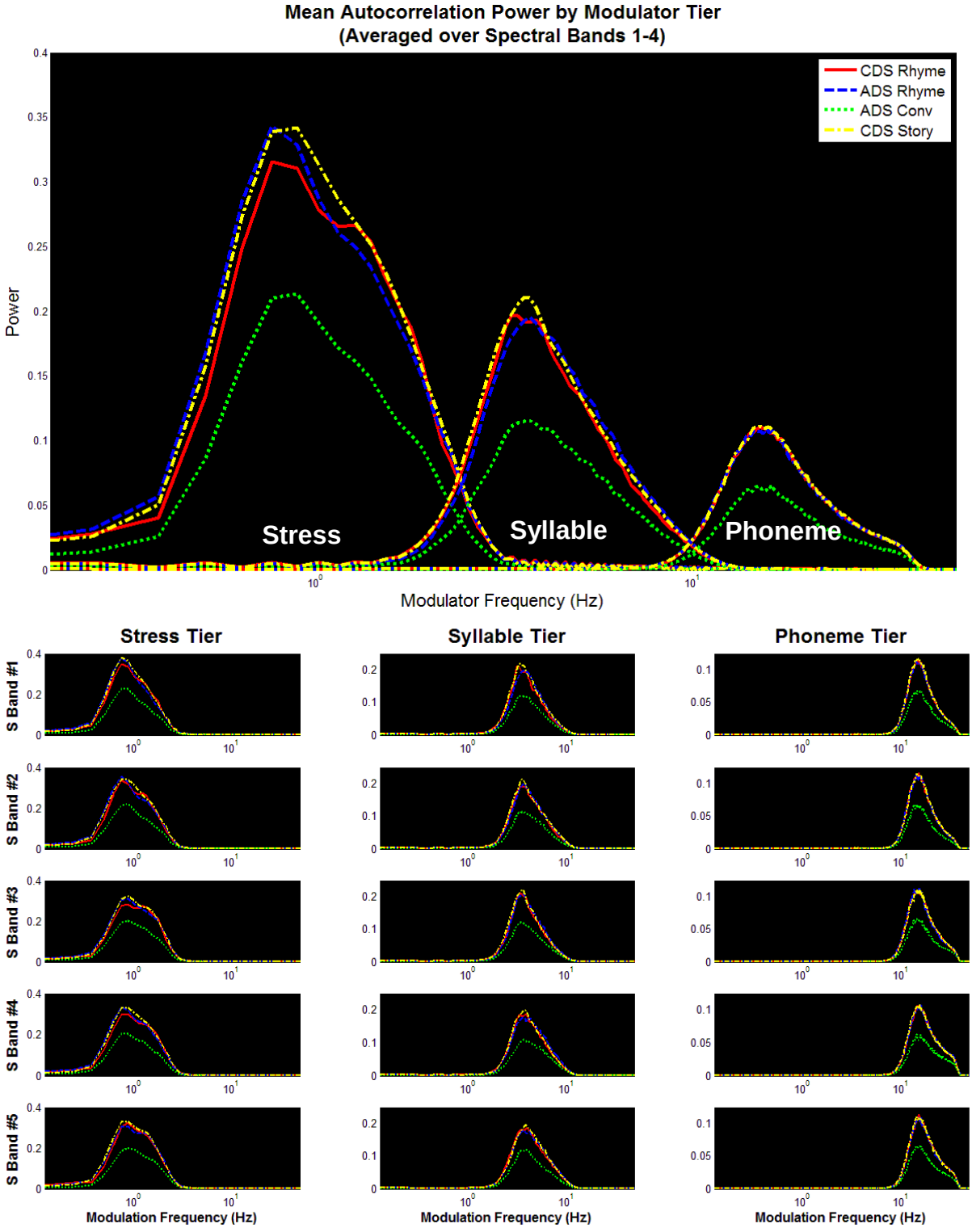
7.2.4.1 Rhythmic Regularity (Periodic Power)

The rhythmic regularity of the three tiers in the AM hierarchy was analysed by computing the autocorrelation function (ACF) of each modulation tier (Stress, Syllable and Phoneme), and then computing the periodic power spectrum of the resulting ACFs. It was expected that when reading nursery rhymes, speaker's utterances should be more rhythmically-regular than when reading narrative stories or producing spontaneous conversation. However, it was not known whether the CDS Story readings would be more or less rhythmically-regular than the ADS Conversation samples.

Figure 7.7 shows the periodic power spectrum obtained for each of the 3 modulation tiers, for each speaking condition. The bottom panel of Figure 7.7 shows the power spectra for each spectral band, and the top panel shows the average power spectrum taken across spectral bands 1-4 (spectral band 5 was excluded from the average since it would contain high frequency sounds like fricatives, which, from Chapter 4, detracted from the overall rhythm of the utterance).

As predicted, both 'Rhyme' speech samples (CDS Rhyme in red and ADS Rhyme in blue) contained much higher periodic power than the non-poetic ADS Conversation samples (green). This higher periodicity was observed for all 3 modulation tiers. However, surprisingly, child-directed Story samples (in yellow) showed *as much* periodicity as the nursery rhyme readings for all 3 modulation tiers. This suggested that speakers were 'unconsciously' patterning their utterances rhythmically when addressing children, even when the material they were reading was non-poetic. It is remarkable that this CDS-related rhythmic patterning of stories was so strong that it was similar to that of the metrically-regular nursery rhymes.

Figure 7.7. Periodic power for the autocorrelation function of each modulator tier. (Top) Mean power spectrum averaged over spectral bands 1 to 4. (Bottom) Power spectra for each Spectral band (rows), for each modulation tier (columns).



7.2.4.2 Hierarchical Organisation (Peak-Phase Distribution Pattern)

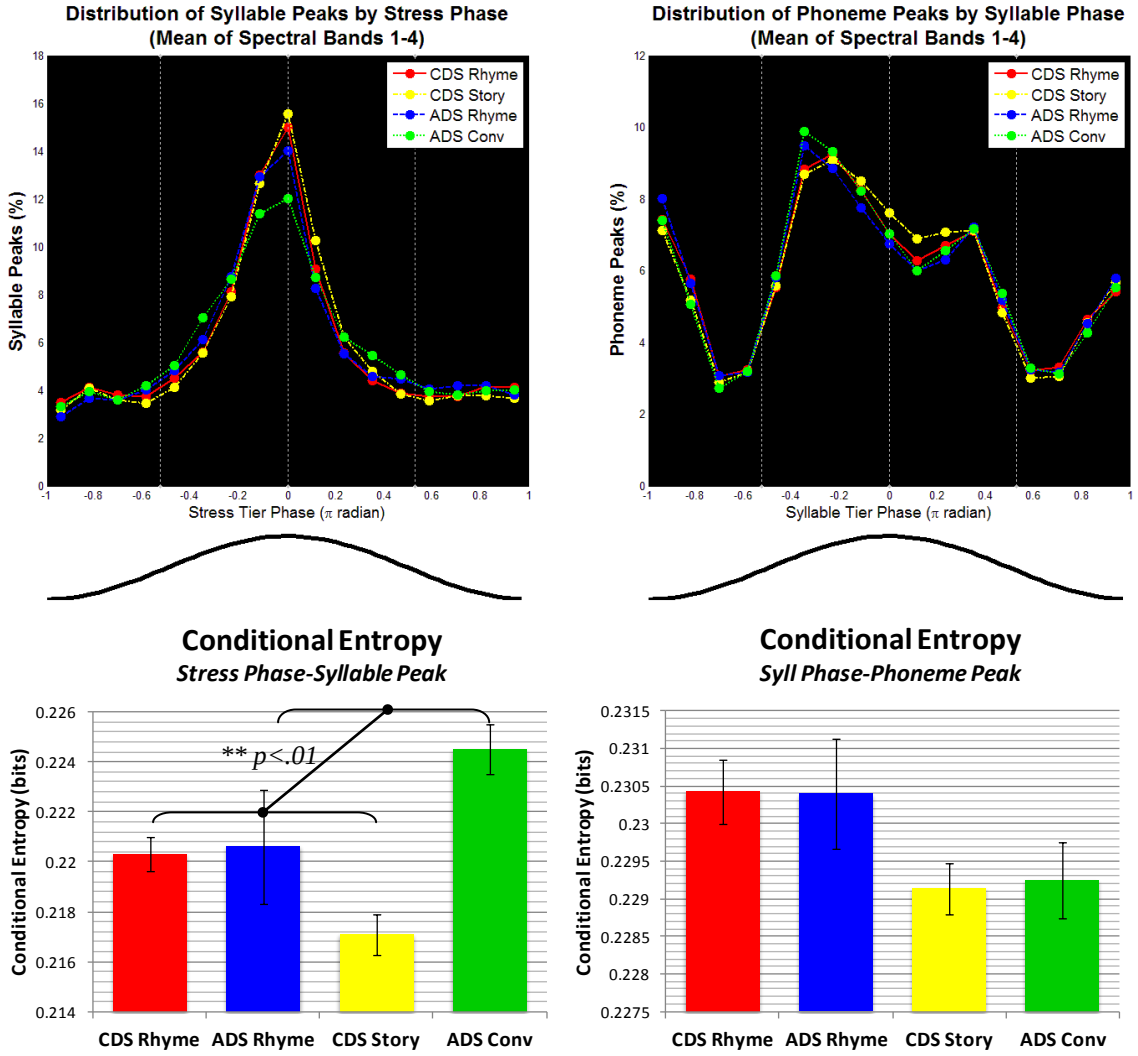
Finally, the hierarchical organisation of CDS and ADS samples was analysed by computing their Syllable peak-Stress phase distributions and Phoneme peak-Syllable Phase distributions. The average peak-phase distribution for each speaking condition is shown in the top panel of Figure 7.8. To compute this average distribution pattern, the distributions for spectral bands 1-4 were averaged together (spectral band 5 was discarded, as was done for the autocorrelation analysis). Plots on the left relate to the Syllable peak-Stress phase relationship, and plots on the right relate to the Phoneme peak-Syllable phase relationship.

Visual inspection of the Syllable peak-Stress phase distribution pattern (top left plot) indicates that the CDS Story distribution had the highest kurtosis ('peakedness'), with the largest proportion of Syllable peaks concentrated around the peak of Stress phase (0π radians). The distribution with the next highest kurtosis was CDS Rhyme, followed by ADS Rhyme and ADS Conversation. As explained in Appendix 7.2, the kurtosis of the distribution pattern is directly related to its entropy, where flat (rectangular) distributions are associated with the highest entropy and distributions with high kurtosis have low entropy.

Next, the conditional entropies (CE) for each average distribution were computed, for each speaking condition. The middle and bottom panels of Figure 7.8 show the computed conditional entropies for each speaking condition as a bar graph (middle panel), and in a table (bottom panel) respectively. As expected from the shape of their distributions, CDS Story had the lowest conditional entropy followed by CDS Rhyme and ADS Rhyme. ADS Conversation had the highest entropy values. Recall that lower CE values indicate stronger hierarchical phase-nesting (i.e. phase in the slower tier exerts a stronger constraint on the occurrence of peaks in the faster tier).

To analyse these differences statistically, a repeated measures ANOVA was conducted on the Stress phase-Syllable peak conditional entropy scores with Style (CDS or ADS) and Material (Rhyme or Narrative) as factors. Results indicated a strong significant main effect of Style ($F(1,5) = 18.90$, $p<.01$), with child-directed speech samples significantly lower in conditional entropy than adult-directed speech samples. There was no significant effect of Material ($F(1,5) = 0.07$, $p=.81$), but there was a significant interaction between Style and Material ($F(1,5) = 6.96$, $p<.05$).

Figure 7.8. (Top) Hierarchical distribution of peaks for each modulator tier with respect to the phase of the upper tier. The left plot shows the distribution of Syllable peaks with respect to Stress phase. The right plot shows the distribution of Phoneme peaks with respect to Syllable phase. The distributions shown are the mean distributions across of spectral bands 1-4. (Middle) Corresponding conditional entropy scores for the distribution pattern of each speech corpus. Distributions with higher kurtosis have a lower entropy while distributions with lower kurtosis have a higher entropy. Errorbars show the standard error across 6 speakers. (Bottom table) Mean CE scores for each speech corpus.



CE (bits)	Stress Phase-Syll Peaks	Syllable Phase-Phon Peaks
CDS Rhyme	0.2203	0.2304
ADS Rhyme	0.2206	0.2304
CDS Story	0.2171	0.2291
ADS Conv	0.2245	0.2292

A Tukey HSD post-hoc analysis of the Style x Material interaction indicated that Narrative materials showed a stronger CDS-ADS difference than Rhyme materials. This was not surprising since, as previously noted, participants may have found it difficult to produce nursery rhymes in an adult-directed manner rather than a child-directed manner. Although the difference between ADS Rhymes and CDS Rhymes was very small (CDS : 0.2203 vs ADS : 0.2206), this difference was in the predicted direction, with CDS Rhymes showing lower conditional entropy than ADS Rhymes.

The same repeated measures ANOVA was then conducted with conditional entropy scores for the Phoneme peaks-Syllable phase distribution (right column in Figure 7.8). This time however, none of the main effects (Style or Material) were significant, and the interaction between Style and Material was also non-significant. Moreover, in paired-samples t-tests, none of the pairs of conditions (speech corpora) showed significant differences. Therefore hierarchical phase-nesting in CDS was significantly greater than ADS for the Stress-Syllable relationship, but not for the Syllable-Phoneme relationship.

In the Syllable peak-Stress phase analysis, the largest difference in entropy was between CDS Story and ADS Conversation corpora. Since the CDS story material was scripted (participants read out a printed text), while the ADS conversation material was unscripted and spontaneous, it is possible that 'scripting' was a confound in the experimental design. For example, it might be argued that the entropy differences observed here could reflect the degree of syntactic organisation and well-formedness in scripted versus spontaneous utterances, rather than child- or adult-directedness per se. However, if the degree of structure and organisation in the utterance was the main determinant of entropy, one would expect nursery rhymes to show the most structure and therefore the lowest entropy. Instead, child-directed stories had a significantly lower entropy than the tightly structured CDS nursery rhymes (paired t-test, $t(5) = 8.49$, $p<.001$), indicating that entropy scores were being driven by a factor other than regular syntactic structure.

Moreover, all 6 speakers produced ADS spontaneous utterances that were well-formed and grammatically correct. Although their conversation material was not scripted, the topics were familiar to participants (e.g. their hobbies) and they had had time to plan their utterances before the recording session began. Therefore, even though the ADS conversation samples were unscripted, they were still well-structured utterances. Finally, even when the speech material was matched exactly, and fully scripted (as with ADS and CDS Rhymes),

entropy differences still occurred in the predicted direction, although these differences were small and not statistically significant. The influence of scripting can be investigated further in a new experiment, however that is beyond the scope of this thesis.

7.3 RESULTS SUMMARY & DISCUSSION

There were strong similarities as well as clear differences between child- and adult-directed speech samples. In terms of similarities, the PCA analyses indicated that both CDS and ADS samples were well-represented by a dimensionally-reduced 5x3 spectro-temporal structure. This suggests that the 5x3 spectro-temporal structure identified in this thesis may be a fairly robust and ubiquitous way to represent the spectro-temporal variation inherent in the speech signal. Even when speakers are consciously modifying the way they speak, these changes occur *within* the bounds of the 5x3 spectro-temporal bands, rather than changing the boundaries between the bands entirely.

For example, in the spectral domain, CDS samples showed a 'boost' in RMS power and PCA component loading that was well-located to spectral band 3 (middle spectral frequencies ~1200 Hz). Since vowels sounds commonly contain energy in this spectral region, this middle-frequency boost is interpreted as indicating a greater emphasis on vowel sounds in child-directed speech. This vowel emphasis was associated with both increased loudness (RMS power), as well as stronger co-modulation. Such exaggeration of vowel sounds in CDS is thought to be a didactic device for teaching language to children, since speech directed to *pets* does not show evidence of vowel hyperarticulation (Burnham et al, 2002).

In the modulation rate domain, CDS Story and CDS Rhyme samples were both strongly periodic at all 3 Stress, Syllable and Phoneme modulation rates. While ADS Rhymes also showed strong periodicity, ADS Conversation samples did not. Therefore, child-directed speech possessed strong periodic regularity even when the spoken material itself was non-poetic. In contrast, ADS speech for non-poetic material (ADS Conversation) had a much lower periodic regularity. Finally, in the hierarchical peak-phase distribution analysis of the 3 modulation rate bands, CDS showed stronger hierarchical phase-nesting than ADS (as

measured using conditional entropy). However, this only occurred for the Syllable peak-Stress phase relationship.

Child-directed utterances could be more rhythmically regular in order to facilitate speech segmentation by the young listeners. If words are regularly-stressed, and syllables are produced at regular intervals, the word boundaries would also be more predictable in time, allowing children to more accurately locate these important boundaries. Similarly, the lower Stress-Syllable entropy (and stronger phase-nesting) in child-directed speech could also indicate that prosodic boundaries are more frequently marked by stress in CDS. According to the S-AMPH model, syllable peaks occurring near the peak of the Stress cycle are 'Strong' stressed syllables, while Syllable peaks that occur near the trough are 'weak' syllables. By this view, the highly-kurtotic distribution associated with the lower entropy of CDS Story samples implies that more syllables (proportionately) are stressed in stories than in nursery rhymes.

In nursery rhymes, the strict metrical structure imposes constraints on the placement of syllable stress. For example, in duple beat rhymes, stress typically occurs every 2, 4 or 8 syllables, but cannot occur on consecutive syllables. In continuous prose however, speakers do not have such constraints. Therefore, they would be 'allowed' to place stress on consecutive syllables for the purpose of emphasis. In the CDS Story samples, they may be doing this relatively often (e.g. "LI-ttle PIG, LI-ttle PIG, LET ME COME IN!"). According to this explanation, CDS Story samples would have the highest proportion of stressed syllables, followed by the nursery rhymes. ADS conversation samples would have the lowest proportion of stressed syllables. Therefore, if young children were to use a metrical segmentation strategy in which strong syllables were perceived as potential word boundaries (Cutler & Norris, 1988), they would find many more word boundaries in CDS than in ADS using this method. This suggests that in CDS, adult speakers could be helping child listeners with the task of speech segmentation by strongly signaling word and phrase boundaries through stress, and by placing syllables and words at rhythmically-regular intervals. Such word boundary exaggeration could help the child to segment words from the speech stream more easily, facilitating speech comprehension and new vocabulary acquisition. Therefore, the rhythmic adaptations seen in CDS are especially suited for the language needs and abilities of the young listener.

Another point to address is why the difference in entropy between CDS and ADS only occurs for the Stress-Syllable interface, and not for the Syllable-Phoneme interface. If the speech segmentation explanation is accepted, then it follows that young children do not need to segment phonemes from syllables for speech comprehension (although children taught alphabetic orthographies do learn to do this later on as they become literate readers, see Ziegler & Goswami, 2005). Therefore adults do not need to emphasise phonemes to children, and the Syllable-Phoneme phase relationship and distribution remains the same across both CDS and ADS samples. If adults exaggerate what is necessary for speech segmentation in CDS, this suggests that child-directed speech in different languages may show different properties, depending on the cues used for segmentation in that language. For example, children learning syllable-timed languages like French, Spanish or Italian may indeed benefit from exaggeration in the Syllable-Phoneme relationship, since stress is not as strong a cue for word segmentation in these languages. Therefore, one might predict a lower entropy in the Syllable-Phoneme distribution in French, Spanish and Italian CDS as compared to ADS.

Finally, in the entropy and autocorrelation analyses, ADS Rhyme was not substantially different from CDS Rhyme. However, in the Spectral PCA analysis, Spectral PCA loading patterns were clearly differentiated for ADS Rhyme and CDS Rhyme samples. This suggests that speakers were not as successful in differentiating ADS nursery rhymes from CDS nursery rhymes in terms of rhythmic structure, and relied more on changing the spectral structure (e.g. pitch) of their utterances. Indeed, several participants commented that the natural rhythmic structure of the nursery rhymes was very strong, and they found it hard to suppress this in the ADS rendition. Therefore, the ADS rendition of nursery rhymes was only partially successful since it was neither truly devoid of CDS characteristics, nor fully representative of typical adult speech. Therefore, in future experiments comparing ADS to CDS, it might be better to use neutral sentences that are not inherently biased toward a child or adult audience.

8 SPEECH RHYTHM PERCEPTION AND PRODUCTION IN DEVELOPMENTAL DYSLEXIA

Developmental dyslexia is associated with phonological difficulties and also with rhythmic difficulties in speech and music tasks (e.g. Huss et al, 2011). In speech, rhythm-bearing syllable and prosodic stress patterns are associated with slow amplitude modulations (AM) in the speech envelope. Consequently, dyslexics' rhythm deficits may be associated with impaired perception and production of these slow AMs in the speech envelope. Recall that in Section 1.11 of the Introduction, the syllable stress experiment conducted by Leong et al (2011) had linked participants' auditory sensitivity to onset rise times to deficits in syllable stress perception. The amplitude rise time parameter measured in previous studies (e.g. Goswami et al, 2002; Goswami et al, 2011; Huss et al, 2011) may be thought of as the upward-going half of the oscillatory AM cycle (i.e. $-\pi$ to 0 radians phase). Since the rise times of tone stimuli used by Goswami and colleagues varied between 15ms to 300ms, the corresponding AM rates for these rise times would be 1.7-33 Hz (taking the full AM cycle length to be twice the rise time length). For the original AMPH hierarchy, this range of AM rates included the Stress, Syllable and Subbeat tiers, as well as half of the Fast tier. For the S-AMPH hierarchy, the 1.7-33 Hz range covers the Stress and Syllable tiers, as well as almost the entire Phoneme tier. Therefore, in this study, dyslexics' speech rhythm perception and production was measured in each of these AM tiers, to see if the dyslexic problem could be more specifically pinpointed to a particular AM tier or tiers. Accordingly, this would implicate speech processing of those specific linguistic unit or units.

In this chapter, 3 AM-based speech rhythm perception and production experiments are presented. Each of these experiments used the same 4 metrically-regular nursery rhyme sentences as experimental stimuli, and the same group of dyslexic and non-dyslexic adults as participants. In each experiment, the AMPH or S-AMPH model was used as an analysis framework alongside traditional methods of analysis. In Experiment 1, perception of rhythm in speech was examined, where the rhythm was provided by AM patterns only (i.e. tone-vocoding). In Experiment 2, motor entrainment (tapping) to AM patterns in speech was examined. In Experiment 3, rhythmic production of speech was investigated, with emphasis placed on the AM patterns in the produced utterances. To examine the specific relationship between individual differences in rhythm perception and production, and reading-related

skills, a battery of tasks measuring other cognitive correlates of reading was also administered. These included standard abilities tests (IQ, memory), phonological awareness measures and psychoacoustic threshold measures for detecting change in acoustic rise time, frequency, intensity and duration. Reading and spelling were also measured.

8.1 METHODS

8.1.1 PARTICIPANTS

A total of 21 adults with dyslexia (9 M, 12 F), and 22 adults without dyslexia (7 M, 15 F) participated in the study. Dyslexic participants had a formal statement of developmental dyslexia, were native English speakers, and had no other documented learning disabilities. Dyslexic and non-dyslexic control groups were matched for age, verbal and non-verbal intelligence. Participants were recruited by advertisement in the Cambridge Graduate Union Bulletin, and were students at the University of Cambridge. Informed consent was obtained from each participant, and each participant was paid £15 for participating in the study. The consent form and background information form used in this study are shown in Appendix 8.1.

8.1.2 TASK SUMMARY

Each participant completed the 3 speech rhythm tasks, and another 4 sets of tasks. All the tasks used in the study are summarised in Table 8.1.

Table 8.1. Summary of Tasks Used in the Dyslexia Study

Task Battery	**No. of Tasks**	**Names of Tasks**
a. Standardised Ability Tests	3	WASI Block Design & Vocabulary WAIS-R Digit Span
b. Reading and Spelling Tests	4	WRAT Reading and Spelling TOWRE Word & Non-Word Reading
c. Phonological Awareness Measures	3	Spoonerisms RAN Dense & Sparse
d. Psychoacoustic Threshold Measures	4	'Dino' Rise Time, Frequency, Intensity and Duration
e. AM-Based Speech Rhythm Perception & Production Tasks	3	Expt 1 Rhythm Perception Expt 2 Rhythm Entrainment Expt 3 Rhythm Production

To improve the flow of the chapter, a description of the general ability, literacy, phonology and psychoacoustic tasks (a-d) will be followed immediately by the results for these tasks (a-d). After this, the three AM-based rhythm experiments (e) will be described. Each experiment description will be followed directly by the results of that rhythm experiment.

8.2 General Ability, Literacy, Phonology & Psychoacoustic Measures

8.2.1 Task Description

a. Standardised Ability Tests

i. *Non-Verbal and Verbal Intelligence*. All participants completed 2 subscales of the Wechsler Abbreviated Scale of Intelligence (WASI; Weschler, 1999), a nonverbal subscale (Block Design) and a verbal subscale (Vocabulary). In the Block Design task, participants were shown a picture of a geometric shape, and had to construct the shape using individual coloured blocks as quickly as possible. The coloured blocks each had two completely red faces, two completely white faces, and two faces that were half red and half white (split diagonally). Initially the pictures could be constructed using only 4 blocks, but later required 9 blocks. Participants constructed 10 designs in total, and were scored according to the time taken to complete each design (0 to 7 points). In the Vocabulary task, participants were verbally presented with a word such as *"intermittent"*, and were asked to provide a verbal definition of the word. Participants were scored according to a pre-set list of accepted definitions, and could obtain a score of either 0, 1 or 2 points for each of 42 words. The two subscale raw scores were then converted into standardised T-scores according to the age of the participant.

ii. *Auditory short-term memory*. The Weschler Adult Intelligence Scale-Revised forward digit span subtest (WAIS-R; Weschler, 1981) was administered as a measure of auditory short-term memory. In this task, participants heard a sequence of digits and had to repeat the digits back to the experimenter in the same order that they were presented. Initially, the sequence comprised just 2 digits (e.g. *"1 - 7"*), but the sequences grew in length

up to 9 digits (e.g.*"2 – 7 – 5 – 8 – 6 – 2 – 5 – 8 – 4"*). In the presentation phase, the digits were spoken by the experimenter at a pace of 1 digit per second, with a neutral affect. In the recall phase, participants were free to speak as quickly as they wished. Participants scored 1 point for each sequence that they recalled correctly with no mistakes, and the task was stopped after the participant failed on two consecutive trials. The maximum score for this task was 16.

b. Reading and Spelling Tests

Literacy skills were assessed using the untimed Wide Range Achievement Test (Reading and Spelling scales, WRAT-III, Wilkinson, 1993) and the timed Test of Word Reading Efficiency (TOWRE, Single Word Efficiency [SWE] and Phonological Decoding Efficiency [PDE], Torgesen, Wagner, Rashotte, 1999).

In the WRAT Reading test, participants were shown a list of 42 words and had to read the words aloud as clearly as possible, with no restrictions on time. The words were of increasing difficulty (with *"in"* as item 1 and *"terpsichorean"* as item 42). Participant were scored 1 point for each word that they read correctly, with a maximum raw score of 42. In the WRAT Spelling task, participants had to spell a total of 40 words of increasing difficulty (from *"and"* as item 1 to *"vicissitude"* as item 40). Each of these words was presented orally 3 times - one time on its own, one time used in a sentence, and the last time on its own again. For example, for the word *"lucidity"*, participants heard *"Lucidity. We think best in moments of lucidity. Lucidity"*. Participants received 1 point for each word that they spelled correctly, for a maximum total raw score of 40. The raw scores were then converted into standardised scores according to the age of the participant.

In the two TOWRE tasks, participants were presented with a list of words (or non-words), and had to read aloud as many items as they could within 45 seconds. Prior to starting the timed task, participants were given a short practice list of words to read, e.g. *"on, my, bee, old, warm, bone, most, spell"* (TOWRE Word), and *"ba, um, fos, gan, rup, nasp, luddy, dord"* (TOWRE Non-word). The words in the actual test list were sorted in order of increasing length, going from 1-syllable words at the beginning to 3-syllable words at the end. Participants were told to read the words quickly but clearly. They received 1 point for each item that was correctly read. The maximum score for TOWRE Word was 104, and the maximum score for TWO Non-word was 63.

c. Phonological Awareness Measures

i. Spoonerisms. This task was drawn from the Phonological Assessment Battery (PhAB; Fredrickson, ed., 1997). There were two parts to the task. In the first part, participants heard 10 single words presented orally by the experimenter. They were asked to replace the onset phoneme(s) of each word with a different phoneme(s). For example, the experimenter said *"cot with a /g/ gives..."* and the correct reply from the participant would have been *"got"*. In the second part of the task, participants heard 10 *pairs* of words instead of single words. Participants were asked to swap the onset phonemes of the pair of words (e.g. for *"sad cat";* the participant responded *"cad sat"*). Participants received 1 point for each correct response in the first part of the task, and a maximum of 2 points for each correct response in the second part of the task. Therefore scores on this measure were out of a possible 30 points.

ii. RAN (Rapid Automatized Naming). Two versions of an object RAN task designed originally for children were administered. One version was based on pictures of objects whose names resided in dense phonological neighbourhoods (RAN Dense: Cat, Shell, Knob, Thumb, Zip). The other version was based on pictures of objects whose names resided in sparse phonological neighbourhoods (RAN Sparse: Web, Dog, Fish, Cup, Book). Participants were shown a sheet of paper with the same pictures repeated 50 times, arranged in a grid. In each case, they were asked to name the pictures in order, as quickly and accurately as possible. It was expected that words from dense neighbourhoods would take longer to produce. Performance was timed, and the final score was the time taken (in seconds) for participants to complete naming all the pictures in the grid.

d. Psychoacoustic Threshold Measures

These 'Dino' tasks were designed to measure participants' ability to discriminate small acoustic changes in a single auditory dimension. The four auditory dimensions measured were rise time, frequency, intensity and duration. Acoustic changes in these four dimensions cue prosodic stress in speech. For example, stressed syllables are characterised by higher pitch, higher intensity, longer duration and longer rise times as compared to unstressed syllables (Leong et al, 2011). The auditory tasks were programmed for this study by Martina Huss, and were originally designed to be used with children. The name of the tasks ('Dino') derives from the cartoon animals used in the presentation of the acoustic stimuli. The tasks were designed to assess participants' threshold for a just-noticeable difference (JND) on each

dimension, via an adaptive staircase procedure (Levitt, 1971). In this procedure, participants initially heard stimuli that had a large acoustic difference. If these were correctly discriminated, the acoustic gap on subsequent trials was narrowed (making the trials more difficult) until the participant gave an incorrect response (a 'reversal'). At this point, an easier trial (with a larger difference) was presented. This procedure therefore adapted to the performance of the individual by shifting the difficulty of trials up or down. In the Dino task, a combined 2-up 1-down[29] and 3-up 1-down procedure was used; after 2 reversals (i.e. incorrect responses), the 2-up 1-down staircase procedure changed into 3-up 1-down. The step size was halved after the 4th and 6th reversal so that difficulty would increase in smaller steps as the participant neared his or her JND threshold. A test run typically terminated after 8 response reversals or alternatively after a maximum of 40 possible trials was completed. Four attention trials were randomly presented during each test run, using the maximum contrast of the respective stimuli in each auditory task. The threshold score achieved was calculated using the mean of the last four reversals. All psychoacoustic stimuli were presented binaurally at 74 dB SPL using Sennheiser HD 580 headphones.

i. Amplitude Envelope Onset (Rise Time) Task (1 Rise). This was a rise time discrimination task in AXB format. Three 800 ms tones were presented on each trial, with 500 ms ISIs. Two (standard) tones had a 15 ms linear rise time envelope, 735 ms steady state, and a 50 ms linear fall time. The third tone varied the linear onset rise time logarithmically, with the longest rise time being 300 ms. Participants were introduced to three cartoon dinosaurs. It was explained that each dinosaur would make a sound and that the task was to decide which dinosaur's sound was different from the other two and had a softer rising sound (longer rise time). As an integral part of the software programme, feedback was given after every trial on the accuracy of performance. Schematic depiction of the stimuli can be found in Richardson et al. (2004).

ii. Frequency task. This was a frequency discrimination task delivered in a 2IFC format. The standard was a pure tone with a frequency of 500 Hz presented at 74 dB SPL, which had a duration of 200 ms. The maximum pitch difference between the stimuli presented in this task was 60 Hz. Participants were introduced to two cartoon elephants. It was explained that each elephant would make a sound and that the task was to decide which elephant's sound was higher in pitch.

[29] Difficulty is shifted up after 2 consecutive correct responses, but shifted down after just one incorrect response.

iii. Intensity task. This was an intensity discrimination task also delivered in a 2IFC format. The standard was a pure tone with a frequency of 500 Hz presented at 74 dB SPL, which had a duration of 200 ms. The intensity of the second tone ranged from 54 to 74 dB SPL. Participants were introduced to two cartoon mice. It was explained that each would make a sound, and the task was to decide which sound was softer.

iv. Duration task. This was a duration discrimination task delivered in an AXB format. It was explained that each dolphin would make a sound and the task was to decide which sound was different in length (longer) as compared to the other two. The (two) standard sounds had a duration of 125 ms. The third tone varied linearly in duration from 125 ms to 250 ms. All sounds were pure tones at 500 Hz and had a 5 ms rise and fall time.

8.2.2 RESULTS

Participants' performance on all the non-rhythm tasks (a-d) are shown in Table 8.2. One-way ANOVAs were used to compare the scores of control and dyslexic participants. Table 8.2 shows the group means for each test, and the results of the one-way ANOVA for group differences, with significant differences highlighted in blue. For each variable, where the assumption of homogeneity of variances is violated (Levene's test $p<.05$), the Welch's statistic and p-value is reported instead. As shown in Table 8.2, control and dyslexic groups were matched for age, verbal and non-verbal IQ. As expected, dyslexic participants performed significantly more poorly than controls on all reading, spelling and phonological measures. They also showed a significant short-term memory deficit. These results confirmed that our recruited dyslexic cohort did indeed have significant reading and phonological problems. For the psychoacoustic threshold measures, dyslexic participants showed a significantly lower sensitivity (higher threshold) for intensity detection, but not for detection of rise time, frequency or duration.

Table 8.2. Summary of Task Results (a-d) and One-Way ANOVA Tests

Task			**Controls**	**Dyslexics**	**One-way ANOVA**			
					F(1,41)	*p-value*	Welch's statistic	*p-value*
(a) Age & General Ability	**Age** (years)	Mean (SD)	24.08 (2.45)	22.90 (2.93)	2.08	*0.157*	-	-
	WASI Non-Verbal IQ (standardised T-score)	Mean (SD)	70.59 (4.14)	70.57 (3.03)	0.00	*0.986*	-	-
	WASI Verbal IQ (standardised T-score)	Mean (SD)	62.09 (7.86)	62.04 (4.71)	0.00	*0.983*	-	-
	Digit Span (score out of 16)	Mean (SD)	13.14 (2.01)	10.33 (1.71)	**24.16**	***0.000****	-	-
(b) Reading & Spelling	**WRAT Spelling** (standardised score)	Mean (SD)	116.45 (6.07)	104.71 (6.67)	**36.49**	***0.000****	-	-
	WRAT Reading (standardised score)	Mean (SD)	115.59 (5.34)	110.81 (6.44)	**7.05**	***0.011****	-	-
	TOWRE Word Reading (words read in 45 seconds)	Mean (SD)	99.73 (6.27)	88.14 (11.15)	17.84	*0.000*	**17.41**	***0.000****
	TOWRE Non-word Reading (words read in 45 seconds)	Mean (SD)	59.55 (3.10)	46.76 (7.62)	52.88	*0.000*	**51.11**	***0.000****

Task			Controls	Dyslexics	One-way ANOVA			
					F(1,41)	*p-value*	Welch's statistic	*p-value*
(c) Phonological Measures	**Spoonerisms** (score out of 30)	Mean (SD)	28.50 (1.41)	26.10 (2.05)	**20.33**	***0.000****	-	-
	RAN Sparse (time in seconds, faster = better)	Mean (SD)	23.32 (3.39)	26.42 (4.83)	**6.00**	***0.019****	-	-
	RAN Dense (time in seconds, faster = better)	Mean (SD)	24.51 (3.92)	28.58 (5.75)	**7.41**	***0.009****	-	-
(d) Psychoacoustic Threshold Measures	**Dino Rise Time** (threshold up to 39, lower = better)	Mean (SD)	7.57 (6.32)	6.65 (6.65)	0.35	*0.558*	-	-
	Dino Frequency (threshold up to 39, lower = better)	Mean (SD)	8.20 (6.94)	9.80 (6.42)	0.62	*0.436*	-	-
	Dino Duration (threshold up to 39, lower = better)	Mean (SD)	7.46 (2.00)	9.29 (5.75)	1.99	*0.166*	1.92	*0.178*
	Dino Intensity (threshold up to 39, lower = better)	Mean (SD)	4.95 (1.63)	6.21 (2.05)	**4.99**	***0.031****	-	-

p<.05, **p<.01,*p<.001*

8.3 AM-BASED SPEECH RHYTHM PERCEPTION & PRODUCTION TASKS

8.3.1 MATERIALS

Four nursery rhyme sentences were used for all three rhythm experiments. These were the same sentences that had earlier been used for the tone-vocoding experiment in Chapter 3 (see Table 3.1). To recap, all sentences were 8 syllables in length and had an alternating 'S-w' rhythm pattern. Two sentences ('Mary Mary' and 'Simple Simon') had a trochaic stress pattern such as 'S-w-S-w-S-w-S-w' while the other two sentences ('St Ives' and 'Queen of Hearts') had an iambic stress pattern such as 'w-S-w-S-w-S-w-S'. In the perception and entrainment experiments (Experiments 1 & 2), these sentences were produced by a female native British English speaker who was speaking in time to a 4 Hz (syllable rate) metronome beat. Therefore the four sentences were perfectly metrically-regular, with syllables occurring every 250 ms, and stressed syllables occurring every 500 ms. The sentences had a duration of around 2s. In the third production experiment (Experiment 3), participants produced these four sentences themselves in time to a metronome beat.

8.3.2 EXPERIMENT 1 : RHYTHM PERCEPTION (TONE VOCODER) TASK

8.3.2.1 Task Description

The aim of this task was to investigate AM-based speech rhythm perception in dyslexia. Therefore, the four nursery rhyme sentences were tone-vocoded using different AM tiers, and participants were asked to identify the sentences on the basis of the rhythm pattern that they heard. By systematically presenting specific AM tiers and combinations of AM tiers, deficits in rhythm perception could be accurately pinpointed to a problem with a specific AM rate or rates.

This task was delivered during the tone-vocoding experiment previously reported in Chapter 3, where the 1-channel vocoder condition previously reported comprised half of the overall session. In the second part of the session, a 29-channel vocoder was also used to generate *intelligible* stimuli (29 ERB_N-spaced channels spanning 100 Hz - 7250 Hz), and

both sets of stimuli (1-channel and 29-channels) were presented to control[30] and dyslexic participants. Therefore, the design of Experiment 1 followed an AM tier (5) x Phase Shift (3) x Demodulation Method (2) x Channels (2) x Group (2) design. As this experiment was conducted at the time when the original AMPH model was being developed, the AM tiers used for vocoding came from the original AMPH 5-tier AM hierarchy (rather than the new 3-tier S-AMPH hierarchy). The procedures used for tone-vocoding and phase-shifting were described in Chapter 3, Section 3.1.4. For the 29-channel stimuli, these vocoding and phase-shift procedures were applied to each individual channel, and the channels were equalised to 70dB before being summed together in the final stimulus.

As described in Chapter 2, the five different AM tiers or tier combinations used for vocoding were 1) Stress only; 2) Syllable only; 3) Sub-beat only; 4) Stress+Syllable and 5) Syllable+Sub-beat. Each of these AM combinations was presented in three phase shift conditions : 1) No Shift ; 2) 1π radians-shifted and 3) 2π radians-shifted. Recall that for AM tier pairs (e.g. Stress+Syllable and Syllable+Sub-beat), phase-shifting involved shifting the slower AM with respect to the faster AM. Fewer phase-shifted stimuli (1π radians or 2π radians) were presented than non-phase-shifted versions (0 radians) to allow participants to maintain a strong representation of the correct metrical pattern for each nursery rhyme. Thus, participants heard the normative (0 radians) version five times for each nursery rhyme, but they only heard each of the phase-shifted variants (1π radians or 2π radians) twice.

Phase-shifted and normal (0 radians) stimuli were presented within the same experimental block in a randomised fashion. Stimuli that were vocoded using MFB-produced AMs and PAD-produced AMs were presented in separate experimental blocks. 29-channel and 1-channel stimuli were also presented in separate blocks. 29-channel stimuli were always presented before 1-channel stimuli as these stimuli were intelligible and 'easier', allowing participants to get used to listening to the tone-vocoded stimuli[31]. However, the order of presentation for MFB or PAD stimulus blocks was counterbalanced across participants. This gave a total of 720 trials divided into 4 blocks over the entire experiment (5 AM tier combinations x 9 phase variants [5 x 0π radians, 2 x 1π radians, 2 x 2π radians] x 4 nursery

[30] The 22 control participants included in this experiment were a subset of the 23 participants from the 1-channel vocoder experiment in Chapter 2. One participant was removed so that the control and dyslexic groups would be more closely matched in age.

[31] In a pilot experiment, the 1-channel stimuli were presented first to some participants (in a counterbalanced procedure), but their performance was at chance and participants complained that the task was too difficult and were less motivated to complete the experiment.

rhymes per block, x 2 demodulation method blocks x 2 vocoding channel blocks). The task set-up was as described in Chapter 3, Section 3.1.3.

8.3.2.2 Results of Rhythm Perception (Tone Vocoder) Task

Participants were scored in two ways. First, an Accuracy score was obtained which corresponded to the number of sentences that were correctly identified. Next, a 'Rhythm Pattern' (RP) score was computed, which assessed whether participants had correctly identified the trochaic or iambic pattern of the sentence. As the Accuracy and RP scores in the 1-channel vocoder condition were normally-distributed (Kolmogorov-Smirrnov test, $p>.05$), these were analysed using parametric statistics. However, performance in the 29-channel condition was near ceiling because these sentences were intelligible. Since the scores were not normally distributed for this condition, (Kolmogorov-Smirrnov test, $p<.05$), non-parametric statistics were used instead. Recall that the design of the experiment involved 5 factors. These were AM tier, Phase Shift, Demodulation Method, Channels and Group. Therefore, the main analysis involved comparing the first 3 factors (AM tier, Phase Shift and Method) across Groups. This analysis was repeated for 1-Channel and 29-Channel conditions using parametric and non-parametric statistics respectively. In the following section, the results of the 1-Channel condition are presented first, followed by the 29-Channel condition.

a. 1-Channel Vocoder

The mean Accuracy and RP scores for each experimental condition and group for the 1-Channel vocoder condition are shown in Table 8.3. These results were analysed using a Repeated Measures ANOVA, taking AM tier (5 tiers), Phase Shift (3 shifts) and Method (PAD or MFB) as within-subjects factors, and Group (Control or Dyslexic) as the between-subjects factor.

For Accuracy scores, the results indicated no significant main effect of Group ($F(1,41) = 1.33$, $p=.26$) and no significant interactions between any factor and Group. There were, however, as expected, significant main effects of AM tier ($F(4,164) = 6.07$, $p<.001$) and Phase ($F(2,82) = 19.3$, $p<.001$), and a significant interaction between AM tier and Phase ($F(8,328) = 4.65$, $p<.001$) with general trends similar to those reported in Chapter 3 (Sections 3.2.1 and 3.2.2). These trends are not described further here as the focus of this analysis was on Group differences. Surprisingly, there was a significant effect of Method ($F(1,41) = 9.89$,

p<.01) which was not present in the analysis with controls only in Chapter 3. Here, participants showed higher accuracy for PAD stimuli as compared to MFB stimuli, but there was no interaction between Group and Method ($F(1, 41) = 1.39$, $p=.24$).

For RP scores, there was again no significant main effect of Group ($F(1,41) = 1.19$, $p=.28$) and no significant interactions between any factor and Group. For the other factors, only Phase showed a significant main effect ($F(2,82) = 19.84$, $p<.001$), whereas both AM tier and Method were not significant. In the critical Group comparisons, dyslexics and controls did not differ significantly in any comparison for their performance on 1-Channel vocoded stimuli. Therefore statistically, both groups performed equally on all 5 AM tiers, and showed an equal phase-shift effect.

b. 29-Channel Vocoder

The mean Accuracy and RP scores for each experimental condition and group for the 29-Channel vocoder condition are shown in Table 8.4. The Accuracy and RP scores from this condition were analysed using a non-parametric Mann-Whitney U test. Since it was not possible to perform 60 separate tests (which would require a severe Bonferroni correction and loss of power), the Accuracy and RP scores were averaged into one composite score, and PAD and MFB scores were also averaged (based on the 1-channel finding that controls and dyslexics did not show a different pattern of responses across the two demodulation methods). This produced 15 new variables - composite scores for 5 AM tiers and 3 Phase Shifts. Figure 8.1 plots these new composite scores by group for each AM tier and phase shift condition. Accordingly, 15 Mann-Whitney U tests were conducted to test for differences between groups, and a Bonferroni-corrected p-value of 0.003 was used for these tests. Out of the 15 comparisons, only 2 comparisons produced a p-value of less than 0.05 ($p = 0.026$ and $p = 0.028$). These two comparisons are marked with a (#) in Figure 8.1, and correspond to the Stress+Syllable and Syllable+Subbeat AM combinations in the no-phase-shift condition. However, these p-values did not survive the Bonferroni correction, as they were not less than 0.003.

Figure 8.1. Composite scores for each AM tier/tier combination and phase shift condition. Each subplot shows a different AM tier/combination, and the x-axis shows the phase shift. Controls are shown in blue and dyslexics are shown in red. Errorbars indicate the standard error. (#) Indicates a group difference significant at the p<.05 level, but not at the Bonferroni-corrected p< .003 level.

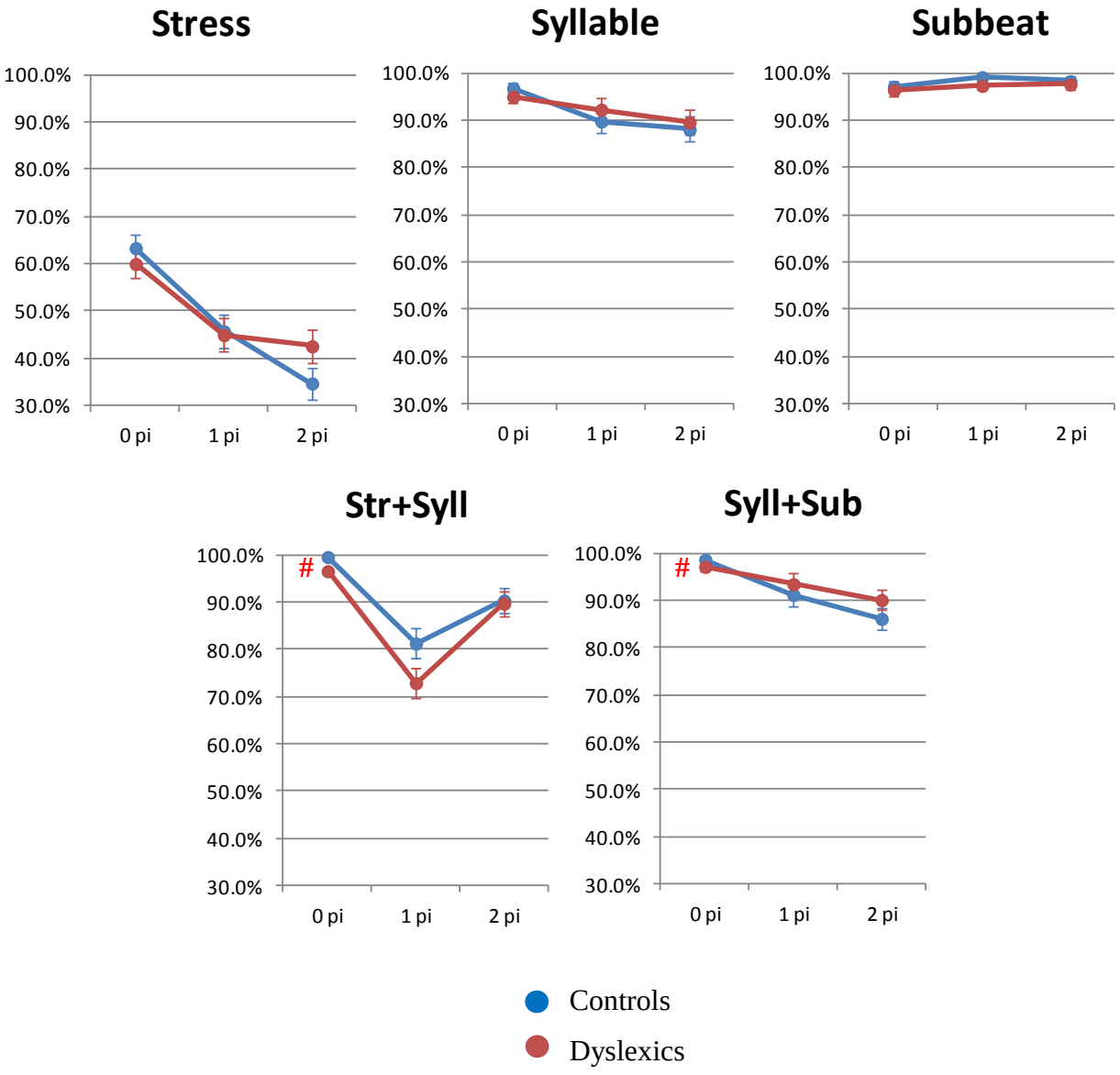


Table 8.3. Mean Accuracy and RP Scores in the 1-Channel Vocoder Condition

Method	**Phase-Shift (radians)**	**AM Tier**	**Accuracy Scores**		**RP Scores**	
			Controls	*Dyslexics*	*Controls*	*Dyslexics*
PAD	0 π	Stress	35.2%	30.5%	64.4%	62.7%
		Syllable	35.8%	35.1%	59.9%	58.9%
		Subbeat	31.6%	31.8%	54.8%	53.5%
		Str+Syll	42.2%	33.9%	70.3%	59.5%
		Syll+Sub	33.8%	31.4%	57.9%	54.4%
	1 π	Stress	22.7%	19.6%	41.5%	42.9%
		Syllable	32.4%	27.9%	58.0%	48.6%
		Subbeat	31.8%	33.4%	58.0%	58.1%
		Str+Syll	18.8%	24.7%	40.9%	44.3%
		Syll+Sub	29.0%	32.3%	59.1%	55.4%
	2 π	Stress	27.3%	28.6%	49.4%	56.0%
		Syllable	31.3%	28.7%	59.9%	56.8%
		Subbeat	20.5%	31.2%	49.4%	55.3%
		Str+Syll	36.9%	30.9%	59.7%	54.3%
		Syll+Sub	32.4%	31.3%	56.8%	59.2%
MFB	0 π	Stress	27.2%	29.8%	62.7%	59.7%
		Syllable	31.6%	30.9%	62.3%	58.9%
		Subbeat	28.2%	28.6%	56.8%	56.1%
		Str+Syll	38.9%	34.8%	68.6%	65.2%
		Syll+Sub	36.8%	30.0%	63.9%	56.5%
	1 π	Stress	19.3%	22.6%	47.2%	47.5%
		Syllable	28.4%	25.1%	51.7%	48.7%
		Subbeat	31.8%	19.6%	57.5%	46.6%
		Str+Syll	26.7%	20.8%	50.0%	44.0%
		Syll+Sub	26.1%	29.7%	48.9%	54.6%
	2 π	Stress	27.6%	18.0%	60.1%	54.5%
		Syllable	25.6%	26.8%	44.9%	46.1%
		Subbeat	24.4%	23.0%	43.2%	52.3%
		Str+Syll	38.6%	30.8%	71.0%	60.7%
		Syll+Sub	27.3%	30.3%	53.4%	54.7%

Table 8.4. Mean Accuracy and RP Scores in the 29-Channel Vocoder Condition

Method	**Phase-Shift (radians)**	**AM Tier**	**Accuracy Scores**		**RP Scores**	
			Controls	*Dyslexics*	*Controls*	*Dyslexics*
PAD	0 π	Stress	54.2%	55.1%	74.2%	72.6%
		Syllable	95.0%	93.0%	96.6%	95.9%
		Subbeat	69.3%	61.1%	80.9%	77.7%
		Str+Syll	96.8%	94.3%	98.6%	97.1%
		Syll+Sub	98.9%	94.5%	99.1%	95.0%
	1 π	Stress	39.8%	41.9%	61.4%	65.1%
		Syllable	84.7%	76.2%	89.8%	82.1%
		Subbeat	70.5%	63.2%	81.8%	77.0%
		Str+Syll	48.9%	51.8%	62.5%	66.1%
		Syll+Sub	67.0%	56.5%	73.3%	66.1%
	2 π	Stress	31.3%	33.6%	55.7%	62.3%
		Syllable	85.2%	76.6%	88.1%	82.7%
		Subbeat	59.1%	53.7%	72.2%	67.6%
		Str+Syll	84.1%	74.2%	88.1%	84.4%
		Syll+Sub	76.1%	71.9%	83.5%	81.0%
MFB	0 π	Stress	55.9%	52.2%	70.8%	67.7%
		Syllable	96.4%	94.0%	97.3%	96.0%
		Subbeat	96.6%	95.7%	97.7%	97.1%
		Str+Syll	99.3%	96.2%	99.8%	96.9%
		Syll+Sub	98.6%	97.1%	98.6%	97.1%
	1 π	Stress	35.8%	36.9%	55.7%	53.0%
		Syllable	88.6%	89.9%	90.9%	94.6%
		Subbeat	98.9%	97.0%	99.4%	97.6%
		Str+Syll	79.0%	69.3%	83.5%	76.4%
		Syll+Sub	89.8%	92.9%	92.6%	94.0%
	2 π	Stress	21.0%	29.2%	48.4%	56.0%
		Syllable	86.9%	88.7%	89.2%	90.5%
		Subbeat	98.3%	97.6%	98.3%	97.6%
		Str+Syll	86.9%	86.7%	93.8%	92.7%
		Syll+Sub	84.3%	89.2%	87.9%	91.0%

In both the 1-Channel vocoder and 29-Channel vocoder tasks, there were no significant differences in performance between controls and dyslexics. One reason could be that participants were relatively old and highly compensated dyslexics. However, it is interesting to note that even in the 29-channel condition where the sentences were intelligible, participants still showed a phase-shift effect for the Stress+Syllable AM combination (see bottom left subplot in Figure 8.1). Compared to the no-shift baseline, performance dropped for the 1π radians shift and recovered for the 2π radians shift. Therefore, even when phonetic information was available, participants' judgments were still influenced by the rhythm pattern of the sentence. Participants found it harder to identify the sentence when its prosodic pattern was incongruent to what they expected (i.e. 1π radians shift), and easier to identify the sentence when its prosodic pattern was congruent with their expectations (i.e. 2π radians shift). Moreover, this drop-recovery phase-shift pattern was only observed for the Stress+Syllable AM combination, lending further support for the proposal that rhythm information is specifically carried by these two key rates of amplitude modulation.

c. Correlations Between Speech Rhythm Perception, Reading & Phonology

Even though there were no group differences, the theory predicts that relationships should exist between phonological awareness and participants' perception of speech rhythm patterns. Therefore, individual differences in performance on the 29-channel vocoder task were correlated with participants' reading, phonology and psychoacoustic scores. For this analysis, 29-channel vocoder composite scores were used, for the no-phase-shift condition, since some of these scores had showed differences between groups (prior to the Bonferroni correction). Table 8.5 shows the resulting correlations, where significant correlations are marked in blue. The only significant correlation with reading was for the Stress+Syllable AM tier combination, which was one of the two AM conditions with group differences under a significance value of p=.05. Here, TOWRE non-word reading was significantly correlated with performance for Stress+Syllable AM tiers ($r = 0.30$, $p<.05$).

For the phonological measures, two AM tiers showed significant correlations. Performance on the Syllable AM tier was significantly negatively correlated with RAN Sparse times, where better vocoder performance was related to *faster* picture naming ($r = -0.33$, $p<.05$). Performance on the Stress+Syllable AM tier combination was significantly negatively correlated to both RAN Sparse and RAN Dense times ($r = -0.35$, $p<.05$ for both),

where better vocoder performance was also related to faster picture naming. If Stress and Syllable AMs together carry prosodic stress patterns in speech, it is not surprising that relationships to reading and phonology are strongest for this AM combination, since prosodic stress perception is related to reading and phonology (Whalley & Hansen, 2006; Leong et al, 2011; Goswami et al, 2010).

Table 8.5. Correlations between Composite scores for each AM tier (no phase shift condition), and performance in reading, phonology and psychoacoustic measures. Correlations are reported over the full group of participants, df = 41.

	Task	**AM Tier (no phase shift)**				
		Str	*Syl*	*Sub*	*Str+Syl*	*Syl+Sub*
Reading & Spelling	WRAT Spelling	0.07	0.08	0.16	0.16	0.07
	WRAT Reading	-0.02	0.09	-0.01	0.03	0.08
	TOWRE Word Reading	0.25	0.26	0.22	0.13	0.12
	TOWRE Nonword	0.22	0.20	0.15	***0.30**	0.23
Phonology	Spoonerisms	-0.01	0.12	0.02	0.09	0.20
	RAN Sparse	-0.20	***-0.33**	0.01	***-0.35**	-0.18
	RAN Dense	-0.28	-0.22	-0.03	***-0.35**	-0.20
Psychoacoustic	Dine Rise Time	-0.08	0.09	0.12	0.02	0.09
	Dino Frequency	***-0.38**	-0.28	-0.24	-0.12	0.04
	Dino Duration	-0.22	****-0.49**	***-0.32**	***-0.31**	****-0.51**
	Dino Intensity	-0.04	***-0.35**	0.01	-0.27	****-0.45**

**p<.05, **p<.01, ***p<.001*

For the psychoacoustic measures, the strongest correlations existed for *duration* discrimination, where duration thresholds were significantly related to performance on Syllable, Subbeat, Stress+Syllable and Syllable+Subbeat AM tier combinations. These correlations were negative because *lower* psychoacoustic thresholds indicate *better* discrimination. Only performance for the single Stress tier was unrelated to duration thresholds, but was instead significantly negatively related to Frequency thresholds (r = -0.38, p<.05). Two of the AM conditions were related to Intensity discrimination (Syllable and Syllable+Subbeat), but none were related to Rise Time discrimination. Therefore,

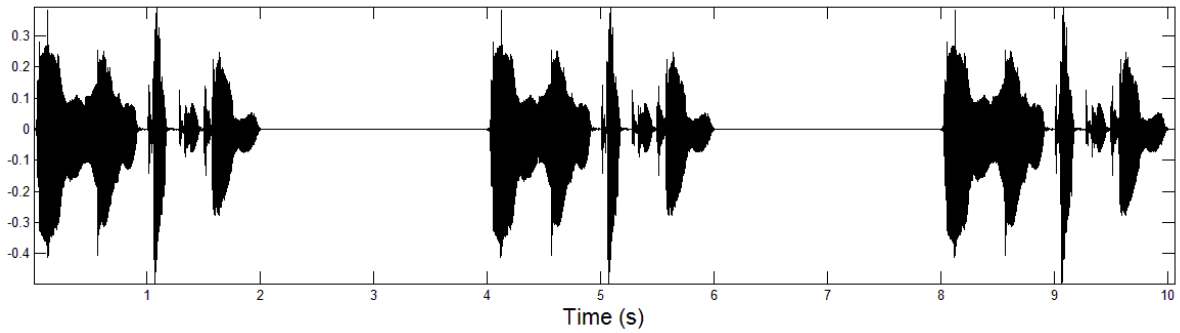
individual differences in perception of speech AMs at most rates appeared to be primarily related to duration discrimination. Bearing in mind that the speech stimuli in the 29-channel vocoder condition were intelligible (not pulse patterns), this suggests that changes in acoustic duration may be important both for speech rhythm perception, as well as for general speech intelligibility.

8.3.3 Experiment 2 : Rhythm Entrainment Task

8.3.3.1 Task Description

In this task, participants heard the original nursery rhyme sentence (*not* tone-vocoded) and were asked to tap along to the rhythm of each nursery rhyme sentence. Each sentence was repeated three times, with a silent gap between repetitions that was equal to the length of that sentence. Silence was inserted between sentence presentations so that participants would have to actively find the beat every time the sentence was presented, rather than relying on remembering the beat if the presentations had been continuous. Therefore, their beat-finding cognitive processes would be actively engaged throughout the entire experimental trial. Figure 8.2 shows an example of the three repetitions of the sentence *"Mary Mary quite contrary"* as presented to participants. Here, the length of the original sentence was 2.01s, and this was the length of silence inserted between repetitions of the sentence.

Figure 8.2. Example of a single trial for the sentence 'Mary Mary'.



Participants were instructed to begin tapping as soon as they heard the sentence begin. They could continue tapping through the silent periods, but were told to aim to come back in on time with the next presentation of the sentence. Therefore, the emphasis of the task was on timing their taps correctly to the beat of each sentence, every time they heard it, with a

'maintenance period' in-between sentence presentations. No instructions were given as to the rate of tapping, but all participants spontaneously tapped according to the *stress* rate of the sentence (i.e. 2 Hz), rather than trying to tap on every syllable. As it was expected that participants would take some time to entrain to the rhythm of the sentence, only taps from the second and third presentations of the sentence were used for analysis, and taps from the first presentation were discarded. The 2 trochaic sentences were presented first before the 2 iambic sentences, as the trochaic sentences were easier to track rhythmically. However, the order of presentation within the pairs of trochaic and iambic sentences (i.e. 'Mary Mary' first or 'Simple Simon' first) was counterbalanced across participants. Circular tests and analyses were conducted using the Matlab Toolbox for Circular Statistics (Berens, 2009).

It is important to note that the sentences used here did not contain an audible metronome beat, but were recordings of rhythmic speech produced to a metronome beat - that is, speech with a clear beat. Therefore, this experiment tested entrainment to the *acoustic carriers* of rhythmic beats in speech (e.g. AM patterns), corresponding to the 'p-centres' in speech. Rhythmic entrainment to a pure metronome beat (not speech) has previously been tested in dyslexia. In these studies, adult dyslexic individuals showed greater anticipation for the beat (Wolff, 2002) and greater inter-tap variability (Thomson et al, 2006). However, to the author's knowledge, this is the first study investigating motor entrainment to *speech* rhythm in dyslexia. Therefore, the results of this experiment have direct implications for dyslexics' perception of p-centres, rhythm and prosody in speech.

8.3.3.2 Results of Rhythm Entrainment Task

The results of this task were analysed using conventional measures (e.g. tapping intervals) as well as using the 5 x 3 AM hierarchy representation from the S-AMPH model.

a. Tapping Intervals & Tapping to Vowel Onsets

For this task, participants tapped along to the beat of metrically-regular nursery rhyme sentences. Since the stress rate of the target sentence was 2 Hz (4 Hz syllable rate, stress on alternate syllables), it was expected that participants would tap at this stress rate, generating tapping intervals of ~500 ms. The results were as predicted. On average over the four nursery rhyme sentences, controls had a mean tapping interval of 520 ms (SD = ±38ms) while dyslexics had a mean tapping interval of 512 ms (SD = ±16 ms). In an independent samples t-

test, there was *no* significant difference between controls and dyslexics in terms of their mean interval of tapping (t=.93, *p*=.*36*). Therefore, both control and dyslexic participants generated taps at appropriate intervals. Surprisingly, the dyslexics were *less* variable than controls in their tapping intervals (SDs of 16 ms vs 38 ms). This result suggested that both control and dyslexic participants were successfully entraining to the stress *rate* of the sentences. However, the tapping interval only indicates the average rate of tapping, and not whether controls and dyslexics were actually early or late with respect to the stress beat (i.e. p-centres) in the acoustic signal.

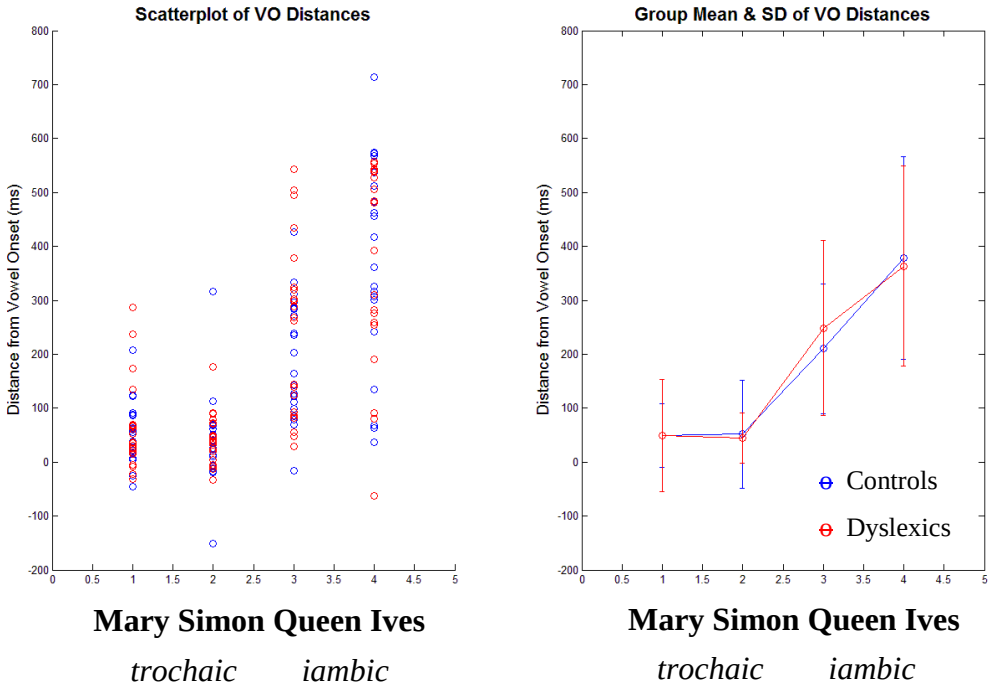
To investigate this, participants' taps were analysed in terms of their distance from linguistic p-centre markers. Since p-centres are thought to be located near[32] the onsets of vowels in stressed syllables (Allen, 1972), the vowel onsets of the stressed syllables in each sentence were identified, and participants' tap distance from each respective vowel onset was measured. Recall that there were four stressed syllables, and therefore four taps were made per sentence. In this analysis, the first two more variable taps for each sentence were discarded to ensure that the taps used for analysis reflected a stable 'entrained' state. Therefore, for each nursery rhyme sentence, four taps were taken for each participant - the last two taps from the last two repetitions of that sentence. The scatterplot of tapping distances from the vowel onset (averaged over the four taps) for each participant and nursery rhyme is shown in Figure 8.3.

From visual inspection of Figure 8.3, it is clear that participants' tapping was near to the vowel onsets (within 50 ms) for the two trochaic sentences ('Mary Mary', 'Simple Simon'), but much more distant and variable for the two iambic sentences ('Queen of Hearts', 'St Ives'). Indeed, some of the taps for the iambic sentences were so far distant from the current vowel onset (e.g. over 500ms) that they overalapped with the next vowel onset. It is possible therefore, that some of these iambic taps could have been responses made to the following vowel onset rather than the current one. In the initial screening process, only the first 4 taps made within the period of each sentence were taken, which helped to minimise this possibility. However, due to the continuous and circular nature of the data, it is difficult to distinguish between a very late response to the current vowel onset, and a very early anticipatory response to the next vowel onset. Consequently, the iambic data can be

[32] Note however that the exact location of the p-centre with respect to the vowel onset is influenced by the length of the initial consonant cluster of the syllable, and the length of the syllable coda. Since it was not possible to determine the exact p-centre for each different stressed syllable in the four nursery rhyme sentences, the vowel onset was used as a proxy marker of beat (p-centre) location for all the stressed syllables.

interpreted in two ways. First, participants could have been around 200 ms ('Queen of Hearts') and 380 ms ('St Ives') late with respect to the current vowel onset. Alternatively, the taps could have actually been anticipatory responses to the next vowel onset. In this case, participants would have been around 300 ms and 120 ms early for 'Queen of Hearts' and 'St Ives' respectively (assuming that the next vowel onset was exactly 500ms away). In reality, the measured iambic responses probably comprised a mixture of these two types of responses (delayed reactions and anticipations).

Figure 8.3. Individual scatterplot (left) and group means (right) of the timing of taps (in ms) with respect to the stressed vowel onset. Controls are shown in blue and dyslexics in red. Errorbars show the standard deviation of the tap timings.



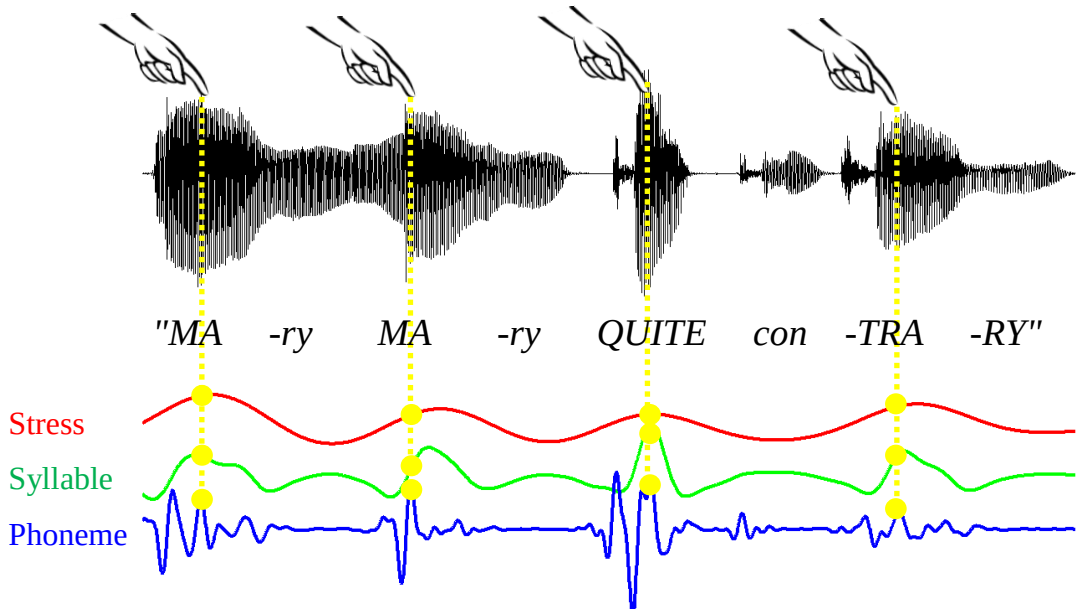
Since it was expected that dyslexics should be poorer than controls in detecting and tapping along to the beat of a sentence, group differences in tapping were examined. A repeated measures ANOVA was conducted taking nursery rhyme as the within-subjects factor and group as the between-subjects factor. As expected, there was a large main effect of nursery rhyme ($F(3,123) = 60.8$, $p<.0001$), with participants tapping closer to the vowel onsets for trochaic than for iambic rhymes. However, there was no main effect of group, and no significant interaction between rhyme and group. This result appeared to indicate that dyslexics were just as good as controls in entraining to vowel onsets in trochaic sentences, but equivalently, they were just as poor at entraining to vowel onsets in iambic sentences.

Therefore, contrary to prediction, the traditional analyses of tapping behaviour (tapping intervals and tapping distance from p-centre markers) appeared to indicate that there were *no* differences between controls and dyslexics in rhythmic entrainment to metrical speech.

b. Tapping with Respect to AM Phase

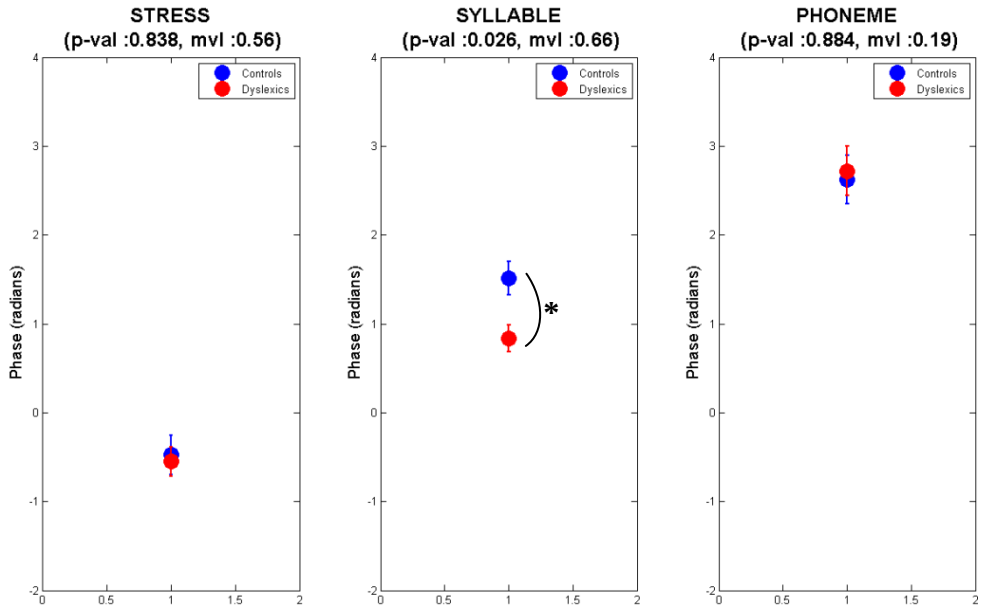
However, these traditional analyses only measured tapping behaviour according to one major rhythmic timescale (or tactus) in the speech stimulus - the stress beat rate. It is possible that listeners can entrain to *more than one* rhythmic timescale at the same time, since taps can also be produced at a faster syllable rate. Therefore, tapping behaviour should be measured according to *all* possible tactus levels in the speech stimulus. To do this, participants' taps were evaluated at Stress, Syllable and Phoneme tactus levels, using the *phase* of each AM tier as the dependent measure. The phase of each AM was used as the measure because phase values vary within a limited and well-defined range (-π radians to π radians), unlike raw amplitude values. Therefore, for each tap, the concurrent phase of the Stress AM, Syllable AM and Phoneme AM at the point of the tap was recorded for each of the five spectral bands in the speech stimulus (Figure 8.4 illustrates this process for one spectral band).

Figure 8.4. Example of possible tap sequence for the sentence "Mary Mary quite contrary". Hand icons represent the occurrence of the 4 taps. Tap timings are analysed according to the phase at the point of occurrence (yellow dot) for each AM tier.



The circular mean phase was then taken across the five spectral bands for each AM tier. As before, only the final two taps for the last two sentence repetitions were included in the final analysis. The average of these four phase values was taken for each nursery rhyme sentence, and a grand average across the four nursery rhyme sentences[33] was obtained for each participant and AM tier. The grand average phase values for each AM tier are shown in Figure 8.5, broken down by participant group.

Figure 8.5. Grand mean tapping phase for each group and AM tier. Controls are shown in blue and dyslexics in red. Coloured dots show the group means and errorbars indicate the group circular standard error of the mean.



First, the phase dispersion of the two groups was checked to ensure that there was a sufficient concentration in phase values within each group to validate a test of group differences. This phase dispersion was expressed as a vector length between 0 and 1, and the minimum acceptable length for a valid test was 0.45. The mean vector lengths were large for Stress and Syllable tiers (0.56 and 0.66 respectively, averaged across both groups), but not for the Phoneme tier (0.19). Therefore, participants' taps were well concentrated around a particular phase value for Stress and Syllable tiers, but were randomly distributed with respect to Phoneme phase. This suggests that participants were entraining their taps to Stress

[33] A grand average across the 4 nursery rhymes was taken in order to reduce variability and increase the phase concentration for each AM tier. If the nursery rhymes were analysed separately, not all would produce sufficiently concentrated phase vectors. This would result in valid tests for some rhymes but not for others, leading to difficulties in interpretation.

and Syllable phase, but not to Phoneme phase. Group differences were then analysed using a circular Watson-Williams test for equal means (the circular equivalent of an ANOVA test). The results of the Watson-Williams test[34] showed a significant difference between groups for Syllable phase ($p < .05$), but not for Stress phase ($p = .84$) (see Figure 8.5). Since phase values were well concentrated for the Syllable tier, the group difference observed here is valid.

Integrating the results from the previous conventional analyses on tapping intervals and timing together with the current findings on tapping phase, these findings suggest that dyslexics are able to entrain reliably to the beat in a speech signal, producing taps at an appropriate interval. However, dyslexics entrain to a different 'temporal anchor point' or p-centre for syllables in the speech signal, entraining their taps to an earlier phase of the Syllable AM than controls.

This difference in syllable beat detection for dyslexics could mean that they have an altered perception of syllable and stress patterns in speech, which could lead to altered phonological representations of these patterns. To investigate this, a circular-linear correlation analysis was conducted to see if individual differences in phase of tapping were related to performance in reading, phonology and psychoacoustic discrimination. Table 8.6 shows these correlations for participants' mean Stress, Syllable and Phoneme phase of tapping, where significant correlations are marked in blue.

Performance in reading and spelling were strongly correlated to participant's *Syllable* phase of tapping. There were significant correlations between Syllable tapping phase and WRAT Spelling ($r = 0.49$, $p < .01$), as well as both TOWRE reading measures ($r = 0.40/0.42$, $p < .05$). In addition, there was also a significant correlation between Stress phase of tapping and WRAT Spelling ($r = 0.41$, $p < .05$). For phonology, the only significant correlation was between Syllable phase of tapping and Spoonerisms ($r = 0.50$, $p < .01$), and there were no significant correlations with psychoacoustic thresholds. It is striking that the vast majority of correlations with reading and phonology occurred for Syllable tapping phase (where there was a significant difference between controls and dyslexics). In contrast, there were no significant correlations with Phoneme tapping phase (which had not shown consistency across individuals).

[34] A non-parametric circular test for equal medians (equivalent to the Kruskal-Wallis test) also showed similar results, with an even larger significance value for the group difference in Syllable phase ($p = .009$).

Therefore, the Syllable phase 'anchor point' used by participants to entrain their tapping was indeed related to reading and phonology, where a later entrained tapping phase was associated with better reading scores. If the Syllable tapping phase is related to the p-centre locations perceived by listeners, then these findings may indicate that dyslexics perceive different (earlier) p-centre loci in speech, as compared to controls. The correlation results further suggest that such syllable timing differences have implications for reading and phonology, even in well-compensated adult dyslexics.

Table 8.6. Correlations between participant's Stress, Syllable and Phoneme phase of tapping with performance in reading, phonology and psychoacoustic measures. Correlations are reported over the full group of participants, df = 41.

	Task	**Modulator Tier (tapping phase)**		
		Stress	*Syllable*	*Phoneme*
Reading & Spelling	WRAT Spelling	***0.41**	****0.49**	0.07
	WRAT Reading	0.19	0.34	0.25
	TOWRE Word Reading	0.20	***0.40**	0.08
	TOWRE Nonword Reading	0.15	***0.42**	0.27
Phonology	Spoonerisms	0.28	****0.50**	0.11
	RAN Sparse	0.27	0.35	0.18
	RAN Dense	0.19	0.36	0.35
Psychoacoustic	Dine Rise Time	0.26	0.26	0.14
	Dino Frequency	0.11	0.18	0.09
	Dino Duration	0.34	0.14	0.22
	Dino Intensity	0.05	0.36	0.34

**p<.05, **p<.01*

8.3.4 EXPERIMENT 3 : RHYTHM PRODUCTION TASK

8.3.4.1 Task Description

In this final task, participants were asked to speak aloud each of the four nursery rhyme sentences in time to a 2 Hz metronome beat, repeating each sentence five times before moving on to the next sentence. For each of the four nursery rhyme sentences, only the last three (out of five) repetitions were used in the analysis. The metronome beat was presented binaurally via headphones, at a sound level that was comfortable for participants. Participants were allowed to practice producing the sentences in time to the beat beforehand, and the recording commenced only after they indicated that they were satisfied that they could produce the sentences successfully. Participants produced the trochaic sentences first ('Mary Mary' and 'Simple Simon') followed by the iambic sentences ('Queen of Hearts', 'St Ives').

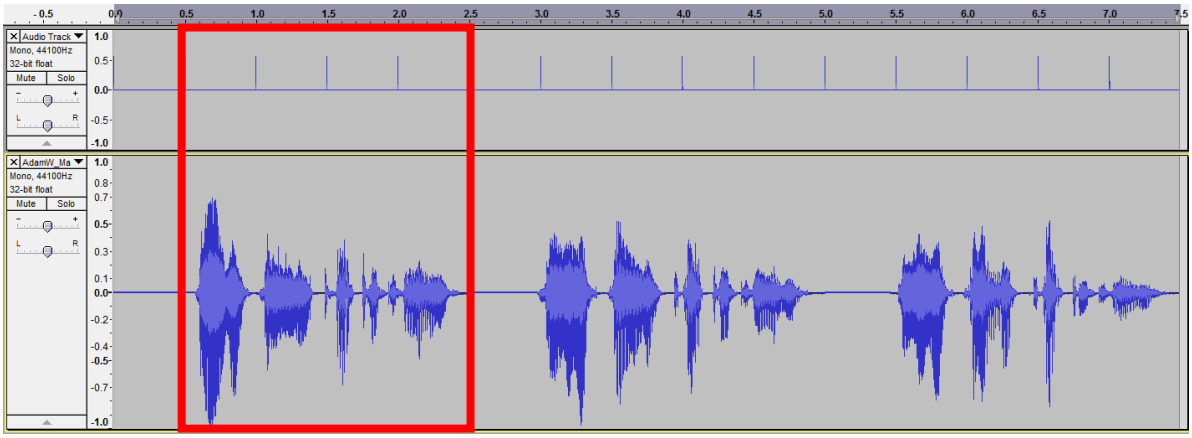
Although this task is described last for the purposes of the flow of the discussion, in the actual study, this rhythm task was always completed first out of the 3 so that the utterances produced by participants would be spontaneous and not affected by the examples that they subsequently heard in the perception and entrainment tasks. To preserve this spontaneity, no explicit instructions were provided to participants as to how many syllables they were supposed to fit into each metronome beat.

8.3.4.2 Results of Rhythm Production Task

For this task, participants were required to repeat the four nursery rhyme sentences in time to a 2 Hz metronome beat. The vast majority of participants spontaneously produced two syllables per beat instead of one syllable per beat, although they were not explicitly instructed to do so. Figure 8.6 shows an example of an utterance produced by a dyslexic participant, where the sentence of 8 syllables was uttered to fit within 4 metronome beats. This suggested that participants preferred to impose a regular *stress* rate on their utterances rather than a regular *syllable* rate, timing every alternate stressed syllable to the beat instead of every syllable. This behaviour is consistent with the proposal that English is a stress-timed language (Abercrombie, 1967; Pike, 1945). However, it is also possible that participants

chose this faster 4 Hz syllable rate of speaking (as compared to a 2 Hz syllable rate) because it was closer to their spontaneous speaking rate.

Figure 8.6. Example of the nursery rhyme sentence "Mary Mary quite contrary" produced by a dyslexic participant, uttered three times. The vertical tick marks in the top part of the figure indicate the pacing metronome beats. The bottom part of the figure shows the waveform of the utterance. Each iteration of the sentence (8 syllables) was spoken to fit within 4 metronome beats(red box).



A small number of 2 controls and 2 dyslexics did spontaneously choose to produce 1 syllable per beat instead of 2 syllables per beat, using this slower rate of production across all four sentences. A further 3 controls also used this slower rate of production for either one or two out of the four sentences. All of these more-slowly-produced 'syllable-timed' utterances were excluded from the analysis. It is also worth noting that equal numbers of controls and dyslexics consistently produced these 'syllable-timed' variants across the 4 sentences (although more controls produced this sporadically). Therefore, the rate or timing preference of participants was not a factor that differed significantly between groups in this experiment.

Each of the four nursery rhyme sentences was analysed separately, resulting in between 17-20 controls and 19 dyslexics per nursery rhyme sentence after the 'syllable-timed' utterances were removed. Recall that each nursery rhyme sentence was repeated 5 times, but only the last 3 repetitions were included in the analysis. For each sample, the timings of the onsets of the 8 syllable vowel nuclei were manually determined. From these vowel onset timings, vowel onset-to-vowel onset intervals were computed by subtracting the time of the current vowel onset from the time of the next vowel onset, resulting in 7 vowel-to-vowel intervals. These 7 vowel-to-vowel intervals were then averaged (across the 7 intervals and 3

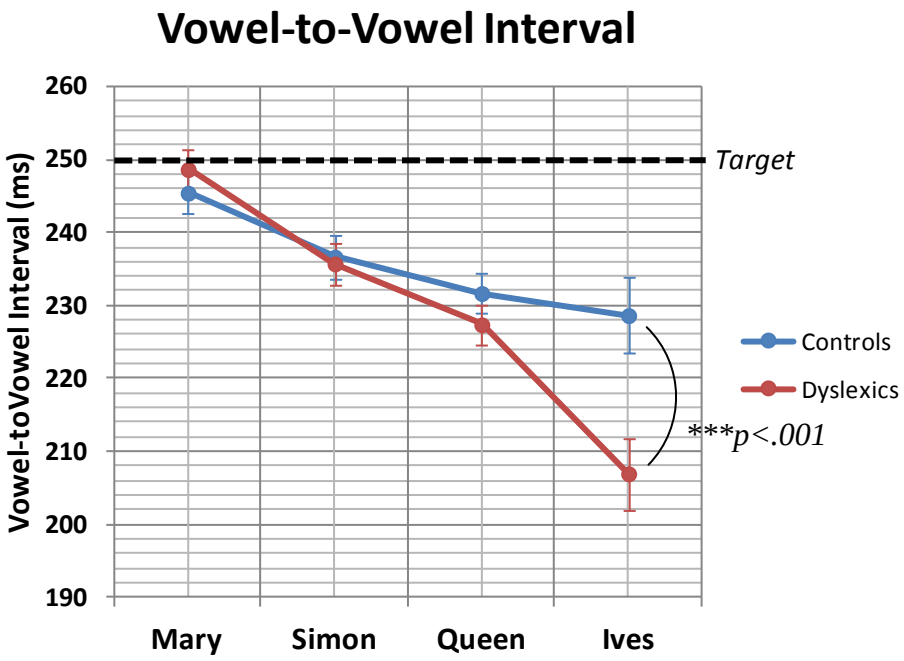
sentence repetitions) to produce a mean vowel-to-vowel interval for each participant and nursery rhyme sentence. The resulting vowel-to-vowel interval was analogous to the inter-tap interval computed for the Tapping data in Experiment 2.

The time difference between vowel onsets and metronome beats was not used as a measure because in some case (e.g. for iambic rhymes discussed in the next section), the pace of participants' utterances was quite far off from the metronome beat, leading to ambiguities in identifying which vowel onset corresponded to which metronome beat. Also, if participants were producing syllables with regular intervals, but at a different rate from the metronome beat, the time difference between the vowel onsets and the metronome beat would increase more and more as the utterance progressed. If these differences were measured, the result would lead one to believe that the utterance was not rhythmically-regular when in fact the utterance was regular, but at a different rate from the metronome. Therefore in this analysis, syllable vowel rate is measured (i.e. vowel-to-vowel interval) instead of the absolute vowel-metronome time difference.

a. Vowel-to-Vowel Interval

The vowel-to-vowel interval is a proxy indicator of syllable length and therefore syllable rate. Since the beat interval of the metronome was 500 ms (2 Hz), and participants uttered 2 syllables per beat (4 Hz), the ideal vowel-to-vowel interval should have been 250 ms. As shown in Figure 8.7, both control and dyslexic participants were close to this ideal syllable length for the trochaic rhyme 'Mary Mary'. However, for the iambic rhymes 'Queen of Hearts' and 'St Ives', syllable lengths grew shorter for both groups. The dyslexics, in particular, shortened their syllable lengths for 'St Ives' drastically to under 210 ms. To analyse these results, a repeated measures ANOVA was conducted with Nursery Rhyme as the within-subjects factor, and Group as the between-subjects factor. There was a significant main effect of Nursery Rhyme ($F(3,102) = 39.2$, $p<.0001$) with vowel intervals getting shorter from 'Mary Mary' to 'Simple Simon' to 'Queen of Hearts' to 'St Ives', but there was no main effect of Group ($F(1,34)=2.62$, $p=.15$). However, there was a large significant interaction between Nursery Rhyme and Group ($F(3,102) = 7.84$, $p<.0001$) indicating that groups differed in their pattern of performance across the 4 nursery rhymes. To investigate this interaction further, a Tukey HSD post hoc test was conducted. Results of the post-hoc test revealed significant differences between groups only for the iambic nursery rhyme 'St Ives' ($p <.001$). This difference is marked on the graph in Figure 8.7.

Figure 8.7. Mean vowel-to-vowel interval (in ms) for each nursery rhyme and group. The ideal target interval was 250 ms, this is marked on the graph with a dotted line. Controls are shown in blue and dyslexics in red. Errorbars indicate standard error.



The nursery rhyme 'St Ives' was metrically more complex than the other three nursery rhymes because many participants were unsure of how to assign stress on the first three syllables *"As I was..."*. According to the original nursery rhyme, these should have been spoken with a 'w-S-w' pattern. However, a significant number of participants in both control and dyslexic groups were unfamiliar with the original rhyme, and chose to produce a 'S-w-w' pattern instead for these first three syllables (e.g. *"AS i was..."*). Despite this difficulty in metrical patterning, controls still produced the syllables in 'St Ives' with an interval that was close (~20 ms) to the target interval of 250 ms (i.e. an error of 8.5%). However, dyslexics appeared to 'lose track' of the metronome beat for 'St Ives', producing syllables that were more than 40 ms shorter than the target interval (i.e. an error of 17.2%). This trend of dyslexics producing syllables that were too short was also present for the nursery rhyme 'The Queen of Hearts', but the group difference here did not reach statistical significance.

It is interesting to note that for both control and dyslexic groups, there was an overall decrease in vowel-to-vowel interval from trochaic and iambic rhymes (i.e. the significant main effect of Nursery Rhyme). It is possible that this overall increase in speaking rate reflected a change in strategy by participants. For example, for the simpler trochaic rhymes,

participants could have been trying to entrain every stressed syllable to the external beat (i.e. every two syllables). As the metrical complexity of the rhymes increased, taking up more mental resources, participants could have switched to a simpler strategy of only aiming to produce every four syllables in time to the beat. For example, for the rhyme 'St Ives', many participants incorrectly attempted to impose a Strong starting syllable on the sentence, producing the pattern *"AS i was going TO st ives"* (stressed syllables in CAPS). This incorrect pattern contained only two stressed syllables instead of four. In this case, participants could have been focused on synchronising just these two stressed syllables to the beat, while disregarding the timing of the intervening unstressed syllables. These unstressed syllables would then naturally contract to a shorter un-paced length, producing the overall decrease in vowel-to-vowel interval observed in these results.

However, while both groups showed a decrease in vowel-to-vowel interval across the 4 rhymes, the specific pattern of performance across the rhymes differed between the groups (i.e. the Rhyme x Group interaction). While dyslexics performed more poorly on the iambic rhymes, they performed as well (or possibly even better) for the trochaic rhymes. For example, for the sentence 'Mary Mary', the dyslexic group was even closer to the 250 ms target than the control group (although the difference of ~3 ms was not statistically significant). Hence the deficit for adult dyslexics was not a simple case of being unable to entrain to a beat *per se*. Rather, dyslexics appeared to struggle specifically with producing iambic-patterned rhymes that had a more complex metrical structure (or were metrically ambiguous), showing poorer control over syllable timing in this situation. As the participants in this study were highly-compensated adult dyslexics, it is possible that younger children could show a stronger production deficit, even with simple trochaic patterns.

Problems with speech rhythm production could indicate that there are problems with motor co-ordination (e.g. between the speech articulators) in dyslexic adults. Unfortunately, in this study there were no other measures of motor co-ordination to test this hypothesis. However, problems with producing rhythmic speech could also arise from a poor *phonological* representation of the rhythmic p-centres of syllables, making it difficult to synchronise the production of these syllables to an external beat. In this case, one's success in rhythmic speech production would be related to the quality of one's phonological representations. To test this, individual differences in vowel-to-vowel intervals were correlated with participants' reading, phonological and psychoacoustic scores. The results of

the correlation are shown in Table 8.7. Significant correlations are highlighted in blue and bold.

Table 8.7. Correlations between spoken vowel-to-vowel onset intervals, and performance in reading, phonology and psychoacoustic measures. Correlations are reported over the full group of participants, df = 34.

	Task	**Nursery Rhyme Sentence**			
		Mary	*Simon*	*Queen*	*Ives*
Reading & Spelling	WRAT Spelling	-0.23	0.01	0.25	0.15
	WRAT Reading	-0.14	-0.01	0.26	0.12
	TOWRE Word Reading	-0.25	-0.07	0.28	***0.37**
	TOWRE Nonword Reading	-0.22	-0.05	0.24	***0.36**
Phonology	Spoonerisms	-0.10	0.10	0.29	****0.51**
	RAN Sparse	0.23	0.22	0.09	-0.04
	RAN Dense	0.09	0.06	-0.07	-0.13
Psychoacoustic	Dino Rise Time	-0.18	-0.31	-0.16	-0.04
	Dino Frequency	***-0.39**	****-0.44**	*****-0.59**	***-0.33**
	Dino Duration	0.02	0.00	-0.31	-0.13
	Dino Intensity	-0.09	-0.04	-0.19	-0.17

**p<.05, **p<.01, ***p<.001*

Individual differences in vowel-to-vowel timing for the nursery rhyme 'St Ives' were significantly correlated to speeded TOWRE word and non-word reading, as well as to Spoonerisms scores. Therefore, performance in rhythmic speech production was indeed related to the quality of phonological representations (i.e. Spoonerisms task), and therefore to reading. As 'St Ives' was the only nursery rhyme where controls and dyslexics differed significantly in vowel interval, it was not surprising that performance on the other rhymes was uncorrelated to reading and phonology. Also, speaking performance for 'St Ives' was significantly correlated to *timed* measures of reading (TOWRE), but not to untimed measures (WRAT). In the TOWRE, participants read aloud as many words as they could within a time limit of 45s. Typically, participants read at a constant rhythmic rate, and their reading fluency at this rhythmic rate determined their final score. In the experimental rhythmic speaking task, dyslexics were less successful than non-dyslexics in controlling their syllable timing for the nursery rhyme 'St Ives'. This inaccuracy in syllable timing could possibly also underlie

dyslexics' lower reading fluency in the TOWRE tasks, explaining the significant correlation between these tasks.

Finally, across all four nursery rhymes, there was a strong correlation between rhythmic speaking performance, and participants' psychoacoustic thresholds for *frequency* discrimination. No other acoustic dimension was related to rhythmic speaking performance (as measured by vowel-to-vowel intervals). This finding is surprising since the vowel interval is a temporal measure, and one would expect a relationship to timing-related dimensions such as duration or rise time. Nonetheless, the results were unequivocal in indicating a strong relationship to pitch (frequency) discrimination rather than to temporal discrimination. Moreover, this relationship was seen for all 4 nursery rhymes, not just the nursery rhyme that had shown a group difference ('St Ives'), indicating that the pitch relationship was fundamental to the task itself, irrespective of the stimulus.

Why would participants' ability to discriminate pitch be related to how well they can speak in time to a beat? A possible explanation is that these factors are linked by musical ability or experience. That is, speaking in time here could have involved similar cognitive mechanisms as singing or playing an instrument in time - with the difference that musical performance also involves a control of pitch, not just rhythm. According to this view, participants who were more experienced with pitch and rhythm control in music should also perform better on this speaking task. To test this hypothesis, speaking performance was correlated with participants' music experience, as well as with their pitch discrimination ability. For this analysis, participants were assigned scores depending on their number of years of music experience. Participants who had no music experience were given a score of 1, participants with up to 5 years of music experience were given a score of 2, and participants with more than 5 years of music experience were given a score of 3. As predicted, there was a significant relationship ($p<.05$) between music experience and speaking performance for 3 out of the 4 rhymes ('Simple Simon', $r = 0.37$; 'Queen of Hearts', $r = 0.37$; and 'St Ives', $r = 0.39$). There was also a strong negative relationship between Dino Frequency scores and music experience ($r = -0.50$, $p<.01$), indicating that participants with more music experience had lower (better) discrimination thresholds for pitch. Therefore, the pitch correlation with rhythmic speaking performance could have been mediated by music experience. Fortunately, control and dyslexic groups did not differ on music experience, therefore the group differences observed in this task should not have been due to different exposure to music.

b. S-AMPH Envelope-Based AM Measures

In the previous traditional analysis of vowel-to-vowel onsets, an important difference was revealed between controls and dyslexics in terms of their syllable timing for complex iambic metrical patterns. However, it would also be interesting to analyse the *rhythmic regularity* of participants' utterances, as well as the specific Strong-weak *prosodic patterns* that they produced. This could reveal disorders in timing on different speech levels, and disorders in prosodic patterning. It would also be interesting to analyse the overall modulation statistics present in the envelope of dyslexic and non-dyslexic utterances, which could reveal more long-term, stochastic differences in prosodic organisation. To do this, envelope-based AM measures from the S-AMPH model were applied to the speaking data.

Recall that the final data used for analysis comprised 3 repetitions per nursery rhyme sentence per participant, and slower 'syllable-timed' utterances were excluded. For the AM analysis, 5 x 3 AM hierarchies were derived from the amplitude envelopes of all the speech samples. Three AM-based indices were then computed. First, the autocorrelation function and periodic power for each modulation tier (Stress, Syllable and Phoneme) was computed. Second, the prosodic strength index (PSI) or stress pattern for each sample was computed. Finally, the hierarchical phase relationship between modulation tiers in the AM hierarchy was computed.

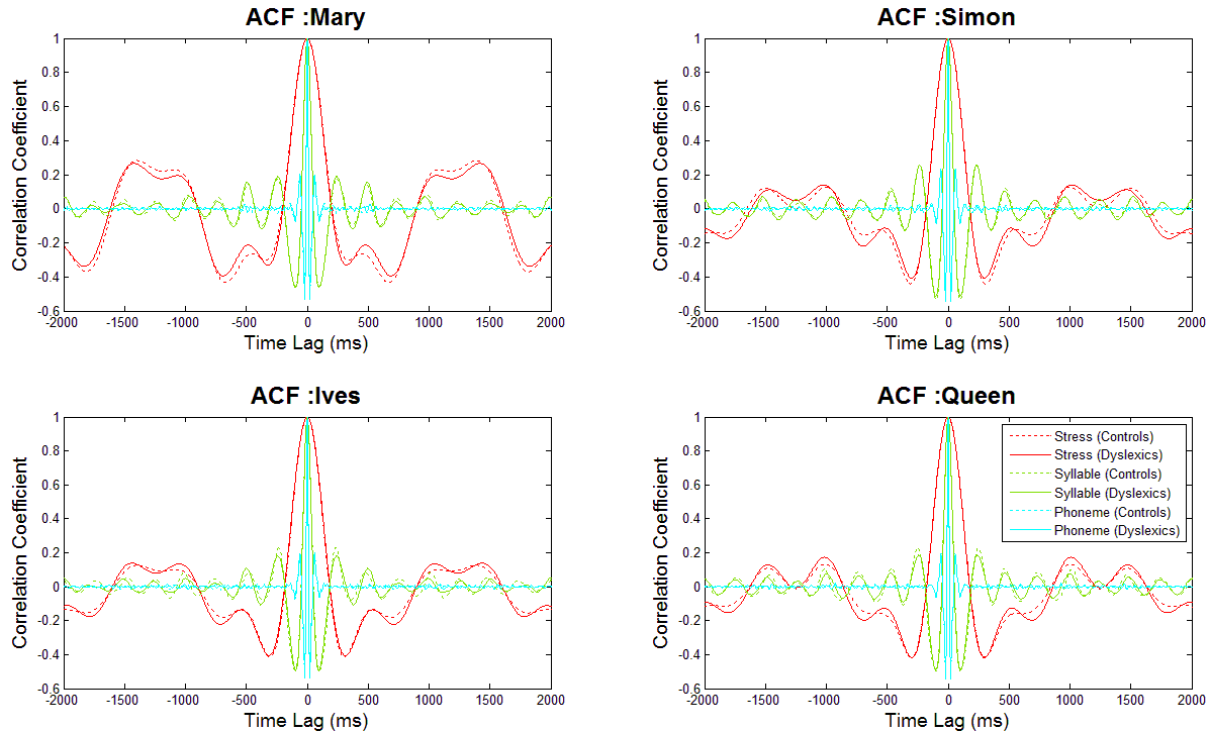
(1) Periodic Power

For each of the 3 S-AMPH modulator tiers (Stress, Syllable, Phoneme), the autocorrelation function (ACF) of that tier was computed. Since the autocorrelation function computes the correlation of the signal with itself at different time lags, it is a measure of periodicity within the signal, and can be used to detect patterns that repeat over time. In the context of speech AMs, the regular stress and syllable patterns of the nursery rhyme sentences should have created regularly repeating modulation patterns at specific time intervals (e.g. syllables every 250 ms). These repeating modulation patterns should be visible as peaks in the ACF at corresponding time lags (e.g. syllable-related peaks in the ACF at +/- 250 ms, and integer multiples of this value). If participants produced these stress and syllable patterns very regularly (i.e. isochronously), then the ACF peaks would be large. Conversely, if participants were more variable in their timing of stress and syllable production, the corresponding ACF peaks would be smaller. Therefore, the ACF indicates both the

prominent *rates* of periodic regularity within the utterance, as well as the *strength* of this regularity.

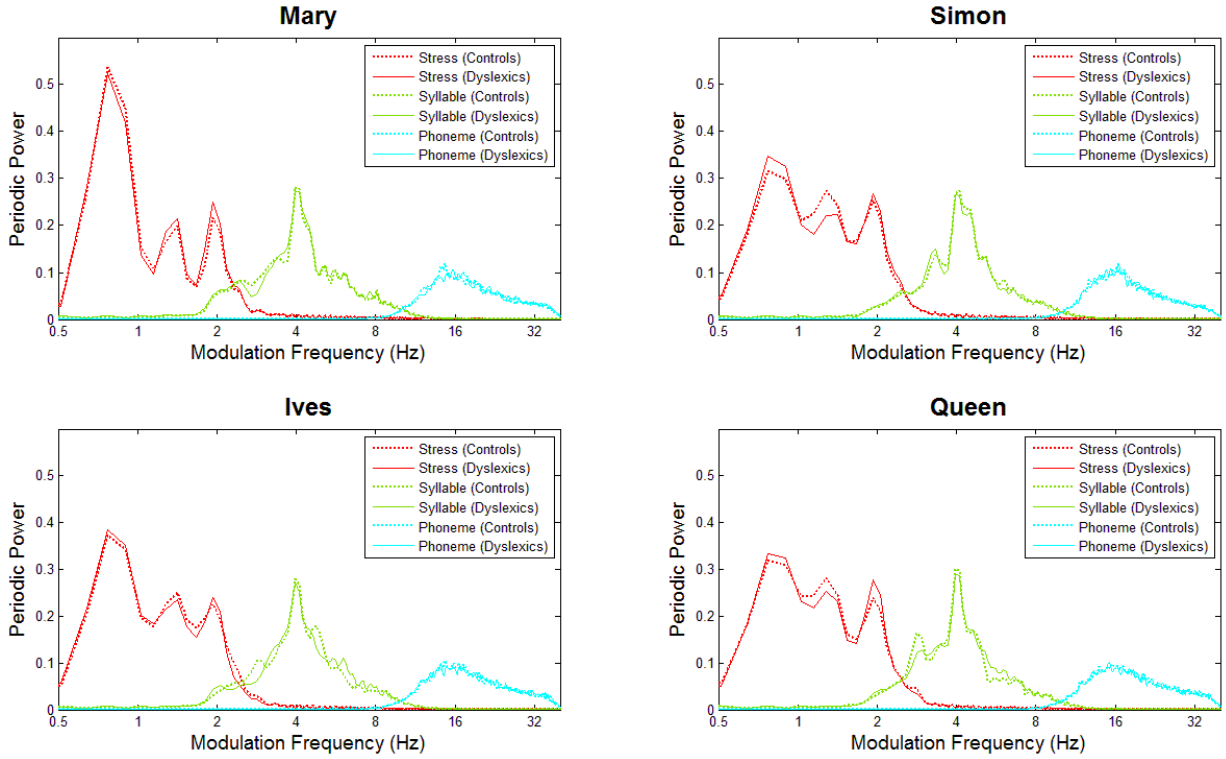
The ACF for each modulation tier and nursery rhyme sentence, averaged over the 5 spectral bands, is shown in Figure 8.8. From visual inspection of Figure 8.8, ACF peaks for the Syllable modulator (green) did indeed occur at +/-250 ms, 500 ms, etc, indicating that controls and dyslexics were indeed producing syllables at a regular rate of ~4 Hz. Similarly, the Stress modulator (red) for all four nursery rhymes showed a small peak in the ACF at ~500 ms (2 Hz) corresponding to the metronome beat, as well as later peaks corresponding to multiples of this metronome rate. For the nursery rhyme 'Mary Mary' (top left subplot), there were also particularly strong peaks in the Stress ACF at +/- 1000 ms and +/- 1500 ms, indicating additional stress patterns occurring every 4 syllables and 6 syllables respectively. These additional stress accents could correspond to patterns such as *"**MA**-ry ma-ry **QUITE** con-tra-ry"* or *"ma-ry ma-ry quite con-**TRA**-ry"* that were produced by different participants.

Figure 8.8. Control and dyslexic mean ACFs for each nursery rhyme and modulation tier. ACFs are averaged over the 5 Spectral bands. The Stress modulator is shown in red, the Syllable modulator in green and the Phoneme modulator in cyan blue. Controls are plotted in a dotted line and dyslexics in a solid line.



To more accurately quantify the amount of periodic regularity in these rhythmic patterns, the Fourier transform of the ACFs was taken. The Fourier transform computes the periodic 'power' at each time lag, for all the possible time lags or modulation rates, effectively computing the power spectral density of the modulator[35]. Figure 8.9 shows the computed periodic power of each modulator for each nursery rhyme sentence (again averaged over the 5 spectral bands).

Figure 8.9. Periodic power in the autocorrelation function for each modulator tier and nursery rhyme. The average power across the 5 spectral bands was taken. Modulation frequency is plotted logarithmically on the x-axis. The 3 modulators are plotted in different colours, dyslexics are shown in the solid line and controls in the dotted line.



For the Stress modulator, all participants produced a peak in modulation power at 2 Hz (corresponding to the metronome beat rate). However, there were also two other prominent peaks - a larger peak at ~0.75 Hz and a smaller peak at ~1.5 Hz (corresponding to lags of 1333 ms and 666 ms). The larger, slower peak at ~0.75 Hz is likely to have arisen from the additional stress patterns in the ACF that occurred every 1000 ms-1500 ms. For the

[35] By the Wiener-Khinchin theorem, the Fourier transform of the autocorrelation function is equivalent to the power spectral density of the signal.

smaller power peak at ~1.5 Hz, since there was no corresponding peak in the ACF at ~666 ms, this suggests that the 1.5 Hz power peak was simply a harmonic of the larger 0.75 Hz peak and should be ignored. Therefore, the power spectrum of the Stress modulator suggests that two key stress rates were present in participants' utterances - the 2 Hz metronome rate, and a slower rate of ~0.75 Hz corresponding to additional stress every 4 or 6 syllables. For the Syllable modulator (green), there was a clear peak in periodic power across the four nursery rhymes at 4 Hz, which was consistent with participants uttering two syllables per 2 Hz metronome beat. For the Phoneme modulator, the peak in periodic power occurred around 16 Hz, an integer multiple of the peak Syllable periodic rate of 4 Hz. This suggested that approximately 4 Phoneme modulator peaks occurred for every Syllable peak in the envelope.

From visual inspection, periodic power at low modulation frequencies ~0.5 Hz and 1.5 Hz in 'Simple Simon' and 'Queen of Hearts' appeared to show a small difference between controls and dyslexics. However, these differences were not statistically-significant in an independent samples t-test. Therefore, for all 3 modulator tiers and all 4 nursery rhyme sentences, the envelope periodic power spectrum of both groups were similar.

This result was surprising given that in the manual analysis of vowel onsets, dyslexics had shown a clear difference in vowel-to-vowel onset timing for iambic rhymes, which should have been reflected in this analysis as dyslexics having a faster peak Syllable periodic rate than controls. However, since the periodic power of the modulators was calculated over the entire spoken sentence, it is possible that small specific differences (e.g. in vowel onset timing) could have been missed. Also, the ACF periodicity measure here is not sensitive to the exact *pattern* (e.g. phase patterns) of the modulator, only to whether that pattern repeats itself over time. Therefore, a sine wave of a given amplitude will have the same periodic power even if its pattern is shifted in phase by 1 pi radians, inverting all peaks to trough and vice versa. Therefore, in order to examine the specific prosodic stress *patterns* produced by controls and dyslexic, the prosodic strength index of each utterance was computed.

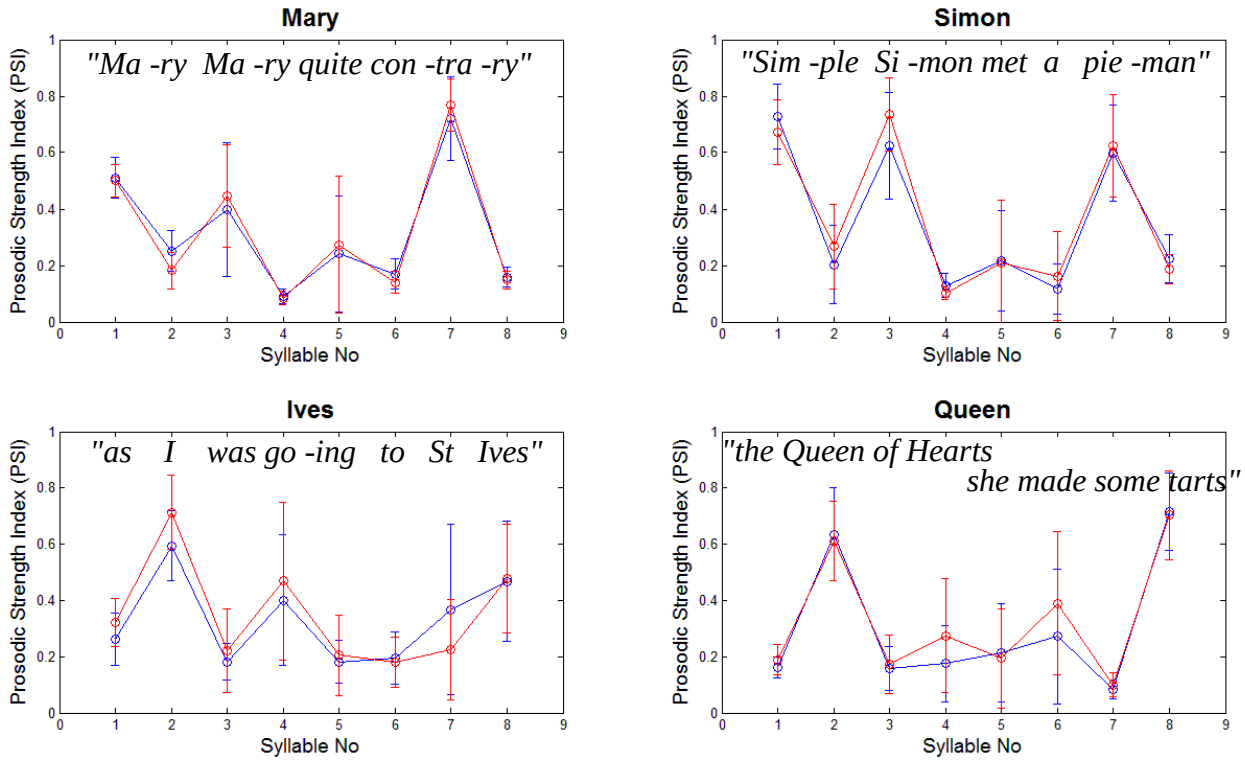
(2) Prosodic Strength Index (Stress Pattern)

Recall from Chapter 5, Section 5.3.2 that the S-AMPH Prosodic Strength Index (PSI) converts the phase relationship between the Stress modulator and individual Syllable peaks into a measure of syllable prominence. By plotting the PSIs of consecutive syllables in an utterance, this allows one to quantify and analyse the syllable stress pattern contained within

that utterance. For this analysis, the manually annotated vowel onsets were used as markers for syllables, instead of automatically-detected peaks. The PSI values were computed for each utterance, and averaged over the three repetitions for each sentence.

The group mean PSI value for each of the four nursery rhyme sentences was then computed, as shown in Figure 8.10, where dyslexics are plotted in red and controls in blue. If control and dyslexic participants differed in the Strong-weak prosodic patterns that they produced, this would be evident as a different PSI pattern across the syllables of each sentence. From visual inspection of the figure, the PSI pattern followed the expected trochaic or iambic stress pattern of each sentence. Moreover, both groups produced similar PSI patterns for each nursery rhyme. For example, both 'Mary Mary' and 'Simple Simon' (top row in the figure) showed a Strong-weak alternation pattern in PSI values, while 'St Ives' and 'Queen of Hearts' (bottom row) showed the opposite weak-Strong alternation pattern.

Figure 8.10. Group mean PSI values for the four nursery rhyme sentences, averaged over the 3 repetitions for each sentence. High PSI values indicate 'Strong' stressed syllables while low PSI values indicate 'weak' unstressed syllables. Controls are shown in blue and dyslexics in red. Error bars indicate the standard deviation.



The PSI scores were entered into a repeated measures ANOVA taking nursery rhyme (4) and syllable number (8) as within-subject factors, and group (2) as the between-subjects factor. The results of the ANOVA revealed no significant main effect for Group ($F(1,34)=2.40$, $p=.13$), although dyslexics had slightly higher PSI values overall (0.34 for dyslexics vs 0.33 for controls). There were also no significant interactions between Group and nursery rhyme, or Group and syllable number. Although the three-way interaction between Group, nursery rhyme and syllable number approached significance ($F(21, 714)=1.52$, $p=.065$), post-hoc Tukey analysis did not indicate any significant difference between controls and dyslexics for individual syllables within the rhymes. Therefore, the controls and dyslexics appeared to have produced statistically-equivalent Strong-weak prosodic patterns in their utterances. So far, therefore, the envelope-based AM measures indicated *no* difference between controls and dyslexics in terms of the overall periodic regularity of their utterances, as well as the specific prosodic patterns they produced.

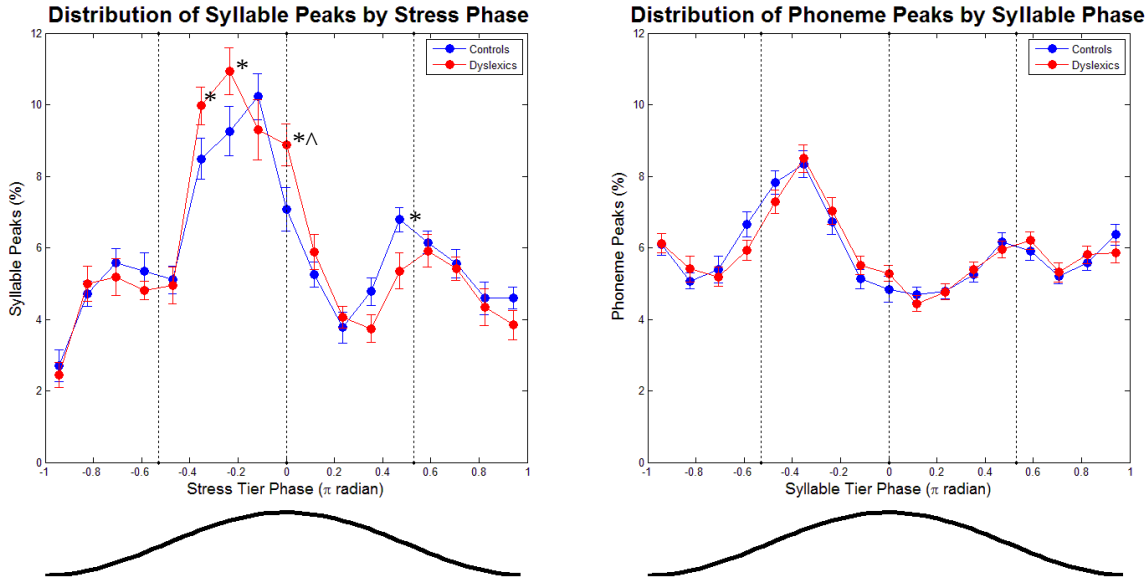
(3) Peak-Phase Distribution within the AM Hierarchy

In a final step, the peak-phase distribution within the modulation hierarchy was computed to examine the overall statistical relationship *between tiers* in the modulation hierarchy. Assuming that peaks corresponded to significant events (e.g. vowel nuclei or bursts of fricative consonant energy), this peak-phase distribution pattern would allow one to see how the timing of key speech events at one timescale was locally organised with respect to speech events at another timescale. For example, within a given stress foot, how early or late did participants tend to produce stressed or unstressed vowel nuclei? Or, within a given syllable, how early or late were the onsets and codas? This hierarchical analysis represents a marked shift away from isochrony (the focus of the earlier autocorrelation analysis in Section 8.3.4.2 b(1)). Instead the focus is on local events occurring within the relative timescale of other events. In this analysis, 'phase' represents this relative timescale (e.g. 1 oscillatory cycle represents 1 stress foot), and 'peaks' represent the occurrence of local speech events (e.g. vowel nuclei).

For this analysis, peaks were detected automatically for both Syllable and Phoneme tiers (since only the syllable vowel onsets had been annotated previously and not individual phonemes). The concurrent phase of the next slowest tier for each peak was then recorded (i.e. Stress phase for Syllable peaks, and Syllable phase for Phoneme peaks). These phase values were then binned into 17 equally-spaced bins between $-\pi$ and π radians. Since the

distribution patterns for each rhyme were quite variable, the four rhymes were averaged into a grand mean distribution, as shown in Figure 8.11. Appendix 8.2 shows the individual peak-phase distributions for each of the 4 nursery rhymes. From visual inspection of Figure 8.11, the distribution pattern for Syllable peaks with respect to Stress phase (left subplot) appears to differ between controls and dyslexics, whereas the distribution pattern of Phoneme peaks with respect to Syllable phase (right subplot) appears similar.

Figure 8.11. (Left) Grand mean distribution of Syllable tier peaks with respect to Stress phase. (Right) Grand mean distribution of Phoneme tier peaks with respect to Syllable phase. Controls are shown in blue, dyslexics in red. The distribution was calculated over 17 phase bins for each Spectral band, and the average over all 5 Spectral bands is shown. Errorbars show standard error, (*) indicates a significant difference between groups in the Fisher LSD post-hoc test, (^) indicates a significant difference in the Newman-Keuls post-hoc test.



To test for group differences, two repeated measures ANOVAs were conducted with Syllable peak distribution and Phoneme peak distribution as the dependent variables. For both analyses, Phase bin was the within-subjects factor, and Group was the between-subjects factor. For the Syllable peak distribution analysis, there was a significant effect of Phase ($F(16,592)=38.08$, $p<0.0001$), but no significant main effect of Group ($F(1,37)=0.95$, $p=.34$). However, there was a significant interaction between Group and Phase ($F(16,592)=1.8243$, $p=.025$), indicating that controls and dyslexics showed a different Syllable peak-Stress phase

distribution *pattern*. To analyse these differences further, Fisher LSD and Newman-Keuls post-hoc tests were conducted. Both post-hoc tests indicated that dyslexics had a significantly higher percentage of peaks at 0π radian Stress phase (marked in Figure 8.11 with [*^]). In addition, the Fisher LSD post-hoc test also indicated significant group differences around -0.3π radians and 0.5π radians (marked in Figure 8.11 with [*]). Dyslexics had a higher percentage of peaks at -0.3π radians, and a lower percentage of peaks at 0.5π radians.

These distribution differences indicate that dyslexics tended to produce more Syllable peaks around the peak of the Stress modulator. Conversely, dyslexics tended to produce less Syllable peaks on the downward slope of the Stress modulator. According to the AMPH and S-AMPH models, syllable peaks that occur near the peak of the Stress modulator correspond to stressed syllables. Therefore, dyslexics appear to be producing a higher proportion of stressed syllables as compared to controls. This observation is consistent with the higher PSI values for dyslexics seen in the previous section (although the PSI difference was not statistically significant). Secondly, the highest percentage of Syllable peaks for dyslexics occurred at -0.24π radians, which was *earlier* than controls, who showed the most Syllable peaks at -0.12π radians. Therefore, not only were dyslexic Syllable peaks more concentrated around the peak of the Stress modulator, but these 'Strong' syllable peaks also tended to occur slightly earlier in Stress phase than the Syllable peaks of controls. If Syllable peaks are taken to indicate vowels, then this difference in distribution pattern could indicate that dyslexics tend to produce stressed syllable vowels slightly *earlier* within the prosodic stress foot - a difference in syllable timing and organisation.

In contrast to these group differences for the Syllable peak distribution, there were no significant Group effects or interactions for the Phoneme peak distribution. Therefore, any differences in speech timing or hierarchical organisation between controls and dyslexics appeared to be specific to the slower Stress- and Syllable AM rates. The results of this peak-phase analysis illustrate that while dyslexics utterances may have a similar periodicity profile and prosodic pattern as compared to controls, their local speech events actually have a subtly different temporal organisation. Specifically, stressed syllables tend to occur slightly earlier within the prosodic stress foot, and a higher proportion of syllables may be stressed. Consistent with the tapping data, these differences in temporal organisation specifically implicate Syllable timing, rather than Phoneme timing, as disordered in dyslexia.

8.4 CHAPTER SUMMARY & DISCUSSION

Developmental dyslexia is associated with phonological difficulties and also with rhythmic difficulties in speech and music tasks. In speech, rhythm-bearing syllable and prosodic stress patterns are associated with slow amplitude modulations (AM) in the speech envelope. Consequently, dyslexics' rhythm deficits may be associated with impaired perception and production of these slow AMs in the speech envelope. Here, dyslexic rhythm perception and production were investigated in 3 AM-based rhythm experiments. Across all three experiments, systematic group differences in Syllable-related timing and prosodic organisation were observed when AM-based measures were applied. Individual differences in Syllable-timing were also the most strongly related to performance in phonological and reading measures. By contrast, Phoneme-related timing appeared to be the same in both groups, and was not related to differences in phonology or reading.

In the first rhythm perception experiment, dyslexic and control participants identified nursery rhyme sentences that had been tone-vocoded using different AM tiers and tier combinations. These tiers were derived from the original 5-tier AM hierarchy in the AMPH model. In the perceptual experiment described in this chapter, participants were presented with sentences that had been vocoded using 29 ERB_N-spaced spectral channels, yielding highly intelligible speech. For the 29-channel stimuli, the performance of dyslexics lagged behind that of controls for the Stress+Syllable and Syllable+Subbeat AM combinations. Although not statistically-significant at the $p<.003$ (Bonferroni corrected) level, these differences were the closest to statistical significance ($p<.05$). Since the result was not significant, firm conclusions cannot be drawn. However, it is interesting to note that dyslexics performed on par with controls when presented only with Stress AM, Syllable AM or Subbeat AMs on their own. A small gap in perfomance only emerged when two AM rates were presented in combination (either Stress+Syllable or Syllable+Subbeat), suggesting that dyslexics may not benefit as much as controls when information at the Syllable rate is combined with information at another AM rate. By contrast, there were no significant group differences when the sentences were vocoded with a 1-channel vocoder, again supporting the view that dyslexic deficits in AM perception only emerged when *multiple* streams of spectral and temporal information had to be combined.

In the second and third rhythm experiments, participants had to tap along to and produce metronome-timed speech respectively. For both experiments, a traditional analysis

was applied followed by an AM-based analysis. For the AM-based analysis, the S-AMPH model was used and speech AMs were divided accordingly into the three key modulation rates : 'Stress' (~2 Hz), 'Syllable' (~4 Hz) and 'Phoneme' (~20 Hz) rates. For both experiments, differences emerged between control and dyslexic participants, with dyslexic participants showing highly specific differences in syllable timing and organisation when AM-based analyses were applied.

In the tapping experiment, participants tapped along to the beat of the four metrically-regular (trochaic or iambic) nursery rhyme sentences. The sentences had a 2 Hz (500 ms) stress rate where every other syllable was stressed. Traditional analysis revealed that the mean tap interval for dyslexics was 512 ms, which was close to the 'ideal' tapping rate, and was not significantly different from controls (520ms). Moreover, the standard deviation of dyslexics' tap intervals was even *lower* than that of the controls (±16 ms vs ±38 ms). There were also no significant difference between controls and dyslexics when the taps were analysed with respect to stressed vowel onsets. These results appeared to indicate that dyslexics had no problems when entraining to rhythmic patterns in speech.

However, when the taps were analysed at *multiple* tactus levels using an AM-based analysis, a specific group difference in Syllable tap timing was revealed. Dyslexics preferentially entrained to an earlier portion of the Syllable AM cycle, 0.7 radians (~1/9 of a cycle) ahead of controls. Therefore, dyslexics were producing taps that were highly regular in interval, but altogether shifted earlier in time, because they were aligned with a different portion of the speech signal. This result suggests that dyslexics perceived an earlier 'p-centre' as compared to controls. This finding is consistent with that reported by Wolff (2002), who found that dyslexic adolescents tended to anticipate the beat more than their non-dyslexic peers when tapping to a metronome. Here, the same effect is demonstrated for rhythmic beats in speech, and the dyslexic 'anticipation' effect is shown to be specific to Syllable-rate modulation patterns in speech.

In the third and final experiment, participants spoke the four nursery rhyme sentences in time to a 2 Hz metronome beat. While both control and dyslexic utterances contained similar metrical patterns, detailed analyses again revealed disruptions in syllable timing for dyslexics. For the more metrically-challenging iambic sentences such as 'St Ives', the syllable vowel-to-vowel interval for dyslexics was significantly shorter than for controls, indicating poorer control of syllable timing. Although the envelope-based AM measures indicated no differences between controls and dyslexics in terms of AM periodicity and overall prosodic

patterning, there was a subtle difference in the way that Syllable peaks were distributed with respect to Stress phase. Across the four nursery rhyme sentences, dyslexics tended to produce stressed syllable vowels slightly *earlier* within the prosodic stress foot, and they also tended to produce stressed syllables more often than controls. Therefore, in both rhythm entrainment and production experiments, even highly-compensated adult dyslexics showed differences in speech syllable timing and prosodic organisation.

Finally, across all three experiments, individual differences in performance on the rhythm-based tasks were related to participants' performance in reading, spelling and phonology. In the vocoder perceptual experiment, performance for Stress+Syllable vocoded AMs was the most strongly-related to reading and phonology. In the tapping experiment, Syllable AM phase of tapping was significantly related to spelling, reading and phonology. Finally, in the production task, syllable vowel-to-vowel intervals for the most challenging nursery rhyme (St Ives) were again strongly related to reading and phonology. Therefore, individual differences in AM-based *Syllable* timing were the most consistently related to reading and phonology. Differences in syllable timing perception could affect the way that dyslexics segment continuous speech into syllables and words, leading to altered or incomplete phonological representations as compared to controls. For example, if dyslexics perceive a syllable to begin earlier, they may also perceive it to end earlier, and therefore fail to encode the coda of the syllable completely. These altered or incomplete phonological representations could then make it difficult for dyslexics to acquire the letter-sound correspondences necessary for learning to read an alphabetic orthography.

Since neuronal oscillations in the theta range are thought to entrain to syllable patterns in speech (Luo & Poeppel, 2007), the dyslexic differences in syllable timing could stem from altered neuronal activity in the theta band. Moreover, since neuronal oscillations in the auditory cortex are hierarchically-nested (Lakatos et al, 2005), altered theta (syllable rate) activity could also be an indirect consequence of atypical delta (stress rate) activity, which could explain why prosodic organisation (e.g. syllable vowel timing *within the stress foot*) also appears to be disrupted in dyslexia.

PART IV CONCLUSIONS & DISCUSSION

In both child-directed speech and dyslexia, envelope-based AM measures provided novel insights into the data. For example, analysis of the Syllable peak-Stress phase distribution patterns revealed that CDS was associated with lower conditional entropy than ADS. Analysis of the periodicity of modulation tiers revealed that surprisingly, even child-directed readings of narrative stories were strongly rhythmic in nature. These differences indicated shifts in the *temporal* organisation and structure of speech in order to accommodate the needs and abilities of the child listener, perhaps to facilitate speech segmentation. In the dyslexia case study, AM-based analysis uncovered subtle differences in the Syllable phase of tapping between dyslexics and controls, as well as differences in their hierarchical organisation of produced Stress and Syllable modulation patterns. These deficits in syllable timing and temporal organisation had been predicted in theory (e.g. Goswami, 2011), but had been difficult to uncover using conventional methods of speech analysis.

A key benefit of using envelope-based measures (e.g. based on the S-AMPH model) for rhythm-based measurement is that they represent a move away from traditional durational measures of isochrony. Instead, the focus is on local relationships between different rates of amplitude modulation (or different beat tactus levels), allowing for more subtle differences to emerge. Moreover, although the S-AMPH model was developed to represent normal speech rhythm perception, its parameters may be modified to produce 'abnormal' speech rhythm perception, such as that seen in dyslexia. Therefore, the S-AMPH model could also potentially be a useful tool in understanding the etiology of speech rhythm disorders.

PART V :

FINAL DISCUSSION & CONCLUSION

Chapter 9 : Final Discussion & Conclusion

CHAPTER OVERVIEW

In this thesis, the over-arching aim was to develop an explanatory account of speech rhythm based on the dynamic amplitude modulation (AM) cues in the speech envelope. Toward this end, two AM Phase Hierarchy (AMPH) models were developed and evaluated. Apart from being useful rhythm-measurement systems, these models also possess *explanatory power* about the causal mechanisms underlying amplitude-based rhythm perception. For example, it was demonstrated in the tone-vocoder experiment in Chapter 3 that human listeners rely on phase relationships between Stress and Syllable rates of amplitude modulation to infer Strong-weak syllable patterning. The AMPH models capture this feature of human rhythm perception and instantiate it in as a formal computational scheme (e.g. the Stress Phase Code/Prosodic Strength Index), so that predictions about perceived rhythm can be generated about any speech sequence. Consequently, when rhythm perception and production go *awry*, as demonstrated in Chapter 8 for adults with developmental dyslexia, these deficits can be understood in terms of mis-specified Stress-Syllable phase relationships. Conversely, when speech rhythmicity is *enhanced*, as demonstrated in Chapter 7 for child-directed speech, this enhancement can also be explained in terms of strengthened Stress-Syllable phase relationships (i.e. tighter hierarchical nesting). Therefore, the two AMPH models are 'models' in the sense that they instantiate the relationship between the amplitude cues in the speech envelope and the speech rhythm patterns perceived by the listener.

At the heart of both AMPH models are two major conceptual innovations - the AM hierarchy and the Stress Phase Code (or Prosodic Strength Index, PSI). The core features of these conceptual innovations are discussed further in Section 9.1 The main contribution of this thesis is that the AMPH models are *methodological innovations* as amplitude-based accounts of speech rhythm. These are complementary to (and not in competition with) previous duration-based accounts of speech rhythm. The findings in this thesis also have wider *theoretical* implications for possible neural mechanisms of speech rhythm perception, phonological development, and mechanisms of speech prediction and attention allocation. These wider theoretical implications are discussed in Section 9.2. Finally, future possible research questions are discussion in Section 9.3.

9 FINAL DISCUSSION & CONCLUSION

9.1 AMPH MODELS AS METHODOLOGICAL INNOVATIONS FOR AMPLITUDE-BASED RHYTHM DETECTION

9.1.1 THE AM HIERARCHY

The AM hierarchy is a novel way of representing modulation patterns in the speech envelope in accordance with the linguistic prosodic hierarchy (Chapter 2, Section 2.1). Ascending tiers in the AM hierarchy capture the modulation patterns generated by linguistic units of increasing grain size, such as phonemes, syllables and stress feet. For example, peaks in the Syllable AM tier commonly correspond to syllable vowel nuclei (Chapter 5, Section 5.2). This hierarchically-nested representation reveals the unique activity at each linguistic level (or tier), as well as the relationships *between* linguistic levels (tiers). For the purposes of rhythm detection, the Stress AM and Syllable AM tiers contain the most relevant prosodic information (as shown in the tone-vocoding experiment in Chapter 3, Section 3.2.1). However, other AM tiers in the hierarchy are also expected to contain important speech information, such as phonetic cues for speech intelligibility (e.g. Rosen, 1992).

The AMPH and S-AMPH models used AM hierarchies of different origin and composition. In the theory-led AMPH model, the 5-tier AM hierarchy was strictly theoretically-defined on the basis of prior literature (Chapter 2, Section 2.4). In the data-led S-AMPH model, the 3-tier AM hierarchy emerged solely from the modulation statistics of the speech envelope, free from theoretical constraints (Chapter 4, Section 4.4). Yet in both cases, the resulting AM hierarchies showed important similarities, suggestive of a convergence between theory and data (modulation statistics). Specifically, the two most important AM tiers for speech rhythm detection - Stress AM and Syllable AM tiers - were conserved across both AM hierarchies, albeit with a difference in the modulation bandwidth for the Syllable tier. This crucial similarity allowed the AMPH and S-AMPH models to operate on essentially the same basis for identifying 'Strong-weak' syllable stress patterns (i.e. the AMPH Stress Phase Code was equivalent to the S-AMPH Prosodic Strength Index).

The idea that the amplitude envelope consists of a nested hierarchy of AMs, each transmitting a different type of speech information is an appealing one. Such an AM hierarchy would have useful properties that could facilitate speech processing, as discussed later in Section 9.3. However, in this thesis, the AM hierarchies are merely convenient ways of representing the modulation information in the speech envelope. It is not claimed that these AM hierarchies are invariant *structural* components of the speech envelope, arising from some inherent physiological constraint or mechanism. Before such a structural claim can be made, several important criteria must be satisfied. First, there must be empirical evidence for *functional separation* of speech modulation information into tiers. Second, there must be empirical evidence for *hierarchical nesting* between these tiers. Third, it must be demonstrated using some objective measure (e.g. based on maximum likelihood or entropy) that the hierarchical representation is the most optimal, stable, or parsimonious representation of modulation information in the envelope. Fourth, and most importantly, a plausible physiological source (or sources) that would be capable of producing such hierarchical patterning must be identified.

In this thesis, there is evidence to satisfy the first and second criteria. Functional separation of AM tiers was demonstrated in the modulation rate PCA analysis (Chapter 4, Section 4.4) and from the results of the tone-vocoder experiment (Chapter 3, Section 3.2.1). Hierarchical nesting was demonstrated in the non-uniform peak-phase distribution patterns of Syllable peaks with respect to Stress phase, and Phoneme peaks with respect to Syllable phase (Chapter 5, Section 5.3, Chapter 7, Section 7.2.4.2). However, more research is required to address the third and fourth criteria. Nonetheless, it should be mentioned that a plausible candidate for generating these hierarchical AM patterns could be the motor articulators, whose actions themselves are strongly co-ordinated in time over different timescales (e.g. Kelso et al, 1986; Saltzman & Byrd, 2000).

9.1.2 The Stress Phase Code (or Prosodic Strength Index)

The Stress Phase Code (or PSI) is a simple computational scheme for deriving 'S-w' prosodic rhythm patterns from the phase relationship between Stress and Syllable AM tiers in the AM hierarchy. This algorithm captures the intuition that syllable prominence is *relative* (e.g. Liberman & Prince, 1977) by making use of Stress phase (a cyclical, relative measure) to code for syllable prominence. Human listeners also rely on this phase relationship when

making rhythm judgments (Chapter 3, Section 3.2.2), suggesting that the Phase Code/PSI is correctly capturing an aspect of human rhythm perception. Furthermore, the Stress-Syllable phase relationship can also be treated as a dependent variable in experimental analysis. For example, dyslexic individuals entrain to a different Syllable phase in speech, and also *produce* speech with a different Stress-Syllable phase-peak distribution (Chapter 8).

As a method for automatic stress transcription, the PSI method has an accuracy of ~90% for metronome-timed speech and ~70% for freely-produced speech (Chapter 6, Section 6.2). This compares favourably with the performance of other methods of automatic stress detection, where an accuracy level of around 65% is achieved when only amplitude cues are used (Silipo & Greenberg, 1999).

Therefore, the Stress Phase Code or PSI operates in a psychologically-valid fashion, and performs adequately for prosodic stress transcription. The first factor (psychological validity) is of little importance for a method whose sole aim is automatic stress transcription, since accurate results may also be obtained by completely 'non-psychological' means (e.g. machine learning). However, in this thesis, the Stress Phase Code and PSI form part of a larger psychological account of amplitude-based rhythm perception. Therefore, the fact that human listeners also rely on the Stress-Syllable phase relationship for determining rhythm patterns is *psychologically* significant. Accordingly, mis-alignments or enhancements in the Stress-Syllable phase relationship, such as those that occur in dyslexia (Chapter 8, Section 8.3.4.2 (b3)) and child-directed speech (Chapter 7, Section 7.2.4.2), can also be inferred to have psychological and functional significance. In this sense, the Stress Phase Code and PSI are methodological innovations for *psychological* inquiry into speech rhythm perception.

9.1.3 POSSIBLE IMPROVEMENTS TO THE AMPH MODELS

The S-AMPH model addressed two of the major short-comings of the original AMPH model. However, further improvements can still be made in these areas :

Methodological Improvements Different methods could be used to extract the AM hierarchy other than those used in this thesis. For example, wavelet analysis, or empirical mode decomposition (Huang et al, 1998), which is especially suited for nonstationary processes, may be useful. The parameters used in the model (e.g. the mathematical function used to compute the PSI, or the PSI threshold) could also be refined, for example by systematically evaluating the outcomes when different parameter values are used (e.g. via a grid search). Different statistics could also be computed from the AM hierarchy, for example the phase-power nesting between different tiers in the hierarchy, or scale-invariant properties across the hierarchy.

Combining Amplitude, Duration and Pitch Cues to Rhythm. The AMPH models are based solely on amplitude changes in the speech envelope, and do not incorporate other cues to speech rhythm - such as duration and pitch. To form a more holistic representation of speech rhythm, these amplitude-based cues could be combined with duration and pitch-based cues in a hybrid model of speech rhythm. For example, coupled oscillator models are similar *in principle* to the AMPH models, but are designed to capture duration or timing changes rather than amplitude changes in speech. A hybrid duration-amplitude model could be envisaged where the hierarchical tiers represent amplitude modulation on different timescales, but the pattern of modulation is modelled on the behaviour of coupled oscillators.

Incorporating Effects of Learning. The AMPH models are based solely on 'bottom-up' acoustic information, and do not incorporate any 'top-down' effects of learning or prior experience. Therefore, these models can only simulate 'naive' rhythm perception. In view of this, it may be interesting to develop a version of the AMPH that is capable of 'learning' speech rhythm patterns. For example, prior knowledge of rhythm patterns could be used to smooth or categorise the input data, so that the rhythm pattern is 'interpreted' through the lens of experience (e.g. Bayesian learning). Such a model could then be used to model developmental changes in rhythm perception as children acquire more experience with the dominant rhythm patterns in their native language.

9.2 WIDER IMPLICATIONS OF THESIS FINDINGS

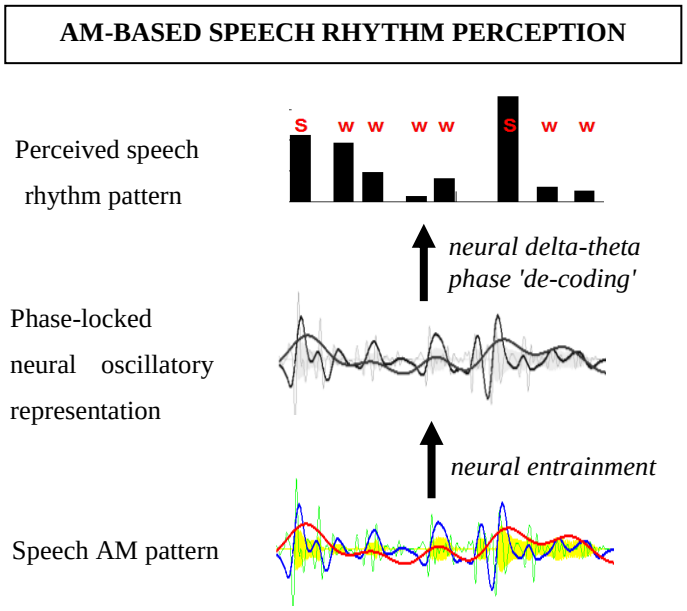
In this section, the wider implications of the findings in this thesis for speech perception are discussed.

9.2.1 A POSSIBLE NEURAL OSCILLATORY REPRESENTATION OF SPEECH RHYTHM

Speech is most commonly described in terms of spectral changes, such as formant patterns and transitions. The AMPH and S-AMPH models proposed in this thesis take a less common approach, describing speech in terms of its amplitude modulation (AM) structure. The AM hierarchies used in these models provide a potential mechanistic link between speech elements of the linguistic prosodic hierarchy (i.e. feet, syllables, phonemes) and neuronal oscillatory architecture (i.e. delta, theta, gamma rates) as noted by multi-time resolution models of speech perception (e.g. Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012). Amplitude modulation activity in speech, particularly in the syllable range, can generate robust neural entrainment with high temporal and spectral fidelity (e.g. Pasley et al, 2012, Luo & Poeppel, 2007, Aiken & Picton, 2008). Theoretically therefore, the AM patterns in the speech envelope could represent linguistic units on the one hand, and drive (entrain) neural activity on the other hand.

For example, the Stress-Syllable phase relationships observed in the speech envelope may also elicit patterns of phase-locked activity between corresponding delta and theta oscillatory rates. This delta-theta phase-locked pattern could then form part of the *neural representation* of speech rhythm, being driven by Stress-Syllable relationships in the speech envelope. This is an exciting prospect, because it suggests that the syllable detection and phase-decoding mechanisms proposed in the AMPH models could also be used by neuronal oscillations in the brain to track syllable patterns (via theta oscillations) and infer rhythm (via theta-delta phase relationships), as illustrated in Figure 9.1. Therefore, the AMPH models represent a 'neuro-plausible' simulation of naive human speech rhythm perception, via the neural mechanism of oscillatory entrainment.

Figure 9.1. Illustration of AM-based speech rhythm perception via neural oscillatory entrainment to speech AM patterns.



9.2.2 Hierarchical Processing of Speech on Multiple Timescales

As discussed in Section 9.1, a major innovation of the AMPH models is that speech is represented as an *hierarchy* of AMs at different rates, where each AM rate transmits a different type of speech information such as prosodic stress, syllable pattern or phonetic contrast. The 3 modulation tiers form a *nested* hierarchy because each tier governs the activity of its 'daughter' (faster) tier. For example, the phase of the Stress AM constrains the peak activity of the Syllable AM, so that the distribution pattern of Syllable peaks with respect to Stress phase is non-uniform, with Syllable peaks occurring more frequently in some Stress phase regions than others (e.g. Chapter 5, Section 5.3). Likewise, at the next level of the hierarchy, the distribution pattern of Phoneme peaks is also constrained by Syllable phase (Chapter 7, Section 7.2.4.2). This hierarchical nesting of speech AMs is reminiscent of the proposed hierarchical nesting of neuronal oscillations in the auditory cortex (Lakatos et al, 2005). In the macaque auditory cortex, the amplitude of theta (syllable-

rate) oscillations is modulated by delta phase, and the amplitude of gamma (phoneme-rate) oscillations is modulated by theta phase[36].

In this thesis, the focus was on the two slower tiers within the AM hierarchy - Stress and Syllable tiers - as carriers of speech rhythm information. However, the faster Phoneme tier also transmits important speech information, such as phonetic cues to manner of articulation, voicing, and vowel identity (Rosen, 1992). As speech unfolds dynamically in real time, these different types of speech information are presented concurrently to the listener. To capture both slow and fast information, speech analysis has been proposed to occur on multiple timescales (e.g. Poeppel, 2003). However, these multiple streams of speech information must eventually be bound together into a single percept. This requires that information from each stream be correctly temporally aligned with other streams. The format of the AM hierarchy supports both these functions - speech sampling on multiple timescales, and temporal alignment/binding.

For example, speech could be sampled on different timescales according to the different tiers in the AM hierarchy (eg. Stress, Syllable, Phoneme), resulting in 3 discrete information streams. The information in each stream could also be re-combined correctly, making using of the hierarchically-nested phase alignment between tiers (e.g. Syllable AM peaks are aligned with 'high peak probability' regions of Stress AM phase). Therefore, a phase-nested AM hierarchy could be a useful way to represent the various types of information in speech. If the brain entrains to this AM hierarchy, it could generate an equivalent *neural* phase-nested hierarchy where each neural oscillatory rate carries speech information on a different timescale. This neural phase-nested hierarchy (e.g. of delta, theta and gamma rates) could likewise allow the brain to encode different types of speech information separately, yet maintain their temporal alignment (e.g. Giraud & Poeppel, 2012). Therefore, the AM hierarchy may support neural encoding of speech information on multiple timescales.

[36] In Lakatos et al's study, a delta phase of 0.6π radians was associated with the highest theta amplitude, while a delta phase of -0.5π radians was associated with the lowest theta amplitude. In view of this, it is interesting to note that in naturally-produced speech relative to metronome-timed speech, the distribution of Syllable peaks shifts *forward* in Stress phase so that most Syllables peaks now occur around -0.2π and 0.5π radians, rather than around 0π and ±π radians (Chapter 5, Section 5.3.1). This forward shift could occur because speakers naturally take advantage of the neural theta enhancement that occurs around delta 0.6π radians in the auditory cortex.

9.2.3 AM Patterns as the Basis for Generating Predictions & Allocating Attention

9.2.3.1 Predicting Prosodic Stress

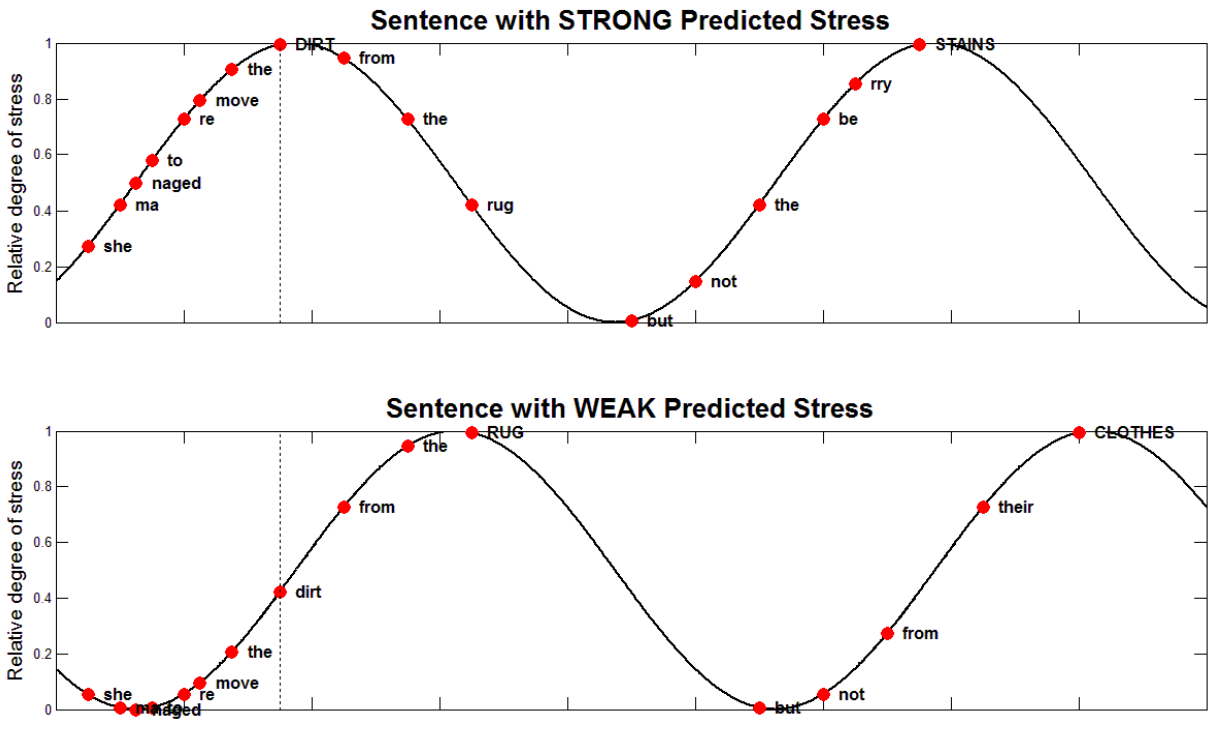
We now turn to discussing a potential role for AM patterns in facilitating 'real-time' speech processing via 'online' stress prediction. In phoneme monitoring tasks, reaction times to phonemes in stressed syllables are faster than reaction times to phonemes in unstressed syllables (Pitt & Samuel, 1990; Cutler, 1976; Cutler & Foss, 1977; Shields et al, 1974). This facilitatory effect is not entirely due to the greater acoustic salience of stressed syllables. Rather, participants also appear to be able to anticipate or *predict* future stress, and greater anticipation for stressed syllables as compared to unstressed syllables contributes to the reduction in reaction time. Cutler (1976) specifically examined this stress prediction effect, using the intonational contour of the preceding sentence to generate a strong or weak prediction about future syllable stress on a target word. In her experiment, participants were told to respond to a target phoneme like /d/ in the word 'dirt'. She then recorded two versions of a sentence containing the target word. Both versions of the sentence contained the target word in the same location, but they differed in terms of whether the word 'dirt' received a strong stress emphasis (shown in CAPS) or not, as shown below :

STRONG stress : *"She managed to remove the DIRT from the rug, but not the berry stains"*
WEAK stress : *"She managed to remove the dirt from the RUG, but not from their clothes."*

The word 'dirt' was then removed (spliced out) from each sentence, and replaced with a neutral token, so that the acoustic features of the target word in both sentences would be identical. However, the intonational contour of the initial 7 syllables *before* the occurrence of the target word (*"She managed to remove the.."*) was left intact, and continued to cue either a future stressed or unstressed target word respectively. Therefore, any reaction time differences to the target phoneme /d/ would not be due to acoustic differences in the target word itself, but solely due to prediction effects generated by the preceding intonational contour of the first 7 syllables. Remarkably, Cutler managed to measure a significant 40 ms difference in reaction time between the two sentences, demonstrating that merely the *prediction* of an upcoming stressed syllable was sufficient to cause a measurable drop in reaction time.

Although Cutler did not measure the precise intonational differences between the two versions of the sentence, it was noted that the relative pitch, duration and amplitude assigned to each word all differed markedly between the sentences. Therefore, it patterns of change in pitch, amplitude and duration across the first 7 syllables all cued whether a stressed syllable was imminent, or further away. Figure 9.2 illustrates how participants could be generating their stress predictions based on the shape of the intonation contour.

Figure 9.2. Illustration of stress prediction using the intonation contour (or Stress AM phase). The oscillatory black lines represent the intonation contour (e.g. pitch or amplitude contour), red dots indicate syllables. The x-axis represents time, and the y-axis represents the relative degree of stress. The vertical dotted line marks the location of the target word ('dirt'). This target word is predicted to have strong stress in the top sentence, but weaker stress in the bottom sentence. Participants' predictions are only based on the first 7 syllables (dots) before the target word (dotted line).



Here, the intonation contour (e.g. representing either pitch or intensity) of each sentence used by Cutler is depicted as following an oscillatory cycle (shown as a black line). Syllables in each sentence (denoted as red dots) fall at different points on the contour, where syllables at the top of the contour receive the most stress, and those at the bottom are the least stressed. Note that in the figure, the exact timing of each of the first 7 syllables is identical for

both sentences (i.e. they have the same x-coordinates in the graph), the only difference is in the phase of the intonation contour that the syllables occupy. In the first sentence, the first 7 syllables occupy the upward-going slope of the intonation contour, but in the second sentence, the syllables occupy the trough of the intonation contour. Therefore, if participants were tracking the on-going phase of the intonation contour, the pattern of the first 7 syllables in the top (strong prediction) sentence would lead participants to predict that the next syllable was very likely to be stressed, since the phase of the seventh syllable ('the') was already very near the peak of the contour. Conversely, the pattern of the first 7 syllables in the bottom (weak prediction) sentence would indicate that a stressed syllable was still some way away, since the 7th syllable remained near the base of the intonation contour.

Note that this method of stress prediction does not require isochrony in syllable timing, or in stress timing. Rather, all that is required to make the prediction about stress status is the phase trajectory of the intonation contour during the initial 7 syllables, and an assumption that the target word 'dirt' will continue the trend (following the current trajectory of the oscillatory cycle). Of course, the sentence is unfolding over time, so participants might also have to predict *when* the target word will occur, in addition to how stressed it is likely to be. Listeners could estimate this timing based on preceding syllable rate, and this temporal prediction should not differ between the two sentences in this example. Also, the perfectly sinusoidal intonation contour used in this example might not be realistic. However, in the context of the AMPH and S-AMPH models, this intonation contour could correspond to the Stress AM pattern, since the phase of the Stress AM also indicates prosodic prominence. Although the Stress AM is not perfectly sinusoidal (except in metrical speech), the phase of the Stress AM does tend to change smoothly over time and phase jumps are rare. Therefore, it may be possible to make short-term stress predictions based on the local phase trajectory of the Stress AM, as illustrated in Figure 9.2.

Why is it important to be able to generate predictions about up-coming stress and prominence? One suggestion is that this helps the listener to allocate his or her attentional resources more effectively. According to this argument, attention is preferentially allocated to stressed words during speech processing because stressed words are likely to contain important semantic content (for example, the contrastive stress applied in Cutler's sentences). If the aim of speech processing is to quickly infer the speaker's meaning with minimum processing effort, it would be parsimonious to allocate precious attentional resources disproportionately toward portions of speech that are high in 'information' content (i.e.

stressed words) at the expense of portions of speech that are low in 'information' content (i.e. unstressed words). Therefore, the faster reaction times to predicted stressed words may also be accompanied by greater *attention* to these predicted stressed words.

This predictive attention hypothesis could be investigated in an EEG paradigm using Cutler's (1976) original task, but looking for attentional modulation of EEG components (such as the P1 and N1) elicited by the target word. Moreover, if greater attentional resources are directed toward words that the listener expects to be stressed, there might also be secondary effects on memory and learning for these words. For example, words that are *predicted* to be stressed rather than unstressed could be more strongly encoded in memory, and better recalled by the listener. Conversely, if a listener is unable to generate these stress predictions accurately (e.g. due to poor tracking of Stress AM phase patterns or the pitch contour), he or she would benefit less from these anticipatory effects. This could lead to an overload on attentional resources and slower speech processing because he or she is trying to attend to and remember everything, rather than selectively attending only to 'important' (stressed) speech information.

9.2.3.2 Predicting Speech Timing

A related concept to stress prediction is the concept of *rhythmic* prediction, or temporal expectancy. Here, the emphasis is on being able to predict *when* speech information is likely to occur, not just whether it will be stressed or unstressed. Such temporal expectancy has also been formally linked to attentional allocation, for example by Jones et al (2002) in her Dynamic Attending Theory. According to this theory, listeners engage in 'anticipatory attending', which is a *temporal* shift of attention that anticipates (expects) the onset of a sound. Unlike many top-down models of attentional orienting, here, the build-up of temporal expectations is purely stimulus-driven. In a typical paradigm, listeners may be asked to compare a standard tone with a test tone, where the test tone occurs at either an expected or unexpected time. To create temporal expectations, the standard and test tones are separated by a sequence of 'distractor' tones (of different pitch) that are all presented with the same inter-stimulus interval (ISI). The test tone is then presented with either the same ISI (fulfilling temporal expectations), or with a shorter or longer ISI (violating temporal expectations). Participants typically show the best performance in pitch discrimination when the ISI of the test tone matches the ISI of the distractors, and performance is poorer when the

test tone is either presented 'too early' or 'too late'. Jones argues that the differential performance on the task reflects the fact that participants' attention is being directed toward a specific future point of time by rhythmic expectancies generated over the regular distractor ISIs. Stimuli that occur at the expected time benefit from temporal attentional focus and are processed more effectively. Stimuli that do not occur at the expected time miss out on these attentional benefits. Interestingly, this effect persists even when participants are told explicitly to ignore the timing of the tones and focus exclusively on pitch discrimination (a non-temporal dimension), suggesting that temporal expectations can develop automatically and involuntarily.

Do such rhythmic expectancies also form in natural speech, heightening our attention at predicted times, and enhancing speech processing at these times? Cutler and others (Cutler, 1986; Cutler & Foss, 1977; Martin, 1972) have proposed that successive stressed syllables in continuous speech could act like a metrical or rhythmic grid, where the temporal regularity from one stressed syllable to the next allows the listener to predict the future occurrence of the next stressed syllable. In other words, it is argued that listeners form temporal predictions on the basis of durational isochrony between stressed syllables in an utterance, allowing these stressed syllables to be more efficiently processed. However, the idea of durational isochrony in speech has now largely been discredited (e.g. Dauer, 1983; Arvaniti, 2009). Therefore, if temporal predictions do form in speech, they must either be based on non-isochronous stress intervals, or on other speech features apart from stress intervals.

How much anisochrony can listeners cope with before rhythm detection becomes impossible? Madison & Merker (2002) found that listeners could tolerate an average of 8.6% deviation in stimulus anisochrony before they were no longer able to find a regular pulse in a tone sequence. In the context of their study, this meant that for a tone sequence with an 'ideal' ISI of 600 ms, the actual ISI of tones could vary by up to ±50 ms (i.e. between 550 ms to 650 ms) before the rhythmic pulse was lost. A tolerance of close to 10% suggests that human listeners are fairly forgiving of anisochrony in rhythm detection. However, the average variability of interstress intervals in English far exceeds this margin of toleration. For example, Dauer (1983) reported that the average interstress interval in speech was around 450 ms, with a standard deviation of approximately 150 ms or 33%. These figures suggest that interstress intervals would not be perceived as rhythmically regular because their variability is simply too high for a regular pulse to be detected. By extension, it would be

difficult for listeners to generate strong rhythmic predictions if they were not able to find a rhythmic 'pulse' in the first place.

However, Dauer's figures reflect the statistical distribution of interstress intervals over a large speech corpus, and the high standard deviation could partly be due to differences in speaking rate within and between utterances. If one considered just 2 or 3 consecutive stress intervals from the same utterance, these might not have such a large variation, and it might be possible for listeners to form short-term temporal expectations 'on the fly' based on these local stress intervals. There is also anecdotal evidence that temporal expectancies can be generated in speech, and that speech timing is not completely unpredictable. For example, in the use of 'comedic timing', the speaker consciously manipulates listeners' expectancies in order to deliver a punch line with the maximal effect. Also, when two speakers are deeply engaged in conversation, both speakers can be synchronised to the extent that one speaker finishes the sentence of another speaker without breaking the flow of conversation, which must require both semantic and temporal prediction. However, much more empirical study is required to determine the basis of these effects.

A second possibility is that listeners generate temporal expectancies in speech using sources of information *external* to the speech signal. Unlike rhythmic expectancy (which is based on finding a regular pulse pattern), these temporal expectancies are not based on rhythm, but on strong associative or causal relationships that predict the occurrence of speech events within a narrow window of time. Perhaps the strongest non-acoustic predictor of speech events is the *visual* temporal information from the movement of the articulators (e.g. lips and jaw), or from other motor gestures. For example, the onset of mouth opening typically preceeds the onset of auditory speech information by around 200 ms (Schroeder et al, 2008). Since there is a physical causal link between the mouth opening and speech sounds being produced, this visual cue is highly reliable for predicting the onset of speech information[37]. To explain how this prediction mechanism could work, Schroeder and colleagues (e.g. Schroeder et al, 2008) proposed that the early incoming visual information causes a phase-resetting of low-frequency neuronal oscillations in the auditory cortex into an

[37] Moreover, the shape and size of the mouth will also typically correlate with the type of speech sound that will be produced. There is evidence that listeners do generate expectations based on this auditory-visual relationship, because violations of the relationship will typically produce 'fusion' illusions (rather than resulting in the visual information being ignored). For example, in the McGurk effect (McGurk & MacDonald, 1976), when presented with an image of a speaker saying '/ga/' that is paired with a soundtrack of the syllable '/ba/', listeners commonly report hearing '/da/' instead. This illusion occurs because the listener is attempting to reconcile what he predicts he should hear (based on the visual cues) with what he actually hears. In normal conditions, viewing the face of the speaker also increases the intelligibility of spoken communication (Sumby & Pollack, 1954).

optimal phase of excitability. This phase-resetting results in speech information being processed more effectively because the brain is 'ready' to receive the auditory information by the time it actually arrives 200 ms later. Moreover, this temporal prediction mechanism does not depend on speech being rhythmic because no matter how erratic the utterance, visual articulatory information will always reliably precede the corresponding auditory information at every point. However, this auditory-visual mechanism can only predict speech events up to 200 ms in advance, or approximately 1 syllable in advance. If human listeners do indeed predict the timing of consecutive stressed syllables in normal speech (i.e. about 450 ms in advance), as argued by Cutler and others (Cutler, 1986; Cutler & Foss, 1977; Martin, 1972), then other mechanisms will be required.

Therefore, while temporal prediction in speech could bestow advantages in speech processing via attention 'pre-allocation' (e.g. Jones' dynamic attending theory), it is not clear what acoustic cues and mechanisms could be used by the human auditory system to generate temporal predictions beyond ~200 ms. Nonetheless, the generation of temporal predictions using patterns and regularities in the acoustic signal has been proposed to be a fundamental function of the auditory system (e.g. Winkler et al, 2009). AM patterns in speech could be a useful source of such temporal patterns and regularities, and future study could reveal long-range AM patterns and relationships that could be used for such temporal prediction.

9.2.4 'PHONOLOGY' AS STORED SPECTRO-TEMPORAL PATTERNS

If the brain can represent the AM patterns in speech in fine detail (e.g. Pasley et al, 2012), this raises the possibility that these encoded AM patterns also constitute part of our mental representations of words and speech sounds, or 'phonology'. According to Port's theory of 'rich phonology' (e.g. Port, 2007, 2008, 2010), words are not stored in memory as abstract combinations of phones or phonemes (i.e. an alphabetic-like representation). Rather, words are represented in rich, concrete detail, capturing the full range of spectro-temporal variation in the acoustic signal. Moreover, Port argues that word representations in memory also include other 'episodic' detail, such as how the word was said, and who said it (i.e. prosody, speaker identity, etc). Therefore, Port rejects the traditional linguist's view that words are made from building blocks of phonemes, arguing that this view is a cultural artifact of having learned an alphabetic script for language. If Port is correct, this has important implications for language learning. Infants learning a new language would be capturing the

full range of spectro-temporal detail available in the speech signal - including formant patterns and AM patterns. For example, rather than representing the word *"cat"* as a combination of three discrete and invariant phonemes [k], [æ] and [t], infants could be encoding a single complex and continuous spectral-temporal pattern that would be different each time they heard the word being uttered (i.e. because of a different context or speaker). Each of these different exemplars of the spoken word *"cat"* would then be stored as separate entries in the infants' memory. Therefore, the next time the infant heard an utterance that sounded like *"cat"*, the infant could compare the new utterance (in full spectro-temporal detail) to all the previously stored exemplars of the word *"cat"*, and thereby determine how similar the new utterance was to the stored examples.

Port's view of language and phonology is controversial, but it does address why invariant phoneme cues have not been identified in the speech signal (e.g. Pisoni, 1997; Stevens, 1980) - such invariance may not be required. The question he raises, 'what is phonology?', is also a timely one. With advances in neuroimaging technology, human mental representations have become accessible for experimental study, and are no longer merely abstract theoretical constructs. However, there are two major considerations for Port's theory of rich phonology. First, are humans capable of generating and storing such complex spectro-temporal representations in the first place? Second, even if humans can generate such complex representations, why would we choose to do so if a more parsimonious encoding format will suffice?

To address the first question, Port points to findings from memory research regarding 'exemplar memory'. He cites in particular a finding by Palmeri et al (1993) in a word recognition task. Here, participants were asked to recognise verbally-presented words, where the word lists were either read by the same speaker, or up to 20 different speakers. Participants' memory performance was poorer when there were 2 different speakers, as compared to the just 1 speaker, but importantly, performance did not drop any further even when the number of different speakers was increased up to 20. Therefore, listeners were not making specific word-speaker associations (which should have caused a proportional drop in performance with the number of speakers). Rather participants' appeared to be automatically retrieving detail about the speakers' voice along with the word being spoken. This voice detail had to be suppressed when there were 2 or more speakers, but not when there was just 1 speaker. Therefore the uniform drop in performance across all different-speaker conditions was taken to reflect this suppression activity, and evidence that speaker voice was

automatically encoded along with the spoken word. Other evidence in support of 'rich' phonological encoding comes from neuroscience. For example, Pasley et al (2012) and Ding & Simon (2012) both recently demonstrated successful reconstruction of the speech temporal envelope from intracranial EEG and MEG recordings respectively. This suggests that the brain can indeed track the spectro-temporal variation in the acoustic signal with great detail and fidelity.

However, even if the brain is *capable* of encoding such fine spectro-temporal detail, does it actually do so when representing and storing speech sounds? To answer this question, it is helpful to look at infant phonological development. At 6-8 months of age, English-learning infants are able to discriminate phoneme contrasts (spectro-temporal patterns) that are used in English (/ba/-/da/), as well as native-American contrasts that are *not* used in English (/ki/-/qi/). However, by 10-12 months of age, English-learning infants can no longer hear the native-American contrast, even though native-American babies of the same age have no trouble making this discrimination (Werker & Tees, 1984). Therefore, although younger infants initially encode the full spectro-temporal patterns of all speech sounds (native or non-native), older infants apparently only selectively encode spectral-temporal contrasts that are used in their native language. This suggests that while our phonological representations *can* be rich in detail, such detail is only retained in memory when it is required to make discriminations in our native language. In other words, our phonological representations are indeed rich, but they are also parsimonious. Therefore, phonological development involves learning which spectro-temporal features are important and should be elaborated in memory, and which features are unimportant and should be ignored.

A final point to consider is whether the *quality* of mental phonological representations changes over time. For example, if very young children can only perceive a limited amount of spectro-temporal detail, then this would suggest that their mental representations of speech sounds would be likewise limited in scope. In this case, children's phonology would comprise only rough spectro-temporal 'sketches', rather than possessing the richness of detail envisaged by Port. One way to test this is to present children with speech that contains only spectral changes (e.g. sine-wave speech), or only amplitude changes (e.g. vocoded speech). In such a study, Nittrouer et al (2009) compared the performance of 7-year-old English-speaking children with native English-speaking adults, and adults who spoke English only as a second language (L2). They presented each group with sine-wave speech, and 4- and 8-channel noise-vocoded speech, using 4-word sentences as stimuli. They found that for sine-wave

speech, children performed on par with native adults, and performed better than L2 adults. However, for both 4- and 8-channel noise-vocoded speech, children could only perform at the level of L2 adults, and lagged behind native adults. These results were taken to indicate that children learned to encode spectral changes in speech *before* they learned to encode amplitude-related changes, which may require more protracted development. However, it should be noted that in this experiment, only speech intelligibility was tested, not sensitivity to rhythm and prosody, which could show a different pattern of development.

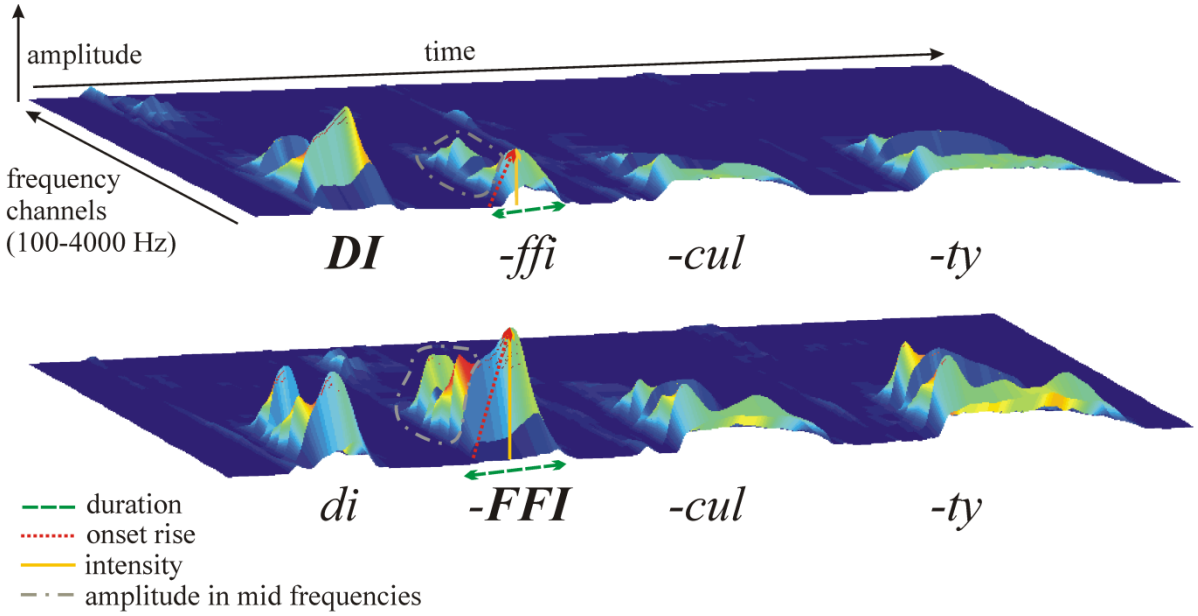
9.2.5 IMPLICATIONS FOR DISORDERS IN LANGUAGE DEVELOPMENT

This slower development of sensitivity to amplitude-related changes in the speech envelope may explain why it is particularly vulnerable in language disorders such as developmental dyslexia and specific language impairment (SLI, e.g., Goswami, 2011; Corriveau & Goswami, 2009). For example, children with developmental dyslexia show specific deficits when asked to detect amplitude-related changes, both in pure tones, as well as and in speech syllables (Goswami et al, 2002; Goswami et al, 2011; Huss et al, 2011). These difficulties in discriminating amplitude envelope 'rise time' are found in dyslexia across many different languages, including Chinese, Dutch, English, Finnish, French, Hungarian and Spanish (Goswami et al., 2002; Richardson et al., 2004; Muneaux et al., 2004; Hämäläinen et al., 2005; Suranyi et al., 2009; Poelmans et al., 2011).

In the context of speech, the amplitude 'rise time' typically refers to the syllable onset, and is measured as the time (in ms) taken for the syllable to reach its peak amplitude. When the syllable stress pattern of a word is artificially manipulated (e.g. 'DIfficulty' vs 'diFFIculty'), the stressed version of a given syllable (e.g. 'FFI') typically has a *longer* rise time than its unstressed version (e.g. 'ffi'). This is illustrated in Figure 9.3 (reproduced from Leong et al, 2011), where the syllable onset slope is shown in a diagonal dotted red line, and the rise time is the horizontal distance covered by this line along the x-axis (time). A longer rise time means that the stressed version of the syllable takes a relatively longer time to reach its peak amplitude than the unstressed version of the same syllable. In Leong et al (2011), dyslexic and non-dyslexic adults were presented with pairs of these stress-manipulated words, and had to indicate whether the two words in the pair had the same (e.g. "DIfficulty" vs "DIfficulty") or a different (e.g. "DIfficulty" vs "diFFIculty") syllable stress pattern. Dyslexics were found to be significantly poorer than their non-dyslexic peers at making this

syllable stress discrimination. Moreover, dyslexics' performance on the stress discrimination task was significantly predicted by their psychoacoustic sensitivity to amplitude *rise time*, but not by their sensitivity to either intensity or frequency (which also differed acoustically between stressed and unstressed syllables). Therefore, although there were multiple acoustic cues to syllable stress in the speech stimuli, the dyslexic deficit (in Leong et al, 2001) appeared to be associated specifically with the amplitude rise time cue.

Figure 9.3. Amplitude envelope across spectral frequencies for the word "difficulty" produced with stress on the first or second syllable. Reproduced from Leong et al (2011).



In the follow-up study presented in this thesis (Chapter 8), a new cohort of adult participants was recruited, and dyslexic stress perception (and production) were examined with specific reference to the AM patterns in speech. This time, the psychoacoustic profile of dyslexic participants indicated that they had *no* significant problems with amplitude *rise time* discrimination, but instead had significant problems with *intensity* discrimination (see Chapter 8, Section 8.2.2). This was the opposite pattern to that observed for participants in the original syllable stress study (Leong et al, 2011). However, across both studies, the dyslexic deficit pertained to amplitude (intensity) discrimination, and there were no group differences in discriminating other auditory parameters, such as frequency or duration. Therefore, even in highly-compensated dyslexic adults (participants were University of Cambridge undergraduates), acoustic problems in amplitude discrimination could still be detected.

Recall that the amplitude rise time can also be viewed as the upward-going portion of the AM cycle (i.e. -π to 0 radians phase). The psychoacoustic 'Dino' test for rise time perception used by Goswami and colleagues spans a broad range of rise times, equivalent to 1.7-33 Hz in AM rate. Therefore, one of the key aims of the dyslexia study in this thesis was to try to narrow down the dyslexia deficit to a particular AM tier or tiers (thereby implicating speech processing of a particular linguistic unit or units). In Chapter 8, the results of the three rhythm perception and production experiments were consistent. Dyslexics repeatedly displayed problems with the Syllable and Stress AM tiers, but not with the Phoneme AM tier. In the tone-vocoder rhythm perception experiment, dyslexics performed more poorly (although not significantly so) when the Syllable AM was combined with another AM tier (Stress or Subbeat). In the tapping experiment, dyslexics tapped at a significantly different phase only with respect to the Syllable AM tier, not the Stress or Phoneme tier. In the rhythm production experiment, dyslexics' syllable timing was disordered. Their distribution of Syllable AM peaks with respect to Stress AM phase also showed a different hierarchical organisation as compared to non-dyslexics, but their distribution of Phoneme AM peaks with respect to Syllable AM phase was normal.

In the AMPH and S-AMPH models proposed in this thesis, Stress and Syllable AM rates play a crucial role in syllable detection (e.g. identifying vowel nuclei) and prosodic stress assignment. Therefore, the observed dyslexic deficit at these slower AM rates is highly consistent with their previously documented problems with syllable stress perception (Kitzen, 2001; Goswami et al, 2010; Leong et al, 2011). According to the temporal sampling framework put forward by Goswami (2011), the phonological problem in dyslexia may be attributed to a fundamental problem with neural oscillatory phase-locking at slower syllable (theta) and stress (delta) rates. Consistent with this suggestion, it has been demonstrated that inefficient neural phase locking at the delta rate to sinusoidal amplitude-modulated noise indeed characterises individuals with developmental dyslexia (Hamalainen et al., 2012). If inefficient neural phase locking to *speech* AMs at a delta or theta rate (i.e. Stress or Syllable AM tiers) is also found (as predicted by the behavioural deficits observed in this thesis), this result would provide further support for Goswami's temporal sampling hypothesis. According to this explanation, poor neural phase-locking at neural delta and theta rates would result in faulty or impoverished representation of Stress and Syllable AM patterns in the speech signal. This in turn would lead to deficits in syllable timing and prosodic stress perception, producing difficulties with speech segmentation. Over the course of development, the

cumulative effects of these syllable timing and speech segmentation problems would result in dyslexic children having altered or incomplete phonological representations.

9.2.6 Potential Educational Applications

The AM-based work in this thesis also points to several possible ways in which language development may be supported or remediated. For example, in the nursery rhyme production task in Chapter 8 (Section 8.3.4) dyslexic adults were found to struggle with the more complex *iambic*-patterned nursery rhymes, while performing on par with their peers for the simpler *trochaic*-patterned nursery rhymes. This is consistent with previous findings by de Bree et al (2006) in which dyslexic children struggled more with the imitation of non-words with *irregular* prosodic stress patterns. This suggests that phonological training for dyslexic children should focus on more complex, infrequent or irregular metrical patterns, such as iambs ('w-S'), in order to strengthen their phonological representations of these more difficult metrical patterns.

Another possible remediation strategy for children with language disorders could be increasing their exposure to child-directed speech. The analysis of child-directed speech in Chapter 7 indicated that both nursery rhymes and non-poetic storybook readings possessed a tightly-nested hierarchical modulation structure when delivered in a child-directed speaking style. The presence of such strong hierarchical patterning between Stress and Syllable AM rates in the acoustic signal might be expected to generate equally strong *neural* hierarchical nesting between equivalent delta and theta oscillatory rates. Therefore, if indeed neural oscillatory phase-locking at delta and theta rates is impaired in dyslexia, repeated exposure to speech stimuli that elicit *strong* delta-theta phase-locking could train the underlying neural oscillatory networks to generate a more robust phase-locking response even to normal speech. Practically, this 'neural training' would be no more onerous than reading a storybook or nursery rhymes to the dyslexic child in a lively child-directed manner.

The difference in modulation structure between child-directed and adult-directed speech also has a possible interesting application in the area of artificial speech enhancement. If child-directed speech does indeed have beneficial effects on language learning, then the modulation statistics and spectro-temporal properties of speech could be artificially *enhanced* so that it is more 'child-directed' in nature. Adults spontaneously alter their speech patterns to be more child-directed when they are speaking to children. However, not all adults are

equally successful at producing these enhancements. For example, the former Children's Laurette, Michael Rosen, is far more skilled and successful at producing prosodically-enhanced child-directed speech than the average adult. This skill is evidenced by his ability to captivate child and adult audiences alike with memorable and enjoyable poetry readings. If the modulation statistics of his speech patterns were studied, these could provide an example of the 'optimal' child-directed speech template. This template could then be used to enhance the 'child-directedness' of other speech samples. Such child-directed speech enhancement could be used in a variety of educational settings, for example in the making of children's audiobooks, or in computerised phonological training games for children with language disorders. A CDS-enhancement hearing aid could even be designed for use by children with language disorders, so that any speech they hear is automatically prosodically enhanced in 'real-time'. Such acoustic enhancement could also be tailored to facilitate speech segmentation and to emphasise the important phonological patterns in speech, thereby supporting the phonological development of children with language disorders.

9.3 Future Research Questions

Finally, there are several other possible ways in which the current research on AM-based speech rhythm could be extended.

Speech Rhythm Differences Across Languages. The debate about language 'rhythm classes' has tended to centre around a fairly narrow definition of rhythm. Typically, researchers have compared durational differences at the segmental level. This means that rhythm differences arising from other cues (e.g. amplitude, pitch) or combinations of these cues, have been ignored. Consequently, no single 'rhythm-metric' method has been wholly successful in describing differences in speech rhythm across languages (see Arvaniti, 2009). The S-AMPH, alongside other new amplitude-based methods (e.g. Todd, 1994; Arvaniti, 2012) could shed new light on this debate by revealing cross-language differences in rhythm arising from *amplitude* changes. Also, traditional 'rhythm-metrics' typically focus on only one linguistic timescale (i.e. segments), rather than on the relationship between speech information at different timescales[38]. The S-AMPH model could reveal any differences in modulation pattern *across* different timescales. For example, in 'stress-timed' languages like

[38] Although Patel (2008) has suggested that a comparison between syllable and stress variability could be helpful

English, the critical phase relationship determining rhythm may involve Stress and Syllable AM tiers. However, for 'syllable-timed' languages, the Syllable and Phoneme AM tiers may be more instrumental in determining rhythm instead. The specific AM tier or tiers that specify the rhythm patterns of a given language could depend on the phonological unit(s) used by that language as the basic unit of rhythm. It is also possible that in fact all human languages can be arranged along a continuum as being 'more-or-less stress-timed' (Dauer, 1983). In this case, the number and configuration of tiers within the AM hierarchy that are used to specify rhythm should systematically predict the location of a given language on this continuum.

Modelling Individual Differences in Rhythm Perception. Previous studies have demonstrated a relationship between sensitivity to rhythm and prosody in speech and reading achievement (e.g. Goswami et al, 2010, Leong et al, 2011). In this thesis, it is also demonstrated that highly-compensated individuals with development dyslexia perceive and produce speech rhythm patterns differently from their non-reading-disabled peers (Chapter 8). It could be interesting to see if the AMPH model can be modified to produce 'deficits' in rhythm perception, similar to those observed in dyslexia. This could shed light on the possible mechanisms underlying rhythm deficits in these individuals.

Speech Rhythm 'Profiling'. The AMPH model(s) could be used to quantify or measure speech rhythm differences in the utterances of individuals with speech production problems, such as dysarthria. For example, one could see if patients with lesions at basal ganglia, cerebellar[39] or motor lesion sites produce distinctly different speech rhythm 'profiles', based on different characteristic changes to the AM hierarchy. One could also generate these speech rhythm profiles for different types of expressed emotions, or different speaker accents, for the purposes of emotion or speaker recognition.

Neural Basis of Rhythm Perception. A clear prediction has been made regarding a possible neural mechanism for rhythm perception (e.g. Figure 9.1). Specifically, this involves entrainment of hierarchically-nested neuronal oscillations to the AM hierarchy. This prediction remains to be tested empirically. Finally, experienced listeners attend selectively to information in the speech signal, and their sensitivity to acoustic information has been 'tuned' with experience. For example, infants gradually learn to 'ignore' non-native phoneme contrasts (Werker & Tees, 1984), while becoming increasingly sensitive to native prosodic

[39] The basal ganglia have been associated with 'beat-based' rhythm, while the cerebellum is associated with 'duration-based' rhythm (Grube et al, 2009; Grahn, 2009).

patterns (e.g. Jusczyk et al, 1999). It could be interesting to investigate the neural correlates of such neural 'tuning' to speech rhythm patterns during early language development, in infants.

9.4 FINAL CONCLUSION

Overall, both AMPH and S-AMPH models succeeded in the overall aim of providing an amplitude-based account of speech rhythm perception. This account was shown to be have psychological validity and to possess explanatory power. The AMPH and S-AMPH models also represent methodological advancements for speech rhythm measurement. As analytical tools for syllable detection and prosodic stress transcription, the models produce an acceptable level of accuracy. In practical applications of the AMPH models to experimental research, the models were found to be useful in uncovering subtle differences in temporal structure, and in highlighting the psychological variables associated with these differences.

Clearly, there are exciting avenues for future research into AM-based speech rhythm. It is hoped that the work in this thesis will draw attention to the amplitude envelope as a potentially rich source of information about speech rhythm patterns, and as a worthy subject for further empirical study.

BIBLIOGRAPHY

Abercrombie, D. (1967): *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 13367–13372.

Aiken, S.J. & Picton, T.W. (2008). Human cortical responses to the speech envelope. *Ear & Hearing*, 29, 139-157.

Albin, D.D. & Echols, C.H. (1996). Stressed and word-final syllables in infant-directed speech. *Infant Behavior & Development*, 19, 401-418.

Allen, G. (1972). The location of rhythmic stress beats in English: an experimental study. *Language and Speech*, 15, 72-100.

Armitage, S.E., Baldwin, B.A., & Vince, M.A. (1980). The fetal environment of sheep. *Science*, 208, 1173-1174.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66, 46-63.

Arvaniti, A. (July, 2012). Rhythm and timing. Unpublished paper presented at the 13th Conference on Laboratory Phonology, Stuttgart, Germany.

Bacon, S. P., & Grantham, D.W. (1989). Modulation masking: Effects of modulation frequency, depth and phase. *Journal of the Acoustical Society of America*, 85, 2575-2580.

Baken, R. J. (1987). *Clinical Measurement of Speech and Voice*. London: Taylor and Francis Ltd

Barbosa, P.A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence*, pages 163-166.

Barnes, S., Gutfreund, M., Satterly, D., & Wells, G. (1983). Characteristics of adult speech which predict children's language development. *Journal of Child Language*, 10, 65-84.

Berens, P. (2009). CircStat: A Matlab toolbox for circular statistics. *Journal of Statistical Software*, 31, Issue 10. http://www.jstatsoft.org/v31/i10

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, 4, 247-260.

Broen, P.A (1972). The verbal environment of the language-learning child. *Monograph of the American Speech and Hearing Association*, 17.

Bryant, P.E., Bradley, L., Maclean, M., & Crossland, J. (1989). Nursery rhymes, phonological skills and reading. *Journal of Child Language*, 16, 407-428.

Burnham, D.K., Kitamura, C. & Vollmer-Conna, U.S. (2002). What`s new pussycat? On talking to babies and animals. *Science*, 296, 1435-1435.

Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., Knight, R.T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313, 1626–1628.

Chi, T., Gao Y., Guyton, M., Ru, P., & Shamma, S. (1999). Spectrotemporal Modulations and Speech Intelligibility, *Journal of the Acoustical Society of America*, 106, 2719-2732.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.

Clarke, E.F. (1999). Rhythm and Timing in Music. In D. Deutsch (Ed.), *Psychology of Music*, 2nd Edition (pp. 473-500). New York: Academic Press.

Cogan, G., & Poeppel, D. (2011). A mutual information analysis of neural coding of speech by low-frequency MEG phase information. *Journal of Neurophysiology*, 106, 554-563.

Cole, R. A. & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and Production of Fluent Speech.* Lawrence Erlbaum Associates, Hillsdale, NJ, 133-163.

Corriveau, K., & Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairment: Tapping to the beat. *Cortex*, 45, 119-130.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31,139-148.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26, 145–171.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, 20, 55-60.

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech*, 29, 201-220.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, 20, 1-10.

Cutler, A. & Norris, D.G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

Dau, T., Puschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system: II. Simulations and measurements. *Journal of the Acoustical Society of America*, 99, 3623-3631.

Dau,T., Kollmeier, B., & Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *Journal of the Acoustical Society of America*, 102, 2892-2905.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *Journal of the Acoustical Society of America*, 102, 2906–2919.

Dauer, R. (1983). Stress-timing and syllable timing revisited. *Journal of Phonetics*, 11, 51-62.

de Bree, E., Wijnen, F., & Zonneveld, W. (2006). Word stress production in three-year-old children at risk of dyslexia. *Journal of Research in Reading*, 29, 304–317.

DeCasper, A. J., & Fifer,W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, 208, 1174–1176.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, 9,133–150.

Dehaene-Lambertz, G., Dehaene, S. & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants, *Science*, 298, 201-205.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M. & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences USA*, 103, 14240-14245.

de Jong, N. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behaviour Research Methods*, 41, 385-390.

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. *Proceedings of the International Congress of Phonetics Science.* (pp.471-474). Barcelona.

di Cristo, A. (1998). Intonation in French. In Hirst, D. & Di Cristo, A. (eds). *Intonation Systems : A Survey of Twenty Languages*, pp. 195-218.

Ding, N. & Simon, J.Z. (2012). The emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences,* 109, 11854-11859.

Divenyi, P.L., & Hirsh, I.J. (1978). Some figural properties of auditory patterns. *Journal of the Acoustical Society of America*, 65, 1369-l 385.

Drullman, R., Festen, J.M., & Plomp, R. (1994a). Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95, 1053-1064.

Drullman, R., Festen, J.M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95, 2670-2680.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). Perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, 36, 202–225.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171, 303-306.

Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z., & Shamma, S.A. (2004). Dynamics of precise spike timing in primary auditory cortex. *Journal of Neuroscience*, 24, 1159–1172.

Elliott, L.L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *Journal of the Acoustical Society of America*, 66, 12-21.

Elliott, T.M., & Theunissen, F.E. (2009) The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, 5(3): e1000302. doi:10.1371/journal.pcbi.1000302

Eriksson, A. (1991). *Aspects of Swedish Speech Rhythm*. University of Goteborg, Goteborg.

Ewert, S. D., & Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations *Journal of the Acoustical Society of America*, 108, 1181–1196.

Ewert, S. D., Verhey, J. L., & Dau, T. (2002). Spectro-temporal processing in the envelope-frequency domain. *Journal of the Acoustical Society of America* ,112, 2921–2931.

Fechner, G.T. (1860). *Elemente der Psychophysik. Leipzig: Breitkopf und Härtel*, 2, p. 559 (Reprinted, Bristol: Thoemmes Press, 1999).

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, S. & Bates, E. (2006). *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Second edition. Brookes.

Ferguson, C.A. (1977). Baby talk as a simplified register. In C.E. Snow & C.A. Ferguson (Eds.), *Talking to Children*. Cambridge: CUP.

Ferguson, C.A., & Debose, C.E. (1977). Simplified registers, broken language, and pidginization. In A. Valdman (Ed.), *Pidgin and Creole Linguistics*. Bloomington: Indiana University Press.

Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: Is the melody the message? *Child Development*, 60, 1497-1510.

Fernald, A. & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, 20, 104-113.

Fernald, R. D. & Gerstein, G.L. (1972) Response to cat cochlear nucleus neurons to frequency and amplitude modulated tones. *Brain Research*. 45, 417-435.

Fredrickson, N., Frith, U., & Reason, R. (1997). *Phonological assessment battery* (Standardised ed.). Windsor: NFER-Nelson.

Friederici, A., Friedrich, M., & Christophe, A. (2007). Brain responses in 4-month-old infants are already language-specific. *Current Biology*, 17, 1208–1211.

Fry, D. B. (1954). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 26, 138.

Fujisawa, Y., Minematsu, N., & Nakagawa, S. (1998). Evaluation of Japanese manners of generating word accent of English based on a stressed syllable detection technique. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Vol 98, pp. 3103-3106.

Fullgrabe, C., Stone, M.A., & Moore, B.C. (2009). Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task (L). *Journal of the Acoustical Society of America*, 125, 1277-1280.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2:130. doi: 10.3389/fpsyg.2011.00130

Ghitza, O. & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66:113–126.

Giraud, A.L. & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15, 511-517.

Giraud, A.L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., & Kleinschmidt (2000). Representation of the temporal envelope of sounds in the human brain. *Journal of Neurophysiology*, 84, 1588-1598.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data, *Hearing Research*, 47, 103-138.

Gleitman, L., & Wanner, E. (1982). Language acquisition: The state of the art. In E. Wanner, & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3–48). Cambridge, UK: Cambridge University Press.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, 82, 88-105.

Goswami, U. (2008). The development of reading across languages. *Annals of the New York Academy of Sciences*, 1145, 1-12. doi: 10.1196/annals.1416.018

Goswami, U. (2010). A psycholinguistic grain size view of reading acquisition across languages. In N. Brunswick, S. McDougall & P. Mornay-Davies (Eds). *The Role of Orthographies in Reading and Spelling*. Hove: Psychology Press.

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences*, 15, 1 3-10.

Goswami U, Fosker T, Huss M, Mead N & Szücs D. (2011). Rise time and formant transition duration in the discrimination of speech sounds: The ba-wa distinction in developmental dyslexia. *Developmental Science*, 14, 34-43.

Goswami, U., Gerson, D., & Astruc, L. (2010). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading and Writing*, 23, 995-1019.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S.K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences*, 99, 10911–10916.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (eds.). *Laboratory phonology* (Vol. 7, pp. 515–546). Berlin: Mouton de Gruyter.

Grahn, J.A. (2009). The role of the basal ganglia in beat perception: neuroimaging and neuropsychological investigations. *Annals of the New York Academy of Sciences*, 1169, 35-45

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 159–176.

Greenberg, S. (2006). A multi-tier framework for understanding spoken language. In S. Greenberg & W. Ainsworth (eds.), *Understanding speech: An auditory perspective* (pp. 411–434). Mahweh, NJ: LEA.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech - a syllable-centric perspective. *Journal of Phonetics*, 31, 465-485.

Grube, M., Cooper, F., Chinnery, P. & Griffiths, T. (2009). Dissociation of duration-based and beat-based auditory timing in cerebellar degeneration. *Proceedings of the National Academy of Sciences*, 107, 11597–11601.

Gueron, J. (1974). The meter of nursery rhymes: An application of the Halle-Keyser theory of meter. *Poetics*, 12, 73-111.

Hämäläinen, J., Leppänen, P., Torppa, M., Müller, K., & Lyytinen, H. (2005). Detection of sound rise time by adults with dyslexia. *Brain & Language*. 94, 32–42.

Hämäläinen, J.A., Leppänen, P.H.T., Eklund, K., Thomson, J., Richardson, U., Guttorm, T.K., Witton, C., Poikkeus,, A-M., Goswami, U., & Lyytinen, H. (2009). Common variance in amplitude envelope perception tasks and their impact on phoneme duration perception and reading and spelling in Finnish children with reading disabilities. *Applied Psycholinguistics*, 30, 3, 511-530.

Hämäläinen, J.A., Rupp, A., Soltész, F., Szücs D, Goswami U. (2012). Reduced phase locking to slow amplitude modulation in adults with dyslexia: An MEG study. *Neuroimage*, 59, 2952–2961.

Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language*, 27, 689-705.

Hanson, K. (2006). Meter. In K. Brown (Ed), *Encyclopedia of Language & Linguistics* (Second Edition), Elsevier, Oxford, pp. 616-619.

Harsin, C. A. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics*, 59, 243-251.

Hayes, B. (1995). *Metrical stress theory: principles and case studies*. Chicago: University of Chicago Press.

Hillenbrand, J.M., Getty, L.A., Clark, M.J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.

Hirst, D.J., (2006). Prosodic aspects of speech and language. In Keith Brown, (ed.), *Encyclopedia of language and linguistics. 2nd edition*, 539-546. Oxford: Elsevier.

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior & Development*, 32, 262–274.

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomics Bulletin Review*, 7, 504–509.

Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *Journal of the Acoustical Society of America,* 85, 1676-1680.

Houtgast, T., & Steeneken, H., (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77, 1069-1077.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. In Van den Broecke, M.P.R. and Cohen, A. (eds.*), Proceedings of the Tenth International Congress of Phonetic Sciences*. Dordrecht: Foris.

Howell, P. (1988a). Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics*, 43, 90-93.

Howell, P. (1988b). Prediction of P-center location from the distribution of energy in the amplitude envelope: II. *Perception and Psychophysics*, 43, 99.

Howitt, A. (2000). Automatic syllable detection for vowel landmarks. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., & Liu, H. H. (1998). The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Nonstationary Time Series Analysis. *Proceedings of the Royal Society of London A*, 454, 903–995.

Huss, M., Verney, J., Fosker, T., Mead, N. & Goswami, U. (2011). Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex*, 47, 674-89.

Jittiwarangkul, N., Jitapunkul, S., Lukaneeyanavin, S., Ahkuputra, V., & Wutiwiwatchai, C. (1998). Thai syllable segmentation for connected speech based on energy. In *Proceedings of the 1998 Asia-Pacific Conference on Circuits and Systems, pp 169-172, IEEE.*

Johnson, E., & Seidl, A. (2008). At eleven months, prosody still outranks statistics. *Developmental Science*, 11, 1–11.

Jones, M. R., Moynihan, H. MacKenzie, N., & Puente, J. (2002) Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science,* 13, 313-319.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress patterns of English words. *Child Development*, 64, 675–687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159–207.

Kalinli, O., & Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *Proceedings of Interspeech, 2007*, pp. 1941-1944.

Kalinli, O. (2011). Syllable segmentation of continuous speech using auditory attention cues. In *Proceedings of Interspeech, 2011*, pp. 425-428.

Kayser, C., Montemurro, M., Logothetis, N., & Panzeri, S. (2009). Spike-phase coding boosts and stabilizes the information carried by spike patterns. *Neuron*, 61, 597-608.

Kelso, J., Saltzman, E., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14, 29-59.

Kitzen, K. R. (2001). *Prosodic sensitivity, morphological ability and reading ability in young adults with and without childhood histories of reading difficulty.* Doctoral dissertation, University of Columbia, 2001. Dissertation Abstracts International, 62 (02), 0460A.

Klein, W., Plomp, R., & Pols, L. C. W. (1970). Vowel spectra, vowel spaces and vowel identification, *Journal of the Acoustical Society of America*, 48, 999–1009.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency adds little. *Journal of the Acoustical Society of America*, 118, 1038–1054.

Kuijpers, C., Coolen, R., Houston, D., & Cutler, A. (1998). Using the head-turning technique to explore cross-linguistic performance differences. *Advances in Infancy Research* (Vol.12, pp. 205–220). Stamford, CT: Ablex.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., Schroeder, C.E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94, 1904 –1911.

Langner, G., & Schreiner, C.E. (1988). Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *Journal of Neurophysiology*, 60, 1799-1822.

Lee, C., & Todd, N. (2004). Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora, *Cognition*, 93, 225-254.

Leong, V., Hamalainen, J., Soltesz, F., & Goswami, U. (2011). Rise time perception and detection of syllable stress in adults with developmental dyslexia. *Journal of Memory and Language*, 64, 59-73.

Lerdahl, F.,& Jackendoff, R.(1983). *A generative theory of tonal music*. Cambridge, MA:MIT Press.

Leroy, F., Glasel, H., Dubois, J., Hertz-Pannier, L., Thirion, B., Mangin, J-F., & Dehaene-Lambertz, G. (2011). Early maturation of the linguistic dorsal pathway in human infants.. *Journal of Neuroscience*, 31, 1500-1506.

Levitt, H. (1971). Transformed up–down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49, 467–477.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8, 249-336.

Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J. & Chauvel, P. (2004) Temporal envelope processing in the human left and right auditory cortices. *Cerebral Cortex*, 14, 731-740.

Lorenzi, C., Soares, C., & Vonner, T. (2001). Second-order temporal modulation transfer functions, *Journal of the Acoustical Society of America*, 110, 1030–1038.

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biolology*, 8(8): e1000445. doi:10.1371/journal.pbio.1000445

Luo, H., & Poeppel, D. (2007) Phase patterns of neuron**al** responses reliably discriminate speech in human auditory cortex. *Neuron*, 54,1001–1010.

Maclean, M., Bryant, P.E., & Bradley, L. (1987). Rhymes, nursery rhymes and reading in early childhood. *Merrill-Palmer Quarterly*, 33, 255-282.

MacPherson, S. (1930). *Form in Music*. London: Joseph Williams Ltd.

Madison, G., & Merker, B. (2002) On the limits of anisochrony in pulse attribution. *Psychological Research*, 66, 201–207.

Mampe, B., Friederici A. D., Christophe A., & Wermke K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, 19, 1994–1997.

Marcus, S.M. (1981). Acoustic determinants of Perceptual center (P-center) location. *Perception & Psychophysics*, 30, 247-256.

Marín-Padilla M, & Marín-Padilla T (1982) Origin, prenatal development and structural organization of layer I of the human cerebral (motor) cortex A Golgi Study. Anatomy *and Embryology*, 164, 161–206.

Marshall, L., Brandt, J.F., Marston, L.E., & Ruder, K. (1979). Changes in number and type of errors on repetition of acoustically distorted sentences as a function of age in normal children. *Journal of the American Auditory Society*, 4, 218-225.

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 79, 487-509.

Mermelstein, P., & Kuhn, G.M. (1974). Segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 55, 22(A).

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America,* 58, 880-883.

McAuley, J.D., Jones, M.R., Holub, S., Johnston, H.M., Miller, N.S. (2006). The time of our lives: lifespan development of timing and event tracking. *Journal of Experimental Psychology: General*, 135, 348–367.

McDermott, J.H., & Simoncelli, E.P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron,* 71, 926-940.

McGurk H., & MacDonald J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.

Minematsu, N., Fujusawa, Y., & Nakagawa, S. (1999). Automatic detection of stressed syllables in English words using HMM and its application to prosodic evaluation of pronunciation efficiency. In *IEICE Transactions on Information and Systems, Pt 2.* Vol J82-D-2, pp. 1865-1876.

Møller, A.R. (1974). Responses of units in the cochlear nucleus to sinusoidally amplitude-modulated tones. *Experimental Neurology*, 45, 104–117.

Montemurro, M.A., Rasch, M.J., Murayama, Y., Logothetis, N.K., & Panzeri, S. (2008). Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Current Biology,* 18, 375-80.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*, 6th Ed. (Emerald, Bingley, UK).

Moore, J. K. (2002). Maturation of human auditory cortex: Implications for speech perception. *The Annals of Otology Rhinology & Laryngology*, Supplement, 189, 7–10.

Moore, J.K., & Linthicum, F.H. (2007). The human auditory system: A timeline of development. *International Journal of Audiology*, 46, 460-478.

Morton, J., Marcus, S. M., & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review*, 83, 405–408.

Muneaux, M., Ziegler, J.C., Truc, C., Thomson, J. & Goswami, U. (2004). Deficits in beat perception and dyslexia: Evidence from French. *Neuroreport*, 15, 1255-1259.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science*, 15, 133–137.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24 , 756–766.

Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54, 283-299.

Nittrouer, S. (2006). Children hear the forest (L). *Journal of the Acoustical Society of America*, 120, 1799-1802.

Nittrouer, S., Lowenstein, J.H., & Packer, R.R. (2009). Children discover the spectral skeletons in their native language before the amplitude envelopes. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1245-1253

Nørgaard Jørgensen, R., Dale P.S., Bleses, D. & Fenson, L. (2010). CLEX: A cross-linguistic lexical norms database. *Journal of Child Language*, 37 , 419-428. http://www.cdi-clex.org/

Nursery Treasury (2010). Essex, UK : Miles Kelly Publishing.

Obleser, J. & Weisz, N. (in press). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex*. DOI: 10.1093/cercor/bhr325

O'Dell, M., Lennes, M. & Nieminen, T. (2008) Hierarchical levels of rhythm in conversational speech. Proceedings of Speech Prosody 2008, 355-358.

O'Dell, M., Lennes, M., Werner, S. & Nieminen, T. (2007) Looking for rhythm in conversational speech. Proceedings of the International Congress of Phonetic Sciences 2007, 6.-10.8.2007, Saarbrücken, Germany.

O'Dell, M. & Nieminen, T. (1999) Coupled oscillator model of speech rhythm. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. Bailey (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences*, Volume 2, pages 1075–1078. University of California, Berkeley.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology, Learning Memory, and Cognition,* 19, 309–328.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15, 1191-1253.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., & Chang, E.F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, 10(1): e1001251. doi:10.1371/journal.pbio.1001251

Patel, A. (2008). *Music, Language, and the Brain.* NY: Oxford University Press.

Patel, A.D., Iversen, J.R., & Rosenberg, J.C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of Americ*a, 119, 3034-3047.

Patel, A.D., Löfqvist, A., & Naito, W. (1999). The acoustics and kinematics of regularly-timed speech: A database and method for the study of the P-center problem. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1999, Volume 1, pp.405-408.

Peelle, J.E., Gross, J., & Davis, M.H. (in press). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*.

Pfitzinger, H., Burger, S., & Heid, S. (1996). Syllable detection in read and spontaneous speech. In *ICSLP 86 Proceedings, Fourth International Conference on Spoken Language*, Volume 2, pp 1261-1264. IEEE.

Pike, P. (1945). *The intonation of American English.* Ann Arbor: University of Michigan.

Pisoni, D. B. (1997). Some thoughts on normalization in speech perception. In K. Johnson, & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.

Pitt, M. A., & Samuel, A.G. (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 564-573.

Plomp, R. (1983a). The role of modulation in hearing. In Klinke, R. and Hartmann , R. (Eds.), *Hearing: Physiological bases and psychophysics*. Springer-Verlag, Berlin, pp. 270-276.

Plomp, R. (1983b). Perception of speech as a modulated signal. *Proceedings of the 10th International Congress of Phonetic Sciences*, Utrecht, 29-40.

Poelmans, H., Luts, H., Vandermosten, M., Boets, B., Ghesquière P., & Wouters, J. (2011). Reduced sensitivity to slow-rate dynamic auditory information in children with dyslexia. *Research in Developmental Disabilities*, 32, 2810-2819.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, 41, 245-255.

Pols, L.C.W., Tromp, H.R.C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53, 1093–1101.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P center phenomenon. *Journal of Phonetics*, 17, 175-192.

Ponton, C.W. & Eggermont, J.J. (2001). Of kittens and kids: Altered cortical maturation following profound deafness and cochlear implant use. *Audiology &Neurotology* , 6, 363-380.

Ponton, C. W., Moore, J. K.., & Eggermont, J. J. (1996). Auditory brain stem response generation by parallel pathways: Differential maturation of axonal conduction time and synaptic transmission. *Ear & Hearing*, 17(5):402-410.

Port, R. (2003) Meter and speech. *Journal of Phonetics*, 31, 599-61.

Port. R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, 25, 143- 170.

Port, R. (2008) All is prosody: Phones and phonemes are the ghosts of letters. Keynote address for Prosody2008 in Campinas, Brazil. Appeared in conference proceedings.

Port, R. (2010) Rich memory and distributed phonology. *Language Sciences*, 32, 43-55.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, 288, 349-351.

Ramus, F., Nespor, I., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265-292.

Ratner, N.B. (1984). Patterns of vowel modification in mother–child speech. *Journal of Child Language*, 11, 557-578.

Rees, A. & Møller, A.R. (1983). Response of neurons in the inferior colliculus of the rat to AM and FM tones. *Hearing Research*, 10, 301–330.

Repp, B. (2005). Sensorimotor synchronization: a review of the tapping literature. Psychon Bull Rev 12:969–992.

Richardson, U., Thomson, J., Scott, S.K. & Goswami, U. (2004). Auditory processing skills and phonological representation in dyslexic children. *Dyslexia*, 10, 215-233.

Roach, P.J. (1982). On the distinction between "stress-timed" and "syllable-timed" languages, in D.Crystal (Ed.) *Linguistic Controversies* ,pp. 73-79. London, Edward Arnold.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 336, 367-373.

Rosenblum, M.G., Pikovsky, A.S., Kurths, J., Schäfer, C., & Tass, P. (2001). Phase Synchronization: From Theory to Data Analysis. In: *Handbook of Biological Physics, Elsevier Science*, A.J. Hoff (Ed.), Vol. 4, Neuro-informatics and Neural Modeling, F. Moss and S. Gielen (Eds), Chapter 9, pp. 279-321.

Sachs, J., Brown, R., & Salerno, R. (1976). Adult's speech to children. In W. von Raffler Engel & Y. Lebrun (Eds.), *Baby talk and infant speech* (pp. 240-245). Lisse: Peter de Riddler Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926-1928.

Saltzman, E., & Byrd, D. (2000). Task dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19, 499-526.

Sargent, D., Li, K., & Fu, K. (1974). Syllable detection in continuous speech. *Journal of the Acoustical Society of America*, 55, 410.

Schane, S.A. (1979). The rhythmic nature of English word accentuation. *Language*, 55, 559–602.

Scholes, P. (1977). "Metre" and "Rhythm", in The Oxford Companion to Music, 6th corrected reprint of the 10th ed. (1970), revised and reset, edited by John Owen Ward. London and New York: Oxford University Press

Schreiner, C.E. & Urbas, J.V. (1986). Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, 21, 227-241.

Schroeder, C.E.& Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, 32, 9-18.

Schroeder,C.E., Lakatos,P., Kajikawa,Y., Partan,S., & Puce,A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12, 106-113.

Scott, S. K. (1993). *P-Centres in speech: an acoustic analysis*. Unpublished PhD thesis, University College London.

Scott, S.K. (1998). The point of P-centres. *Psychological Research*, 61, 4-11.

Sek, A., and Moore, B. C. J. (2002). Mechanisms of modulation gap detection. *Journal of the Acoustical Society of America*, 111, 2783–2792.

Sek, A., and Moore, B. C. J. (2003). Testing the concept of a modulation filter bank: The audibility of component modulation and detection of phase change in three-component modulators. *Journal of the Acoustical Society of America*, 113, 2801-2811.

Selkirk, E.O. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry*, 11, 563-605.

Selkirk, E.O. (1984). *Phonology and syntax. the relation between sound and structure*. Cambridge, Ma.: MIT Press.

Selkirk, E.O. (1986). *On derived domains in sentence phonology*. Phonology Yearbook 3:371–405.

Shallice, T., Fletcher, P., Frith, C.D., Grasby, P., Frackowiak, R.S.J., & Dola, R.J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature*, 368, 633 - 635.

Shannon R.V., Zeng, F-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

Shastri, L., Chang, S., & Greenberg, S. (1999). Syllable detection and segmentation using temporal flow neural networks. In *Proceedings of the 14th International Congress of Phonetic Sciences,* pp. 1721-1724.

Shepard, R. N. (1972) Psychological representation of speech sounds. In E. E. David and P. B. Denes (Eds.) *Human Communication: A unified view.* New York: McGraw-Hill, 67–113. Measuring perceptual distance from a confusion matrix.

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, 102, 250-255.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: how important is directed speech? *Developmental Science*, 15, 659-673.

Silipo, R., & Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. "The Phonetics of Spontaneous Speech," ICPhS-99, San Francisco, CA, August.

Snowling, M.J. (2000). *Dyslexia* (2nd Edition), Oxford, U.K : Blackwell Publishers.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of Acoustical Society of America*, 68, 836–842.

Stone M.A., Moore B.C.J. (2003). Effect of the speed of a single-channel dynamic range compressor on intelligibility in a competing speech task. *Journal of Acoustical Society of America*, 114, 1023-1034.

Stone M.A. & Moore B.C.J. (2007). Quantifying the effects of fast-acting compression on the envelope of speech. *Journal of Acoustical Society of America*. 121, 1654-1664.

Sumby, W.H. & Polack, I. (1954) Perceptual amplification of speech sounds by visual cues. *Journal of Acoustical Society of America*, 26, 212–215.

Surányi, Z., Csépe, V., Richardson, U., Thomson, J.M., Honbolygó F., & Goswami, U. (2009). Sensitivity to rhythmic parameters in dyslexic children: a comparison of Hungarian and English. *Reading & Writing*, 22, 41–56.

Tamburini, F. (2003). Automatic prosodic prominence detection in speech using acoustic features: An unsupervised system. In *Proceedings of Eurospeech*, Vol 1, pp 129-132.

Tass, P., Rosenblum, M.G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., Schnitzler, A., & Freund, H.-J. (1998). Detection of n:m phase locking from noisy data: application to magnetoencephalography. *Physical Review Letters*, 81 , 3291-3294.

The puffin baby and toddler treasury (1998). London, England : Puffin.

This little puffin (1991). London, England : Puffin Books.

Thomson, J. , Fryer, B., Maltby, J. & Goswami, U. (2006). Auditory and motor rhythm awareness in adults with dyslexia. *Journal of Research in Reading*, 29, 334-348.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science*, 33, 839-879.

Tilsen, S. & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*, 124, EL34-39.

Todd, N.P.M. (1994). The auditory "primal sketch": a multiscale model of rhythmic grouping. *Journal of New Music Research*, 23, 25–70.

Todd, N.P.M. & Brown, G.J. (1996). Visualization of rhythm, time and metre. *Artificial Intelligence Review*, 10, 253–273.

Todd, N.P.M., Lee, C.S. & O'Boyle, D.J. (1999). A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research*, 28, 5–29.

Todd, N.P.M., O'Boyle, D.J., & Lee, C.S. (2002). A sensorimotor theory of beat induction and temporal tracking. *Psychological Research*, 66, 26–39.

Torgesen, J.K., Wagner, R.K., & Rashotte, C.A.. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.

Treves, A. & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7, 399–407.

Turner, R.E. (2010). *Statistical models for natural sounds*. Doctoral dissertation, University College London. Retrieved from : http://www.gatsby.ucl.ac.uk/~turner/Publications/Thesis.pdf

Turner, R.E. & Sahani, M. (2011). Demodulation as Probabilistic Inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 2398-2411.

Villing, R.(2010). *Hearing the moment: Measures and models of the perceptual centre.* Doctoral dissertation, National University of Ireland Maynooth. Retrieved from : http://eprints.nuim.ie/2284/1/Villing_2010_-_PhD_Thesis.pdf

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised.* New York, US : The Psychological Corporation.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. San Antonio, TX: The Psychological Corporation.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*,7, 49-63.

Whalley, K, & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading*, 29, 288-303.

Wilkinson, G.S. (1993). Wide range achievement test 3. Wilmington, DE: Wide Range.

Winkler, I., Denham, S.,L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, 13, 532-540.

Wolff, P.H. (2002). Timing precision and rhythm in developmental dyslexia. *Reading and Writing: An Interdisciplinary Journal*, 15, 179–206.

Xie, Z., & Niyogi, P. (2006). Robust acoustic-based syllable detection. In *Proceedings of Interspeech, 2006.*

Yehia, H., Kuratate T., & Vatikiotis-Bateson E. (2002). Linking facial animation, head motion, and speech acoustics, *Journal of Phonetics*, 30, 555-568.

Zhang, Y., & Glass, R. (2009). Speech rhythm guided syllable nucleus detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp.3797-3800. IEEE.*

Ziegler, J., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 1, 3-29.

Zion Golumbic E.M., Poeppel, D., Schroeder, C.E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain & Language*, 122, 151-161.

LIST OF APPENDICES BY CHAPTER

Maturation of the Auditory Cortex and Effects on Speech Perception

Human newborns enter the world with a relatively mature cochlea and brainstem. By the 29th fetal week (third trimester of pregnancy), click-evoked auditory brainstem potentials are already present, indicating that information is being conducted through brainstem pathways (Ponton, Moore & Eggermont, 1996). By contrast, the auditory cortex is immature at birth and undergoes a protracted period of postnatal development. To investigate the developmental changes occurring in the human auditory cortex at a histological (cell and tissue) level, Moore (2002) used immunohistochemical (antibody) staining of axonal filaments in postmortem brain tissue. These filaments are produced by neuronal axons at the time when they begin to function, and their production immediately preceeds myelination and rapid conduction. Hence, the immonolabelling of such axonal filaments marks the onset of function in a neuronal system. Based on the timing and pattern of expression of these neurofilaments, Moore (2002) reported that the maturation of the human auditory cortex appeared to occur in three developmental stages, each characterised by a different axonal system coming 'online'. These three periods are the perinatal period (third trimester to 4 months), the early childhood period (4.5 months to 5 years), and the late childhood period (5 years to 12 years).

Perinatal period (third trimester to 4 months). During this period, mature axons are only present in the most superficial layer of the cortex, the marginal layer. In this superficial layer, axons run parallel to the cortical surface for long distances, contacting the apical dendrite tips of neurons from deeper layers of the cortex. It is thought that marginal layer neurons drive the activity of these deeper cells, promoting their structural and functional maturation (Marin-Padilla & Marin-Padilla, 1982). However, the vast majority of the auditory cortex remains immature at this stage. Consequently, the auditory discrimination abilities demonstrated by infants under 4 months of age most likely derive from the analytical abilities of their mature cochlea and brainstem. And these abilities are remarkable. Newborn infants already show memory and recognition for voices (DeCasper & Fifer, 1980) and stories (DeCasper & Spence, 1986), and are able to differentiate languages with different rhythmic properties (Nazzi et al, 1998). At 1 month of age, infants are not only capable of making fine phonetic discriminations of voice onset time (VOT), but show categorical perception of VOT in a manner similar to adults (Eimas et al, 1971). By 2 months of age,

infants show sensitivity to syllable structure (Bertoncini & Mehler, 1981), and by 4 months of age, the neural mismatch responses of German and French infants already begin to reflect the prosodic patterns of their native language (Friederici et al, 2007). In terms of measurable brain responses during this stage, axons projecting from the inferior colliculus (brainstem) to the medial geniculate (thalamus) may contribute to the generation of the middle latency cortical potentials like the Po-Na complex, which is measurable at the time of birth (Moore & Linthicum, 2007). Longer-latency evoked potentials like the N2 and MMN are also measurable in infants at this age (e.g. Friederici et al, 2007). These slower components are likely to be generated by the marginal layer afferents in the cortex which are thin and only lightly myelinated.

Early childhood period (4.5 months to 5 years). From around 4.5 months to 1 year, thalamocortical afferents begin to develop. This system of axons carries input from the lower levels of the auditory system (ear and brainstem) to deep layers of the auditory cortex (layers 4,5 & 6). This network of afferents becomes progressively denser up to around 5 years of age. Commensurate with this, the Pa (peak latency 25-30 ms) and P1 (80-100ms) components become increasingly detectable in early childhood years, with the latency of the P1 gradually shortening over time. These components appear to be generated by activity in the deeper layers of the auditory cortex (Moore & Linthicum, 2007). At around this time, when deep thalamocortical afferents are beginning to develop, infants begin to show specialisation or 'tuning' for sound patterns in their native language. For example, 9-month-old English-learning infants begin to listen longer to trochaic prosodic foot patterns that are most common in English (Jusczyk et al, 1993). While younger English infants readily discriminate foreign phoneme contrasts such as the native-American /ki/-/qi/, by 10 months of age, English infants no longer make this discrimination (Werker & Tees, 1984). Hence, the thalamic input to the deeper cortical layers appears to play a role in 'tuning' the auditory system to respond preferentially to relevant (e.g. native) speech sounds, and to ignore irrelevant (e.g. foreign) speech sounds.

Late childhood period (5 to 12 years). At around 5 years of age, mature axons begin to appear in superficial cortical layers 2 & 3, reaching an adult-like density by 11-12 years. These axons represent cortico-cortical connections such as commisural axons that interconnect the cerebral hemispheres, and association fibres that connect different areas of the cortex. Coincident with the maturation of these upper superficial cortical layers, the N1 wave begins to appear by around age 6 to 8 (Ponton & Eggermont, 2001). During this last

stage of development, the ability to perceive speech in noise improves markedly (Elliott, 1979). Children are also become better able to maintain speech perception in conditions of binaural switching, interruption, filtering or spectral degradation (Marshall et al, 1979). Hence, maturation of layers 2 & 3 of the auditory cortex appear to be associated with robust coding of speech information in adverse conditions.

However, this account of the relatively late maturation of the auditory cortex is not universally accepted. For example, Dehaene-Lambert et al (2002, 2006) and colleagues (Leroy et al, 2011) argue instead for the early maturation of linguistic pathways, pointing to fMRI evidence that even 3-month-old infants show activation in the temporal lobes, inferior and dorsolateral frontal areas when engaged in a speech task. However, these neuroimaging results should be interpreted with caution because the presence of BOLD activation in a brain area does not necessarily imply an active role in online speech processing. Rather, according to Moore's account, neurons in the auditory cortex may simply be passively receiving information from subcortical areas in order to stimulate their development. Moreover, a closer examination of fMRI results indicates that the pattern of activation in infants is not fully adult-like in the first few months of life. For example, in the Dehaene-Lambert et al (2002) study, 3-month-old infants were played 20s spoken sentences either forward or backwards. Both types of sentences elicited a broad pattern of activation over the temporal lobes, with greater activation over the left than right temporal lobe. However, when the activation pattern for forward speech was compared to backward speech, infants showed greater activation in the left angular gyrus (parietal lobe), but not in the temporal lobe for forward as compared to backwards speech. In contrast, adults showed greater activation for forward speech over the left superior temporal sulcus (temporal lobe). The parietal activation in infants is consistent with a memory retrieval explanation, since a similar region is activated in adults when performing memory retrieval of words (Shallice et al, 1994). Hence, at 3 months of age, infants appeared to differentiate forwards and backwards speech on the basis of stimulus familiarity (engaging the parietal cortex), rather than on the basis of linguistic potential (which should engage the auditory cortex).

The infant fMRI evidence thus indicates that while the auditory cortex does indeed respond to auditory input from as early as 3 months of age, its activation pattern does not differ in response to speech with an intact or reversed temporal structure. This is consistent with the explanation of an immature auditory cortex that is receiving input, but whose response is as yet 'un-tuned' to linguistically-relevant or meaningful sounds. As such, it is

debatable whether or not the auditory cortices play any active role in processing speech at this early stage. To shed light on this issue, it would be interesting to conduct a comparative fMRI study of infants younger and older than 4.5 months of age, to look at the patterns of neural activation associated with language-specific 'tuning' (eg. loss of non-native categorical phoneme discrimination). According to the late development hypothesis, patterns of auditory cortical activation in younger infants should show no difference for native and non-native phoneme contrasts. However, in older infants, native phoneme contrasts should elicit stronger activation over temporal cortices than non-native phoneme contrasts.

Journal of Memory and Language 64 (2011) 59–73

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Rise time perception and detection of syllable stress in adults with developmental dyslexia

Victoria Leong, Jarmo Hämäläinen, Fruzsina Soltész, Usha Goswami *

Centre for Neuroscience in Education, University of Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 6 April 2010
revision received 17 August 2010

Keywords:
Dyslexia
Syllable stress
Rise time
Rhythm
Phonology

ABSTRACT

Introduction: The perception of syllable stress has not been widely studied in developmental dyslexia, despite strong evidence for auditory rhythmic perceptual difficulties. Here we investigate the hypothesis that perception of sound rise time is related to the perception of syllable stress in adults with developmental dyslexia.
Methods: A same-different stress perception task was devised and delivered to a sample of 40 adults in two formats, one using pairs of identical 4-syllable words and one using pairs of two different 4-syllable words. Auditory perception of rise time, frequency and intensity, and phonological awareness, phonological memory and reading were also measured.
Results: We show that adults with dyslexia performed significantly more poorly in both versions of the stress perception task. Individual differences in the perception of rise time were linked to the accuracy of performance.
Conclusions: To our knowledge this is the first direct demonstration of syllable stress perception deficits in dyslexia. The accurate perception of intonational patterning and rhythm may be critical for the development of the phonological lexicon and consequently for the development of literacy. Even high-functioning compensated adults with dyslexia show impairments in speech processing.

Introduction

Developmental dyslexia is a neurodevelopmental condition found across languages, for which the cognitive hallmark is impaired phonological processing (Snowling, 2000; Ziegler & Goswami, 2005). Evidence that this hallmark "phonological deficit" is related to impaired basic auditory processing has been accumulating during the last decade, in studies of both alphabetic and non-alphabetic languages. The auditory parameter most consistently found to be impaired has been perception of the amplitude envelope onset (rise time), or its correlate, amplitude modulation depth (Corriveau, Pasquini, & Goswami, 2007; Goswami, Fosker, et al., 2010; Goswami, Gerson, & Astruc, 2009; Goswami et al., 2002; Goswami, Wang, et al., 2010; Hämäläinen, Leppänen, Torppa, Muller, & Lyytinen, 2005; Hämäläinen, Salminen, & Leppänen, in press; Hämäläinen et al., 2009; Lorenzi, Dumont, & Fullgrabe, 2000; Muneaux, Ziegler, Truc, Thomson, & Goswami, 2004; Pasquini, Corriveau, & Goswami, 2007; Richardson, Thomson, Scott, & Goswami, 2004; Rocheron, Lorenzi, Fullgrabe, & Dumont, 2002; Suranyi et al., 2009; Thomson, Fryer, Maltby, & Goswami, 2006; Thomson & Goswami, 2008). Behaviourally, rise time is most closely associated with the perceptual experience of speech rhythm and stress (Hoequist, 1983; Morton, Marcus, & Frankish, 1976). However, to date, there has been no investigation of the possible relationship between basic auditory processing of rise time and the perception of syllable stress in spoken words in dyslexia. A clear prediction of the "rise time" theory of developmental dyslexia (Goswami et al., 2002) is that the perception of syllable stress should be impaired in individuals with dyslexia, and that individual differences in rise

* Corresponding author. Address: Centre for Neuroscience in Education, 184 Hills Rd., Cambridge CB2 8PQ, United Kingdom. Fax: +44 1223 767602.
E-mail address: ucg10@cam.ac.uk (U. Goswami).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.09.003

time perception should predict the severity of any impairment in perceiving syllable stress.

Despite the lack of direct evidence for a stress perception impairment in dyslexia, recent studies using reiterative speech tasks are consistent with the prediction that individuals with developmental dyslexia should be impaired in perceiving syllable stress. Regarding adults with developmental dyslexia, Kitzen (2001) developed a reiterative speech task in which each syllable in a word was converted into the same syllable (here DEE). This enabled distinctive phonetic information in words and phrases to be removed while retaining the stress and rhythm patterns of the originals. Kitzen converted film and story titles into "DeeDees", so that (for example) "Casablanca" became DEEdeeDEEdee (STRONG weak STRONG weak, or SWSW). Adolescent participants with dyslexia heard a tape-recorded DeeDee sequence while viewing three alternative (written) choices, for example "Casablanca", "Omega Man" and "The Godfather". Kitzen found that her participants with dyslexia were significantly poorer in this choice task than age-matched controls. She also reported that performance in the DeeDee measure was significantly associated with syllable and phoneme segmentation skills, and with word reading abilities and reading comprehension. In logistic regression analyses carried out to predict group membership (dyslexic versus control), the DeeDee measure was a highly significant predictor of group status (along with syllable segmentation and rapid naming measures). All three measures together predicted group membership with 97% accuracy (phoneme segmentation was not a significant predictor). However, one drawback with this study was the use of written response choices for participants who had difficulties in processing written language.

Goswami et al. (2009) developed two DeeDee tasks suitable for children with dyslexia, which avoided reading demands (see also Whalley & Hansen, 2006). In their tasks, children saw a picture of a "famous name" familiar to British participants (such as the English footballer David Beckham) or pictures corresponding to familiar film and book titles (such as Harry Potter). Familiarity with the pictures was assessed in a pretest. During the experimental sessions, the children were asked to select which of two "DeeDee" phrases that they heard matched the picture. For example, the correct match for "Harry Potter" was DEEdeeDEEdee (SWSW). Goswami et al. reported that the children with dyslexia (who were aged on average 12 years) performed significantly more poorly than age-matched controls in both the 'Film and Book Titles' and 'Famous Names' DeeDee tasks. Performance in the DeeDee tasks was also a significant predictor of reading development in the sample, for example individual differences in the 'Famous Names' task accounted for 25% of unique variance in reading accuracy after controlling for age and IQ. DeeDee perception predicted reading even when phonological awareness (performance in a rhyme oddity task) was additionally controlled (still accounting for 16% of unique variance, $p < .001$). Finally, individual differences in measures of the auditory perception of rise time predicted unique variance in the reiterative speech task.

One drawback of reiterative speech tasks is that they require participants to derive an abstract representation of the stress patterning of a particular utterance rather than to perceive the stress patterns in the utterance directly. Studies of direct stress perception in non-dyslexic adults have used a variety of experimental paradigms, including visual and auditory lexical decision, shadowing tasks, speech gating tasks, and word recognition of mis-stressed words (see Cutler (2005), for a recent review). As discussed by Cutler (2005), prior information about stress patterning does not seem to facilitate lexical access in English, although in some studies stress information helps to resolve lexical competition. For example, Cooper, Cutler, and Wales (2002) showed using a fragment priming task that information about syllable stress helped listeners to assign initial syllables to source words such as *admiral* versus *admiration*. The adults heard sentences like "The speech therapist said." and then had to make a lexical decision about the target words (e.g., admiral/admiration). The auditory primes were fragments of complete words pronounced with either first syllable stress ("ADmir") or third syllable stress ("admir" from *admiration*). Cooper et al. reported that a fragment like "ADmir" activated *admiral* more than *admiration*, while a fragment like "admir" activated *admiration* more than *admiral*. Their conclusion was that English adults do make use of suprasegmental information in recognising spoken words. Slowiaczek (1990) asked participants to listen to spoken words that were mixed with white noise and were presented with either correct stress (e.g., SPECulative) or incorrect stress (specUlative). Participants had to write down what they heard and were credited for accurate word recognition. Slowiaczek found no effects of mis-stressing in this recognition task. However, when she asked participants to shadow what they heard in a subsequent experiment, there was an effect of mis-stressing on response *speed*. Participants were slower to produce the mis-stressed words, suggesting that lexical stress is coded as part of the phonological representation.

As the cognitive difficulty in developmental dyslexia lies in the accurate neural representation of the phonological information in words, stress perception may be expected to play an important role in the development of well-specified phonological representations. English is a free-stressed language, as prominence may occur on different syllables, falling at different positions in different words (as in "orNATE" for the isolated word versus "ORnate BALcony" for continuous speech). Studies of early phonological development in English suggest that infants and very young children adopt a primarily lexical strategy to stress placement, that is they learn stress as part of the phonological representation of a particular word (e.g., Klein, 1984). However, many English words used with infants and young children follow a strong–weak pattern (*mummy*, *daddy*, *baby*, *doggie*), and so it is possible that template learning plays a role in the development of knowledge about stress. In general, strong syllables are louder and longer than weak syllables, and have a higher pitch (frequency). Jusczyk, Houston, and Newsome (1999) reported that infants could segment words with strong–weak patterns by 7½ months of age, but appeared to mis-segment words following a weak–strong pattern. For example, if the infants heard a sentence such as "*her guitar is too fancy*", they segmented "*taris*" as a

plausible word (treating "*taris*" rather than "*guitar*" as familiar during the dishabituation test). By 10½ months of age, infants did not make these mistakes. Sensitivity to the predominant stress patterns of English words is clearly important for segmenting words and syllables from the speech stream, and therefore for phonological representation (see also Echols, 1996; Mattys & Juscyk, 2001).

Recent theories of developmental phonology have also suggested an important role for prosodic sensitivity in explaining phonological development (Gerken, 1994; Pierrehumbert, 2003; Vihman & Croft, 2007). For example, Pierrehumbert (2003) argued for early-acquired "prosodic structures" as the basis for language acquisition, proposing a model based on the acquisition of complex language-specific exemplars from the input that were stored in rich phonetic and prosodic detail (see also Port, 2007). She argued that phonetic perception is dependent on the prosodic context. Indeed, stress perception studies with both children and adults have suggested that target phonemes are detected more efficiently when they are in stressed syllables (e.g., Mehta & Cutler, 1988; Wood & Terrell, 1998). Therefore, current evidence suggests that stress is an integral part of the phonological representations of English words developed by infants, and that phonological development is characterised by an inter-dependency of phonetic and prosodic information.

It thus seems plausible to propose that the phonological difficulties experienced by children and adults with developmental dyslexia must involve reduced sensitivity to stress and intonational patterning as well as reduced sensitivity to phonological units like syllables, onsets, rimes and phonemes. As noted, the auditory correlates of stress are most usually defined as involving amplitude, duration and frequency. Classical theories (e.g., Fry, 1954) accorded fundamental frequency the key role in stress perception, with duration and intensity (amplitude) playing secondary roles. More recent investigations using natural speech have shown that amplitude and duration cues play a stronger role in prosodic prominence than fundamental frequency (Choi, Hasegawa-Johnson, & Cole, 2005; Greenberg, 1999; Kochanski, Grabe, Coleman, & Rosner, 2005). For example, Greenberg (1999) described an automatic prosodic algorithm developed to label stressed and unstressed syllables in a corpus of spontaneous speech. The algorithm depended on three separate parameters of the acoustic signal, duration, amplitude and fundamental frequency. In contrast to classic accounts, Greenberg reported "fundamental frequency turns out to be relatively unimportant for distinguishing between the presence and absence of prosodic prominence... the results indicate that the product of amplitude and duration ... yields the performance closest to ... linguistic transcribers" (p. 172). Similar conclusions were reached by Kochanski et al. (2005) in an investigation of a large corpus of natural speech covering 7 English dialects.

Greenberg (2006) has explicitly linked changes in rise time to prosodic prominence by proposing a theory of how the "energy arc" of speech (the linguistic manifestation of the energy arc is the syllable) is produced by manner of articulation. By this account, the energy contour of the speech signal is an arc rising to a peak in the nucleus of each syllable and then descending. Rise time (the rate of change in intensity or signal energy as the nucleus of the syllable is produced by the articulators) should be particularly critical for stress perception. The specific way in which the arc ascends to the peak depends on whether the syllable is stressed (here more energy is produced) and the phonetic composition of the syllable onset – with more sonorous onsets, speakers take longer to reach the peak. Prosody thus affects both the height and length of the energy contour, and so the amplitude envelope of speech reflects the prosodic properties of speech.

Loudness (amplitude) perception per se is not usually impaired in studies of auditory processing in developmental dyslexia. Rather, perception of the *rate of onset* of changes in amplitude (rise time) is impaired. For example, the different cohorts of children with developmental dyslexia tested by Richardson et al. (2004), Thomson and Goswami (2008) and Goswami et al. (2009) did not exhibit significantly raised auditory thresholds for amplitude compared to age-matched controls in two forms of a two-interval forced choice (2IFC) task. In one version of this intensity threshold task, the children were asked to judge which of two sounds A and B was softer (Richardson et al., 2004). In the second version, the children heard two sequences of five sounds (AAAAA versus ABABA), and had to detect which sequence varied in intensity (Goswami et al., 2009; Thomson & Goswami, 2008). Group thresholds for intensity discrimination were statistically equivalent for children with dyslexia and age-matched controls in all three studies. Nevertheless, individual differences in the ABABA intensity discrimination task were predictive of performance in the "Film and Book Titles" reiterative speech task, an indirect measure of sensitivity to syllable stress, accounting for 18% of unique variance after controlling for age and IQ (Goswami et al., 2009). Similarly, in the Thomson and Goswami (2008) study, intensity discrimination was significantly correlated with performance in a Tempi discrimination task even when non-verbal IQ was controlled (the Tempi task asked children to judge which of two cartoon bears playing trumpets were producing notes at a slower pulse rate, $r = .39$, $p < .01$). Therefore, if outcome measures involve an element of periodicity, as in the DeeDee task and in Tempi detection, intensity discrimination may be a significant predictor of individual differences in addition to rise time. The relationship of intensity discrimination to perceiving syllable stress patterns in multi-syllabic words remains to be tested (although see Foxton, Riviere & Barone, 2010 for an audio-visual stress recognition task in which amplitude perception did play a role in detecting *visual* prosody).

These relationships between simple intensity discrimination and periodicity are consistent with a more recent study of developmental dyslexia using a musical metrical perception task based on simple tunes comprised of strong and weak "beats" (Huss, Verney, Fosker, Fegan, & Goswami, 2010). In this musical study, children with dyslexia aged 10 years and control children were asked to judge whether two short tunes were the same or different in metrical structure. The tunes varied in metrical complexity (e.g., a 6-note tune in duplex time with takt on the first note, versus a 15-note tune in 4/4 time with takt on the

second note). Huss et al. found that the children with dyslexia were impaired in perceiving metrical similarity irrespective of the metrical complexity of the different tunes. The severity of the children's metrical perceptual difficulties was uniquely predicted by performance in only two of the basic auditory processing tasks that were administered, rise time discrimination and intensity discrimination. Pitch and duration thresholds did not predict unique variance in the metrical perception task in block-entry multiple regression equations, despite the fact that metrical dis-similarity depended on inserting longer durations between adjacent musical notes. As metrical structure is a focus of interest in linguistic studies of syllable stress, with metrical structure accorded an important organisational role in determining syllable, word and clausal boundaries, the difficulties of individuals with dyslexia in metrical perception are again consistent with the difficulty hypothesised here in perceiving syllable stress in dyslexia. In fact, metrical perception accounted for 42% of unique variance in reading in the musical metre study, making it a stronger predictor of reading development in this sample of children than phonological awareness.

Accordingly, we assume here that very basic auditory processes are used in perceiving metrical structure in both music and language, and that individual differences in these basic auditory processes affect individual differences in the extraction of periodic structure and accordingly the perception of syllable stress in speech. To test this hypothesis, we measured basic auditory processing in a sample of adults with and without developmental dyslexia, and we also measured stress perception in a same-different task based on 4-syllable words. From our analysis of over 2500 4-syllable words drawn from the CELEX database, we found that 4-syllable words in English most commonly receive primary stress on the second syllable. Forty-four per cent of words (like *maternity* and *ridiculous*) conform to this typical stress template, which can be denoted as '0200'. The remainder of words either received primary stress on the first syllable (24%), as in *difficulty* and *military* (2000 stress template), or on the third syllable (28%), as in *comprehensive* and *interaction*. These words also had secondary stress on the first syllable (1020 stress template).

In the current study, we used only words with first or second syllable stress, that is, 2000 or 0200 template words. We recorded a female British speaker saying tokens of each type of word (2000 or 0200 template) with either correctly or incorrectly placed stress. For example, two tokens of the word *maternity* were recorded, one with correctly placed stress as in *maTERnity* (WSWW) and the other with incorrectly placed stress as in *MAternity* (SWWW). We then paired these tokens in all four possible ways (SWWW–SWWW, WSWW–WSWW, SWWW–WSWW, WSWW–SWWW). Participants were asked to judge whether the two tokens in the pair contained the "same" or "different" stress patterns. We also varied whether the spoken token was the same word in each pair (as in *maternity–maternity*, Experiment 1, thereby keeping segmental phonology constant), or was two different words with matching syllable stress templates (as in *maternity–ridiculous*, Experiment 2, thereby conceptually more similar to the DeeDee task, in that abstract stress templates must be compared to make a judgement). We were interested to see whether participants with dyslexia would find it more difficult to make judgements about shared syllable stress in each experiment. Note that in both experiments, stress pattern similarity can nevertheless be judged "on-line" using the acoustic information in the heard tokens, without recourse to the mental lexicon.

Experiment 1

Method

Participants

Twenty adults with developmental dyslexia (11 male; mean age 25.3 years, range 17.5 years – 41.8 years) and twenty adults without dyslexia (7 male; mean age 26.3 years, range 18.1 years – 38.5 years) participated in the study. Eighteen of the adults with dyslexia had a formal statement of developmental dyslexia, the remaining two participants showed severe literacy and phonological deficits according to our own test battery which was administered to all participants. As phonological deficits were part of the inclusion criteria for the study, it is possible that participants whose difficulties were visual and not phonological were excluded from the sample. All participants had no diagnosed additional learning difficulties (e.g. dyspraxia, ADHD, autistic spectrum disorder, speech and language impairments) and spoke English as a first language. Participant details are shown in Table 1. All participants took part in both Experiments 1 and 2 on separate days, with Experiment 1 being performed first.

Tasks

Standardised ability tests. All participants were given 2 subscales of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), a non-verbal subscale (Block Design) and a verbal subscale (Vocabulary). Literacy skills were assessed using the untimed Wide Range Achievement Test (Reading and Spelling scales, WRAT-III, Wilkinson, 1993). A measure of short-term memory, the Weschler Adult Intelligence Scale-Revised forward digit span subtest was also administered (WAIS-R; Weschler, 1998).

Table 1
Participant details.

Group	Dyslexic	Controls	$F(1, 38)$
Chronological age (years)	25.3	26.3	.32
(*sd*)	(5.6)	(5.2)	
WRAT reading standard score	102.5	114.7	23.44***
(*sd*)	(10.0)	(5.3)	
WRAT spelling standard score	97.8	115.6	38.21***
(*sd*)	(11.3)	(6.2)	
WASI vocabulary subscale *T* score (mean = 50)	63.1	64.7	.64
(*sd*)	(6.3)	(5.9)	
WASI block design subscale *T* score (mean = 50)	59.0	61.0	.85
(*sd*)	(7.4)	(6.3)	
WAIS-R digit span subscale score (out of 16)	10.5	12.2	5.28*
(*sd*)	(2.6)	(2.0)	

* $p < .05$.
*** $p < .001$.

Phonological awareness measures.

i. *Spoonerisms.* This task was drawn from the Phonological Assessment Battery (PhAB; Fredrickson, Frith, & Reason, 1997). Participants heard 10 pairs of words presented orally by the experimenter. Participants were asked to swap the onset phonemes of the pair of words (e.g. for "*sad cat*"; subject responded "*cad sat*"). Scores on this measure were out of a possible 20 points.

ii. *RAN (Rapid Automatized Naming).* Two versions of an object RAN task designed originally for children were administered, one based on pictures of objects whose names resided in dense phonological neighbourhoods (RAN Dense: *Cat*, *Shell*, *Knob*, *Thumb*, *Zip*), and one based on pictures of objects whose names resided in sparse phonological neighbourhoods (RAN Sparse: *Web*, *Dog*, *Fish*, *Cup*, *Book*). Participants were shown a sheet of paper with the same pictures repeated 50 times. In each case, they were asked to produce the names as quickly and accurately as possible. Performance was timed, and the two tasks were combined to give an average RAN score in seconds.

Psychoacoustic tasks. The psychoacoustic stimuli were presented binaurally through headphones at 74 dB SPL. The auditory tasks were presented using an adaptive staircase procedure (Levitt, 1971) with a combined 2-up 1-down and 3-up 1-down procedure; after 2 reversals, the 2-up 1-down staircase procedure changes into 3-up 1-down. The step size halves after the 4th and 6th reversal. A test run typically terminates after 8 response reversals or alternatively after the maximum possible 40 trials. Four attention trials were randomly presented during each test run, using the maximum contrast of the respective stimuli in each auditory task. The threshold score achieved was calculated using the mean of the last four reversals.

i. *Amplitude Envelope Onset (Rise Time) Task (1 Rise).* This was a rise time discrimination task in AXB format. Three 800 ms tones were presented on each trial, with 500 ms ISIs. Two (standard) tones had a 15 ms linear rise time envelope, 735 ms steady state, and a 50 ms linear fall time. The third tone varied the linear onset rise time logarithmically with the longest rise time being 300 ms. Participants were introduced to three cartoon dinosaurs. It was explained that each dinosaur would make a sound and that the task was to decide which dinosaur's sound was different from the other two and had a softer rising sound (longer rise time, this was either sound A or B, never sound X). As an integral part of the software programme feedback was given after every trial on the accuracy of performance. Schematic depiction of the stimuli can be found in Richardson et al. (2004).

ii. *Frequency task.* This was a frequency discrimination task also delivered in an AXB format. The standard was a pure tone with a frequency of 500 Hz presented at 74 dB SPL, which had a duration of 200 ms. The maximum pitch difference between the stimuli presented in this task was 60 Hz. Participants were introduced to three cartoon elephants. It was explained that each elephant would make a sound and that the task was to decide which elephant's sound was higher.

iii. *Intensity task.* This was an intensity discrimination task delivered in a 2IFC format. The standard was a pure tone with a frequency of 500 Hz presented at 74 dB SPL, which had a duration of 200 ms. The intensity of the second tone ranged from 54 to 74 dB SPL. Participants were introduced to two cartoon mice. It was explained that each would make a sound, and the task was to decide which sound was softer. Participant's performance on phonological awareness and psychoacoustic tasks are shown in Table 2.

Syllable stress task. This task was based on 20 4-syllable words with lexical templates that had first syllable stress (2000, such as *caterpillar* and *difficulty*) and 20 4-syllable words with lexical templates that had second syllable stress (0200, such as *maternity* and *ridiculous*). The words were selected from an initial list of more than 2500 4-syllable words with first and second syllable-stress pooled from two linguistics databases (MRC Psycholinguistic Database and CELEX). The words were selected on the basis of syllable structure (no consonant clusters in the first two syllables), spoken and written frequency, and overall familiarity. Words also did not have alternative pronunciations. The full list of stimuli is presented as Appendix A. The words were divided into two lists of 20 words each (each list comprising 10 words with 2000 lexical templates and 10 words with 0200 lexical templates). Participants received one word list in Experiment 1 and the other in Experiment 2, which were given on separate days, with order of presentation of the word lists counterbalanced across participants. The two lists, and the two sets of lexical templates (2000, 0200), were matched as closely as possible for spoken and written frequencies. Mean values for 2000 templates were Cobuild spoken frequency 21.7

Table 2
Group performance in the phonological and auditory tasks.

Group	Dyslexic	Controls	$F(1, 36)$
Spoonerisms[a]	15.2	17.8	7.70[b,**]
(*sd*)	(3.2)	(2.3)	
RAN time in seconds	35.2	30.3	11.84[b,**]
(*sd*)	(4.5)	(4.0)	
Auditory threshold			
1 Rise in ms	63.0	40.3	11.50[**]
(*sd*)	(28.0)	(5.5)	
Frequency in Hz	12.5	9.1	5.20[*]
(*sd*)	(5.5)	(3.5)	
Intensity in dB	2.1	1.9	[c].80
(*sd*)	(0.9)	(0.4)	

[a] Score out of 20.
[b] Degrees of freedom are (1, 34).
[c] Degrees of freedom are (1, 32).
[*] $p < .05$.
[**] $p < .01$.

(*sd* 22) and written frequency 288.6 (*sd* 294.2). Mean values for 0200 templates were Cobuild spoken frequency 15.5 (*sd* 30) and written frequency 224.3 (*sd* 315.2). Neither difference was statistically significant, $F(1, 38)$ for spoken frequency = 0.44, $F(1, 38)$ for written frequency = 0.55.

All items were produced naturally by a native female speaker of British English and recorded for computerised presentation using Audacity and Praat software. Two spoken tokens were recorded for each word. In one token, the speaker emphasised only the first syllable of the word (producing a SWWW stress pattern). In the other token, the speaker emphasised only the second syllable of the word (producing a WSWW stress pattern). This resulted in a total of 80 spoken tokens from 40 words. Word pairs were then created for each trial by combining the two spoken tokens in all four possible ways. The recorded tokens were analysed for mean intensity, duration, amplitude rise time and F0. Mean values for unstressed or stressed first syllables (such as *ma* or *MA* in *maTERnity* and *MAternity* respectively) and stressed or unstressed second syllables (such as *TER* or *ter* in *maTERnity* and *MAternity* respectively) are shown in Table 3. The values shown confirm that the acoustic parameters differed consistently between stressed and unstressed syllables across different words on a paired samples *t*-test. On average, stressed syllables were higher in intensity and pitch, and had longer durations and slower rise times than unstressed syllables. These acoustic differences are illustrated in Fig. 1, which shows a 3D plot of the amplitude envelopes for the word pair *DIfficulty and diFFIculty*. To create the figure, sound stimuli were first bandpass filtered into 12 logarithmically-spaced channels spanning a frequency range from 100 to 4000 Hz. Each frequency channel was then demodulated individually to extract its amplitude envelope. The figure plots time on the *x*-axis, frequency on the *y*-axis, and amplitude on the *z*-axis. Marked with arrows on the plot are duration, onset rise and intensity changes for stressed and unstressed versions of the syllable '*ffi*'. Differences in the frequency profile (circled) are also apparent as the stressed '*FFI*' shows larger amplitudes in mid-frequency channels than the unstressed '*ffi*'.

Table 3
Acoustic parameters of stressed and unstressed syllables (mean across 40 words).

	Stressed	Unstressed	*t*(39)
First syllable manipulated	E.g. MA in *MAternity*	E.g. ma in *maTERnity*	
Median intensity in dB	73.2	71.2	4.89***
(*sd*)	(5.1)	(4.2)	
Duration in ms	181.4	148.1	5.58***
(*sd*)	(61.9)	(51.9)	
Amplitude rise time in ms	94.3	82.5	3.17**
(*sd*)	(35.3)	(33.8)	
Mean F0 in Hz	243.5	209.2	9.77***
(*sd*)	(23.3)	(15.6)	
Second syllable manipulated	E.g. TER in *maTERnity*	E.g. ter in *MAternity*	
Median intensity in dB	72.3	70.1	5.03***
(*sd*)	(4.3)	(4.6)	
Duration in ms	175.3	145.2	5.14***
(*sd*)	(58.9)	(50.3)	
Amplitude rise time in ms	95.5	79.2	3.10**
(*sd*)	(43.3)	(37.4)	
Mean F0 in Hz	241.8	199.3	11.64***
(*sd*)	(22.4)	(14.7)	

** $p < .01$.
*** $p < .001$.

During task presentation, participants simply heard a word pair where two word tokens were presented one after the other. Participants were told to make same-different judgments about the position of syllable stress in the pair (such as *MIlitary – miLItary* [different] or *MIlitary – MIlitary* [same]). There was a 500 ms interval between the words in a pair, and a 2000 ms interval between trials after a response was given. Participants responded by pressing right or left buttons on the keyboard. The side of the same/different buttons was randomised across participants. Participants were told to respond as quickly and accurately as possible after they saw a question mark appear on the screen, which appeared at the end of the second word. Reaction time was recorded as the time from the question mark appearing to the participant's response (correctly-answered trials only). Feedback on the correctness of the response was provided on each trial by showing either a 'happy' smiley cartoon icon (correct response), or a 'pirate' cartoon icon (incorrect response). Apart from the '?' prompt and feedback icons, the computer screen remained blank whenever auditory stimuli were being presented. Prior to starting the experiment, participants received four practice trials. Note that participants were instructed to judge whether the position of stress was on the same or different syllables, not whether the word tokens were correctly or incorrectly pronounced. Participants did not report any difficulty in understanding what judgement was required.

There were four possible types of word pairs which differed in stress position, SWWW–SWWW, WSWW–WSWW, both requiring "same" judgements, and SWWW–WSWW, WSWW–SWWW, both requiring "different" judgements. Examples of these pairs are given in Fig. 1.This factor is referred to as Same/Different Judgement. In Experiment 1, the words were either based on 10 tokens with first syllable stress lexical templates (e.g., *difficulty–difficulty*) or were based on 10 tokens with second syllable stress lexical templates (e.g., *maternity–maternity*). This factor is referred to as First/Second stress template. Combining this factor with the four types of word pairs created 80 trials, which were fully randomised and presented in two 40-trial blocks. The experiment therefore used a $2 \times 2 \times 2$ design (Group $\times$ First/Second $\times$ Same/Different Judgement). The experimental design is summarised in Fig. 2.

Results

Auditory discrimination and phonological awareness data were explored by group to check that assumptions of normality (skew and kurtosis) were met. The SPSS boxplot function was used to check for outliers, and any data points lying farther than three interquartile ranges from

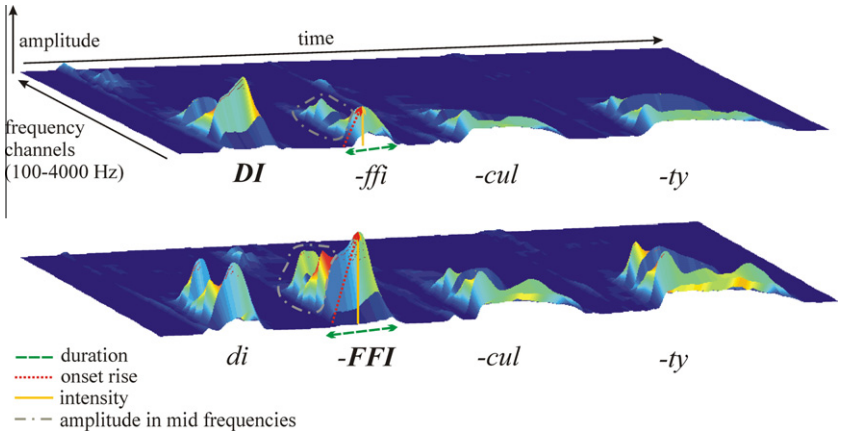


Fig. 1. Amplitude envelope across frequencies for the word *difficulty* produced with stress on the first or second syllable.

	Lexical template (First/Second)	Trial type (Same/Different)	Examples
1	First syllable stress template (2000)	SAME	*DIfficulty – DIfficulty* *diFFIculty – diFFIculty*
2		DIFFERENT	*DIfficulty – diFFIculty* *diFFIculty – DIfficulty*
3	Second syllable stress template (0200)	SAME	*maTERnity – maTERnity* *MAternity – MAternity*
4		DIFFERENT	*maTERnity – MAternity* *MAternity – maTERnity*

Fig. 2. Schematic depiction of design of Experiment 1.

the further edge of the box were removed. Five outlier scores were identified and removed for the auditory processing tasks (2 control scores for 1 Rise, 2 dyslexic scores for Frequency, 1 control score for intensity). Group data for the standardised tasks is provided in Tables 1 and 2. As would be expected given previous work, the participants with dyslexia were significantly less sensitive to auditory rise time and to frequency than their controls, but were not significantly different for intensity discrimination. Participants with dyslexia were significantly impaired in all the reading measures, and were also significantly impaired in the phonology measures. These differences were established using a series of one-way ANOVAs ($N = 40$), and F and p values are reported in Tables 1 and 2.

Mean performance (% correct and reaction time) for making judgements about shared syllable stress in each condition, as well as calculated d' and criterion values are shown in Table 4. Preliminary analyses confirmed that reaction times did not differ between groups and response time is not analysed further. Paired t-tests for d' and c values revealed significant group differences on both measures. Participants with dyslexia showed a significantly lower sensitivity (d') than controls ($t(1, 38) = 2.7$, $p = .01$) on the task. They were also more biased toward giving a 'same' response than controls ($t(38) = -3.2$, $p = .004$). This indicates that participants with dyslexia had more difficulty detecting acoustic differences between two items that were stressed differently, sometimes mistaking them as having the same stress pattern.

In order to check the effects of varying the syllable template, a 2 × 2 ANOVA (Group × First/Second syllable stress) was carried out, taking d' as the dependent variable. As would be expected, this showed a significant main effect of Group, $F(1, 38) = 7.3$, $p = .010$. However, the effect of

Table 4
Group performance on the stress perception task in Experiment 1: Mean% correct, Mean RT, d' and c (*sd* in parentheses).

	% Correct		RT in ms	
	Dyslexic	Control	Dyslexic	Control
First syllable stress template (2000)				
Same judgement	98 (3.0)	98 (4.1)	1085 (233)	1046 (292)
Different judgement	94.8 (8.7)	99 (2.1)	1069 (224)	1040 (312)
Second syllable stress template (0200)				
Same judgement	98.3 (3.4)	98.8 (2.2)	1068 (231)	9882 (286)
Different judgement	92.3 (7.9)	98.5 (2.9)	1044 (207)	1051 (309)
d' (sensitivity)	4.3 (0.6)	4.7 (0.3)		
Criterion (bias)	0.2 (0.3)	0.0 (0.1)		

First/Second stress template was not significant, $F(1, 38) = .81$, $p = .372$, and there was no interaction between First/Second syllable stress and Group, $F(1, 38) = 1.2$, $p = .281$. The results suggest that the participants with dyslexia found it difficult to judge shared stress when an identical item was pronounced with two different stress patterns, whether the stress template was SWWW or WSWW.

In order to examine whether these stress perception difficulties were related to inefficiencies in auditory perception, multiple regression analyses were used. Three 2-step fixed order equations were computed, all entering Group at Step 1 and then either rise time threshold, frequency threshold or intensity threshold at Step 2. The dependent variable in each case was d'. The results are shown in Table 5. As can be seen, rise time discrimination contributed 24% of unique variance to judgements about syllable stress. Frequency and intensity discrimination did not contribute significant unique variance to stress judgements, even though frequency discrimination also differed significantly between the two groups of participants. The data suggest a unique relationship between basic auditory perception of rise time and the accurate perception of syllable stress in speech.

The results from Experiment 1 are thus very consistent with the predictions that were made *a priori* on the basis of experiments using metrical musical perception tasks and reiterative speech tasks with participants with dyslexia. High-functioning adults with dyslexia showed difficulties in the auditory perception of rise time and difficulties in perceiving syllable stress. Individual differences in rise time perception predicted individual differences in stress perception. However, as the two spoken items to be judged were identical, the task was rather easy for all the participants. We therefore repeated the experiment using different real word tokens in the same stress perception task. Using different words increases the cognitive load of the task, as differences in segmental phonology must be ignored, making it likely that abstract stress templates must be extracted and compared. Experiment 2 therefore measures more than stress perception per se, and is conceptually more similar to the reiterative speech (DeeDee) task in requiring a more abstract stress-based comparison.

Table 5
Unique variance (R^2 change) in the syllable stress task in Experiment 1 (d') in 2-step fixed entry regression equations.

Step	Beta	R^2 change
1. Group	−.40	.16*
2. Rise time	−.56	.24**
2. Frequency	−.02	.00
2. Intensity	−.25	.06

Beta = standardized Beta coefficient; R^2 change = unique variance accounted for at each step of the 2-step fixed entry multiple regression equations.
* $p < .05$.
** $p < .01$.

Experiment 2

Participants and tasks were as in Experiment 1, but the stress judgement task was based on pairs of two different words.

Syllable stress task

In Experiment 2, the words were 10 pairs of non-identical tokens created by pairing the 20 items from Experiment 1. Five pairs had first syllable stress templates (2000, e.g., *difficulty–voluntary*), and the other five pairs had second syllable stress templates (0200, e.g., *maternity–botanical*). This factor is referred to as First/Second. The pairs again either had the same stress (SWWW–SWWW or WSWW–WSWW) or different stress (SWWW–WSWW or WSWW–SWWW). This factor is referred to as Same/Different Judgement. Word pairs were presented in both possible orders (e.g. *difficulty–voluntary* and *voluntary–difficulty*). This resulted in a total of $10 \times 2 \times 4 = 80$ experimental trials. The experiment was again based on a $2 \times 2 \times 2$ design (Group × First/Second × Same/Different Judgement). Fig. 3 shows a schematic depiction of the design of Experiment 2, and also provides examples of the word pairs used. Word pairs were selected to have similar spoken frequencies. Appendix B provides the full list of word pairs presented.

As this second syllable stress task was substantially more difficult for participants, we added filler items containing novel pairings to discourage the use of memory strategies. These filler items comprised 20 additional easy 'catch' trials containing pairs of the same word (e.g. DIfficulty–DIfficulty as in Experiment 1), and 20 additional trials containing novel pairings of words with different lexical stress templates (e.g. DIfficulty–deMOcracy). These

	Lexical template (First/Second)	Trial type (Same/Different)	Examples
1	First syllable stress template (2000)	SAME	*DIfficulty – VOluntary* *diFFIculty – voLUNtary*
2		DIFFERENT	*DIfficulty – voLUNtary* *diFFIculty – VOluntary*
3	Second syllable stress template (0200)	SAME	*maTERnity – boTAnical* *MAternity – BOtanical*
4		DIFFERENT	*maTERnity – BOtanical* *MAternity – boTAnical*
5		Catch trials (not included in analysis)	*DIfficulty-DIfficulty* *DIfficulty-deMOcracy*

Fig. 3. Schematic depiction of design of Experiment 2.

novel pairs were included to reduce the likelihood that participants would use a strategy of relying on memory for the exact word pairs that had already been presented, rather than making judgments based on the actual stress pattern of the words. These 40 extra trials were not included in the analyses. There were thus 120 trials in total in Experiment 2, fully randomised and presented in 5 blocks of 24 trials each.

Results

Mean performance (% correct and reaction time) in each condition, and overall d' and criterion values are shown in Table 6. As can be seen, control performance on average was above 80% correct for all conditions, but the participants with dyslexia performed at a much lower level. Reaction time was again very similar across groups, and no differences by Group in response times were found in preliminary analyses. Response time is not analysed further. Paired *t*-tests for d' and c values revealed significant group differences for sensitivity, but not for criterion bias. Participants with dyslexia again showed a significantly lower sensitivity (d') than controls ($t(38) = 5.9$, $p < .001$). However, there was no significant difference in the response bias of both groups, indicating that neither group was more biased toward giving a 'same' or 'different' response. The d' measure from Experiment 1 was highly correlated with the d' measure from Experiment 2 ($r = 0.56$, $p < .001$).

In order to explore the effects of the experimental manipulations, a 2×2 ANOVA (Group $\times$ First/Second

Table 6
Group performance on the stress perception task in Experiment 2: Mean% correct, Mean RT, d' and *c* (*sd* in parentheses).

	% Correct		RT in ms	
	Dyslexic	Control	Dyslexic	Control
First syllable stress template (2000)				
Same judgement	64 (13.9)	88 (10.8)	2100 (674)	1783 (669)
Different judgement	59.8 (17.7)	85.3 (18.8)	2303 (765)	1832 (686)
Second syllable stress template (0200)				
Same judgement	68.3 (13.3)	86.5 (12.4)	2089 (817)	1787 (665)
Different judgement	51.8 (19.0)	82.3 (19.2)	2311 (794)	1936 (772)
d' (sensitivity)	1.2 (0.9)	3.2 (1.2)		
Criterion (bias)	0.1 (0.3)	0.0 (0.3)		

stress) was again carried out, taking d' as the dependent variable. The ANOVA showed a significant main effect of Group, $F(1, 38) = 38.1$, $p = .000$, but no significant main effect of First/Second stress, $F(1, 38) = 2.0$, $p = .161$, and no significant interaction between First/Second stress × Group, $F(1, 38) = .02$, $p = .898$. Overall, as in Experiment 1, Experiment 2 found significantly less accurate performance by individuals with dyslexia, irrespective of the stress judgement (SWWW, SWSS) required.

To explore whether individual differences in basic auditory processing contributed to individual differences in making judgements about syllable stress when two different words had to be compared, multiple regression analyses were again used. Three 2-step fixed order equations were again computed, again entering Group at Step 1 and rise time threshold, frequency threshold or intensity threshold at Step 2. The dependent variable was d'. The results are shown in Table 7. As can be seen, rise time discrimination contributes 5% of unique variance to the accuracy of judgements about syllable stress, a finding which approached significance ($p = .07$). Neither frequency discrimination nor intensity discrimination contributed unique variance (0% and 1% respectively). As d' was significantly related in the two experiments, we also present analyses for average d' in Table 7. Average d' is a measure of stress sensitivity across the two experiments combined. As Table 7 shows, rise time was the only significant predictor of individual differences in making judgements about syllable stress, even when Group was controlled as a factor.

Finally, we were interested in the relationships between performance in the stress perception tasks (assessed via d' in Experiments 1 and 2, and the average d' measure) and performance in the literacy, phonology and language measures. The full correlation matrix is shown in Table 8. Table 8 shows that prosodic sensitivity as measured by the stress perception tasks is significantly related to individual differences in reading, spelling, phonological skills and RAN. The correlations suggest that stress processing is related to both phonological and literacy performance in this sample, although the direction of causation cannot be assessed. However, it is possible to use logistic regression to predict each individual's group membership (control or dyslexic) on the basis of their performance on these different measures. Therefore, a backwards stepwise logistic regression analysis was conducted. The regression model was initialised with four predictor variables – reading, phonology (Spoonerisms), average d' across both syllable stress experiments, and rise time threshold. As will be recalled, the groups differed significantly on all four of these variables. In the backwards method, predictors that do not contribute a significant change to the likelihood ratio statistic are removed sequentially until only significant predictors remain in the model. Table 9 shows the results from this first set of logistic regressions. Only two predictors for group membership were retained in the final model – syllable stress and reading. Of these, syllable stress was the stronger predictor, contributing a larger change to the likelihood ratio statistic. In contrast, phonology and auditory perception were not retained in the model as significant predictors of group membership. Having identified syllable stress perception and reading as the strongest predictors for group membership, a second stepwise logistic regression was conducted using only these variables. Syllable stress (average d') was entered as the first step since this was the strongest predictor in the backward model. Reading was entered as the second step. As shown in Table 10, syllable stress alone correctly predicted group membership for 80% of participants. Adding reading to the regression model improved the accuracy of predictions to 87.5%. Overall, these data suggest that stress perception is a more persistent discriminator of dyslexic difficulties than phonological or auditory measures, at least when participants are high-performing and well-compensated dyslexics, as was the case for the current sample.

Table 7
Unique variance (R^2 change) in the syllable stress task in Experiment 2 (d', see 7A) and in both experiments combined (average d', see 7B) explained by the basic auditory processing measures in 2-step fixed entry regression equations.

	Beta	R^2 change
7A		
1. Group	−.69	.48***
2. Rise time	−.25	.05[a]
2. Frequency	−.06	.00
2. Intensity	.09	.01
7B		
1.Group	−.68	.46***
2. Rise time	−.37	.10**
2. Frequency	−.05	.00
2. Intensity	.02	.00

Beta = standardized Beta coefficient; R^2 change = unique variance accounted for at each step of the 2-step fixed entry multiple regression equations.
** $p < .01$.
*** $p < .001$.
[a] $p = .07$.

Table 8
Raw correlation matrix for Experiment 1 d', Experiment 2 d' and the average d' measure.

	Experiment 1 d'	Experiment 2 d'	Av. d'
Age	.09	.15	.14
NVIQ	.18	.33*	.31*
VIQ	.35*	.14	.22
Rise thresh	−.62**	−.53**	−.61***
Frequency thresh	−.12	−.30	−.28
Intensity thresh	−.30	−.02	−.09
Spoonerisms	.11	.57***	.50**
RAN	−.35*	−.46**	−.47**
Reading	.39*	.53**	.54***
Spelling	.27	.53**	.51**
Digit span	.32*	.50**	.50**

Note: Expt = Experiment; NVIQ = non-verbal IQ (standard score on WASI Blocks subtest); VIQ = standard score on WASI Vocabulary subtest; Reading/spelling = reading/spelling standard score on Wide Range Achievement Test, Spoonerism = No. correct on spoonerisms task, RAN = naming speed averaged across dense and sparse object RAN, Digit span = standard score on WASI digit span test.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 9
Backwards stepwise (likelihood ratio) logistic regression for participant group membership using reading, phonology, syllable stress and rise time threshold as predictors.

Step	Predictors	*B*	Exp *b*	Change in −2 log likelihood if variable removed	Model R^2 (Nagelkerke)	Overall % correct predictions (%)
1.	Stress (Av. *d*′)	−2.12*	.12	6.56*	.68	81.8
	Reading	−.13	.88	2.28		
	Rise time	.05	1.1	.76		
	Spoonerisms	.10	1.1	.20		
2.[a]	Stress (Av. *d*′)	−1.93*	.15	6.80**	.68	81.8
	Reading	−.11	.90	2.28		
	Rise time	.06	1.1	1.08		
3.[b,c]	Stress (Av. *d*′)	−2.10*	.12	9.83**	.66	84.8[d]
	Reading	−.12*	.89	3.44		

B = regression coefficient, significance calculated using Wald statistic; exp *b* = change in odds ratio; Model R^2 = total variance accounted for by the model at each step.
* $p < .05$.
** $p < .01$.
[a] Variable removed on step 2 = spoonerisms.
[b] Variable removed on step 3 = rise time.
[c] Model $\chi^2(2) = 22.31$, $p < .001$.
[d] Correct predictions for controls = 81.3%, dyslexics = 88.2%.

Table 10
Stepwise logistic regression for participant group membership using syllable stress and reading as predictors.

Step	Predictors	*B*	Exp *b*	Model R^2 (Nagelkerke)	Overall % correct predictions (%)
1.	Stress (Av. *d*′)	−2.47**	.09	.58	80.0
2[a]	Stress (Av. *d*′)	−2.08*	.13	.69	87.5[b]
	Reading	−.16*	.86		

Note: *B* = regression coefficient, significance calculated using Wald statistic; exp *b* = change in odds ratio; Model R^2 = total variance accounted for by the model at each step.
[a] Model $\chi^2(2) = 29.22$, $p < .001$.
[b] Correct predictions for controls = 85.0%, dyslexics = 90.0%.
* $p < .05$.
** $p < .01$.

Discussion

We proposed here that very basic auditory processes may be required to perceive periodic structure in speech, following the multi-tier framework for understanding spoken language proposed by Greenberg (2006). On the basis of our prior data with children with dyslexia, we also proposed that individual differences in basic auditory processing of rise time may affect the development of metrical language processing skills such as the perception of spoken syllable stress. Given the importance of accurate prosodic perception for phonological development (Pierrehumbert, 2003), and the well-documented phonological deficits found in developmental dyslexia, we expected difficulties in stress perception in adult individuals with dyslexia. Consistent with this hypothesis, the same-different judgement task designed here to measure stress perception in adults was indeed found to be performed less accurately by adults with developmental dyslexia. This difficulty was consistent across two experiments, whether adults were making a judgement about an identical lexical item repeated twice (*maternity–maternity*), or about two different lexical items (*maternity–ridiculous*). This suggests that individuals with dyslexia are impaired in the detection of acoustic prominence in speech.

In addition, correlational analyses demonstrated that individual differences in the accuracy of stress perception were associated with individual differences in rise time discrimination, for both the "easy" (Experiment 1) and the "difficult" (Experiment 2) versions of the stress perception task, as well as for performance averaged across the two experiments (average *d*′). These relationships are consistent with data from previous studies utilising both indirect stress sensitivity paradigms (such as reiterative speech, Goswami et al., 2009), and a metrical perception paradigm involving music (Huss et al., 2010). For both reiterative speech and metrical structure in music, rise time discrimination was also found to be a significant predictor of individual differences in performance accuracy. Although participants with dyslexia in the current study showed poorer auditory discrimination of *both* rise time and pitch, only individual differences in rise time discrimination predicted stress perception. Rise time may be a more important acoustic cue to acoustic prominence than pitch (cf. Greenberg, 2006), as rise time quantifies the change in sound energy (intensity of the signal) produced

by speakers as they articulate the onsets of stressed and unstressed syllables. Intensity discrimination, which was also related to accuracy in the musical metrical perception task used with children by Huss et al. (2010), was not a significant predictor of stress judgements. This makes sense, as the musical sequences in Huss et al.'s study all used the same instrument, and so only intensity and not rise time varied when notes were accented. In speech, both rise time and overall intensity will vary when syllables are accented or stressed.

Performance in the syllable stress task (average d' measure) was also a strong predictor of literacy, predicting group membership with 80% accuracy. This suggests that subtle speech processing difficulties in developmental dyslexia, such as the difficulty with stress perception documented here, persist into adulthood and can be stronger markers than the auditory and phonological difficulties that are markers of dyslexic difficulty in childhood. Although *a priori* there may appear to be little reason to link prosodic sensitivity and written word recognition, significant relations between stress perception and reading have been demonstrated in languages where stress is marked in the orthography, such as Greek (e.g., Protopapas & Gerakaki, 2009). Such demonstrations suggest that the perception of stress patterning in speech (the accurate detection of alternating strong and weak beats) is important for both phonological development and for acquiring literacy.

Studies are just beginning to demonstrate developmental relations between stress perception and reading acquisition, both in languages where stress is marked in the orthography (e.g., Gutiérrez-Palma & Palma-Reyes, 2007, Spanish) and in languages where it is not (Miller & Schwanenflugel, 2008, English). Even though stress is not marked by overt codes such as diacritics in English, there may be subtle orthographic cues to stress (e.g., when a syllable is written with more letters than necessary, it usually signifies that it is stressed, as in DISCUSS versus DISCUS, see Kelly, Morris, & Verrekia, 1998). Regarding phonological development, stress or prosodic patterning has been demonstrated to be an integral part of the phonological representations of individual words that are stored in the mental lexicon during infancy and early childhood (e.g., Curtin, Mintz, & Christiansen, 2005; Pierrehumbert, 2003; Vihman & Croft, 2007). During language acquisition, it appears critical that infants and children can process efficiently the temporal positions of the syllable "beats" in speech and thereby extract prosodic structure. In fact, a recent study with infants showed that statistical learning alone is a limited means of word segmentation. Johnson and Tyler (2010) studied infants' abilities to track transitional probabilities between syllables in an artificial language modelled after that used by Thiessen and Saffran (2003). The infants were aged on average 5.5 and 8 months, and two artificial languages were used, one based solely on 'words' of uniform length (CVCV), and the other based on 'words' that were either CVCV or CVCVCV. The transitional probabilities to 'word boundaries' in each language were the same. While even the 5.5-month-olds could segment 'words' in the uniform language (all CVCV), neither age group succeeded in the language with non-uniform word lengths. Johnson and Tyler (2010) noted that when artificial words are all the same length, a consistent rhythmic (periodic) cue to word segmentation is provided in addition to the transitional probability cues that are the focus of study. They suggested that more attention needed to be given to prosodic cues at the level of whole utterances in early infant word segmentation studies.

For individuals who are less sensitive to auditory cues to stress beats, in particular rise time, there may be reduced sensitivity to the rhythmic structure of speech, and this will have important consequences for developing the high-quality phonological representations of spoken words necessary for the acquisition of literacy. If a causal relationship can be established in future studies, then rhythmic and/or metrical training would be an important intervention for children with dyslexia (see Goswami, in press Huss et al., 2010, for an extended discussion). The place and role of "stress beats" (strong and weak syllables) provides temporal constraints across the different levels (syllable, word, phrase) that require functional co-ordination in speech *production* as well as speech perception (see Cummins & Port, 1998). Hence interventions addressing production as well as perception could be important. Certainly, there is ample developmental evidence that metrical structure (strong versus weak syllables) is related to how children produce words. For example, Gerken (1994) proposed a metrical template account of children's omission of weak syllables when producing multi-syllabic words. As she pointed out, during language acquisition young children are far more likely to omit weak syllables from word-initial positions than word-internal positions. The weak first syllable of a word like *giraffe* or *banana* is more often omitted than the weak second syllable of a word like *tiger*. Utilising a nonword production paradigm based on 4-syllable words, Gerken reported that while children omitted more weak syllables (45%) than strong syllables (11%) overall, their pattern of weak syllable omissions was predicted by the metrical segmentation hypothesis. For SWWS items, the first weak syllable was preserved 59% of the time, compared to 39% for the second weak syllable. However, for WSWS items, the first weak syllable was preserved 41% of the time, compared to 79% of the time for the second weak syllable. Gerken argued that young learners of English rely on metrical production templates. Infants learn rapidly from perceiving English words that they tend to begin with strong syllables, and young children apply this metrical learning to their own word productions. Our data could mean that metrical production templates would be weaker in children with developmental dyslexia.

The data presented support the view that the acoustic parameter of rise time is central to the perception of syllable stress in speech. As noted by Greenberg (2006), rise time is also important for perceiving intonational grouping because of its links with prosody. This has interesting implications for the notion that languages can be grouped into different rhythm classes, such as stress- versus syllable-timed, on the basis of different formulae quantifying consonantal and vocalic variability (e.g., Arvaniti, 2009; Grabe & Low, 2002; Ramus, Nespor, & Mehler, 1999). These

formulae typically depend on durational acoustic differences, but the criteria used to place languages on a rhythmic continuum do not reflect durational variation per se, rather they depend on the extent to which a language has easily-defined prominences or accents (see Dauer, 1983, 1987; and extended discussion in Arvaniti, 2009). As rise time is the critical cue to prominence or stress accent in speech (Greenberg, 1999, 2006; Greenberg, Carvey, Hitchcock, & Chang, 2003), analyses based on rise time may help to describe stress patterning in languages that have been classically difficult to place on rhythmic continua, such as Greek, Italian and Spanish. As Arvaniti (2009) argues, rhythm does not equate to timing, as metrical structure must also be taken into consideration. She defines metrical structure as the alternation of strong and weak elements. By her account, the key acoustic factors contributing to rhythm perception in different languages are grouping and relative prominence, and durational variability plays only a small role in the creation of rhythm. Consistent with Arvaniti's linguistic argument, Huss et al. (2010) did not find that children's duration thresholds were predictive of their performance in the musical metrical task.

In their work on speech production, Cummins and Port (1998) defined rhythm in speech as the hierarchical organisation of temporally co-ordinated prosodic units. They noted that Liberman (1975) originally proposed that speech, music and dance all conformed to the "metrical organisation hypothesis", that all temporally-ordered human behaviour is metrically organised. The centrality of prosodic perception (alternating strong and weak beats) to temporally-ordered language behaviours is supported here by the strong associations found between stress perception, phonology and literacy. If human utterances are structured so that stress beats lie at privileged phases of a higher-level prosodic unit, for example marking word onsets or phrase-level information (Cummins & Port, 1998; Greenberg, 2006), then periodicity is a key organisational principle underlying phonological and intonational structure in human speech. Accordingly, an insensitivity to the auditory parameters (such as rise time) that are critical for the perception of metrical structure would be expected to affect the development of both language and literacy in children, across languages from putatively different rhythm classes (Goswami, Wang, et al., 2010). The current study provides some evidence consistent with this hypothesis.

Acknowledgments

We would like to thank our participants. This research was supported by funding from the Medical Research Council, Grant G0400574, from an EU Framework VI STREP grant, Humans as Analogy Makers, from a Leverhulme Major Research Fellowship to Usha Goswami, and by the Academy of Finland. Requests for reprints should be addressed to Usha Goswami, Centre for Neuroscience in Education, 184 Hills Rd., Cambridge CB2 8PQ, UK.

Appendix A

Word. lists

List 1	List 2
First syllable stress	*First syllable stress*
DIFFICULTY	SECONDARY
VOLUNTARY	MILITARY
COMFORTABLE	AUDITORY
ORGANIZER	CITIZENSHIP
DELICACY	LAVATORY
MONASTERY	FERTILIZER
CAULIFLOWER	DANDELION
CATERPILLAR	MERCENARY
EDUCATOR	PUNISHABLE
CATEGORIZE	PACIFIER
Second syllable stress	*Second syllable stress*
DEMOCRACY	CAPACITY
VELOCITY	RIDICULOUS
HISTORICAL	REMARKABLE
CURRICULUM	DISCOVERY
MAGNIFICENT	FACILITY
DELIVERY	NECESSITY
MATERNITY	PARTICIPANT
BOTANICAL	MANIPULATE
DEBATABLE	MIRACULOUS
HARMONICA	PISTACHIO

Appendix B

Word. pairs in Experiment 2

List 1	List 2
First syllable stress	*First syllable stress*
DIFFICULTY–VOLUNTARY	SECONDARY–MILITARY
COMFORTABLE–ORGANIZER	AUDITORY–CITIZENSHIP
DELICACY–MONASTERY	LAVATORY–FERTILIZER
CAULIFLOWER–CATERPILLAR	DANDELION–MERCENARY
EDUCATOR–CATEGORIZE	PUNISHABLE–PACIFIER
Second syllable stress	*Second syllable stress*
DEMOCRACY–VELOCITY	CAPACITY–RIDICULOUS
HISTORICAL–CURRICULUM	REMARKABLE–DISCOVERY
MAGNIFICENT–DELIVERY	FACILITY–NECESSITY
MATERNITY–BOTANICAL	PARTICIPANT–MANIPULATE
DEBATABLE–HARMONICA	MIRACULOUS–PISTACHIO

References

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66*, 46–63.

Choi, J.-Y., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *Journal of the Acoustical Society of America, 118*, 2579–2587.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language & Speech, 45*, 207–228.

Corriveau, K., Pasquini, E., & Goswami, U. (2007). Basic auditory processing skills and specific language impairment: A new look at an old hypothesis. *Journal of Speech, Language & Hearing Research, 50*, 1–20.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96*, 233–262.

Cutler, A. (2005). Lexical stress. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 264–289). Oxford: Blackwell.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51–62.

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. *Paper presented at the 11th international congress of phonetic sciences* (Vol. 5, pp. 447–450). Tallinn.

Echols, C. H. (1996). A role for stress in early speech segmentation. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 151–170). Mahwah, NJ: Lawrence Erlbaum Associates.

Foxton Riviere, L.-D., & Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition, 115*, 71–78.

Fredrickson, N., Frith, U., & Reason, R. (1997). *Phonological assessment battery* (Standardised ed.). Windsor: NFER-Nelson.

Fry, D. B. (1954). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America, 26*, 138.

Gerken, L. (1994). A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language, 21*, 565–584.

Goswami, U. (in press). Language, music and children's brains: A rhythmic timing perspective on language and music as cognitive systems. In P. Rebuschat et al. (Eds.), *Language and music as cognitive systems*.

Goswami, U., Gerson, D., & Astruc, L. (2009). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading & Writing*. doi:10.1007/s11145-009-9186-6.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., et al. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 10911–10916.

Goswami, U., Fosker, T., Huss, M., Mead, N., & Szücs, D. (2010). Rise time and formant transition duration in the discrimination of speech sounds: The Ba–Wa distinction in developmental dyslexia. *Developmental Science* [Published online March].

Goswami, U., Wang, H. -L. S., Cruz, A., Fosker, T., Mead, N., & Huss, M. (2010). Language-universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese. *Journal of Cognitive Neuroscience*. doi:10.1162/jocn.2010.21453.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.). *Laboratory phonology* (Vol. 7, pp. 515–546). Berlin: Mouton de Gruyter.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication, 29*, 159–176.

Greenberg, S. (2006). A multi-tier framework for understanding spoken language. In S. Greenberg & W. Ainsworth (Eds.), *Understanding speech: An auditory perspective* (pp. 411–434). Mahweh, NJ: LEA.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech – A syllable-centric perspective. *Journal of Phonetics, 31*, 465–485.

Gutiérrez-Palma, N., & Palma-Reyes, A. (2007). Stress sensitivity and reading performance in Spanish, a study with children. *Journal of Research in Reading, 30*, 157–168.

Hämäläinen, J., Leppänen, P. H. T., Torppa, M., Muller, K., & Lyytinen, H. (2005). Detection of sound rise time by adults with dyslexia. *Brain & Language, 94*, 32–42.

Hämäläinen, J., Leppänen, P. H. T., Eklund, K., Thomson, J., Richardson, U., Guttorm, T. K., et al. (2009). Common variance in amplitude envelope perception tasks and their impact on phoneme duration perception and reading and spelling in Finnish children with reading disabilities. *Applied Psycholinguistics, 30*, 511–530.

Hämäläinen, J. A., Salminen, H. K., & Leppänen, P. H. T. (in press). Basic auditory processing deficits in dyslexia: Review of the behavioural and event-related potential/field evidence. *Journal of Learning Disabilities*.

Hoequist, C. A. (1983). The perceptual centre and rhythm categories. *Language & Speech, 26*, 367–376.

Huss, M., Verney, J. P., Fosker, T., Fegan, N., & Goswami, U. (2010). Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex* doi:10.1016/j.cortex.2010.07.010.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science, 13*, 339–345.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Kelly, M. H., Morris, J., & Verrekia, L. (1998). Orthographic cues to lexical stress: Effects on naming and lexical decision. *Memory & Cognition, 26*, 822–832.

Kitzen, K. R. (2001). *Prosodic sensitivity, morphological ability and reading ability in young adults with and without childhood histories of reading difficulty*. (Doctoral dissertation, University of Columbia, 2001). Dissertation Abstracts International, 62 (02), 0460A.

Klein, H. (1984). Learning to stress: A case study. *Journal of Child Language, 11*, 375–390.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency adds little. *Journal of the Acoustical Society of America, 118*, 1038–1054.

Levitt, H. (1971). Transformed up–down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*, 467–477.

Liberman, M. (1975). *The intonational system of English*. Ph.D. thesis, MIT Cambridge, MA. Published by Indiana University Linguistics Club, 1978.

Lorenzi, C., Dumont, A., & Fullgrabe, C. (2000). Use of temporal envelope cues by children with developmental dyslexia. *Journal of Speech, Language and Hearing Research, 43*, 1367–1379.

Mattys, S. L., & Juscyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91–121.

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language & Speech, 31*, 135–156.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly, 43*, 336–354.

Morton, J., Marcus, S. M., & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review, 83*, 405–408.

Muneaux, M., Ziegler, J. C., Truc, C., Thomson, J., & Goswami, U. (2004). Deficits in beat perception and dyslexia: Evidence from French. *NeuroReport, 15*, 1255–1259.

Pasquini, E., Corriveau, K., & Goswami, U. (2007). Auditory processing of amplitude envelope rise time in adults diagnosed with developmental dyslexia. *Scientific Studies in Reading, 11*, 259–286.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning and acquisition of phonology. *Language & Speech, 46*, 115–154.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology, 25*, 143–170.

Protopapas, A., & Gerakaki, S. (2009). Development of processing stress diacritics in reading Greek. *Scientific Studies in Reading, 13*, 453–483.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in speech signal. *Cognition, 73*, 265–292.

Richardson, U., Thomson, J., Scott, S. K., & Goswami, U. (2004). Supra-segmental auditory processing skills and phonological representation in dyslexic children. *Dyslexia, 10*, 215–233.

Rocheron, I., Lorenzi, C., Fullgrabe, C., & Dumont, A. (2002). Temporal envelope perception in dyslexic children. *NeuroReport, 13*, 1683–1687.

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech, 33*, 47–68.

Snowling, M. J. (2000). *Dyslexia* (2nd ed.). Oxford: Blackwell.

Suranyi, S., Csepe, V., Richardson, U., Thomson, J., Honbolygo, F., & Goswami, U. (2009). Sensitivity to rhythmic parameters in dyslexic children: A comparison of Hungarian and English. *Reading & Writing, 22*, 41–56.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Thomson, J. M., & Goswami, U. (2008). Rhythmic processing in children with developmental dyslexia: Auditory and motor rhythms link to reading and spelling. *Journal of Physiology – Paris, 102*, 120–129.

Thomson, J. M., Fryer, B., Maltby, J., & Goswami, U. (2006). Auditory and motor rhythm awareness in adults with dyslexia. *Journal of Research in Reading, 29*, 334–348.

Vihman, M., & Croft, W. (2007). Phonological development: Towards a ''radical" templatic phonology. *Linguistics, 45*, 683–725.

Wechsler, D. (1998). *The wechsler adult intelligence scale* (3rd ed.). London: The Psychological Corporation.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. San Antonio, TX: The Psychological Corporation.

Whalley, K., & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading, 29*, 288–303.

Wilkinson, GS. (1993). *Wide range achievement test 3*. Wilmington, DE: Wide Range.

Wood, C., & Terrell, C. (1998). Poor readers' ability to detect speech rhythm and perceive rapid speech. *British Journal of Developmental Psychology, 16*, 397–413.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholiguistic grain size theory. *Psychological Bulletin, 131*, 3–29.

Modulation Filterbank Parameters

The wideband envelope was extracted from the speech signal by taking the absolute value of the Hilbert transform. The envelope was then passed through the MFB to derive the AM hierarchy. The MFB had the following edge frequencies (in Hz) : 50; BEAT*5; BEAT*1.75; BEAT/1.75; BEAT/5; 0.5 , where 'BEAT' refers to the syllable beat rate computed for that sample. Factors of 1.75 and 5 were used to take into account filter roll-off. Table A shows the resulting edge frequencies for a BEAT rate of 4.04 Hz (as used in the tone vocoding experiment).

Table A. Edge frequencies for MFB

	AM Tier (filter channel)	MFB Bandpass Edges (Hz)	Q_{-6} value
1	Fast	20 - 50	1.2
2	Sub-beat	7 - 20	1.0
3	Syllable	2.3 - 7	1.0
4	Stress	0.8 - 2.3	1.0
5	Slow	0.5 - 0.8	2.2

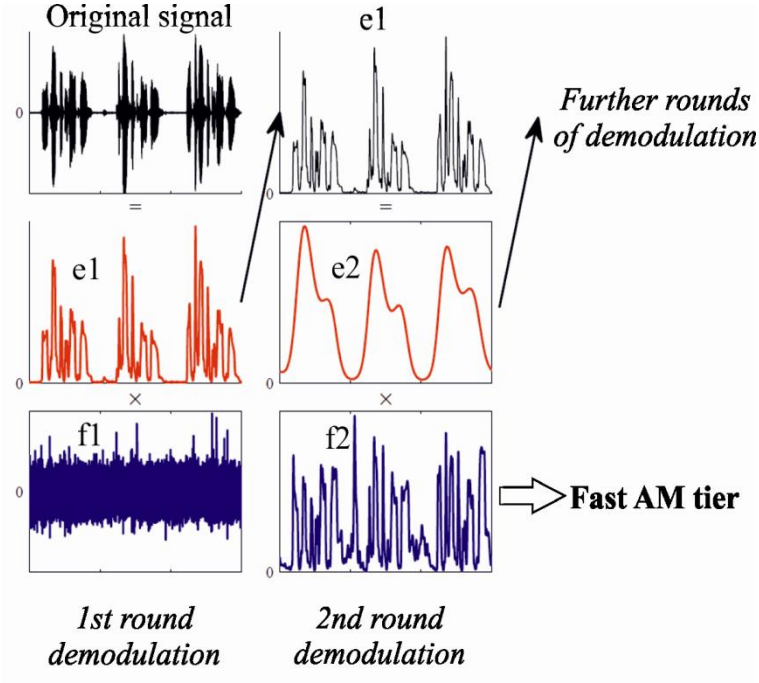
The modulation filterbank comprised a series of adjacent FIR bandpass filters. Each filter had a response of –6 dB at the cross-over edge with its adjacent filter, but –55dB at the cross-over with the next-but-one channel. Response at channel edges was similar across all channels on a logarithmic frequency scale, for both the low-pass and high-pass edges. The time delay introduced by each filter was removed by a suitable time-alignment of the filter output. The MFB is a scaled version of a filterbank originally used for separating wideband speech into audio frequency channels. Here we scaled the channel edge frequencies to be suitable for the modulation frequency ranges of interest. For further technical details of the filterbank design, see Stone & Moore (2003). In determining the spacing of channels in the MFB, consideration was also taken for the filter Q values (centre frequency divided by bandwidth). The Q values of auditory modulation filters are typically assumed to lie between 1–2 (Dau *et al.*, 1997a, b; Ewert and Dau, 2000; Lorenzi *et al.*, 2001; Ewert *et al.*, 2002; Sek and Moore, 2002). The Q values calculated from our MFB (whose bandwidths and centre frequencies were chosen on theoretical grounds) are shown in Table A. With the exception of

the very slowest channel (whose lower edge was artificially limited by the length of our stimuli), the Q values for each channel are consistently close to 1. This value is consistent with Sek & Moore (2003), who found that their human psychophysical data fitted with a modulation filterbank with a Q value of 1 or slightly less.

a. Probabilistic Amplitude Demodulation (PAD) Demodulation Cascade

In this method, the timescale of demodulation can be controlled so that only very fast or very slow modulations are extracted from the signal via a process of Bayesian inference. Hence, the AM hierarchy is derived by applying the PAD method recursively at progressively slower and slower timescales (see Figure a).

Figure a. Illustration of the PAD demodulation cascade



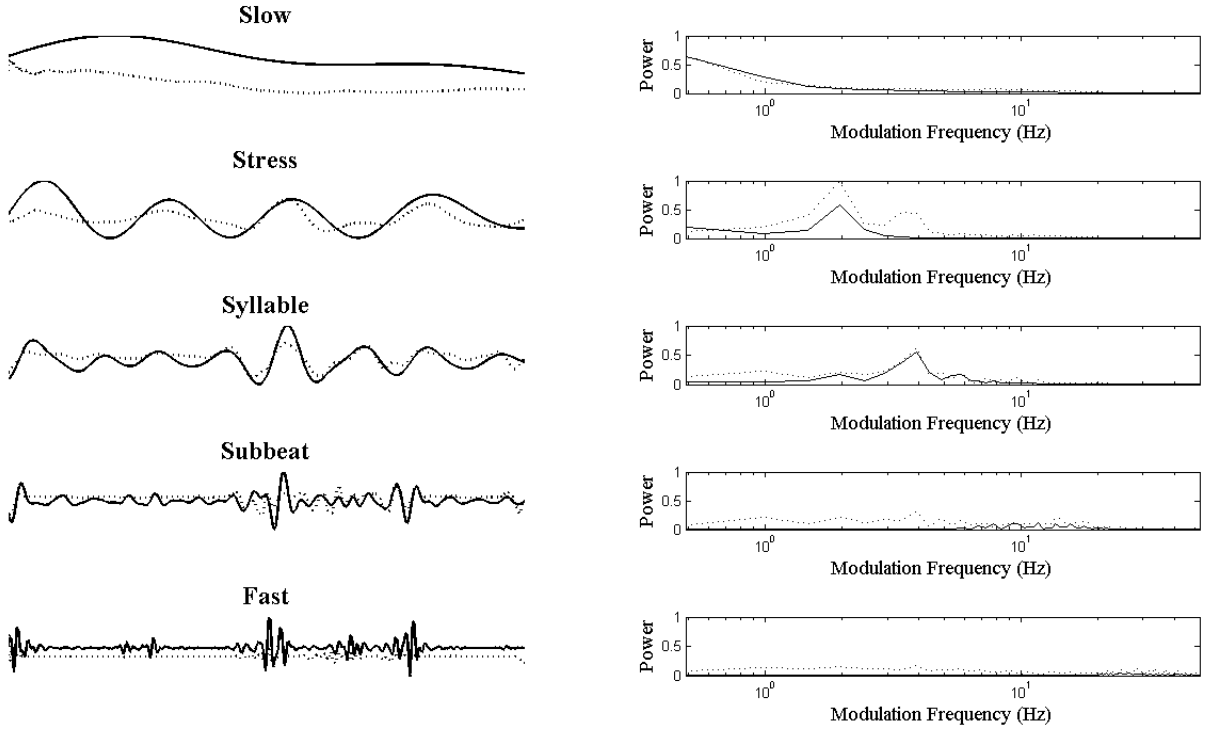
In the first stage, PAD is set to demodulate the original speech signal at the fastest timescale, extracting a fast-varying envelope (e1) and its complementary fine structure (f1, derived by dividing the original signal by the envelope). PAD is then adjusted to a slower time scale and re-applied to the fast-varying envelope (e1), extracting a more slowly-varying envelope (e2) and complementary fine structure (f2). This fine structure contains the remnant fast modulations that are not extracted by the second round of demodulation, and forms the *fast* AM tier. PAD is then applied at an even slower time scale to e2, extracting a third envelope (e3) and fine structure (f3). Accordingly, f3 becomes the *Sub-beat* tier and e3 is demodulated for the fourth time. This process is continued until the final round of demodulation where the final envelope becomes the *slow* tier, and final fine structure becomes the *stress* tier. Since the frequency content of the PAD envelopes is related to the timescale of demodulation in a nonlinear fashion, the parameters of demodulation were

calibrated by hand to ensure that the resulting PAD AM hierarchy provided the best possible match to the MFB hierarchy.

b. Comparison of MFB and PAD AMs

As shown in Figure b, the two sets of AMs were similar in shape. However, all PAD-derived AMs contained a low background level of modulation that was slower than the target rate. This was an inherent property of the PAD algorithm in which the timescale of demodulation limited only the upper tail of the modulation spectrum. Also, PAD achieved less distinct modulation frequency separation, as shown by the spectrum of the 'Stress' tier, which also contained faster modulation at the 'Syllable' rate. Despite these differences, our experimental data indicated that participants did not differ in performance regardless of whether they were hearing PAD- or MFB-derived AM tiers. Moreover, participants frequently noted that they found PAD-derived stimuli more "natural" sounding, indicating that the presence of background slow modulation and the less abrupt frequency roll-off may be an inherent property of natural sounds.

Figure b. Left : MFB (bold) and PAD (dotted) AM tiers. Right : Respective modulation spectra for MFB (bold) and PAD (dotted) AMs.



List of 44 Nursery Rhymes

Nursery Rhyme	**Music Time Signature**	**Rhythmic Meter**
Baa Baa Black Sheep	2/4	Duple
Once I Caught a Fish Alive	4/4	Duple
One Two Buckle my Shoe	4/4	Duple
Old MacDonald Had a Farm	4/4	Duple
Twinkle Twinkle Little Star	4/4	Duple
London Bridge is Falling Down	4/4	Duple
Mary Had a Little Lamb	4/4	Duple
Polly Put the Kettle On	2/4	Duple
Yankee Doodle	2/4	Duple
Peter Peter Pumpkin Eater	4/4	Duple
Mary Mary Quite Contrary	4/4	Duple
Simple Simon Met a Pieman	4/4	Duple
As I Was Going to St Ives	2/4	Duple
The Queen of Hearts	N.A.	Duple
Lucy Lockett	4/4	Duple
Cobbler Cobbler Mend My Shoe	4/4	Duple
Peter Piper	N.A.	Duple
I'm a Little Teapot	4/4	Duple
Sing a Song of Sixpence	2/4	Duple
Wee Willie Winkie	2/4	Duple
Old King Cole	4/4	Duple
The Wheels on the Bus	4/4	Duple
Three Little Monkeys	N.A.	Duple
Grand Old Duke of York	4/4	Duple
Incy Wincy Spider	6/8	Duple
Jack and Jill	6/8	Duple
Humpty Dumpty	6/8	Duple
Ring-a-Ring-a-Roses	6/8	Duple

Row Row Row Your Boat	6/8	Duple
Hickory Dickory Dock	6/8	Duple
Here We Go Round the Mulberry Bush	6/8	Duple
Little Miss Muffet	6/8	Triple
Little Jack Horner	6/8	Triple
Little Boy Blue	6/8	Triple
Curly Locks	6/8	Triple
To Market	6/8	Triple
Pussycat Pussycat	6/8	Triple
Ladybird Ladybird	6/8	Triple
There Was An Old Lady	6/8	Triple
Two Cats of Kilkenny	N.A	Triple
Ride a Cock Horse	3/4	Triple
Orange and Lemons	3/4	Triple
Rock-a-Bye-Baby	3/4	Triple
Lavender's Blue	3/4	Triple

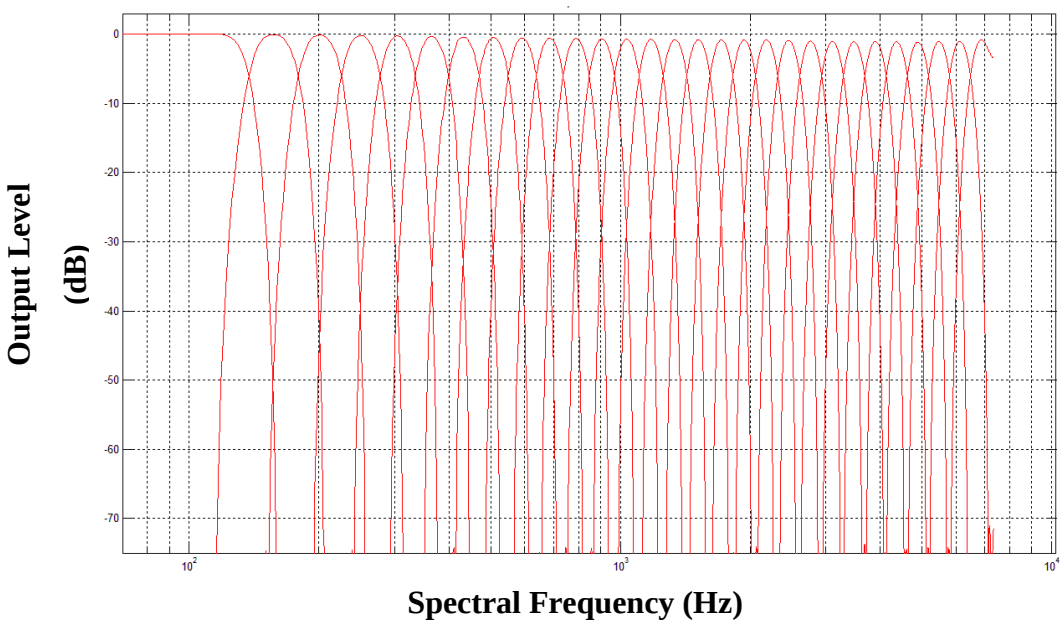
Note that some nursery rhymes with an assigned with a 'Duple' or 'Triple' meter actually had a compound musical time signature, such as 6/8. Compound time signatures consist combinations of duple or triple beats within each bar, for example 6/8 indicates 2 sets of triple beats. Therefore, these rhymes can be uttered to fit a duple meter as well a triple meter, depending on the rate of speaking. In these cases, the decision as to whether a rhyme was 'duple'- or 'triple'-meter was made on the basis of poetic scansion, using the dominant prosodic foot length.

From inspection of the table, 31 out of the 44 (70%) nursery rhymes were assigned a duple meter, while 13 (30%) were assigned a triple meter. The fact that all the nursery rhymes had relatively short prosodic feet (2 or 3 syllables in length) is consistent with Gueron's (1974) analysis of the metrical structure of 130 Mother Goose nursery rhymes. She concluded that all but one of the nursery rhymes had a simple 'Strong (S) - weak (w)' alternating metrical pattern of : (w) S w S (w) S w S (w), with the weak elements in parenthesis omitted in some

rhymes. In Gueron's analysis, 'S' elements were usually realized by a single stressed syllable while 'w' elements were realised by between one to three unstressed syllables. Consequently, the prosodic feet in Gueron's analysis had a maximum length of 4. While the relative frequencies of each type of prosodic foot were not given in the study, the current set of nursery rhyme material indicates a higher incidence of nursery rhymes with shorter (e.g. 2-syllable-long) prosodic feet.

ERB_N-spaced Spectral Filterbank

The spectral filterbank comprised a series of adjacent FIR band-pass filters, reflecting the equivalent rectangular bandwidth (ERB) of cochlea channels in a normal hearing individual. The first channel in the filterbank was low-pass rather than band-pass. Each filter channel had a response of –6 dB at the cross-over edge with its adjacent filter, but –55dB at the cross-over with the next-but-one channel. Response at channel edges was similar across all channels on a logarithmic frequency scale, for both the low-pass and high-pass edges. The time delay introduced by each filter was removed by a suitable time-alignment of the filter output. For further technical details of the filterbank design, see Stone & Moore (2003). The figure below shows the frequency response curves for the 29 channels in the spectral filterbank, where the spectral frequency (x-axis) is on a logarithmic scale. The table below lists the cross-over edges (-6 dB) between the adjacent filter channels.

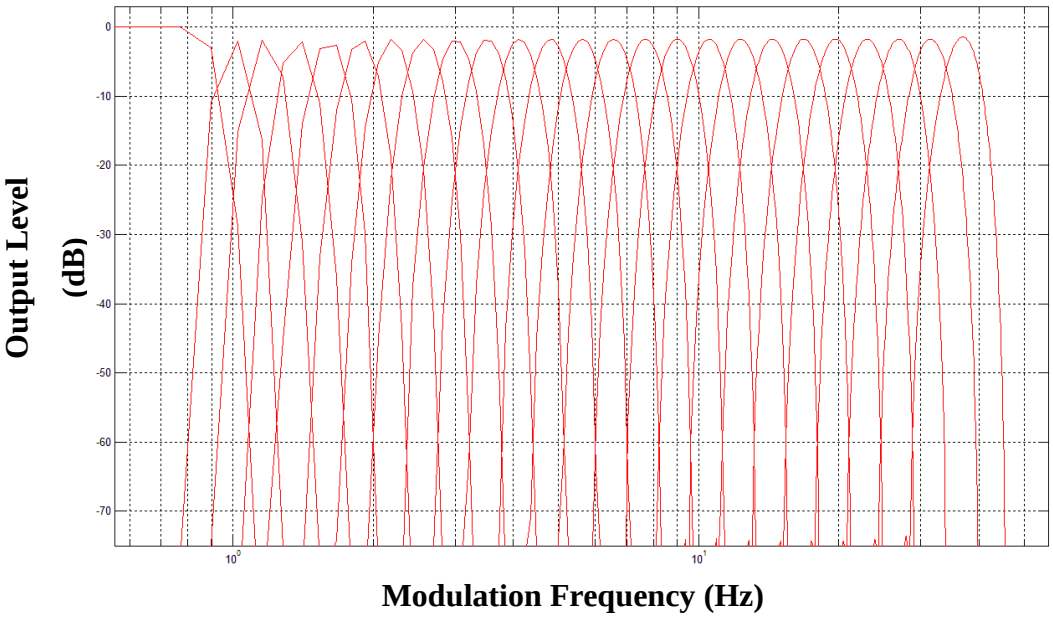


Edge Number	Edge (Hz)
1 *(low-pass)*	100
2	137
3	179
4	225
5	277
6	334
7	398

8	470
9	549
10	638
11	736
12	846
13	969
14	1105
15	1257
16	1426
17	1614
18	1824
19	2057
20	2317
21	2607
22	2930
23	3289
24	3689
25	4135
26	4631
27	5184
28	5800
29	6486
30	7250

24-Channel Modulation Filterbank

The 24-channel modulation filterbank also comprised a series of adjacent FIR band-pass filters, and had the same overall design as the spectral filterbank, but with appropriately-scaled and log-spaced channel edge frequencies. The first channel in the filterbank was a dummy low-pass channel rather than band-pass, and the output from this channel was discarded (leaving 24 channels from 0.9-40 Hz). As was the case for the spectral filterbank, each modulation channel had a response of –6 dB at the cross-over edge with its adjacent filter, but –55dB at the cross-over with the next-but-one channel. The time delay introduced by each filter was removed by a suitable time-alignment of the filter output. The figure below shows the frequency response curves for the 24 channels in the modulation filterbank, where the modulation frequency (x-axis) is on a logarithmic scale. The table below lists the cross-over edges (-6 dB) between the adjacent filter channels.



Edge Number	Edge (Hz)
1 *(low-pass dummy channel, output discarded)*	0.79
2	0.93
3	1.09
4	1.27
5	1.49
6	1.74

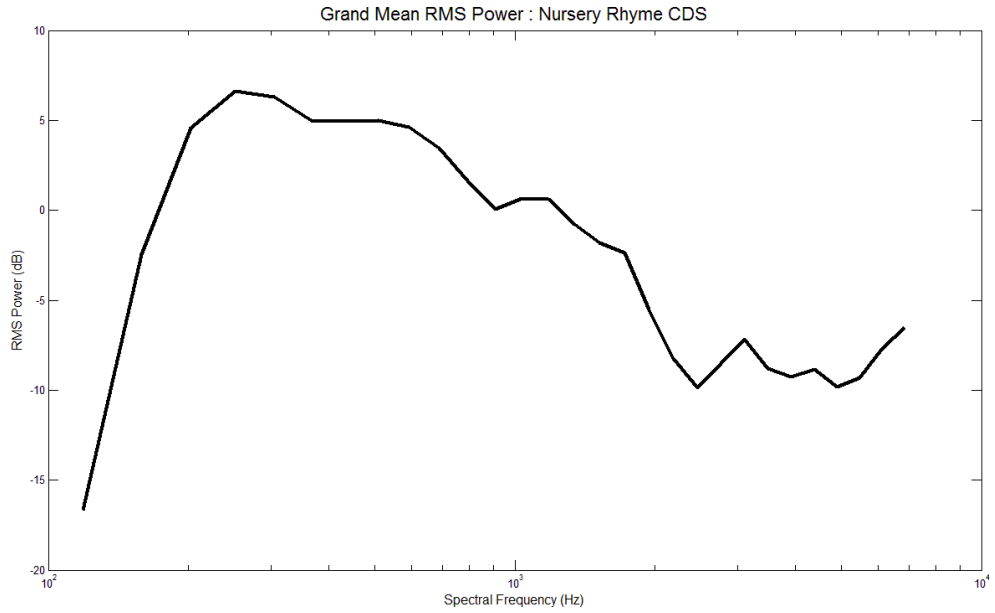
7	2.03
8	2.38
9	2.78
10	3.25
11	3.80
12	4.45
13	5.20
14	6.08
15	7.11
16	8.32
17	9.72
18	11.38
19	13.30
20	15.56
21	18.20
22	21.28
23	24.89
24	29.11
25	34.04
26	39.81

Spectral RMS Power and Spectral Correlation Patterns

RMS Power.

The RMS (root-mean-square) spectral power of the 29 cochlear channels was computed. Since the actual RMS power varied across samples and speakers, for each sample, the average power across all spectral channels was subtracted from each channel, leaving only the difference from the average. This difference was then averaged over samples and speakers. Figure a shows the computed difference RMS power by spectral channel, averaged over all 44 nursery rhymes and 6 speakers. As shown in the figure, RMS power is strongest for low spectral frequencies around 200 Hz, and steadily declines as frequency increases.

Figure a. RMS power across 29 ERB_N-spaced cochlear channels

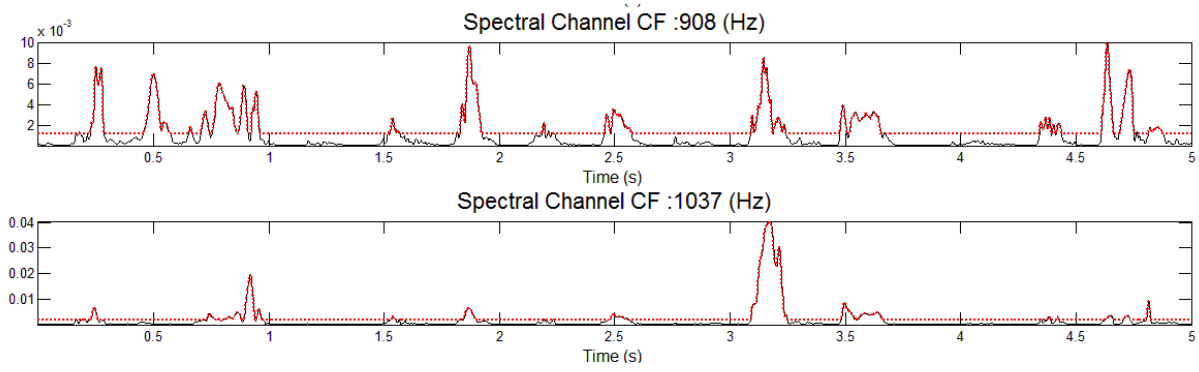


As expected from the classical acoustic phonetics and engineering literatures, the long term power spectrum has a low-frequency maximum with gradual fall towards high frequencies and Figure a confirms this for the present material. Power is generally stronger at lower spectral frequencies, and decreases approximately logarithmically (since the y-axis unit is dB) with increasing spectral frequency, thereby obeying a 1/f law. The drop off in power below ~200-300 Hz corresponds to the approximate lower end of the fundamental frequency for female speakers.

Spectral Correlation.

Next, the cross-(spectral)-channel correlation was computed for the 29 cochlear channels. Before this could be done, a thresholding procedure was employed to remove extraneous low-level modulations in the envelope arising from background noise. Such noise would reduce the true correlation between spectral channels. For each spectral channel, the long-term RMS power was determined. Using a threshold of -16dB from the long-term RMS power of each spectral channel, all portions of the envelope with power above this value were left unchanged. All time periods of the signal with power below RMS -16dB were set to RMS -16dB (henceforth referred to as the 'floor') plus a very small amount (amplitude of 1^{-10}) of random noise. This small amount of noise was added to portions that did not meet the threshold level so that floored sections of the envelope would not be completely flat, as this could artificially elevate correlations between spectral channels in subsequent analyses. This flooring procedure is described in greater detail in Stone & Moore (2007). Figure b shows an example of the original (black) and floored (red, dashed) envelopes for two adjacent spectral (cochlear) channels. Since the human discrimination of intensity and the subjective sensation of loudness varies approximately logarithmically with signal power (Fechner, 1860), the base 10 logarithm of the floored envelope in each band was taken, and this logarithmic envelope was used for the subsequent correlation analysis.

Figure b. Example of floored envelopes in adjacent spectral channels

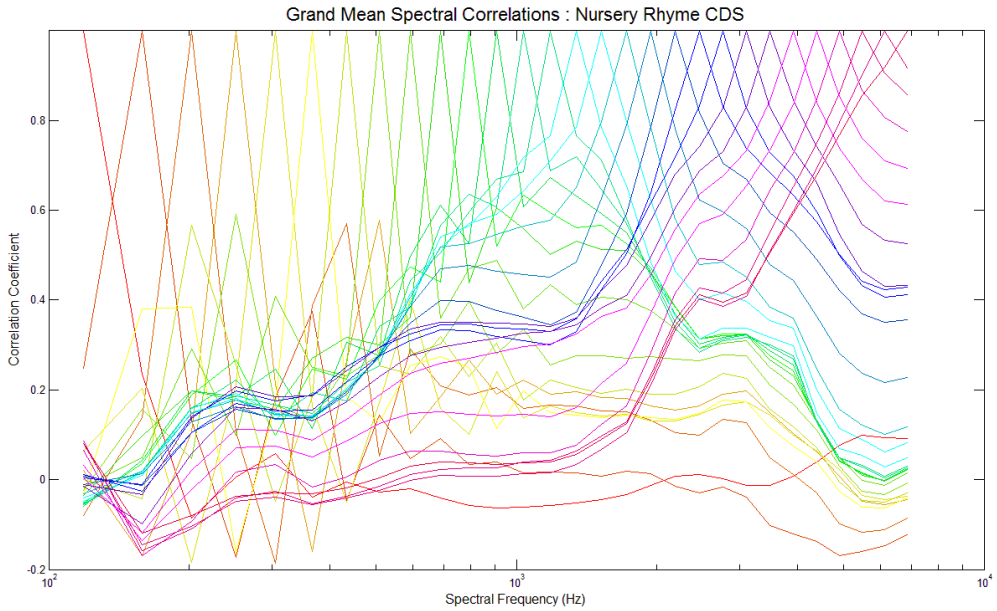


For the correlation analysis, only the unfloored sections of each cochlear channel (i.e. active sections) were correlated with the temporally-corresponding sections of all the other cochlear channels. To do this, floored sections were excised from the target channel, and the remaining unfloored sections were concatenated. This concatenated envelope was then z-scored, and correlated with temporally-corresponding z-scored sections of every other channel (ie irrespective of whether those were floored or unfloored). Thus the temporally-

corresponding sections in other channels could contain floored (silenced) sections as well as unfloored (active) sections, although the target channel itself only contained unfloored (active) sections.

Figure c shows the result of this cross-correlation across cochlear channels (with zero lag), where the mean correlation coefficient over 44 speech samples and 6 speakers is plotted. Visual inspection of the figure shows that mid-frequency spectral channels around 1000 Hz (green-cyan) show the strongest correlation with each other in this general region, and with other spectral channels as well. There is also some evidence of channel 'clustering' for example among green-cyan (~1000 Hz) channels, or among blue (~3000 Hz) channels.

Figure c. Intercorrelations between spectral channels. In the top plot, each coloured line indicates a single cochlear channel.

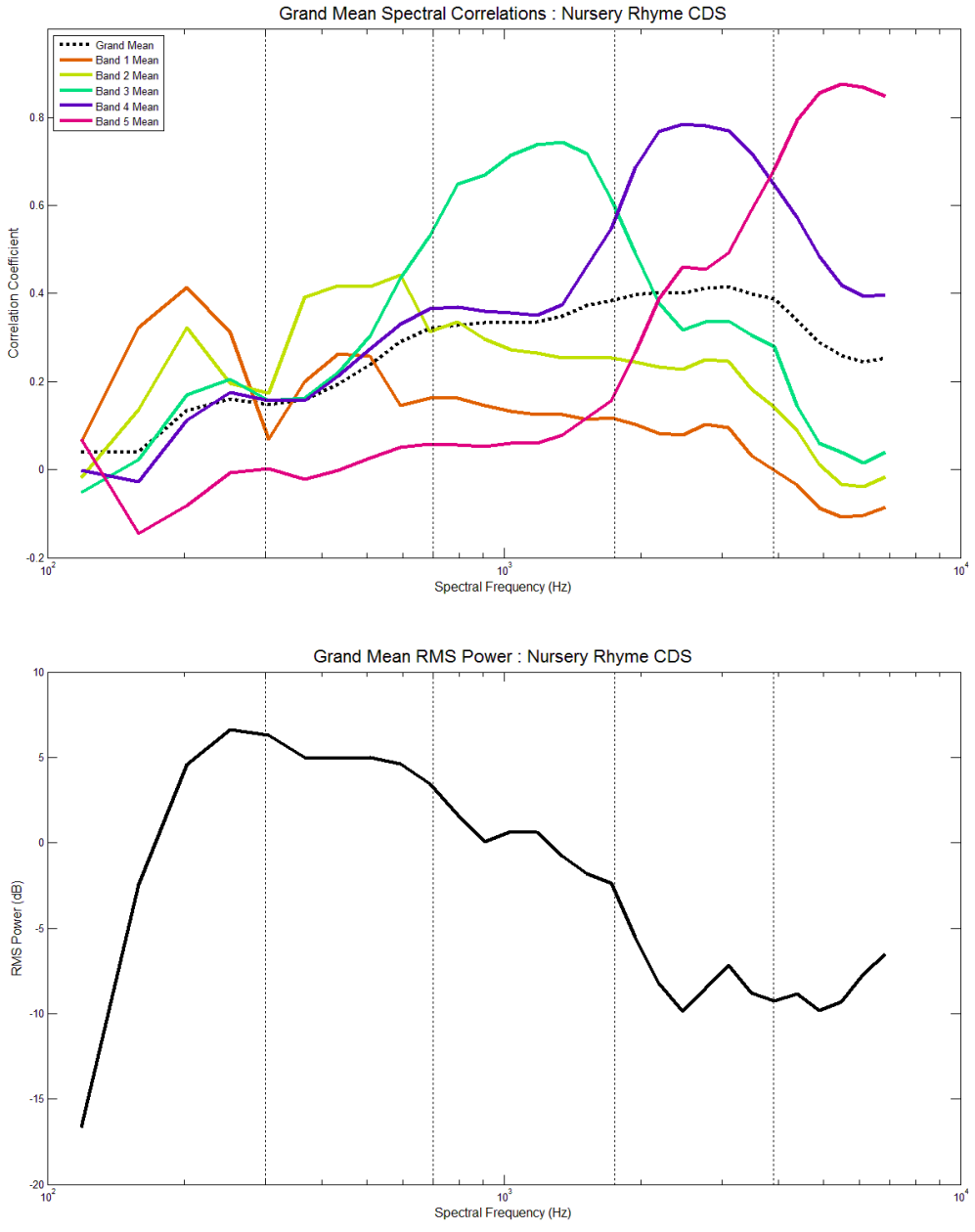


Spectral Correlations by Spectral Band

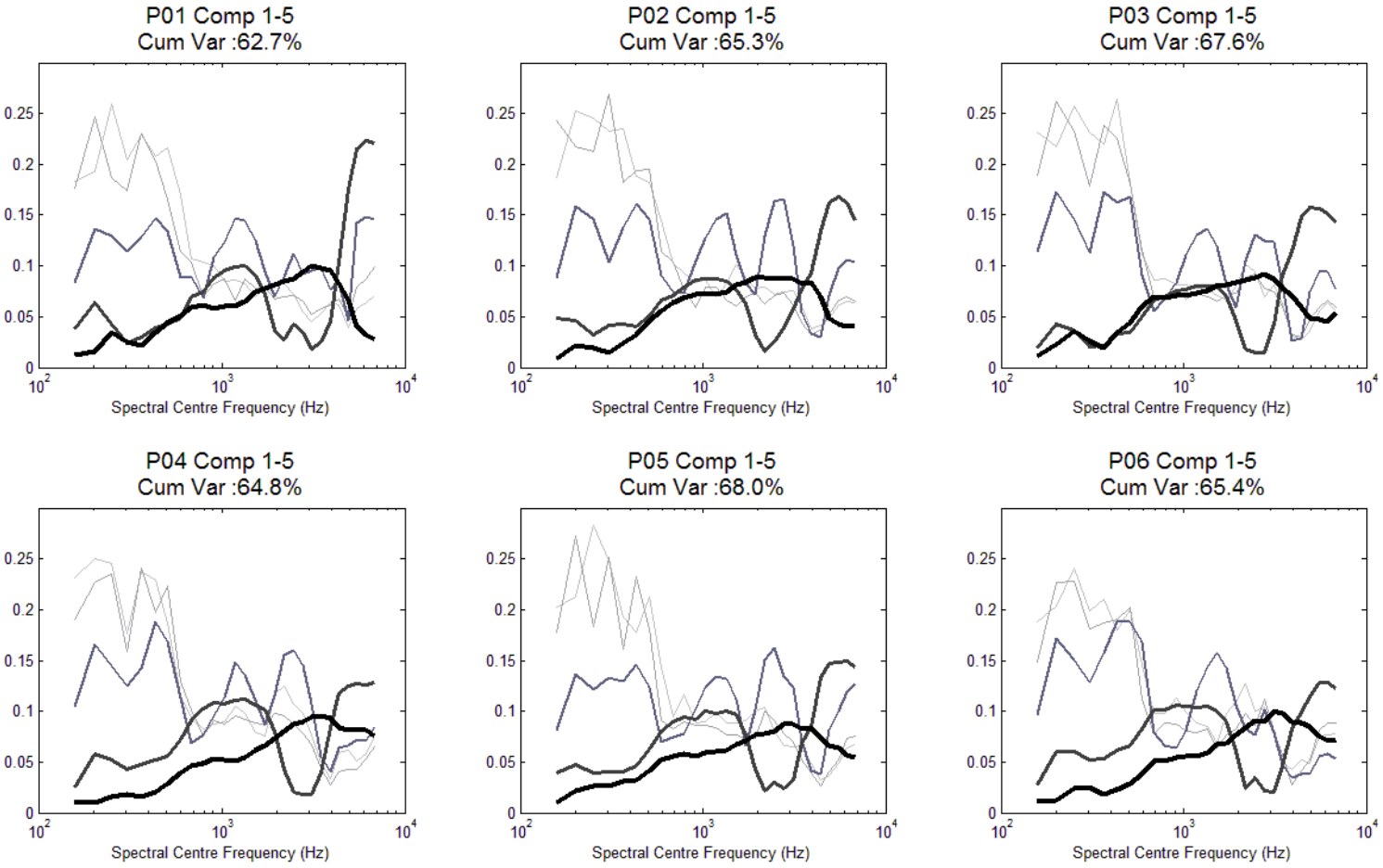
5 spectral bands were identified in Chapter 4 on the basis of the rectified component loading pattern of the top 5 principle components. The top panel of Figure d shows the mean correlation of the cochlear channels in each spectral band (shown as coloured lines) with the channels in other spectral bands. The black dotted line shows the grand mean correlation over all cochlear channels. Note the similarity in pattern between this grand mean correlation, and the loading pattern of principle component 1 in Chapter 4, Section 4.3.2. The bottom panel of Figure d shows the RMS power across the 29 cochlear channels, with the boundaries of the spectral bands superimposed as vertical dotted lines. It is interesting to note that there is an

approximately inverse relationship between RMS power and correlation strength. So Spectral Band 4, which has lowest power, shows strong correlations with adjacent channels (has the highest mean). In contrast, Spectral bands 1 & 2 have high power, but only correlate weakly with other channels. Hence, the RMS power in an audio frequency region is not necessarily a good indicator of its correlation strength with other channels.

Figure d. (top) Mean spectral correlations between each spectral band and the other spectral bands. Each band is shown in a different colour. The black dotted line indicates the grand mean correlation over all spectral bands. The vertical dotted line shows the boundary between spectral bands. (bottom). RMS power of each cochlear channel, with spectral band boundaries overlaid as vertical dotted lines.



Spectral PCA Component Loadings by Speaker



The lines of different thickness indicate different PCA components. More important (lower numbered) components are shown in a thicker line. The loading patterns for the top 5 PCA components shown here are broadly similar across the 6 speakers. The total amount of variance explained by the top 5 components was also similar across speakers, ranging from 62.1% to 68.0%.

Rate Normalisation of Modulator Channels

To remove the effects of the rate disparity between modulation channels, a rate normalisation procedure was developed. This involves taking the unwrapped angular phase of each channel, and normalising its rate of angular change over time with respect to a standard reference channel (i.e. central channel 13 out of 25), as shown in Figure a. Note that when a different reference channel is chosen (eg. channel 20), the result of the subsequent PCA analysis is highly similar. The rate-normalised angular phase for each channel is then multiplied back with the original power of that channel. The result of the procedure is that *rate* differences between channels are removed, but power and relative *phase* differences are retained, as shown in Figure b. This should increase the overall degree of correlation between modulation channels without losing critical differences in temporal patterning.

Figure a. Example of the unwrapped phase angle for original (black) and rate-normalised (red) modulators increasing as a function of time

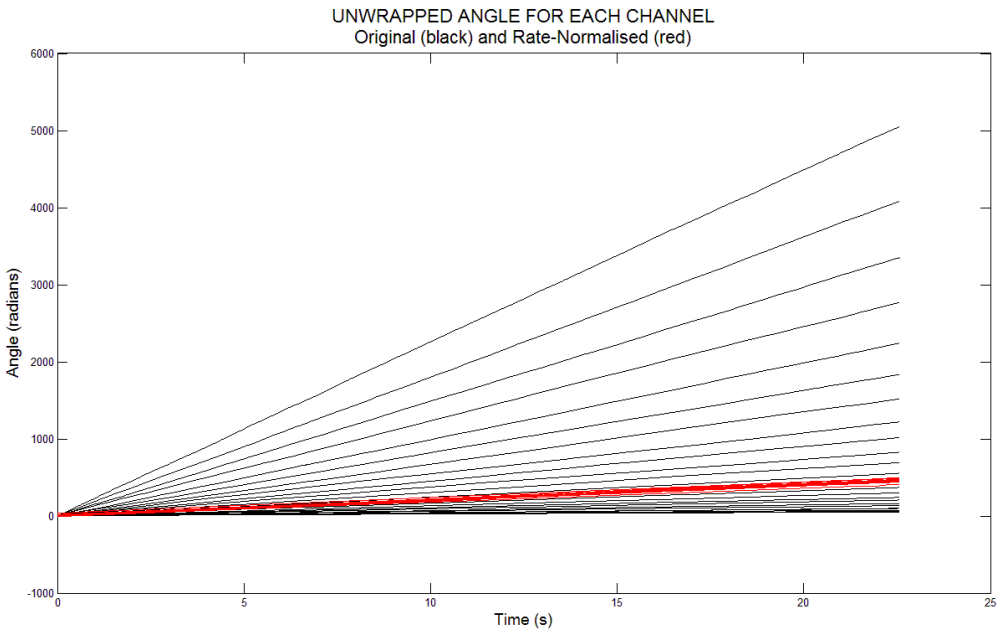
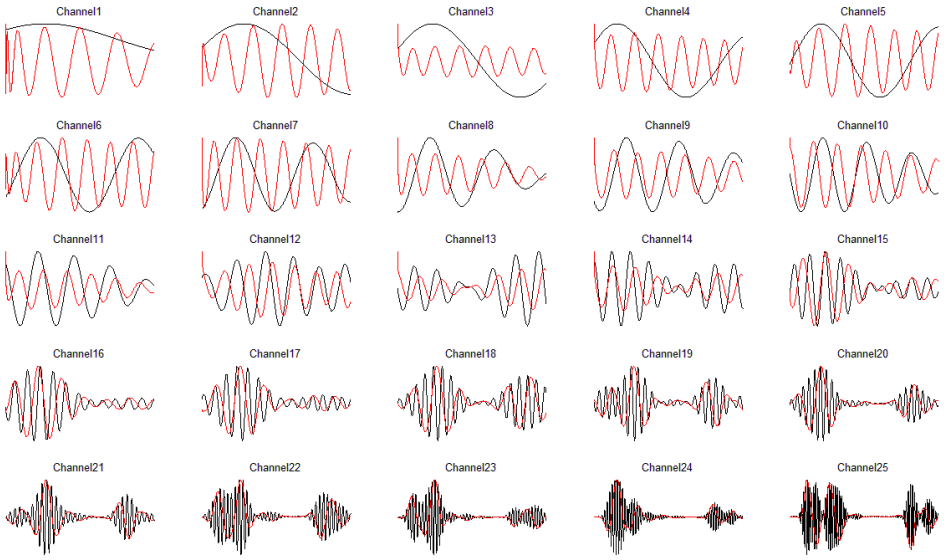


Figure b. Example of original (black) and rate-normalised (red) modulators for each channel.

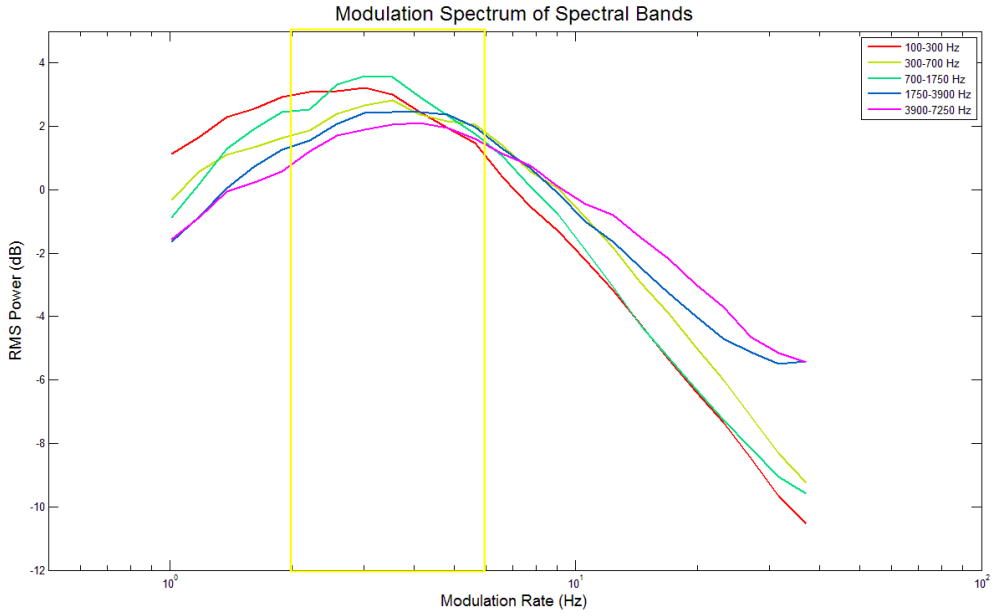


When the PCA procedure was then performed on these rate-normalised modulators, the first 5 components explained on average over 80% of variance (compared to ~30% for the whole modulator), indicating that the top few extracted components provided a representative view of the data. Note that these results are highly similar to those obtained using the power of each channel only.

RMS Power of the Modulation Spectrum for Each Spectral Band

The RMS *difference* power across the modulation spectrum was computed rather than the raw RMS power. This was determined by subtracting the RMS power over all 24 modulation channels from the RMS of each modulation channel. This process was repeated for each Spectral Band. The power differences were then averaged across the 44 nursery rhyme samples and 6 speakers, and these averages are shown in Figure a. For all 5 Spectral bands, there is a clear peak in modulation power between 2-6 Hz. This is consistent with the data of Greenberg et al (2003) on modulation spectrum power (based on the 1-2 kHz band in speech).

Figure a. RMS modulation power across modulation rate for each spectral band.



However, there is a slight trend for the peak in modulation rate power to be higher for spectral bands of higher frequency, as was also observed by Plomp (1983b). For example, while the peak in power for Spectral Band 1 (100-300 Hz, red) occurs around 3 Hz, the peak in power for Spectral Band 5 (3900-7250 Hz, magenta) is between 4-5 Hz. Also the 5 Spectral Bands show opposite trends at the slowest and fastest modulation rates. At the slowest modulation rates (~ 1 Hz), low frequency Spectral Bands (i.e. 1 & 2) have the highest power while high frequency Spectral bands (i.e. 4 & 5) have the lowest power. At the fastest modulation rates (~40 Hz), this trend is reversed as now high frequency Spectral Bands show the highest power and low frequency Spectral bands show the lowest power. This trend can

be interpreted to mean that slow changes in speech energy (e.g. prosodic stress) tend to be represented more strongly at low spectral frequencies while fast changes in speech energy (e.g. noise bursts following the release of stop consonants) are more strongly represented at high spectral frequencies.

List of Nursery Rhyme Sentences Used to Develop Prosodic Indices & Evaluate Models

(Each sentence is 24 syllables long)

Duple Meter

1. Old MacDonald had a farm, E-I-E-I-O. And on that farm he had some cows, E-I-E-I ...
2. Mary had a little lamb its fleece was white as snow. And everywhere that Mary went the lamb was ...
3. Polly put the kettle on, Polly put the kettle on, Polly put the kettle on, we'll all have ...
4. Yankee Doodle came to London riding on a pony. He stuck a feather in his hat and called ...
5. Peter Peter pumpkin eater had a wife and couldn't keep her. Put her in a pumpkin shell and ...
6. Mary Mary quite contrary, how does your garden grow? With silver bells and cockle shells and pre ...
7. Simple Simon met a pieman going to the fair. Says Simple Simon to the pieman, "let me ...
8. Lucy Lockett lost her pocket, Kitty fisher found it. Not a penny was there in it, only ...
9. Cobbler Cobbler mend my shoe, get it done by half past two. Half past two is much too late, get it done ...
10. Peter Piper picked a peck of pickled peppers. A peck of pickled peppers Peter Piper picked ....

Triple Meter

1. Little Miss Muffet sat on a tuffet eating her curds and whey. There came a big spider who sat ...

2. Little Jack Horner sat in a corner eating his Christmas pie. He stuck in his thumb and pulled out ...

3. Little Boy Blue come blow your horn, the sheep's in the meadow the cow's in the corn. Where is the boy who ...

4. Curly locks, curly locks will you be mine? You shall not wash dishes nor feed the swine, but sit on a...

5. To market to market to buy a fat pig. Home again home again dancing a jig. To market ...

6. Pussycat pussycat where have you been? I've been up to London to visit the Queen. Pussycat ...

7. Ladybird ladybird fly away home, your house is on fire and your children are gone. All except ...

8. There was an old Lady who swallowed a spider, that wriggled and wiggled and tickled inside her ...

9. There once were two cats of Kilkenny. Each thought there was one cat too many. So they fought and they fit ...

10. Lavender's blue, dilly dilly, lavender's green. When I am king, dilly dilly, you shall be queen ...

Nursery Rhyme Duration and Rate of Speaking

Appendix 7.1

Nursery Rhyme	Repetitions	Total Sylls	Speaker 1				Speaker 2			
			ADS		CDS		ADS		CDS	
			Length (s)	Rate (syll/s)	Length (s)	Rate (syll/s)	Length (s)	Rate (syll/s)	Length (s)	Rate (syll/s)
Baa Baa Black Sheep	2	76	23.2	3.3	22.6	3.4	20.1	3.8	21.1	3.6
Once I Caught a Fish Alive	2	106	30.2	3.5	34.1	3.1	31.5	3.4	39.2	2.7
One Two Buckle my Shoe	1	69	22.0	3.1	28.0	2.5	23.5	2.9	25.5	2.7
Old MacDonald	2	118	30.3	3.9	36.8	3.2	31.2	3.8	36.2	3.3
Twinkle Twinkle Little Star	2	84	22.8	3.7	31.8	2.6	21.2	4.0	27.3	3.1
London Bridge	1	72	20.9	3.4	27.7	2.6	19.4	3.7	24.5	2.9
Mary Had a Little Lamb	1	111	26.6	4.2	35.5	3.1	26.7	4.2	29.9	3.7
Polly Put the Kettle On	2	102	25.7	4.0	33.9	3.0	21.7	4.7	26.1	3.9
Yankee Doodle	3	87	19.5	4.5	24.3	3.6	17.7	4.9	20.6	4.2
Peter Peter Pumpkin Eater	3	93	21.9	4.2	28.5	3.3	21.1	4.4	21.5	4.3
Mary Mary Quite Contrary	3	90	25.5	3.5	30.5	3.0	24.8	3.6	24.3	3.7
Simple Simon	2	116	29.7	3.9	34.3	3.4	27.8	4.2	30.7	3.8
St Ives	2	96	30.3	3.2	33.4	2.9	26.6	3.6	29.2	3.3
The Queen of Hearts	1	56	15.2	3.7	16.8	3.3	15.3	3.7	15.9	3.5
Lucy Lockett	3	84	21.7	3.9	24.1	3.5	19.8	4.2	20.5	4.1
Cobbler Cobbler	3	84	28.3	3.0	32.3	2.6	24.0	3.5	25.7	3.3
Peter Piper	2	100	25.4	3.9	25.6	3.9	23.4	4.3	22.9	4.4
I'm a Little Teapot	3	96	31.7	3.0	39.5	2.4	27.2	3.5	34.7	2.8
Sing a Song of Sixpence	1	97	24.5	4.0	26.6	3.6	23.3	4.2	25.9	3.7
Wee Willie Winkie	2	84	23.1	3.6	24.8	3.4	22.0	3.8	25.1	3.3
Old King Cole	1	74	16.7	4.4	19.6	3.8	17.3	4.3	19.9	3.7
The Wheels on the Bus	1	54	14.7	3.7	19.8	2.7	17.4	3.1	18.8	2.9
Three Little Monkeys	1	111	27.5	4.0	34.0	3.3	26.8	4.1	33.2	3.3
Grand Old Duke of York	1	62	13.9	4.5	16.7	3.7	14.2	4.4	17.5	3.5
Incy Wincy Spider	2	92	23.2	4.0	30.6	3.0	25.0	3.7	31.8	2.9
Jack and Jill	3	84	21.3	3.9	24.5	3.4	22.1	3.8	25.2	3.3
Humpty Dumpty	3	108	26.6	4.1	30.9	3.5	25.3	4.3	27.1	4.0
Ring-a-Ring-a-Roses	3	69	18.6	3.7	22.6	3.1	16.8	4.1	23.6	2.9
Row Row Row Your Boat	1	81	20.8	3.9	23.1	3.5	19.7	4.1	24.1	3.4
Hickory Dickory Dock	3	84	23.5	3.6	23.2	3.6	20.4	4.1	26.6	3.2
Little Boy Blue	2	106	28.3	3.7	32.1	3.3	27.8	3.8	29.2	3.6
Mulberry Bush	1	92	23.9	3.8	26.9	3.4	21.7	4.2	26.2	3.5
Ride a Cock Horse	2	82	20.8	3.9	23.0	3.6	19.6	4.2	20.4	4.0
To Market	2	126	35.6	3.5	38.3	3.3	30.4	4.1	30.6	4.1
Two Cats of Kilkenny	2	102	27.7	3.7	25.9	3.9	27.1	3.8	26.4	3.9
Pussycat Pussycat	2	84	23.3	3.6	23.1	3.6	20.1	4.2	22.1	3.8
Ladybird Ladybird	2	78	19.5	4.0	22.1	3.5	20.0	3.9	21.6	3.6
There Was An Old Lady	1	72	16.9	4.3	19.5	3.7	16.3	4.4	17.0	4.2
Orange and Lemons	1	88	23.4	3.8	26.1	3.4	20.5	4.3	27.7	3.2
Curly Locks	2	84	22.7	3.7	24.5	3.4	20.9	4.0	22.2	3.8
Rock-a-Bye-Baby	2	74	20.9	3.5	27.0	2.7	19.1	3.9	27.5	2.7
Lavender's Blue	1	72	19.0	3.8	22.3	3.2	16.8	4.3	25.2	2.9
		Means	***23.5***	***3.8***	***27.3***	***3.3***	***22.2***	***4.0***	***25.5***	***3.5***

Grand Mean Rate of Speaking (syll/s)	3.6	3.2
(SE)	*0.32*	*0.11*
	ADS	**CDS**

Appendix 7.1

Nursery Rhyme	Speaker 3				Speaker 4			
	ADS		CDS		ADS		CDS	
	Length (s)	Rate (syll/s)	**Length (s)**	Rate (syll/s)	**Length (s)**	Rate (syll/s)	**Length (s)**	Rate (syll/s)
Baa Baa Black Sheep	21.6	3.5	24.0	3.2	33.8	2.2	27.7	2.7
Once I Caught a Fish Alive	34.8	3.0	34.9	3.0	60.1	1.8	43.9	2.4
One Two Buckle my Shoe	26.1	2.6	27.3	2.5	43.6	1.6	31.3	2.2
Old MacDonald	36.7	3.2	34.7	3.4	59.4	2.0	52.2	2.3
Twinkle Twinkle Little Star	26.9	3.1	28.4	3.0	37.6	2.2	37.1	2.3
London Bridge	25.1	2.9	21.7	3.3	35.2	2.0	31.1	2.3
Mary Had a Little Lamb	32.9	3.4	30.5	3.6	41.1	2.7	46.0	2.4
Polly Put the Kettle On	30.5	3.3	24.7	4.1	32.5	3.1	38.0	2.7
Yankee Doodle	24.0	3.6	23.2	3.8	25.7	3.4	27.1	3.2
Peter Peter Pumpkin Eater	28.1	3.3	27.2	3.4	35.9	2.6	28.1	3.3
Mary Mary Quite Contrary	30.0	3.0	28.2	3.2	36.0	2.5	34.3	2.6
Simple Simon	36.1	3.2	31.2	3.7	48.9	2.4	39.7	2.9
St Ives	34.6	2.8	30.8	3.1	46.8	2.1	37.1	2.6
The Queen of Hearts	17.8	3.1	18.4	3.0	27.2	2.1	21.0	2.7
Lucy Lockett	27.8	3.0	24.0	3.5	35.0	2.4	27.0	3.1
Cobbler Cobbler	32.7	2.6	27.0	3.1	39.0	2.2	36.1	2.3
Peter Piper	31.1	3.2	25.6	3.9	38.5	2.6	24.4	4.1
I'm a Little Teapot	23.1	2.8	24.8	2.6	40.0	2.4	42.3	2.3
Sing a Song of Sixpence	29.4	3.3	24.8	3.9	37.6	2.6	30.8	3.1
Wee Willie Winkie	27.8	3.0	22.9	3.7	34.7	2.4	27.9	3.0
Old King Cole	22.3	3.3	18.2	4.1	26.8	2.8	24.4	3.0
The Wheels on the Bus	21.5	2.5	20.2	2.7	21.7	2.5	24.3	2.2
Three Little Monkeys	34.2	3.2	31.1	3.6	36.9	3.0	39.5	2.8
Grand Old Duke of York	18.0	3.4	16.3	3.8	25.9	2.4	18.9	3.3
Incy Wincy Spider	31.1	3.0	31.7	2.9	30.2	3.0	36.2	2.5
Jack and Jill	31.8	2.6	26.9	3.1	31.0	2.7	28.0	3.0
Humpty Dumpty	38.2	2.8	29.1	3.7	40.6	2.7	36.0	3.0
Ring-a-Ring-a-Roses	25.1	2.7	22.1	3.1	28.6	2.4	24.8	2.8
Row Row Row Your Boat	28.7	2.8	23.8	3.4	35.5	2.3	30.3	2.7
Hickory Dickory Dock	31.0	2.7	24.4	3.4	36.2	2.3	32.6	2.6
Little Boy Blue	37.7	2.8	30.0	3.5	39.7	2.7	36.6	2.9
Mulberry Bush	29.5	3.1	23.2	4.0	35.5	2.6	27.7	3.3
Ride a Cock Horse	24.7	3.3	22.1	3.7	33.0	2.5	25.8	3.2
To Market	47.1	2.7	35.1	3.6	67.3	1.9	35.7	3.5
Two Cats of Kilkenny	34.5	3.0	29.3	3.5	48.0	2.1	39.0	2.6
Pussycat Pussycat	30.2	2.8	22.0	3.8	35.4	2.4	25.9	3.2
Ladybird Ladybird	25.1	3.1	20.6	3.8	28.9	2.7	26.0	3.0
There Was An Old Lady	21.5	3.3	18.1	4.0	22.1	3.3	26.4	2.7
Orange and Lemons	31.3	2.8	21.6	4.1	32.2	2.7	30.4	2.9
Curly Locks	29.2	2.9	22.5	3.7	39.0	2.2	28.7	2.9
Rock-a-Bye-Baby	27.3	2.7	24.4	3.0	29.5	2.5	28.8	2.6
Lavender's Blue	25.8	2.8	18.3	3.9	24.1	3.0	24.2	3.0
	29.1	***3.0***	***25.4***	***3.5***	***36.6***	***2.5***	***31.7***	***2.8***

	Speaker 5				Speaker 6			
	ADS		CDS		ADS		CDS	
Nursery Rhyme	**Length (s)**	Rate (syll/s)	**Length (s)**	Rate (syll/s)	**Length (s)**	Rate (syll/s)	**Length (s)**	Rate (syll/s)
Baa Baa Black Sheep	17.1	4.4	27.2	2.8	20.5	3.7	27.6	2.8
Once I Caught a Fish Alive	27.9	3.8	43.4	2.4	28.3	3.7	40.0	2.7
One Two Buckle my Shoe	20.7	3.3	29.9	2.2	19.6	3.5	25.7	2.7
Old MacDonald	26.5	4.5	32.5	3.6	33.3	3.5	36.1	3.3
Twinkle Twinkle Little Star	26.2	3.2	41.6	2.0	22.1	3.8	35.5	2.4
London Bridge	16.4	4.4	24.3	2.7	20.3	3.5	27.4	2.6
Mary Had a Little Lamb	26.6	4.2	32.7	3.4	27.6	4.0	35.6	3.1
Polly Put the Kettle On	18.9	5.4	24.1	4.2	24.1	4.2	28.1	3.6
Yankee Doodle	16.1	5.4	23.2	3.8	18.2	4.8	27.3	3.2
Peter Peter Pumpkin Eater	20.4	4.6	31.8	2.9	23.0	4.0	34.0	2.6
Mary Mary Quite Contrary	22.8	3.9	37.5	2.4	15.4	3.9	23.5	2.6
Simple Simon	25.7	4.5	40.4	2.9	27.6	4.2	36.9	3.1
St Ives	20.6	4.7	32.6	2.9	27.3	3.5	36.0	2.7
The Queen of Hearts	13.4	4.2	19.2	2.9	12.9	4.3	20.8	2.7
Lucy Lockett	17.1	4.9	21.7	3.9	23.1	3.6	27.4	3.1
Cobbler Cobbler	19.4	4.3	28.9	2.9	24.8	3.4	32.3	2.6
Peter Piper	17.8	5.6	27.0	3.7	29.2	3.4	32.6	3.1
I'm a Little Teapot	20.8	4.6	36.6	2.6	32.0	3.0	37.6	2.6
Sing a Song of Sixpence	21.8	4.4	26.4	3.7	24.0	4.0	30.7	3.2
Wee Willie Winkie	17.0	4.9	23.7	3.5	22.5	3.7	23.9	3.5
Old King Cole	12.7	5.8	18.2	4.0	17.7	4.2	22.4	3.3
The Wheels on the Bus	12.4	4.4	20.5	2.6	17.0	3.2	19.5	2.8
Three Little Monkeys	24.1	4.6	43.3	2.6	25.4	4.4	32.6	3.4
Grand Old Duke of York	11.5	5.4	17.1	3.6	14.9	4.2	17.1	3.6
Incy Wincy Spider	19.7	4.7	38.7	2.4	26.9	3.4	38.0	2.4
Jack and Jill	18.1	4.6	30.2	2.8	24.2	3.5	26.4	3.2
Humpty Dumpty	19.4	5.6	29.9	3.3	26.8	4.0	32.6	3.3
Ring-a-Ring-a-Roses	13.3	5.2	19.7	3.5	16.3	4.2	18.2	3.5
Row Row Row Your Boat	17.2	4.7	21.5	3.8	22.3	3.6	30.5	2.7
Hickory Dickory Dock	17.3	4.9	29.8	2.8	24.9	3.4	29.6	2.8
Little Boy Blue	24.0	4.4	44.4	2.4	27.5	3.9	34.4	3.1
Mulberry Bush	16.4	5.6	26.6	3.5	22.3	4.1	28.1	3.3
Ride a Cock Horse	17.0	4.8	23.2	3.5	20.8	3.9	24.5	3.3
To Market	23.8	5.3	31.0	4.1	35.3	3.6	35.9	3.5
Two Cats of Kilkenny	20.6	5.0	28.2	3.6	30.4	3.4	32.0	3.2
Pussycat Pussycat	16.6	5.1	30.8	2.7	23.7	3.5	24.6	3.4
Ladybird Ladybird	16.8	4.6	27.6	2.8	19.1	4.1	23.5	3.3
There Was An Old Lady	13.5	5.3	26.3	2.7	16.6	4.3	19.8	3.6
Orange and Lemons	17.2	5.1	33.3	2.6	22.5	3.9	25.7	3.4
Curly Locks	19.9	4.2	32.8	2.6	22.2	3.8	28.4	3.0
Rock-a-Bye-Baby	19.2	3.9	32.9	2.2	20.2	3.7	27.1	2.7
Lavender's Blue	16.1	4.5	23.5	3.1	19.1	3.8	22.8	3.2
	19.0	***4.7***	***29.4***	***3.1***	***23.1***	***3.8***	***28.9***	***3.0***

The Conditional Entropy Measure

a. Definitions of Entropy, Conditional Probability and Conditional Entropy

Entropy is a measure of the uncertainty of a random variable. If X is a discrete random variable, and p(x) is its probability mass function for all possible values of X, then its entropy, H(X), can be calculated as :

$$H(X) = -\sum p(x) \log p(x) \qquad (Eq.\ 1)$$

When the logarithm taken is to the base 2, then the unit for entropy is in 'bits'. As an example, one can use this formula to compute the entropy of a fair coin toss, where the probability of heads and tails are both equal at 0.5. In this case,

$$\begin{aligned} H(\text{coin toss}) &= -\left(0.5 \times \log_2(0.5) + 0.5 \times \log_2(0.5)\right) \\ &= 1 \text{ bit} \end{aligned}$$

Therefore, one can infer that there is 1 'bit' of information associated with the event of the coin toss.

When there are 2 random variables, X & Y, we can calculate the *conditional* probability of one event occurring, given that we already know the outcome of the other event. For example, let's say we roll two six-sided dice, and X represents the outcome of the first dice, and Y represents the outcome of the second dice. Now imagine that we see that the first dice has landed on '1' (i.e. X = 1), but the second dice has rolled away under the table and we cannot see its result. What is the probability that both dice landed on '1', given that we already know that the first dice landed on '1'? This question can be re-stated in terms of conditional probability as P(Y=1\X=1) or the probability that Y = 1 *given that* X = 1. In this case, the answer is 1/6, rather than 1/36, which would have been the probability that both die landed on '1' and we did *not* already know the result of the first die.

In a similar way, we can compute the *conditional entropy* of one event, conditional upon another event. This conditional entropy, H(Y\X) is given as :

$$H(Y\backslash X) = -\sum p(x) \sum p(y/x) \log p(y/x) \qquad (Eq.\ 2)$$

where p(x) is the probability distribution of the random variable X, and p(y/x) is the conditional probability distribution of Y, given X.

b. Conditional Entropy of the Peak-Phase Distribution

We can now apply the concept of conditional entropy to the AM hierarchy. Let's say 'peaks' in the Syllable AM represent syllable vowel nuclei, and we would like to track the likelihood of occurrence of these vowel nuclei over a brief period of time (e.g. 50 ms). For example, in one 50 ms window, we may see a Syllable peak, but in another 50 ms window we wouldn't. On average, let's say we see a Syllable peak every 5 such windows or so. Therefore, the probability of there being a Syllable peak in one randomly-selected 50 ms time window is 1/5 (20%). Conversely, the probability of *not* observing a peak in that time window is 4/5 (80%). Therefore, according to Equation 1, the entropy associated with Syllable peak occurrence is :

$$H(\text{peaks}) = -(0.2 \times \log_2(0.2) + 0.8 \times \log_2(0.8))$$

$$= 0.72 \text{ bits}$$

This entropy value expresses how uncertain we are, if we had to make a guess (predict) whether we would observe a Syllable peak in a new 50 ms time window.

Now let's say we want to improve the certainty of our guess about the occurrence of Syllable peaks by making use of other information that is related to the occurrence of Syllable peaks. For example, we might know that Syllable peaks tend to occur most often at specific Stress AM phase regions. Therefore, if we knew the concurrent Stress AM phase during the 50 ms window, this could tell us whether we were likely to see *more* Syllable peaks or *less* Syllable peaks during that period. Of course, the extent to which knowing Stress AM phase reduces our uncertainty depends on how strongly related Stress AM phase is to the occurrence of Syllable peaks in the first place. If there was only a weak relationship between these two variables, then knowing Stress AM phase might not help us very much. On the other hand, if there was a very strong conditional relationship between the two variables (i.e. Syllable peaks only ever occur at one Stress AM phase value), then knowing the Stress AM phase would be *very* helpful in reducing our uncertainty about Syllable peak occurrence.

One way of quantifying the strength of the relationship between the two variables (Syllable peaks and Stress phase) is to compute the conditional entropy of one event given the other, or H(peak/phase). If the conditional entropy (uncertainty) is very low, then the two variables can be said to have a strong relationship. If the conditional entropy is high, then the two variables are only weakly related. This relationship can be visualised by looking at the

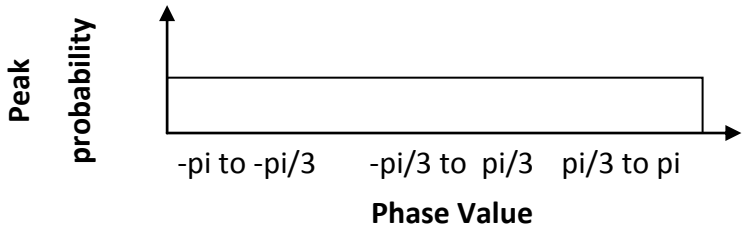
conditional distribution of Syllable peaks given Stress phase, and the conditional entropy value is also computed using this conditional distribution (see Equation 2). Let us now consider the types of conditional peak-phase distributions that could be observed and their implications.

i. Rectangular peak-phase distribution

Under a flat or rectangular peak-phase distribution (see Figure a and Table a), Syllable peaks occur at every Stress phase value with an equal probability. Therefore, knowing the current Stress phase would not change our guess about Syllable peaks in any way. To illustrate this in concrete terms, let's say we randomly sample 15 separate 50 ms time windows from the Syllable modulator. Since Syllable peaks have a base probability of 20% (from the previous page), we can expect 3 out of the 15 windows to contain a Syllable peak, while the other 12 windows would not contain a Syllable peak. If the Syllable peak-Stress phase joint distribution is perfectly flat, this would suggest that the 3 Syllable peaks all occurred at different Stress phases, with no preference toward a specific phase value. Therefore, for ease of illustration, let's divide Stress phase into 3 equal phase bins, and say that 1 Syllable peak occurred in each Phase bin. These frequencies are captured in Table a below, which therefore shows the *joint probability distribution* of Syllable peaks for every Stress phase. In the table, 1 out of the 15 observations contained a Syllable peak that occurred during -pi to -pi/3 Stress phase, another 1 observation contained a Syllable peak that occurred during -pi/3 to pi/3 Stress phase, and a final 1 observation contained a Syllable peak that occurred during pi/3 to pi Stress phase. The remaining 12 observations that did *not* contain a Syllable peak were equally divided among the 3 phase bins (i.e. 4 observations per bin). Notice that the marginal probability of peak or no peak occurrence (the far right column) is the sum of all the individual peak occurrences at each phase value. Notice also that for all 3 Stress phase values (columns), the conditional probability of peak occurrence, P(peak\phase), is exactly the same. Based on this joint probability distribution table, the conditional entropy of Syllable peaks *given* Stress phase can be calculated, which works out to be **0.72 bits**. Notice that this entropy value is exactly the same as the entropy value for peak occurrence when we did not take into account Stress phase. Therefore, when the peak-phase distribution is perfectly flat, knowing the Stress phase beforehand does *not* help us to make a better guess about whether the time window contains a Syllable peak or not. Similarly, this indicates that

the two variables (Stress phase and Syllable peaks) are unrelated, since Stress phase does not exert any constraints on Syllable peak occurrence.

Figure and Table a. Rectangular Peak-Phase Distribution



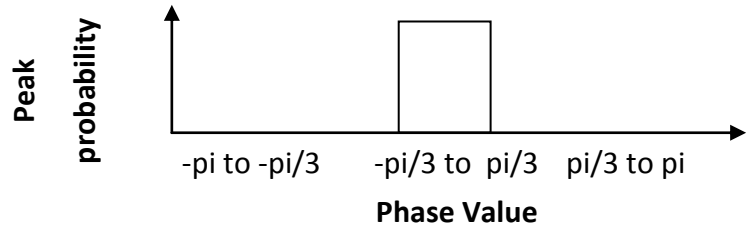
Stress phase	-pi to -pi/3	-pi/3 to pi/3	pi/3 to pi	Marginal Probability (peak or no peak)
Peak	1/15	1/15	1/15	1/15 + 1/15 + 1/15 = 3/15 (20%)
No peak	4/15	4/15	4/15	4/15 + 4/15 + 4/15 = 12/15 (80%)
P(peak\phase)	*(1/15) / (5/15) = 20%*	*(1/15) / (5/15) =20%*	*(1/15) / (5/15) =20%*	

ii. Unimodal peak-phase distribution

Now, let us consider the other extreme case of a perfectly narrow, unimodal peak-phase distribution, such as that in Figure b. In this case, Syllable peaks only ever occur at one Stress phase value. Therefore, if we were to divide the same 3 Syllable peak observations according to the Stress phase bin, all 3 of them would have occurred in the second phase bin, and none would have occurred in the first or third phase bins (see Table b). Accordingly, if we assume that the same total number of observations (5) occurred at each phase value, irrespective of whether they contained a peak or not, this would result in the frequency distribution shown in Table b. Notice that while the marginal probability of peak occurrence (far right column) is the same as in the previous case, the conditional probabilities (bottom row) are now *not* the same across the 3 phase values. The conditional entropy of this unimodal peak-phase distribution can then be computed (assuming log(0) = 0), and works out to be just **0.32 bits**, which is lower than the entropy of the original entropy of the Syllable

peak distribution (0.72 bits) as well as that of the flat peak-phase distribution. This lowered conditional entropy value indicates a stronger relationship between Stress phase and Syllable peaks, where phase is exerting a constraint on the occurrence of Syllable peaks. Also, in this case, knowing about Stress phase beforehand is very helpful in predicting whether or not a random time window contains a syllable peak, because if the Stress phase is between -pi/3 to pi/3, seeing a Syllable peak is quite likely, but if the Stress phase is outside of this range, then we can be sure that no Syllable peaks would occur. Therefore, the *shape* of the peak-phase distribution provides a very strong indication as to the relatedness of the two variables. In general, the sharper (more kurtotic) the distribution, the lower the conditional entropy, and the more related the two variables are.

Figure b and Table b. Unimodal Peak-Phase Distribution



Phase value	-pi to -pi/3	-pi/3 to pi/3	pi/3 to pi	Marginal Probability (peak or no peak)
Peak	0/15	3/15	0/15	0/15 + 3/15 + 0/15 = 3/15 (20%)
No peak	5/15	2/15	5/15	5/15 + 2/15 + 5/15 = 12/15 (80%)
P(peak\phase)	*(0/15) / (5/15) = 0%*	*(3/15) / (5/15) = 60%*	*(0/15) / (5/15) = 0%*	

c. Parameters for Computing Conditional Entropy in the CDS-ADS Study

In the experimental study with child- and adult-directed speech, the conditional entropy measure was used as an index for the strength of the relationship between Stress phase and Syllable peak occurrence, as well as Syllable phase and Phoneme peak occurrence. Here, instead of 3 phase bins, 17 equally-spaced phase bins were used, and conditional entropy was computed as per Equation 2. The number of phase bins affects the maximum entropy of a variable. For example, if just 2 phase bins were used, the maximum entropy would be 1 bit. For 17 phase bins, the maximum entropy increases to 4.1 bits because there are more possible outcomes. Also, the entropy estimated using a limited or finite sample set (such as the speech sentences used in the experimental study) is always an underestimation of the true entropy of the variable (Treves & Panzeri, 1995). Although methods have been developed to estimate and correct for this bias, entropy estimation is still considered problematic by some (Paninski, 2003). Therefore, in this study, it is acknowledged that the entropy values computed are flawed estimates of the 'true' entropy values, and only ***differences*** in entropy observed between the experimental conditions are considered meaningful (not the absolute entropy values themselves). Furthermore, the conditional entropy index is merely used as a measure for quantifying differences in the shape peak-phase distribution (which is the true variable of interest). Therefore, the claims made from the study pertain to the hierarchical peak-phase distribution, rather than to the entropy of speech per se.

An alternative to computing conditional entropy is to compute 'mutual information', or (I(X;Y). The two measures are mathematically-related where the mutual information between two variables is the absolute entropy of one variable minus the conditional entropy between the variables :

$$I(X;Y) = H(X) - H(X\backslash Y) \qquad \textit{(Eq. 3)}$$

For the examples described in this Appendix, the mutual information between Stress phase and Syllable peaks for the rectangular distribution would be 0.72 - 0.72 = 0 bits, and for the unimodal distribution would be 0.72 - 0.32 = 0.4 bits. Therefore mutual information is high when conditional entropy is low, and the two measures are strongly related to each other.

Finally, in the examples described in this Appendix, the time window for Syllable peak observation was (for ease of illustration) set at 50 ms. In the CDS-ADS study, the time

window corresponded to 1 sample at the sampling rate of 1050 Hz, which was a window length of 1.05 ms.

UNIVERSITY OF CAMBRIDGE

Confidential May 2011

Information Sheet

Speech Rhythm Perception in Dyslexia

Dear participant,

Thank you for your interest in this study on speech rhythm perception in dyslexia. The purpose of this study is to understand how rhythmic patterns in speech are processed differently by people with and without developmental dyslexia.

There will be two parts to the experimental session. In the first part, you will be asked to complete a series of short tests for memory, vocabulary and reading. In the second part, you will do several computer-based experiments on rhythm perception. For example, you may be asked to tap along to the rhythm of a song or rhyme, or to match rhythmic patterns with nursery rhymes.

In total, the whole study should take 2.5 hours and you will be paid an honorarium of £20 for your participation.

Confidentiality/Ethical Approval

All data will be identified by a code, with names kept in a locked file. Results are normally presented in terms of groups of individuals and will be presented at conferences and written up in journals. If any individual data were to be presented, the data would be totally anonymous, without any means of identifying the individuals involved. This project has received ethical approval from the Cambridge Psychology Research Ethics Committee (University of Cambridge).

If you would like further information on any of the above, please do not hesitate to contact Victoria Leong at vvec2@cam.ac.uk .

Yours sincerely,

Victoria Leong
PhD Candidate

CONSENT FORM

Speech Rhythms in Nursery Rhymes

Victoria Leong and Prof. Usha Goswami,
Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB
Tel. 01223 333550 Email:vvec2@cam.ac.uk

Have you read the information sheet about the study?	YES/NO
Have you received sufficient information about the study?	YES/NO
Do you understand that you are free to withdraw from the study at any time and without giving a reason for withdrawing?	YES/NO
Do you agree to take part in this study?	**YES/NO**

Name in block letters _____

Signed _____ Date _____

Contact telephone number _____

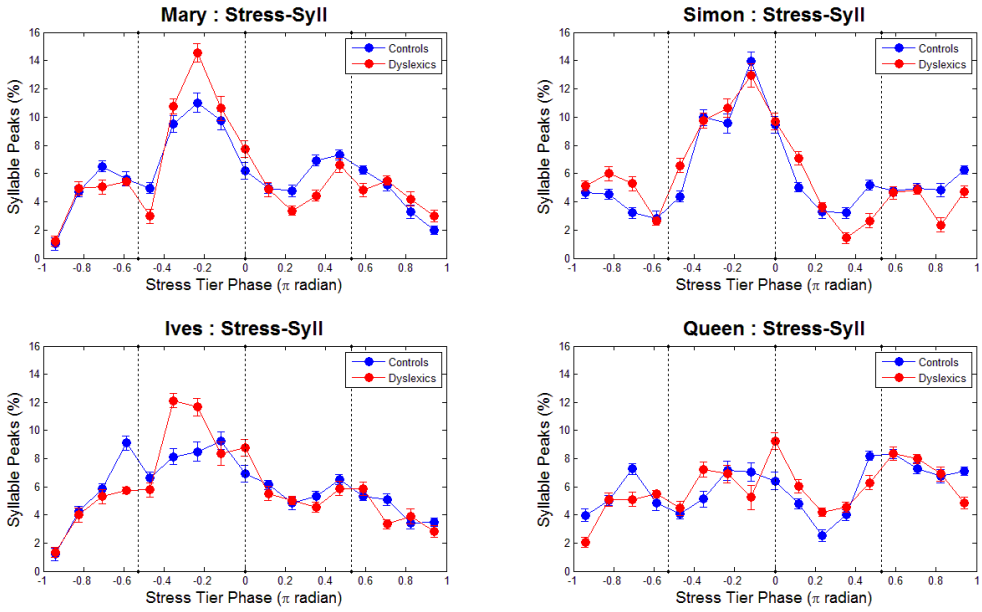
or Email _____

PARTICIPANT BACKGROUND INFORMATION SHEET

GENERAL INFORMATION		
1	Name	
2	Sex	M / F
3	Date of Birth	____ / ____ / 19___
4	Handedness	Right / Left / Ambidextrous
READING DEVELOPMENT		
5	Do you have reading/spelling problems?	Yes / No
5b	*If yes, do you have a formal diagnosis of dyslexia?*	Yes / No
5c	*At what age did you receive your diagnosis?*	_____ yrs
6	Do you have any visual impairments?	Yes / No If yes : _____
7	Do you have any hearing impairments?	Yes / No If yes : _____
8	Do you have any neurological disorders?	Yes / No If yes : _____
9	Do you have (or had as a child) language impairments?	Yes / No If yes : _____
10	Do you have a diagnosis of any other learning or developmental difficulties?	Yes / No If yes : _____
11	Are you a native speaker of English?	Yes / No
11b	*If yes, which country and region are you from?*	_____ , _____
FURTHER INFORMATION		
12	Would you like to receive information on the findings of the present research study?	Yes / No Email : _____
13	May we contact you about participating in our future studies?	Yes / No

Peak-Phase Distributions for Each Nursery Rhyme

(a) Distribution of Syllable tier peaks with respect to Stress phase for each nursery rhyme sentence.



(b) Distribution of Phoneme tier peaks with respect to Syllable phase for each nursery rhyme sentence.