A role for amplitude modulation phase relationships in speech rhythm perception

Victoria Leong[1,a)], Michael A. Stone[2], Richard E. Turner[3], and Usha Goswami[1]

[1]*Centre for Neuroscience in Education, Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB,United Kingdom*

[2]*Auditory Perception Group, Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB,United Kingdom*

[3]*Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*

a) Author to whom correspondence should be addressed. Electronic mail : vvec2@cam.ac.uk

ABSTRACT

Prosodic rhythm in speech (the alternation of 'Strong' (S) and 'weak' (w) syllables) is cued, among others, by slow rates of amplitude modulation (AM) within the speech envelope. However, it is unclear exactly which envelope modulation rates and statistics are the most important for the rhythm percept. Here, the hypothesis that the *phase relationship* between 'Stress' rate (~2 Hz) and 'Syllable' rate (~4 Hz) AMs provides a perceptual cue for speech rhythm is tested. In a rhythm judgment task, adult listeners identified AM tone-vocoded nursery rhyme sentences that carried either trochaic (S-w) or iambic patterning (w-S). Manipulation of listeners' rhythm perception was attempted by parametrically phase-shifting the Stress AM and Syllable AM in the vocoder. It was expected that a 1π radian phase-shift (half a cycle) would *reverse* the perceived rhythm pattern (i.e.trochaic→iambic) whereas a 2π radian shift (full cycle) would *retain* the perceived rhythm pattern (i.e.trochaic→trochaic). The results confirmed these predictions. Listeners judgments of rhythm systematically followed Stress-Syllable AM phase-shifts, but were unaffected by phase-shifts between the Syllable AM and the Sub-beat AM (~14 Hz) in a control condition. It is concluded that the Stress-Syllable AM phase relationship is an envelope-based modulation statistic that supports speech rhythm perception.

(199/200 words)

PACS numbers : 43.71.An, 43.71.Sy, 43.66.Mk, 43.71.Rt

Keywords: Amplitude modulation, amplitude envelope, speech rhythm, language acquisition

I. INTRODUCTION

A. Acoustic cues to speech rhythm

Rhythm commonly refers to an alternating pattern of 'Strong' and 'weak' elements (Schane, 1979; Lerdahl & Jackendoff, 1983). In speech, this rhythmic alternation is expressed as patterns of stressed (Strong, 'S') and unstressed (weak, 'w') syllables, which form rhythmic units known as 'prosodic feet' (motifs comprising of 1 stressed syllable plus 1 or more unstressed syllables). In English, two of the most common prosodic feet are the trochee ('S-w', e.g. 'BA-by') and the iamb ('w-S', e.g. 'gui-TAR'). Adult listeners are able to use speech rhythm patterns to segment the continuous speech signal (Kim et al, 2008), even though the rhythms in normal speech are not perfectly regular in timing (e.g. as compared to music). We know that several acoustic cues contribute to the perception of rhythm, but we do not know whether some of them are more powerful than others, whether there is a hierarchy, and whether some of these cues are valid universally whereas others can be of value in some languages only.

Traditionally, strong and weak syllables were thought to be cued primarily by differences in fundamental frequency (Fry, 1955,1958; Bolinger, 1958). However, more recent studies using larger corpora of natural speech have found that amplitude and duration cues are better discriminators of prosodic stress than fundamental frequency (Greenberg, 1999; Kochanski et al, 2005). Consequently, the focus of the current investigation is on the rhythm contribution of *slow* amplitude modulation (AM) cues that are present within the envelope of speech (i.e. up to 50 Hz, Rosen, 1992), rather than on faster temporal modulations (i.e. 50-500 Hz, Rosen, 1992) that typically carry information about fundamental frequency as well as possibly some information about syllable onsets.

Historically, there has been a particular focus on the role of *syllable*-rate information in speech perception studies (e.g. Miller & Licklider, 1950; Drullman et al, 1994a, Drullman et al, 1994b; Greenberg et al, 2003; Ghitza & Greenberg, 2009). For example, in 'glimpsing' studies in which portions of speech are periodically removed or masked by noise, the resulting intelligibility of the signal is optimal when the interruption rate or 'temporal packaging' corresponds to the syllable rate of speech (Miller & Licklider, 1950; Ghitza & Greenberg, 2009). The syllable rate has also featured prominently in recent *neural* models of speech perception (e.g., Poeppel, 2003; Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012; described further in Section B1). However, the modulation spectrum of the envelope contains AM patterns across a range of modulation rates, and the contribution of supra-syllabic modulation rates toward speech perception is seldom studied (although see Fullgrabe et al, 2009).Specifically, it is not known whether, and how, listeners combine syllable-rate modulation information with the modulation information that occurs at slower rates (e.g. the 'stress' rate), and what listening benefit this combination of rates confers.

It has previously been noted that prosodic rhythm and stress are associated with slow-varying patterns of amplitude modulation (AM) within the speech amplitude envelope (Plomp, 1983; Howell, 1984, 1988a, 1988b; Greenberg et al, 2003; Tilsen & Johnson, 2008; Tilsen & Arvaniti, 2013). Here, the hypothesis that the combination of syllable-rate (~4 Hz) and stress-rate (~2 Hz) modulation information cues the Strong-weak rhythm pattern of spoken sentences is tested. Moreover, it is predicted that the *phase relationship* between syllable-rate and stress-rate modulations determines Strong-weak rhythm patterning in a parametric fashion. The rationale for this phase prediction is explained further in Section B.

Speech rhythms also play an important role in early language acquisition. From as young as 3 months of age, infants are exposed to rhythmically-rich speech when their mothers sing or recite nursery rhymes to them (Trevarthen, 1986, 1987). Children's nursery

rhymes are simple poems with a metrically-regular rhythm, typically comprised of trochees and iambs (Gueron, 1974). By 7.5 months of age, English-learning infants already show a listening preference for the dominant trochaic rhythm pattern (Jusczyk et al, 1993), and are able to use this rhythm pattern as a template for word segmentation (Jusczyk et al, 1999). During early childhood, children's knowledge of nursery rhymes is a strong predictor of their later phonological skill and success in learning to read (Maclean et al, 1987; Bryant et al, 1989). Thus, good perceptual sensitivity to the acoustic cues that transmit speech rhythm is thought to be critical for language development (Whalley & Hansen, 2006; Tierney & Kraus, 2013). Indeed, poor perceptual sensitivity to acoustic rhythm is frequently associated with phonological and reading difficulties in children (Wood & Terrell, 1998; Goswami, 2011).

B. Expressing rhythm as phase relationships within an AM hierarchy

1. Representing the linguistic prosodic hierarchy as an AM hierarchy

In metrical phonology, the prosodic (and rhythmic) structure of speech is conceptualised as an hierarchy, where different hierarchical tiers represent different rhythmic elements such as syllables and stress feet (Selkirk, 1980, 1984, 1986; Liberman & Prince, 1977; Hayes, 1995). Here, the modulation spectrum of the envelope is organised into a 5-tier *AM hierarchy* that captures major rhythmic units from the linguistic prosodic hierarchy. That is, the AM rates that are the most likely to transmit phrasal patterning ('Slow' rate AM), prosodic stress patterning ('Stress' rate AM), syllable patterning ('Syllable' rate AM), sub-syllable patterning ('Sub-beat' rate AM) and phonemic patterning ('Fast' rate AM) are identified. For example, as the average duration of a syllable is ~200 ms, AMs around 3-5 Hz are most likely to relate to syllable-pattern information in speech (Greenberg et al, 2003; Greenberg, 2006). In the simplest rhythmic case where every other syllable is stressed (e.g. a

'S-w' trochaic pattern), the stress rate of modulation should be around half that of the syllable rate (i.e. 2 Hz for a 4 Hz syllable rate). Consistent with this proposal, Dauer (1983) found that the average duration of inter-stress intervals in English was 493 ms, corresponding to a stress rate of ~2 Hz. Accordingly, AMs slower than the 2 Hz rate are likely to correspond to longer rhythmic units such as intonational phrases.

Toward the other end of the modulation spectrum, faster modulations immediately above the 'classic' peak syllable rate of 3-5 Hz should correspond to more quickly-uttered unstressed syllables (~10 Hz, Greenberg et al, 2003). In a rhythmic context, 2 or more such unstressed syllables may be compressed to fit within the 'beat' length of one ordinary syllable, hence the label 'Sub-beat'. For example, in the nursery rhyme sentence "Humpty Dumpty sat on the wall", the syllables "sat" and "on" are compressed together to fit the space of one regular syllable like "Hum". Finally, much faster modulations up to 50 Hz provide phonemic cues to manner of articulation, voicing, and vowel identity (Rosen, 1992).

Recently, Poeppel, Ghitza, Greenberg and others have proposed neural accounts of speech processing that are based on a 'multi-time resolution' of the AM patterns in the speech envelope (multi-time resolution models, e.g., Poeppel, 2003; Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012; although see Obleser et al, 2012 for criticisms of this approach). In multi-time resolution models, neuronal oscillations at different timescales entrain ('phase-lock') to speech modulation patterns on equivalent timescales, so that peaks and troughs in oscillatory activity align with peaks and troughs in modulation within the signal. For example, neuronal oscillatory activity in the Theta band (3-7 Hz) is thought to track syllable patterns in speech. It has been hypothesised (although not empirically demonstrated) that slower oscillatory activity in the Delta band (1-3) Hz tracks phrasal and intonational patterns, such as stress intervals (Ghitza & Greenberg, 2009; Ghitza, 2013). Similarly, fast oscillatory activity in the Gamma band (25-80 Hz) is thought to track quickly-varying phonetic

information, such as formant transitions and voice-onset times, which have timescales in the order of tens of milliseconds. Here, a similar 'multi-timescale' approach to investigating AM-based rhythm cues within the speech envelope is adopted. The contribution of different linguistically-significant AM rates toward listeners' rhythm perception is explored. Based on prior literature, it is predicted that listeners' judgments of rhythm will be determined primarily by the AM patterns at Stress (~2 Hz) and Syllable (~4 Hz) rates. Further, the current approach *extends* the work of Poeppel, Ghitza and Greenberg by proposing that oscillatory *phase relationships* between specific AM rates plays an important role in listeners' perception of speech rhythm. Specifically, it is predicted that the particular pattern of strong and weak syllables that is perceived by listeners (e.g. 'S-w' or 'w-S') depends on the *phase relationship* between Stress and Syllable rates of amplitude modulation. The rationale for the role of *hierarchical phase relationships* (rather than absolute amplitude) in rhythm perception is explained in the next two sections.

2. Relative prominence expressed as oscillatory phase

In the linguistic prosodic hierarchy, rhythm patterns are determined by the *relative* prominence of adjacent elements (Liberman & Prince, 1977; Selkirk, 1984; Hayes, 1995). That is, each element (e.g. syllable) is either stronger or weaker than its neighbour, producing a pattern of strong-weak alternation. Thus, it is not the *absolute* amplitude of a syllable that determines its status as strong or weak, but rather its *relative* amplitude in relation to the other syllables within the stress foot. The concept of oscillatory *phase* perfectly captures this principle of relative prominence within the AM hierarchy. By taking only the phase series of an AM pattern, one is effectively left with its relative pattern of amplitude change, disregarding any fluctuations in overall power. Thus, strong-weak rhythmic alternation can conveniently be expressed in terms of oscillatory phase states. Here, the phase convention

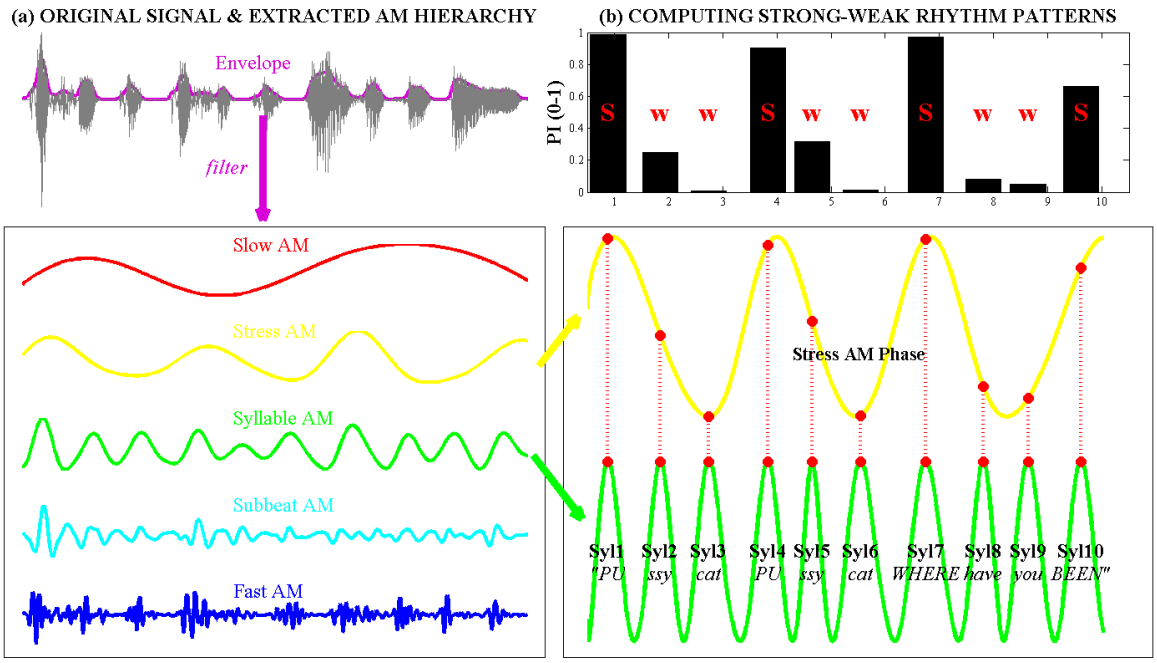
where +/-1π radian refer equally to the oscillatory trough and 0π radians refers to the oscillatory peak is used. Thus, the trochaic [S,w] pattern can be expressed in phase terms as [0π,+/-π], whereas the iambic [w,S] pattern is expressed as [+/-π, 0π].

3. Hierarchical phase relationships

A final property of the linguistic prosodic hierarchy is that rhythmic prominence (e.g. whether a syllable is strong or weak) accrues in a top-down hierarchical manner (Liberman & Prince, 1977; Hayes, 1995). Elements at a higher level in the hierarchy (e.g. stress feet) govern the prominence of elements at a lower level in the hierarchy (e.g. syllables). Thus, the final prominence of a given syllable will depend on both its position within the stress foot and the prominence of that stress foot within the over-arching phrase. Similarly, in the AM hierarchy, the oscillatory phase of a *slower* AM (e.g. the Stress-rate AM) governs the relative prominence of elements represented by a *faster* AM (e.g. the Syllable-rate AM). To illustrate this hierarchical phase-coding of rhythmic prominence, Figure 1 shows the nursery rhyme sentence "Pussycat, pussycat, where have you been?" (stress pattern : 'S-w-w-S-w-w-S-w-w-S'). In this example, the 10 Syllable AM cycles correspond to the 10 uttered syllables in the sentence. For each Syllable AM peak, the concurrent Stress AM phase value (between -π and π radians) was computed, and is indicated with vertical dotted lines in the bottom-right panel of Figure 1. According to the hierarchical phase-coding scheme described earlier, this instantaneous Stress AM phase value should indicate the relative prominence of the underlying Syllable AM peak. For ease of visualisation, the phase values are transformed into a 'prominence index' (PI) in which phase values near the oscillatory peak (0π radians) are assigned values close to 1 (strong), whereas phase values near the trough (-π/π radians) are assigned values close to 0 (weak). To compute the PI, a normal Gaussian probability distribution function (PDF) was used to transform the cyclical phase values (-π to $_{+}$π) into a

linear index (0 to 1). For each phase value, its corresponding probability was computed under the PDF (mean = 0, standard deviation = 1). This computed value was normalised by the maximum probability value (i.e. the probability value obtained for a phase value of 0), to yield PI values ranging from 0 (weak) to 1 (strong). The top right panel of Figure 1 shows the computed pattern of strong and weak syllables resulting from this process ('S-w-w-S-w-w-S-w-w-S'). It becomes apparent that if the phase of the Stress AM was shifted forward by +1π radian (turning peaks to troughs and vice versa) while holding the Syllable AM constant, the computed rhythm pattern would now be inverted as 'w-w-S-w-w-S-w-w-S-w'. Therefore, if the hypothesis that listeners make use of the hierarchical phase-relationships between AMs to perceive rhythm is correct, then it should be possible to manipulate the perceived rhythm by changing the phase-relationship between the Stress AM and the Syllable AM in any given sentence.

Figure 1 (colour online). Computing strong-weak syllable patterns using amplitude modulations in the speech envelope, illustrated with the trochaic (s-w) nursery rhyme sentence "Pussycat pussycat where have you been?". (Left, (a)) The original waveform of the speech signal is shown at the top, with the amplitude envelope superimposed. The envelope is filtered into 5 modulation bands, forming the AM hierarchy shown at the bottom (see Methods section C1). (Right, (b)) Strong-weak rhythm patterns are computed using the Syllable AM and the Stress AM phase. The plotted AM phase values are projected onto a cosine function for ease of visualisation. The 10 Syllable AM cycles correspond to the 10 spoken syllables. The concurrent Stress AM phase at Syllable AM peaks (indicated with dotted lines) is used to compute the prominence index (PI), shown in the bar graph at the top. Syllables with a high PI (near 1) are considered 'strong (s)' and syllables with a low PI (near 0) are considered 'weak (w)'.



C. Rationale and predictions for current experiment

In this paper, the hypothesis that listeners use the *phase relationship* between 'Stress' rate (~2 Hz) and 'Syllable' rate (~4 Hz) AMs within the modulation spectrum of speech to perceive strong-weak rhythm patterns is tested. As described in Section 1.2, these 'Stress' and 'Syllable' AMs are part of a larger AM hierarchy that is used to represent the linguistic prosodic hierarchy in speech. In the current experiment, nursery rhyme sentences with a clear trochaic ('S-w') or iambic ('w-S') rhythm were used. AMs from these sentences were isolated and used to create single-channel tone-vocoded stimuli (to make the envelope-derived AMs audible), while the fine structure was discarded. To test the relative contribution of different AM rates toward the rhythm percept, different AM conditions were used to create a variety of tone-vocoded sentences (e.g. Stress AM only, Syllable AM only, Stress AM + Syllable AM, etc), and listeners performed a rhythm judgment task using these stimuli. From the prior literature, it was expected that listeners' rhythm judgments would be the most accurate when

they were provided with modulation information at the Stress-rate *as well as* at the Syllable rate (i.e. Stress AM + Syllable AM). Next, to assess whether the modulation statistics that influenced the rhythm percept were AM *phase relationships*, *phase-shifted* variants of each vocoded sentence were created. Since oscillatory phase is a circular variable it was expected that the phase-shifts in the stimuli would result in a *circular* pattern of responding, with participants showing *systematic reversals* in perceived rhythm following incremental phase-shifts of 1π radian. That is, a 1π radian phase-shift (half a cycle) should *reverse* the perceived rhythm pattern of a sentence (i.e.trochaic → iambic) whereas a 2π radian shift (full cycle) should *restore* the original rhythm pattern (i.e.trochaic → trochaic), rather than generating even more rhythm distortion as compared to the 1π radian shift. Conversely, if listeners did *not* make use of phase relationships for their rhythm judgments, their judgments should either be unaffected by the phase-shift, or show a *monotonic* degradation in accuracy. As a control, we also included phase-shifted single AM conditions (i.e. Stress only or Syllable only). Our aim was to assess whether listeners were making use of the phase information at any single AM rate only to make rhythm judgments, rather than relying on the phase *relationship* between Stress and Syllable AMs, as we predicted. Accordingly, we predicted that listeners would show no change in rhythm perception as a result of phase-shifts for single AM conditions.

II. METHODS

A. Participants

Twenty-three adults (7 male; mean age 26.0 yrs, range 22.0 years - 37.5 years) participated in the study. The participants were university students or staff who had volunteered for the study, and they received a modest cash payment for their participation.

All participants had no diagnosed auditory, language or learning difficulties and spoke English as a native first language (as confirmed by self-report). Out of the 23 participants, there were 18 British, three American, one Canadian and one Australian. In our initial analysis, we found that there were no significant differences between the performance of British English-speaking and non-British English-speaking participants, so this factor will not be considered further here. 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 initial analyses, the factor of musical training did not affect performance, and so it will not be considered further here.

B. Speech stimuli

Nursery rhyme sentences were used as these are naturally-occurring and rhythmically-rich speech material that have a relevance for early language learning (see Introduction). Four different nursery rhyme sentences, 8 syllables each, were used, representing 2 different prosodic foot patterns ('trochaic', S-w, and ' iambic', w-S), as listed in Table I. The two rhythm patterns were deliberately chosen so that the phase-shift manipulation should result in a simple reversal of these two familiar rhythm patterns (i.e. trochaic → iambic and iambic → trochaic) rather than giving rise to a new and unfamiliar pattern. The sentences were produced by a female native speaker of British English (mid-twenties in age, Southern accent) who was articulating in time to a 4 Hz (syllable rate) metronome beat. The speaker was instructed to produce the rhythm pattern of each nursery rhyme sentence as clearly as possible. Up to 20 takes were recorded for each sentence, and the most representative was used in the experiment. Utterances were digitally recorded using a TASCAM digital recorder (44.1 kHz, 24-bit) in a soundproof chamber, and the metronome was not audible in the final

recording. All stimuli were normalised to 70dB SPL. The difference in peak vowel intensity between 'Strong' and 'weak' syllables across the 4 sentences was 6.3 dB, on average. This figure is well within the normal range of intensity differences observed between stressed and unstressed syllables across speakers (e.g. Fry, 1955 observed vowel intensity differences of up to 10 dB for word tokens like "OBject" and "CONtract").

Insert Table I here

C. Signal processing

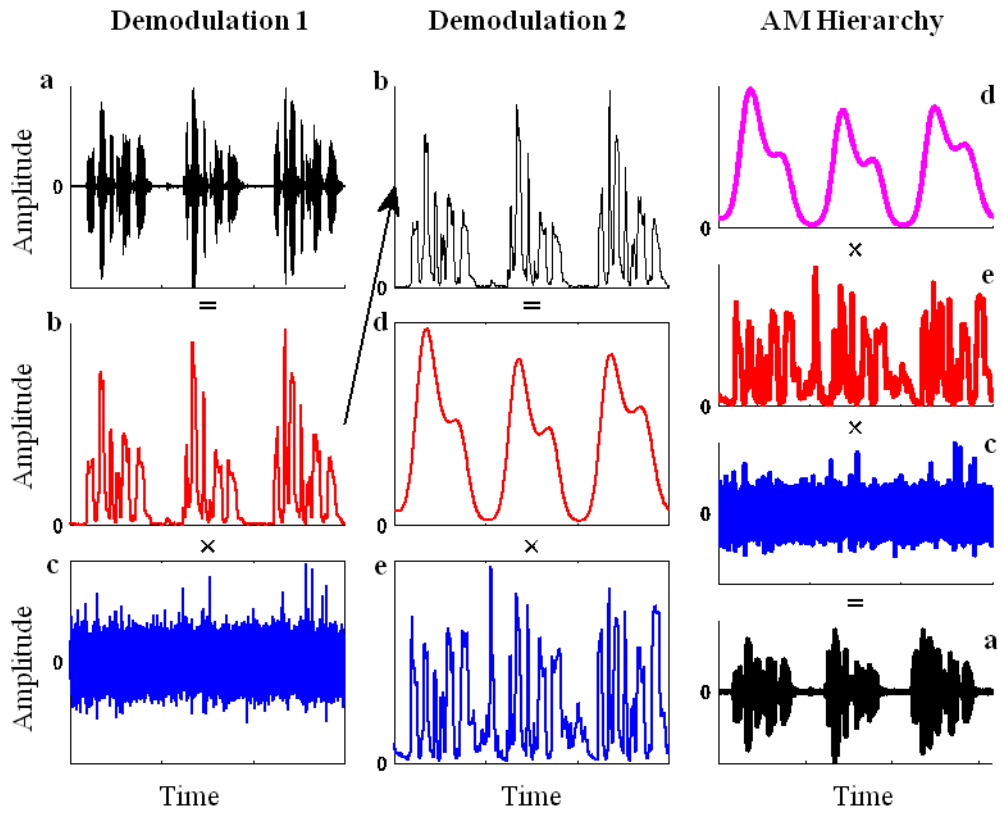
1. The AM hierarchy

A 5-tier AM hierarchy was extracted from the whole-band signal of each nursery rhyme sentence using two different methods. In the first method, the amplitude envelope was extracted from the speech signal of each sentence using the Hilbert transform. This Hilbert envelope was then passed through a modulation filterbank (MFB) comprising 5 adjacent finite-impulse response (FIR) band-pass filters spanning 0.5-50 Hz. Each filter channel had a response of –6 dB at the cross-over edge with its adjacent filter channel, but –55dB at the cross-over with the next-but-one filter channel. This response was the same both at the low-pass and high-pass edges of the channel. The time delay introduced by each filter channel 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. For further technical details of the filterbank design, see Stone & Moore (2003).

The output of the MFB was a set of 5 amplitude modulators (AMs) at 5 different rates, forming the 5-tier AM hierarchy, as shown in the left column of Figure 1. However, artificial modulations may be introduced into the stimuli by the MFB method, since band-

pass filters can introduce modulations near the centre-frequency of the filter, through 'ringing'. Therefore, a second AM-hierarchy extraction method was used as a control. This was Probabilistic Amplitude Demodulation (PAD; Turner & Sahani, 2011; http://learning.eng.cam.ac.uk/Public/Turner/PAD), and does not involve the Hilbert transform or filtering. Rather, the PAD method estimates the signal envelope using a model-based approach in which the signal is assumed to comprise a product of a positive slow envelope and a fast carrier. Bayesian statistical inference is used to invert the model, thereby identifying the envelope which best matches the data and the *a priori* assumptions. In more detail, 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, which may either be entered manually or 'learned' from the data(in the present study, the PAD parameters were entered manually to produce the closest match to the MFB stimuli). 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 white noise, which encodes the fact that it varies more quickly than the envelope. The task of estimating the most appropriate envelope and carrier for the data is then cast in Bayesian terms. Since the problem is ill-posed, the solution takes the form of a probability distribution which describes how probable a particular setting of the envelope and carrier is, given the observed signal (i.e. the 'posterior' distribution p(env,car\data)). PAD summarises the posterior distribution by returning the specific envelope and carrier that have the highest posterior probability and therefore represent the best match to the data. A useful feature of PAD is that this process can be run recursively, using different demodulation parameters each time to recover envelopes with different dominant modulation rates, thereby producing an AM hierarchy (Turner & Sahani, 2007; Turner, 2010), as shown in Figure 2.

Figure 2 (colour online). Example of an AM hierarchy derived by recursive application of PAD. In the first demodulation round (left column), the data, 'a', are demodulated using PAD set to a fast timescale. This yields a relatively quickly-varying envelope ('b') and a carrier ('c'). In the second demodulation round (middle column), the demodulation process is re-applied to the extracted envelope 'b', using a slower timescale than before. This yields a slower daughter envelope ('d') and a faster daughter envelope ('e'). Daughter envelopes 'd' and 'e' form the two tiers of the resulting amplitude modulation hierarchy (right column). Mathematically, these two tiers ('d' & 'e') can be multiplied back with the very first carrier ('c', bottom left) to yield the original signal, 'a'.



All participants heard both MFB-derived and PAD-derived stimuli in the experiment. 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, the observed effects were likely to have arisen from real features in speech rather than filtering artifacts.

The 5 AM tiers in the hierarchy were designated the (1) ‘Slow’ AM tier (0.5-0.8 Hz); (2) ‘Stress’ AM tier (0.8-2.3 Hz), (3) ‘Syllable’ AM tier (2.3-7 Hz), (4) 'Sub-beat' AM tier (7-20 Hz), and (5) ‘Fast’ AM tier (20-50 Hz). As explained in the Introduction, it was intended that each tier in the *AM hierarchy* should capture modulation patterns associated with a different prosodic unit in the *linguistic prosodic hierarchy* (such as syllables or prosodic stress feet). The AM hierarchy was 'syllable-centred' in the sense that the modulation rate of the central 'Syllable' tier was used to determine the bandwidths of the flanking faster and slower tiers. For example, it was intended that the 'Stress' AM tier should capture prosodic stress feet that were either 2 or 3 syllables in length (e.g. trochees or dactyls). Accordingly, the parameters of the 'Stress' filter channel in the MFB were set-up to capture modulations that were between 1/2 to 1/3rd the rate of the modulations passing through the central 'Syllable' filter channel. In our nursery rhyme sentences, the 'Syllable' rate was 4 Hz (following the metronome), so it was intended that the 'Stress' filter channel should comfortably capture modulations occurring at 1.3-2 Hz. In practice, the roll-off of the 'Stress' filter channel was also taken into account, which resulted in the final 'Stress' filter edge frequencies of 0.8 Hz (lower) and 2.3 Hz (upper).

The 'Slow' tier was designed to capture modulations even slower than the 'Stress' rate, for example those corresponding to phrase-level accentual patterns. Therefore, the lower and upper edge frequencies of the 'Slow' filter channel were set at 0.5 Hz and 0.8 Hz respectively (0.5 Hz being the slowest modulation rate that can be resolved in stimuli that are 2s in length, and 0.8 Hz being the bottom edge of the adjacent 'Stress' filter channel).

To define the 'Sub-beat' tier, it was observed that rhythmic speech often contained deliberately-shortened syllables, where the speaker was trying to fit more than one syllable into the space of a single beat (i.e. the syllables were shortened to a 'sub-beat' rate). Typically, between 2 to 3 shortened syllables can be compressed into the space of one normal syllable,

therefore the 'Sub-beat' tier was designed to capture shortened syllables that were between 1/2 to 1/3rd the length of normal syllables. For a regular 'Syllable' rate of 4 Hz, these shortened syllables would be associated with a modulation rate of 8 - 12 Hz. Accordingly, the lower and upper edge frequencies of the 'Sub-beat' filter channel in the MFB were set at 7 Hz and 20 Hz respectively.

Finally, the 'Fast' tier captured modulations above the 'Sub-beat' rate, up to the upper cut-off of 50 Hz, and was expected to contain more quickly-varying acoustic cues to phoneme manner, voicing or vowel identity (Rosen, 1992).

2. Tone vocoding

The extracted AM hierarchy was used to modulate a 500 Hz sine-tone carrier in a single-channel vocoder. Tone vocoding was used rather than noise vocoding because the noise band itself in noise-vocoding introduces inherent fluctuations at multiple rates that can interfere with speech intelligibility (Whitmal et al., 2007). A multi-channel vocoder was not used because it was intended that the sentences should be linguistically completely unintelligible. As the dependent variable in the experiment was how well participants could identify each sentence from a given AM rhythm pattern alone, all other cues to sentence identity should be removed. Therefore, the phonetic fine structure of the signal was intentionally discarded, and the AMs derived from the amplitude envelope were used to modulate the sine-tone carrier, rather than being combined back with the fine structure of the signal. If the original fine structure had been used, additional low-rate modulation information could be introduced during the recombination process as a result of FM-to-AM conversion (Ghitza, 2001, also called 'envelope recovery', Gilbert & Lorenzi, 2006), thereby 'contaminating' the vocoded stimuli. To create single-AM tier 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 3-ms-ramped pedestal at channel RMS power was added prior to combining with the carrier. To create double-AM tier stimuli (e.g. Stress+Syllable), the two AM tiers were first combined via addition (MFB) or multiplication (PAD) before combining with the carrier. The resulting tone-vocoded sentences had clear temporal patterns ranging from "Morse-code" to flutter, but were otherwise completely unintelligible.

3. Phase-shifting

The aim of phase-shifting was to change the phase-relationship between AM tiers to measure whether this changed the rhythm pattern perceived by the listener in a circular fashion (i.e. incremental phase-shifts of 1π radian elicit systematic reversals in perceived rhythm). Since the sentences were either trochaic or iambic in pattern, the aim of phase-shifting was to make trochaic sentences sound iambic, and vice versa. In a 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 a significant change in participants' judgements for these manipulations was not expected.

Due to the use of the metronome, the nursery rhyme sentences were perfectly regular in rhythm structure, therefore their AMs were also highly regular in their temporal pattern of peaks (P) and troughs (t), resembling a pure sinusoid. This enabled phase-shifting to be implemented by cutting and pasting 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', moving 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. Cutting and pasting was chosen as the preferred method because (unlike deletion or silence insertion) this method permitted the retention of all the original information within each sentence, changing only its temporal order. In addition, by using this method, the length of phase-shifted stimuli could be kept the same as the length of the non-phase-shifted stimuli. For the 2π radians shift, the sample length moved was a full period cycle corresponding to a representative single frequency within the bandwidth of the AM tier in question.

Insert Table II here

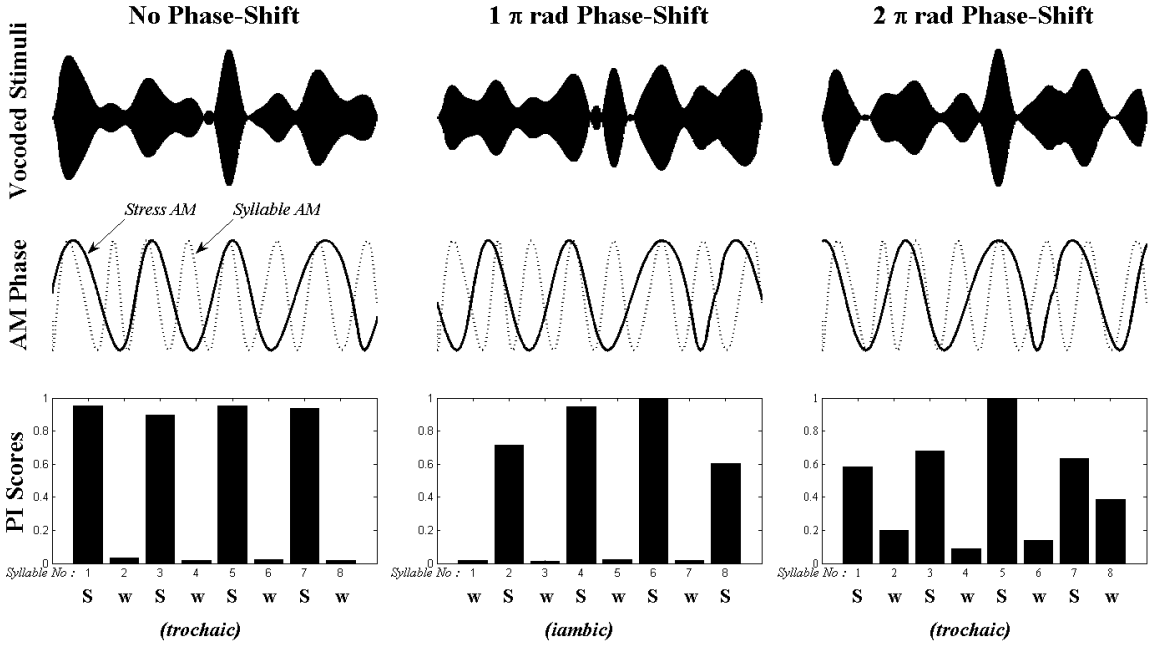
As summarised in Table II, the representative Syllable frequency was determined by finding the component containing the largest RMS power in the 3-7 Hz range of the modulation spectrum for each sample. This turned out to be 4.04 Hz, which was very close to the metronome pacing beat of 4.0 Hz. The principal Sub-beat frequency was determined by taking the mean of 2 and 3 times the Syllable frequency (10.1 Hz), to allow for both double and triple patterns in this tier. For consistency, the representative Stress frequency was also determined by taking the mean of one-half and one-third 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 moved. For a 1π radians shift, the length moved 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. The results of the cut and paste process were checked to ensure that the desired phase changes were produced for all stimuli. Even though

some minor sound artifacts were introduced (e.g. at phase-shift boundaries), the resulting rhythm patterns emerged as predicted, e.g. trochaic (no shift) → iambic (1π shift) → trochaic (2π shift). This is illustrated in Figure 3.

Next, we computed PI scores for the Stress+Syllable stimuli to assess whether the strong-weak rhythm pattern of non-shifted and 1π-shifted stimuli was significantly *different*, but the pattern of non-shifted and 2π-shifted stimuli was the *same*. First, the PI scores for strong and weak syllables in each nursery rhyme sentence were computed separately for each phase-shift condition. We then entered these syllable PI scores into a mixed design repeated measures ANOVA in which Strength (2 levels 'between' factor: strong, weak), Rhythm (2 levels 'between' factor: trochaic, iambic) and Phase (3 levels 'within' factor: no-shift, 1π shift, 2π shift) were the factors. The results of the ANOVA indicated that, there was a significant interaction between Strength and Phase ($F(2,58) = 57.2$, $p<.001$, but no significant interaction between Phase and Rhythm ($F(2,58) = .43$, $p = .65$). That is, Strong and weak syllables differed significantly in the pattern of their PI scores across the 3 different phase-shift conditions. However, trochaic and iambic sentences did not differ in their pattern of phase-shift effects, suggesting that the phase-shift procedure was equally effective for all sentences. To assess the Strength x Phase interaction in greater detail, we conducted a Tukey HSD post-hoc. Post hoc tests revealed that for *both* Strong and weak syllables, there was a significant difference between the PI scores of non-shifted and 1π-shifted syllables ($p<.001$ for both Strong and weak syllables). Similarly, for both Strong and weak syllables, there was a significant difference between the PI scores of 1π-shifted and 2π-shifted syllables ($p<.001$ for both Strong and weak syllables). However, there was *no* significant difference between the PI scores of non-shifted and 2π-shifted syllables (Strong ; $p = .21$; weak : $p = .55$). Therefore, the ANOVA analysis confirmed that the phase-shift procedure had indeed changed the rhythm pattern of Stress+Syllable stimuli as expected. Moreover, although the

non-shifted and 2π-shifted stimuli were substantially different (see Figure 3), they had the same strong-weak rhythm pattern as computed by the prosodic index (PI). Hence, if listeners judged both stimuli as having the same rhythm pattern, this would be evidence that judgements depend on similarity in the key rhythm statistics (phase relationships) rather than on perceptual similarity or familiarity.

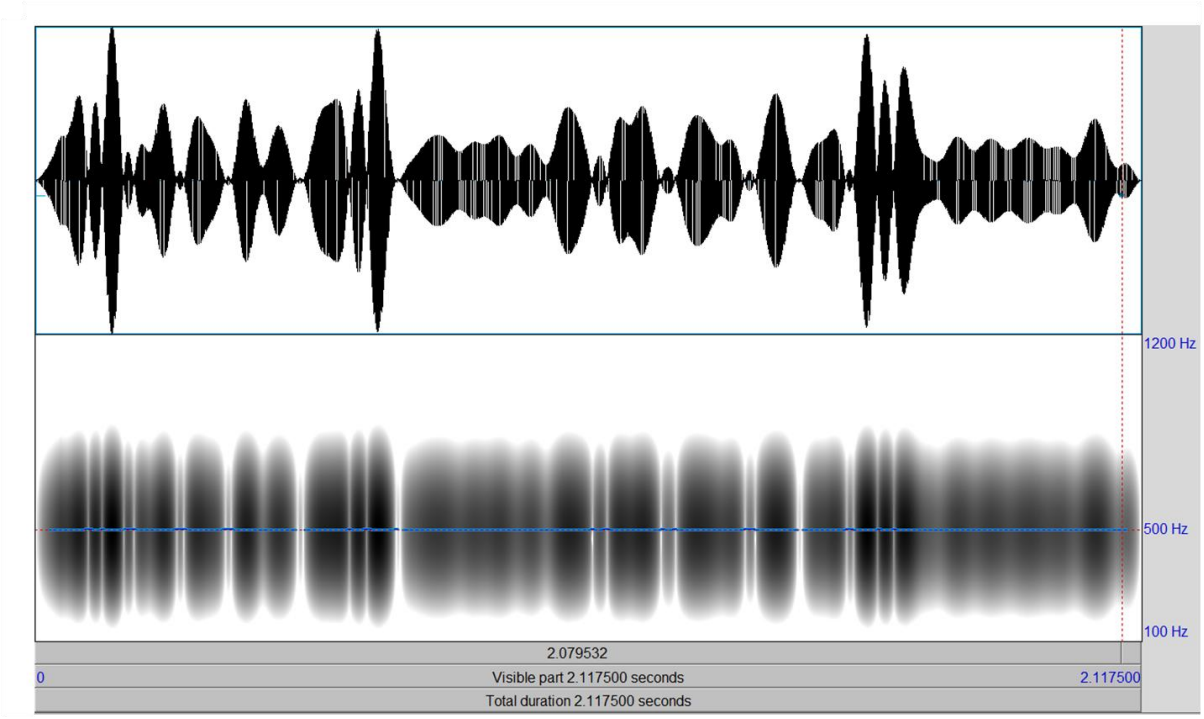
Figure 3. Illustration of the effect of phase-shifting on the rhythm pattern of 'Mary Mary'. (Top row): Tone-vocoded MFB stimuli used in the experiment. (Middle row): Corresponding Stress (bold) and Syllable (dotted) AM phase patterns. Phase values are projected onto a cosine function for visualisation purposes. Only Stress AMs were phase-shifted while Syllable AMs were held constant. (Bottom row) : AM-Based prominence index scores of syllables. Strong syllables ('S') have a prominence value of >0.5, weak syllables ('w') have a prominence value of <0.5.



Finally, to ensure that prosodic pitch cues had been completely eliminated from our stimuli during the vocoding process (thus effectively isolating the contribution of amplitude

cues), we manually checked the spectrogram and fundamental frequency contour of each vocoded stimulus using Praat software. An example of this is shown in Figure 4. Our analysis confirmed that there were indeed no residual pitch variations in our stimuli, all of which had a flat fundamental frequency contour reflecting the 500 Hz sine tone carrier.

Figure 4 (colour online). Example of Praat analysis for the spectral content of the tone-vocoded stimuli. The stimulus shown here is for the non phase-shifted Sub-beat AM vocoded sentence, "simple Simon met a pieman". The top panel shows the sound pressure waveform of the stimulus, the bottom panel shows the corresponding spectrogram (dynamic range = 50 dB) and fundamental frequency contour of the stimulus (a flat line at 500 Hz, see right y-axis).



D. Task

In each trial, participants heard one of four tone-vocoded nursery rhyme sentences (non-phase-shifted and phase-shifted variants were fully randomised). Participants were

asked to indicate which one of the four possible target sentences they thought that they had heard by selecting an appropriate response button (the response button mappings were counter-balanced across participants). Participants were told to respond as accurately as possible, and a time limit was not imposed on responding (i.e. the task was not speeded). Participants' first responses were taken as final (they were not allowed to replay trials), and the programme proceeded onto the next stimulus automatically once a response was received. No feedback was provided to participants regarding the accuracy of their response. 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 sentences as originally spoken, without vocoding. This enabled participants to learn the rhythm pattern of each stimulus, and to become familiar with the response button mapping. After completion of training, participants were asked whether they understood the requirements of the experiment, and whether they were happy to proceed with the actual experiment. If participants indicated any doubt, they were allowed to repeat the training trials. Subsequently, participants performed the task with tone-vocoded stimuli only. The tone-vocoded stimuli retained the temporal pattern of each nursery rhyme sentence, but, as intended, were completely unintelligible. Examples of the experimental stimuli are provided at http://www.cne.psychol.cam.ac.uk/publications-1/. __ 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, as shown in Figure 5. The experimental task was programmed in Presentation and delivered using a Lenovo ThinkPad Edge laptop. Auditory stimuli were presented using Sennheiser HD580 headphones at 70dB SPL, via a UGM96 24-bit USB audio adaptor plugged into the laptop. The experiment was conducted in a soundproof experimental testing room with electromagnetic shielding.

Figure 5 (colour online). 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.

E. Design

The experiment followed an AM condition (5) x Phase Shift (3) x Demodulation Method (2) design. The AM conditions that were used for vocoding were determined after pilot trials. In these trials, the full range of 5 AM hierarchy tiers (Slow, Stress, Syllable, Sub-beat, and Fast) was used. The performance of participants was at chance for the Slow AM tier, and participants 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. Slow and Fast AM tiers were not used. Rather, Stress, Syllable and Sub-beat AM tiers were used in 5 conditions comprising single tier and paired combinations. These conditions were 1) Stress only; 2) Syllable only; 3) Sub-beat only; 4) Stress+Syllable and 5) Syllable+Sub-beat. Each of these AM conditions was presented in three phase shift conditions : 1) No Shift ; 2) 1π radians-shifted and 3) 2π radians-shifted. Fewer phase-shifted stimuli (1π radians or 2π radians) were presented as compared to non-phase-shifted versions (0π radians) because it was intended that participants

should maintain a strong representation of the correct rhythm pattern for each nursery rhyme. Thus, participants heard the normative (0π 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 (0π radians) stimuli were presented within the same experimental block in a randomised fashion. Therefore, each block contained 5 AM conditions x 9 phase variants (5 x 0π radians, 2 x 1π radians, 2 x 2π radians) x 4 nursery rhymes, making 180 trials, all presented in randomised order. Stimuli that were vocoded using MFB-produced AMs and PAD-produced AMs were presented in two separate experimental blocks, giving a total of 360 trials for the entire experiment. Participants were tested in a counterbalanced manner. Half the participants began the experiment with MFB-produced stimuli before switching over to PAD-produced stimuli, while the other half of the participants began the experiment with PAD-produced stimuli before switching over to MFB-produced stimuli.

F. Results analysis

Performance was scored manually in terms of whether participants identified each nursery rhyme correctly (accuracy score). The pattern of *confusion* errors generated by participants is also of interest. If a particular AM condition is providing strong rhythm cues to participants, this should be evidenced by more confusions between rhythmically-similar rather than unrelated nursery rhyme sentences. For example, participants should be more likely to confuse 'Mary Mary' and 'Simple Simon'. By contrast, if a particular AM condition is providing only weak (or no) rhythm cues, then participants should show a random pattern of confusion errors, unbiased by the rhythmic similarity of sentences. Accordingly, participants' confusion errors were captured in 4 x 4 response matrices, which reflected the

distribution of responses given to each stimulus sentence. Moreover, the *degree* of similarity or dissimilarity in the Stress-Syllable AM phase pattern between all four nursery rhymes should be strongly related to the pattern of confusions produced by participants. To test this prediction, we analysed the degree to which similarity in AM modulator *phase* between nursery rhyme sentences predicted participants' confusion patterns. The degree of similarity between the AMs of different sentences was quantified by calculating the Pearson correlation coefficient between the phase series of each possible nursery rhyme sentence pair. This produced a second set of 4 x 4 correlation matrices for each AM combination and for each demodulation method.

III. RESULTS

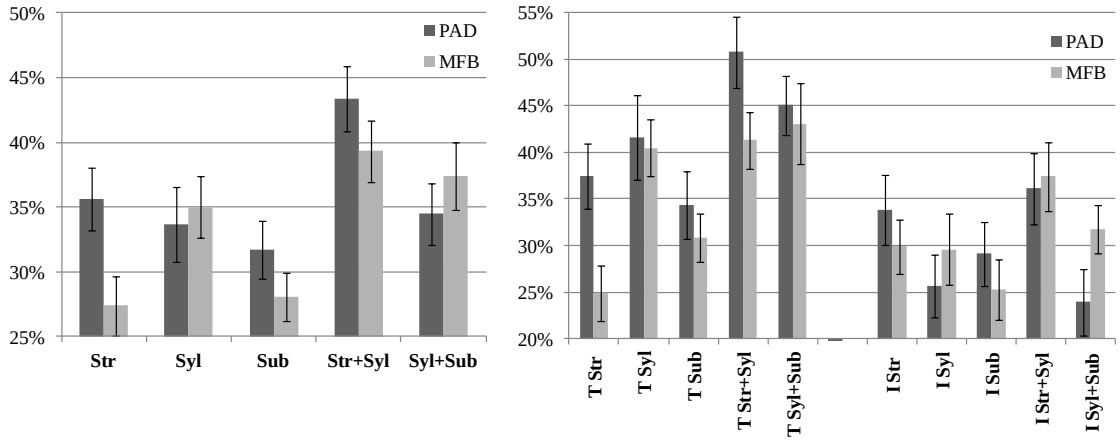
A. Non phase-shifted stimuli

1 Accuracy scores

Participants' accuracy scores over the 4 sentences are shown in Table III. To test which AM condition(s) were the most important for providing rhythm pattern information in perception, listeners' performance across the five AM conditions (non-phase-shifted) was evaluated. Their accuracy scores are shown in Figure 6, broken down by demodulation method (PAD or MFB), and further by sentence rhythm pattern. Participants' accuracy in identifying the correct sentence ranged between 28%-42% (chance = 25%), . The low scores were not surprising given that the sentences were completely unintelligible.

Insert Table III here

Figure 6. Accuracy scores for the five AM combinations, for each Method (PAD and MFB). Error bars indicate standard error. The left subplot shows mean scores over both trochaic and iambic sentences. The right subplot shows a breakdown of scores by sentence rhythm type (T = trochaic, I = iambic).



To compare performance *between* AM combinations, a 2 x 2 x 5 Repeated Measures ANOVA was performed, taking Accuracy as the dependent variable, and Demodulation Method (2), Rhythm Type (2) and AM combination (5) as within-subjects factors. Scores in all conditions were normally distributed ($p > .05$ in Kolmogorov-Smirnov test of normality).There was a significant main effect of AM combination ($F(4,88) = 10.40$, $p < .0001$), and a significant main effect of Rhythm ($F(1,22) = 28.36$, $p < .001$), where participants performed significantly better for trochaic sentences than for iambic sentences. There was also a significant interaction between Rhythm and AM ($F(4,88) = 2.70$, $p < .05$). Tukey HSD post-hoc tests of this interaction revealed that participants performed significantly better for trochaic sentences only in the Syllable AM ($p < .05$) and Syllable+Sub-beat AM ($p < .01$) conditions. Thus, the trochaic listening benefit appeared to apply particularly to AM conditions that included the Syllable AM.

There was no overall difference between demodulation methods ($F(1,22) = 2.87$, $p = .11$) and no interaction between AM condition x Method ($F(4,88) = 1.98$, $p = .10$). However,

there was a significant interaction between Rhythm Type x Method ($F(4,88)=2.7$, $p<.05$). Tukey HSD post-hoc tests of this interaction indicated that the PAD method yielded significant higher sentence identification accuracy for trochaic sentences, but not for iambic sentences ($p<.05$). This confirmed that both PAD and MFB demodulation methods were producing similar patterns of listening performance across the 5 AM conditions in general, but the PAD method provided a specific listening benefit for trochaic-patterned sentences.

The AM condition main effect was analysed further by performing a Tukey HSD post-hoc analysis. Performance with Stress+Syllable AMs was significantly superior to *all* four other AM conditions ($p \leq 0.05$ for all four comparisons). This supported the hypothesis that rhythm pattern identification for the Stress+Syllable AM condition should be reliably better than for any other AM combination tested. Since listeners performed better when hearing Stress+Syllable AMs as compared to Syllable+Sub-beat AMs, the superior performance with Stress+Syllable AMs cannot simply be due to a greater modulation bandwidth being presented to listeners (the bandwidth of the modulation spectrum was actually greater for Syllable+Sub-beat AMs than for Stress+Syllable AMs). It must have been due to the perceptual 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 suggests 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. By contrast, performance with Syllable+Sub-beat AMs was not significantly better than with Syllable AMs alone ($p = 0.92$), indicating that Sub-beat modulations were not providing additional rhythm cues over and above those already present in the Syllable AM. Hence, as predicted, the Stress+Syllable AM condition provided listeners with the most rhythm pattern information.

2 Analysis of confusion errors and acoustic similarity of stimuli

Table IVa shows the 4 x 4 response matrix produced for the non-phase-shifted Stress+Syllable AM condition (responses for PAD and MFB methods were computed separately in the actual analysis but are shown here averaged across the two methods for simplicity of inspection). The response matrix confirms that participants made more confusions within the *same* rhythm pattern than across different rhythm patterns for this Stress+Syllable AM condition. 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 16% of the time each. Table IVb shows the 4 x 4 acoustic similarity matrix produced by cross-correlating the phase series of the Stress+Syllable AM patterns for each possible pair of sentences (for simplicity, the table shows averages across PAD and MFB methods, however in the actual analysis, the data for the two methods were treated as separate entries). As expected, for the Stress+Syllable AM condition, sentences with the same rhythm pattern (e.g. trochaic) showed *positive* correlations in their phase patterns while sentences with different rhythm patterns showed *negative* correlations in their phase patterns.

Insert Table IV here

If participants' confusion errors were driven by the similarity between the AMs that they were hearing, then the acoustic AM similarity matrices should be good predictors for participants' response matrices. To test this, we conducted a Regression analysis with the response matrix as the dependent variable and AM phase similarity as the independent variable, using each cell in the matrix as an observation. This regression analysis was

conducted for all five AM conditions. Table V shows the summary statistics for each regression model, with the adjusted R^2 value indicating the percentage of confusion variance explained.

Insert Table V here

As can be seen, the phase similarity between the sentence AMs were significant predictors of participants' responses for all AM conditions. However, the Stress+Syllable AM condition was the best predictor of response pattern, explaining 57.4% of variance in responding. Furthermore, the amount of variance in responding explained by phase similarity in the Stress+Syllable condition (57.4%) was 12.3% greater than the sum of the variance explained by the Stress phase similarity *plus* the Syllable phase similarity considered independently (18.7% + 26.4% = 45.1%), indicating a super-additive or synergistic effect. In contrast, the Sub-beat AM similarity patterns were only weakly related to participants' responses, only explaining around 9% of variance in responding. Furthermore, when combined with Syllable AMs, the Syllable+Sub-beat Phase condition (32.2%) actually explained *less* variance than the sum of Syllable phase and Sub-beat phase tiers separately (26.4% + 9.4% = 35.8%). This appears to indicate that the temporal information at these two AM rates may compete in the perceptual processing of metrical rhythm.

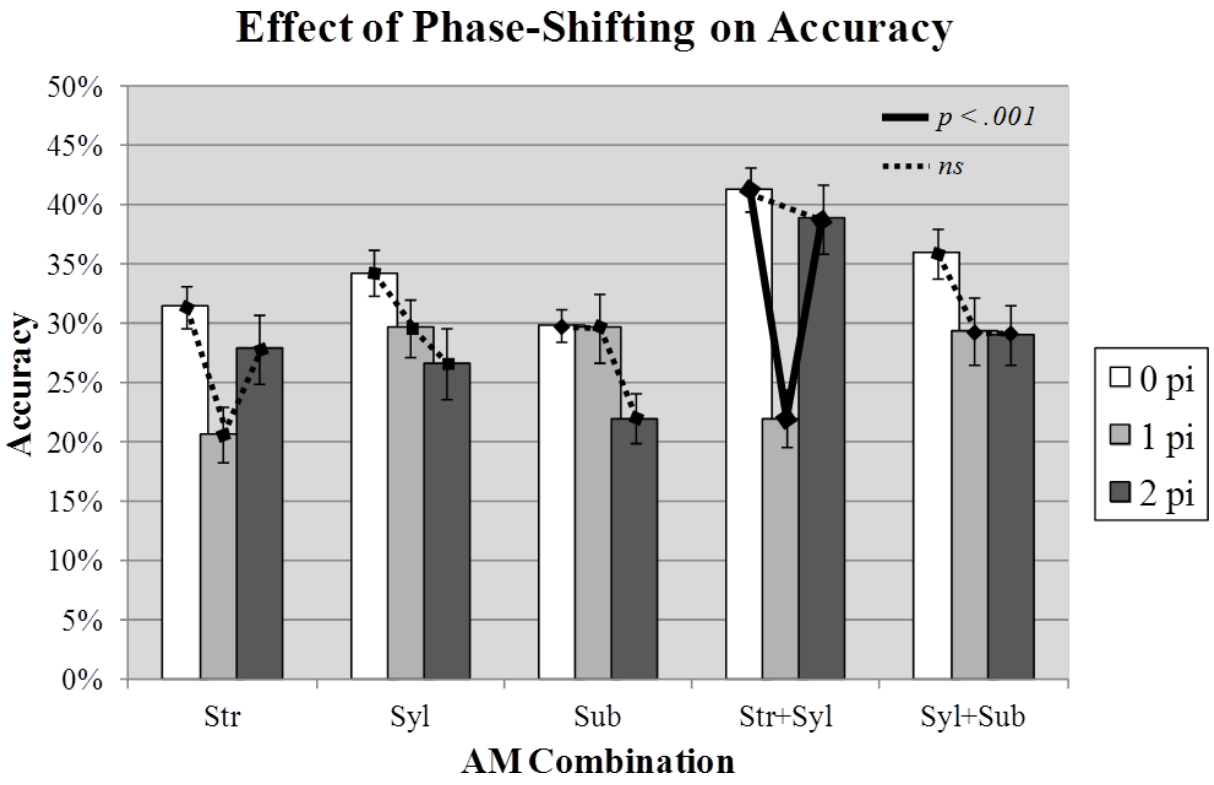
B. Phase-shifted stimuli

1 Accuracy scores

Recall that as an empirical test of AM phase-based rhythm perception, participants were asked to recognise the tone-vocoded nursery rhymes following phase-shifts of 1π radians or 2π radians (described in Section IC).A *circular* pattern of responding had been

predicted specifically for rhythm-bearing AM conditions, and either no change or a monotonic decline in response accuracy for non-rhythm-bearing AM conditions.

Figure 7. Effect of phase-shifting. Accuracy scores averaged across PAD and MFB methods. Error bars indicate standard error.



The effect of phase-shifting on performance accuracy in each AM condition is shown in Figure 7. Visual inspection suggests that the predicted drop in accuracy for a 1π shift and recovery for a 2π shift indeed occurs in the Stress+Syllable AM condition, and possibly also in the Stress-AM condition. A 5 (AM condition) x 3 (Phase shift, 0π , 1π, 2π radians) x 2 (Demodulation Method: PAD, MFB) repeated measures ANOVA was therefore carried out, taking Accuracy as the dependent variable. An interaction between AM condition and phase shift would indicate that a phase-shift effect occurred in some AM conditions, but not in others. The ANOVA showed a significant main effect of AM condition ($F(4, 88)=5.98$, $p < .0001$), 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 ($F(1,22) = 3.71$, $p = .067$). There was also no significant interaction between AM extraction method and phase shift ($F(2, 44) = .16$, $p = .85$), and no significant interaction between AM extraction method and AM condition ($F(4, 88) = .35$, $p = .84$).

The predicted interaction between AM condition and phase shift was significant ($F(4.93,108.35) = 4.78$, $p < .0001$, Greenhouse-Geisser epsilon = 0.62). Therefore, a Tukey-HSD post hoc analysis was used to compare differences between 0π and 1π shifts, and 1π and 2π shifts respectively for each AM condition. Post-hoc testing showed that significant phase-shift effects were limited to the Stress+Syllable AM condition. No other AM condition showed significant changes in accuracy as a result of the phase-shifts. For the Stress+Syllable AM condition, phase-shift effects occurred in the predicted direction. There was a significant drop in accuracy for a 1π-shift coupled with a significant recovery of accuracy for a 2π shift. Additionally, there was no significant difference in performance between 0 and 2π-shifted Stress+Syllable AM stimuli. This shows that the rhythm information in 2π radians phase-shifted stimuli is statistically equivalent to that in non-phase-shifted sentences, which is remarkable given that 2π-shifted stimuli suffered more general acoustic distortion than 1π-shifted stimuli. This circular (rather than monotonic) pattern of response for the Stress+Syllable AM condition is consistent with the hypothesis that rhythm pattern perception is based on the phase-relationship between Stress and Syllable AM rates. Moreover, this circular response pattern is only seen for the combination of Stress and Syllable rates of AM. This is the same AM combination that gave rise to the best rhythm perception performance for non-phase-shifted stimuli.

2 Multi-dimensional scaling (MDS) of confusion errors

Here, listeners' drop in performance for 1π-shifted Stress+Syllable stimuli is interpreted as indicating a systematic change in perceived rhythm pattern (i.e. trochaic → iambic). However, an alternative explanation is that listeners were simply unable to identify the rhythm patterns of these phase-shifted stimuli, and their drop in performance indicated this random uncertainty. Although participants' recovery in performance for 2π-shifted stimuli supports the first (systematic perceptual shift) explanation, it is possible to address the issue more directly by analysing whether participants' perceptual representations of the 4 sentences changed systematically or randomly as a result of the phase-shifts. Consequently, in a final analysis, participant's response patterns in the 0π, 1π and 2π radians phase shift conditions were used as the basis for multi-dimensional scaling (MDS). Recall that previously, 4 x 4 'confusion' matrices had been computed from participants' response data. These 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). MDS is a method that transforms *psychological* proximity (similarity or confusability) into *spatial* proximity (distance), producing 'perceptual maps' that can be used to infer whether participants' rhythm perception had changed systematically or randomly as a result of the phase shifts. In MDS maps, items that are more similar are mapped closer together, whilst items that are more dissimilar are mapped further apart. For this analysis, 2-dimensional representations of MDS solutions were used because 1-dimensional representations provided a poor fit for some matrices (goodness-of-fit 'stress' values >0.1), whilst 2-dimensional representations provided a good fit for all matrices (goodness-of-fit 'stress' values <0.001).

Figure 8 (colour online). MDS solutions for participants' response patterns across 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 (trochaic = solid line; iambic = dashed line). 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.

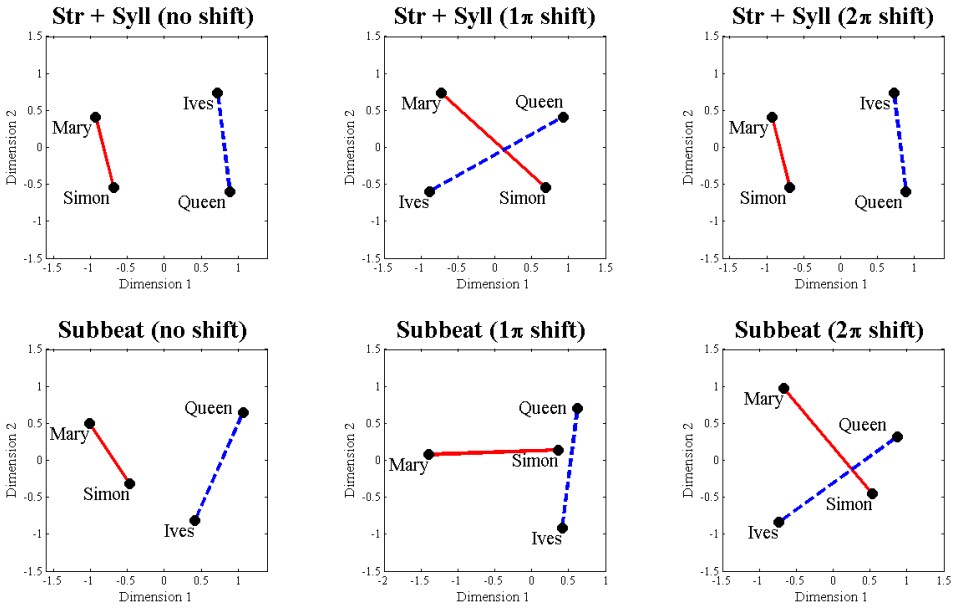


Figure 8 shows the MDS 'perceptual maps' that were obtained for the Stress+Syllable AM condition and the Sub-beat AM condition across the three possible phase-shifts (0, 1π radians, 2π radians). These two AM conditions were selected for display because the data so far indicate that systematic phase-shift effects were present for the Stress+Syllable AM condition but not for the Sub-beat AM condition. The MDS solution confirms that for the Stress+Syllable AM condition in the no-phase-shift condition, participants mapped the two trochaic nursery rhymes ‘Mary Mary’ and ‘Simple Simon’ closer to each other than to the other two iambic nursery rhymes (‘St Ives’ and ‘Queen of Hearts’), which were in turn mapped more closely to each other. Following a 1π radians shift, if there was a systematic change in rhythm perception, each nursery rhyme should move systematically closer to the nursery rhymes in the *opposite* rhythm group, so that the distance between 'Mary' and

'Queen', for example, is now shorter than the distance between 'Mary' and 'Simon'. This indeed occurred, as indicated by the 'crossed' MDS arrangement in Figure 8 (top middle subplot), where rhythmically-similar sentences were now placed on opposite vertices (i.e. further apart) rather than on adjacent vertices (i.e. nearer together) on the map. This systematic MDS pattern strongly indicates that participants did indeed perceive the *opposite* rhythm pattern when stimuli in the Stress+Syllable AM condition were phase-shifted by 1π radians, and were not simply responding randomly. Finally, with a 2π radians shift, the original (non-shifted) similarity map should be restored, as indeed was observed (top right subplot). By contrast, the MDS solutions for the Sub-beat AM condition (bottom panels) did not follow the predicted phase-shift effect or reflect a particular rhythm grouping. Hence participants were *not* using rhythm patterns to group nursery rhymes for the Sub-beat AM stimuli, indeed, they appear to be grouping the nursery rhymes at random when listening to Sub-beat AMs only.

IV. DISCUSSION & CONCLUSION

The goal of the experiment and modelling described here was to better understand how amplitude modulation rates (i.e. Stress & Syllable) and modulation statistics (i.e. phase relationships) from the speech envelope contribute to listeners' perception of speech rhythm. The results of the tone-vocoder rhythm perception experiment were clear. First, the non-phase-shifted data indicated that the Stress+Syllable AM condition transmitted the most rhythm pattern information, since participants were statistically the most accurate at making rhythm discriminations when presented with this AM combination. Second, the phase-shifted data indicated that rhythm discrimination for the Stress+Syllable AM condition was parametrically dependent on the *phase relationship* between the two AM rates. When the

Stress AM was phase-shifted with respect to the Syllable AM by 1π radians and then 2π radians, listeners' performance showed the predicted circular pattern of responding in which rhythm perception was first reversed and then restored (as confirmed by the MDS analysis). No other AM condition showed this pattern of response. Accordingly, it is concluded that the rhythm information contained within the envelope modulation spectrum is primarily located at Stress (~2 Hz) and Syllable (~4 Hz) rates, and that the perception of (English) speech rhythm depends in part on the *phase relationship* between these two key rates of amplitude modulation. It should be noted, however, that while Stress and Syllable AMs alone can provide *sufficient* information to convey a rhythm percept (as demonstrated here), these AMs may not be *necessary* for rhythm perception of natural speech, which contains other complementary acoustic cues to rhythm.

A possible alternative explanation for the observed pattern of responding is that rhythm perception depends solely on the phase pattern at a *single* AM rate (i.e. Stress only or Syllable only), and does not require the combination of modulation information across more than one AM rate. For example, in the Stress+Syllable AM phase-shifted condition, listeners could have based their rhythm judgments solely on the Stress-rate phase pattern, without integrating the concurrently-presented Syllable-rate information. However, if this were the case, then phase-shifting a *single* AM rate (i.e. Stress only or Syllable only) should have been sufficient to induce a significant change in listeners' rhythm perception, which was not observed in our data. Thus, our data support the view that speech rhythm information is not transmitted by the absolute phase information at any given rate alone. Rather, speech rhythm information is transmitted by the phase *relationship* between Stress and Syllable AM rates in the speech envelope.

In real life, adult listeners may track these amplitude modulation patterns when listening to continuous speech in order to aid speech segmentation (Kim et al, 2008; Ghitza,

2013). Specifically, the pattern of amplitude modulation at the Syllable rate could help listeners to infer the approximate location of individual syllables (e.g. by tracking peaks in the modulation pattern), while the concurrent *phase* of modulation at the Stress rate could allow listeners to infer the prosodic strong-weak status of each syllable. A possible neural basis for this AM-tracking mechanism could be neuronal oscillatory entrainment, with phase alignment to these different temporal rates (Giraud & Poeppel, 2012; Ghitza, 2011; Ghitza, 2-13). Consistent with the modelling used here, speech encoding via phase alignment ('re-setting') has already been demonstrated at the syllable rate in adult listeners (Luo & Poeppel, 2007). Further, speech envelope tracking is reduced if syllable-rate fluctuations in the speech signal are removed (Doelling et al, 2014). Doelling and colleagues argued that acoustic landmarks (such as 'auditory edges', or large amplitude rise times) are important for driving the neural entrainment of slow oscillations (< 10 Hz), and pinpointed delta-theta oscillations as of particular importance. This neural work converges with the current data, which has demonstrated the importance of the acoustic Stress-Syllable phase relationship (which drives the neural delta-theta phase relationship) for rhythm pattern perception.

It should be noted that the AM-based approach adopted in this study is complementary to attempts to characterise speech rhythm in terms of *durational* variation (e.g. Ramus et al, 1999; Dellwo & Wagner, 2003; O'Dell & Nieminen, 1999; Barbosa, 2002), and to models in which multiple acoustic cues (duration, F0, intensity) are combined to compute rhythmic prominence (Todd 1994, Todd & Brown 1996; Lee & Todd, 2004). Prior approaches to the study of speech rhythm have focussed particularly on the global durational statistics of the speech signal (e.g., via 'rhythm-metrics': Ramus et al, 1999; Dellwo & Wagner, 2003). Accordingly, previous computational approaches (such as coupled oscillator models of speech rhythm, e.g. O'Dell & Nieminen, 1999; Barbosa, 2002) have focussed on modelling the durational patterning observed in different languages in mechanistic terms. The

AM-based approach adopted in this study offers a complementary perspective. Rather than contradicting these previous attempts to characterise speech rhythm in terms of durational variation, the main contribution of the current work is the identification of a specific AM statistic (the Stress-Syllable phase relationship) which contributes in a parametric way toward English listeners' perception of speech rhythm. The claim is not that AM-based measures of rhythm are superior to duration-based measures. Rather, given that amplitude and duration cues tend to co-vary in speech, a multi-cue approach to speech rhythm analysis (combining duration and amplitude measures) may well achieve the best results (e.g. Silipo & Greenberg, 1997; Kochanski et al, 2005). More generally, the analysis of AM phase relationships could offer useful insights into the acoustic underpinnings of rhythm in different languages.

Stress is not universal across languages (Cutler, 2005). Therefore, the acoustic underpinnings of rhythm may differ between 'stress-timed' languages like English, 'syllable-timed' languages like French (whose speakers appear insensitive to stress; Dupoux et al, 2008), and 'mora-timed' languages like Japanese. Even neonates can detect the rhythmic differences between 'stress-timed' languages like English and 'syllable-timed' languages like Spanish (e.g. Nazzi et al, 1998; Ramus et al, 2000), yet the acoustic basis for these different rhythm typologies remains elusive (e.g. Dauer, 1983; Roach, 1982; Arvaniti, 2009). The AM-based approach adopted in this study could offer fresh insights into the acoustic cues utilised by infants. For example, there may be consistent differences in the way that different languages recruit AMs at different rates to specify rhythm. It is possible that 'stress-timed' languages such as English may rely strongly on Stress-rate amplitude modulations to specify strong-weak rhythm patterns (as shown in this paper), whereas 'syllable-timed' languages may not recruit Stress-rate amplitude modulations to the same extent. Rather, syllable-timed languages may specify speech rhythm patterns more through durational patterning of the Syllable-rate AM. Note that in this case, the rhythm differences between stress-timed and

syllable-timed typologies would arise from the *relative* contribution of Stress- and Syllable-rate amplitude modulation toward the rhythm percept. A difference in the relative dominance of different AM rates in speech may also produce systematic cross-language differences in the entrainment of neuronal oscillatory networks in the cortex (vis-a-vis Giraud & Poeppel, 2012).

Finally, there are some limitations to the current approach which should be noted. In the speech stimuli used for this study, the speaker was articulating the sentences in time to a metronome beat. This introduced durational isochrony into the speech samples, which is not a characteristic of natural speech. This regularisation was necessarily to produce the phase-shifted stimuli used in the rhythm perception experiment. However, it remains to be shown that the AM-based approach and modulation statistics used in this study are therefore applicable to spontaneously-produced un-timed speech (see Leong, 2012, for application of a similar AM analysis method to a larger corpus of naturally-produced [not metronome-timed] speech). Moreover, the phase-manipulation procedure used in the current study was conservative: stimuli were subjected to a modest temporal shift by having a small section of the signal moved from the beginning to the end. In the double-AM conditions (e.g. Stress+Syllable), this procedure produced a completely different pattern of modulation for each phase-shift (see Figure 3) because only 1 AM was phase-shifted before being combined with a second AM. However, in the single-AM conditions (e.g. Stress only, Syllable only), large portions of the phase-shifted stimuli remained intact. This could have led participants to ignore the small temporal shifts at the beginning and end of the sentence, and to rely on a whole-sentence recognition strategy instead. Such a strategy would result in an *underestimation* of the true phase-shift effect in these single-AM conditions. Accordingly, it remains to be demonstrated whether more aggressive phase-shifting procedures - such as computationally replacing the phase of the analytic signal across all time points - would

replicate the results obtained in this study. Future AM-based studies should aim to supplement the current proposal regarding how listeners make use of key modulation statistics in normal speech for prosodic rhythm perception.

ACKNOWLEDGEMENTS

This research was funded by a Harold Hyam Wingate Research Scholarship to VL and by the Medical Research Council, G0902375.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. Phonetica, 66, 46-63.

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.

Bolinger, D. (1958). A theory of the pitch accent in English, Word: Journal of the International Linguisic Association 7, pp. 199–210, reprinted in D. Bolinger, Forms of English: accent, morpheme, order, Harvard University Press, Cambridge, MA.

Bryant, P.E., Bradley, L., Maclean, M., & Crossland, J. (1989). Nursery rhymes, phonological skills and reading. Journal of Child Language, 16, 407-428.

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. (1983). Stress-timing and syllable timing revisited. Journal of Phonetics, 11, 51-62.

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. Proceedings of the International Congress of Phonetics Science. (pp.471-474). Barcelona.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage, 85, 761–768.

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.

Dupoux, E., Sebastian-Galles, N., Navarrete, E. & Peperkamp, S. (2008). Persistent stress "deafness": The case of French learners of Spanish. Cognition, 106, 682-706.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. Journal of the Acoustical Society of America, 26, 138.

Fry, D. B. (1958). Experiments in the perception of stress, Language and Speech, 1, 126–152.

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. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. Journal of the Acoustical Society of America, 110, 1628-1640.

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. (2013). The theta-syllable: a unit of speech information defined by cortical function. Frontiers in Psychology. 4:138. doi: 10.3389/fpsyg.2013.00138

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.

Gilbert, G., & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. Journal of the Acoustical Society of America, 119, 2438-2444.

Giraud, A.L. & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience, 15, 511-517.

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. Trends in Cognitive Sciences, 15, 1 3-10.

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.

Gueron, J. (1974). The meter of nursery rhymes: An application of the Halle-Keyser theory of meter. Poetics, 12, 73-111.

Hayes, B. (1995). Metrical stress theory: principles and case studies. Chicago: University of Chicago Press. 458 pages.

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 (pp. 429-433). Dordrecht, Holland : 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.

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.

Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. Language and Speech, 51(4), 343-359.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.

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. (2012). Prosodic rhythm in the speech amplitude envelope : Amplitude modulation phase hierarchies (AMPHs) and AMPH models. Doctoral dissertation, University of Cambridge, 2012. 359 pages. Available online at : http://www.cne.psychol.cam.ac.uk/pdfs/phds/vleong (date last viewed 24/10/13)

Lerdahl, F.,& Jackendoff, R.(1983). A generative theory of tonal music. Cambridge, MA:MIT Press. 368 pages.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. Linguistic Inquiry, 8, 249-336.

Luo, H., & Poeppel, D. (2007) Phase patterns of neuronal 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.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech, Journal of the Acoustical Society of America. 22, 167–173.

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.

Obleser, J., Herrmann, B., Henry, M. J. (2012). Neural oscillations in speech: don't be enslaved by the envelope. *Frontiers in Human. Neuroscience*, 6:250.10.3389/fnhum.2012.00250

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.

Plomp, R. (1983). Perception of speech as a modulated signal. Proceedings of the 10th International Congress of Phonetic Sciences, Utrecht, 29-40.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication, 41, 245-255.

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, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265-292.

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.

Schane, S.A. (1979). The rhythmic nature of English word accentuation. Language, 55, 559–602.

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. 476 pages.

Selkirk, E.O. (1986). On derived domains in sentence phonology. Phonology Yearbook, 3, 371–405.

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, pp. 67–113. Measuring perceptual distance from a confusion matrix.

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.

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 the Acoustical Society of America, 114, 1023-1034.

Tierney, A., & Kraus, N. (2013). The ability to tap to a beat relates to cognitive, linguistic, and perceptual skills. Brain and Language. 124, 225–231.

Tilsen, S. & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. Journal of the Acoustical Society of America, 124, EL34-39.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. Journal of the Acoustical Society of America, 134, 628-639.

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.

Trevarthen, C. (1986). Development of intersubjective motor control in infants. In M.G. Wade & H.T.A. Whiting (eds). Motor development in children : Aspects of coordination and control (pp. 209-261). Dordrecht : Martinus Nijhoff.

Trevarthen, C. (1987). Sharing makes sense. In R.Steele & T.Treadgold (eds), Language topics - essays in honour of Michael Halliday. Vol 1 (pp. 177-199). Amsterdam : John Benjamin Publishing Company.

Turner, R.E., & Sahani, M. (2007). Probabilistic amplitude demodulation. Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation, pp. 544–551.

Turner, R.E. (2010). Statistical models for natural sounds. Doctoral dissertation, University College London. Available online at : http://www.gatsby.ucl.ac.uk/~turner/Publications/turner-2010.html (date last viewed 24/10/13)

Turner, R.E. & Sahani, M. (2011). Demodulation as Probabilistic Inference. IEEE Transactions on Audio, Speech, and Language Processing, 19, 2398-2411.

Whalley, K, & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. Journal of Research in Reading, 29, 288-303.

Whitmal, N.A., III, Poissant, S.F., Freyman, R.L., Helfer, K.S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. Journal of the Acoustical Society of America, 122, 2376-2388.

Wood, C., & Terrell, C. (1998) Poor reader's ability to detect speech rhythm and perceive rapid speech. British Journal of Developmental Psychology, 16, 397-413.

Table I. List of nursery rhyme sentences and their rhythm pattern

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

Table II. Summary of phase-shifting parameters for each AM tier

AM Tier	**Representative Frequency (Hz)**	**Derivation**	**Full Cycle Length for 2π Shift (ms)**	**Half Cycle Length for 1π Shift (ms)**
Stress	1.68	Mean of 1/2 and 1/3 of Syllable frequency	595	297.5
Syllable	4.04	Highest RMS power within 3-7 Hz of modulation spectrum	248	124
Sub-beat	10.1	Mean of 2 and 3 times Syllable frequency	99	49.5

Table III. Accuracy scores for AM combinations and phase shift conditions. Means shown are averages across PAD and MFB methods. 'SE' refers to standard error.

		Accuracy Scores (%)		
	Phase Shift	*0 rad*	*1π rad*	*2π rad*
AM combinations	Stress only (SE)	31.4 (1.8)	20.7 (2.3)	27.9 (2.9)
	Syllable only (SE)	34.3 (1.9)	29.6 (2.4)	26.6 (2.9)
	Sub-beat only (SE)	29.9 (1.4)	29.6 (2.9)	22.0 (2.0)
	Stress + Syllable (SE)	41.4 (1.8)	22.0 (2.5)	38.9 (2.9)
	Syllable + Sub-beat (SE)	35.9 (2.1)	29.3 (2.8)	29.1 (2.5)

Table IV. Examples of participant response (confusion) matrix for the Stress + Syllable AM band (no phase-shift), and the corresponding AM phase similarity matrix. Values shown are averages across PAD and MFB methods.

(a) Participant response (confusion) matrix. Grand averages for 23 participants are used.

(Values shown in the table are response percentages)			**RESPONSE**			
			Trochaic		**Iambic**	
			Mary Mary	Simple Simon	St Ives	Queen of Hearts
STIMULUS (Sentence presented)	**Trochaic**	Mary Mary	**62.7%**	<u>15.7%</u>	12.7%	8.7%
		Simple Simon	<u>42.7%</u>	**29.1%**	13.1%	14.9%
	Iambic	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

(b) Phase similarity matrix

(Values shown in the table are cross-correlation co-efficients)			**STIMULUS (AM Phase)**			
			Trochaic		**Iambic**	
			Mary Mary	Simple Simon	St Ives	Queen of Hearts
STIMULUS (AM Phase)	**Trochaic**	Mary Mary	1.00	0.42	-0.53	-0.33
		Simple Simon	0.42	1.00	-0.19	-0.03
	Iambic	St Ives	-0.53	-0.19	1.00	0.47
		Queen of Hearts	-0.33	-0.03	0.47	1.00

Table V. Regression statistics for AM Phase similarity matrices predicting the Response Matrix in each AM condition

	Stress Only	**Syllable Only**	**Sub-beat Only**	**Stress + Syllable**	**Syllable + Sub-beat**
F (1,30)	8.13	12.16	4.23	42.76	15.71
p	<.01	<.01	<.05	<.000001	<.001
Adjusted R^2	0.187	0.264	0.094	0.574	0.322

FIGURE CAPTIONS

Figure 1 (colour online). Computing strong-weak syllable patterns using amplitude modulations in the speech envelope, illustrated with the trochaic (s-w) nursery rhyme sentence "Pussycat pussycat where have you been?". (Left, (a)) The original waveform of the speech signal is shown at the top, with the amplitude envelope superimposed. The envelope is filtered into 5 modulation bands, forming the AM hierarchy shown at the bottom (see Methods section C1). (Right, (b)) Strong-weak rhythm patterns are computed using the Syllable AM and the Stress AM phase. The plotted AM phase values are projected onto a cosine function for ease of visualisation. The 10 Syllable AM cycles correspond to the 10 spoken syllables. The concurrent Stress AM phase at Syllable AM peaks (indicated with dotted lines) is used to compute the prominence index (PI), shown in the bar graph at the top. Syllables with a high PI (near 1) are considered 'strong (s)' and syllables with a low PI (near 0) are considered 'weak (w)'.

Figure 2 (colour online). Example of an AM hierarchy derived by recursive application of PAD. In the first demodulation round (left column), the data, 'a', are demodulated using PAD set to a fast timescale. This yields a relatively quickly-varying envelope ('b') and a carrier ('c'). In the second demodulation round (middle column), the demodulation process is re-applied to the extracted envelope 'b', using a slower timescale than before. This yields a slower daughter envelope ('d') and a faster daughter envelope ('e'). Daughter envelopes 'd' and 'e' form the two tiers of the resulting amplitude modulation hierarchy (right column). Mathematically, these two tiers ('d' & 'e') can be multiplied back with the very first carrier ('c', bottom left) to yield the original signal, 'a'.

Figure 3. Illustration of the effect of phase-shifting on the rhythm 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) : AM-Based prominence index scores of syllables. Strong syllables ('S') have a prominence value of >0.5, weak syllables ('w') have a prominence value of <0.5.

Figure 4 (colour online). Example of Praat analysis for the spectral content of the tone-vocoded stimuli. The stimulus shown here is for the non phase-shifted Stress+Syllable AM vocoded sentence, "Mary Mary quite contrary". The top panel shows the sound pressure waveform of the stimulus, the bottom panel shows the corresponding spectrogram and fundamental frequency contour of the stimulus (a flat line at 500 Hz, see right y-axis).

Figure 5 (colour online). 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.

Figure 6. Accuracy scores for the five AM combinations, for each Method (PAD and MFB). Error bars indicate standard error. The left subplot shows mean scores over both trochaic and iambic sentences. The right subplot shows a breakdown of scores by sentence rhythm type (T = trochaic, I = iambic).

Figure 7. Effect of phase-shifting. Accuracy scores averaged across PAD and MFB methods. Error bars indicate standard error.

Figure 8 (colour online). MDS solutions for participants' response patterns across 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 (trochaic = solid line; iambic = dashed line). 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.