Situational influences on rhythmicity in speech, music, and their interaction

Sarah Hawkins, Centre for Music and Science, University of Cambridge, sh110@cam.ac.uk

To the memories of Dennis Klatt 1938-1988 and Kenneth Stevens 1924-2013, with gratitude

Key words
rhythmic entrainment; linguistic structure; communicative function; P-centres; dynamic attending theory; human interaction

Abstract 182 words (max 200)
Brain processes underlying the production and perception of rhythm indicate considerable flexibility in how physical signals are interpreted. This paper explores how that flexibility might play out in rhythmicity in speech and music. There is much in common across the two domains, but there are also significant differences. Interpretations are explored that reconcile some of the differences, particularly with respect to how functional properties modify the rhythmicity of speech, within limits imposed by its structural constraints. Functional and structural differences mean that music is typically more rhythmic than speech, and that speech will be more rhythmic when the emotions are more strongly engaged, or intended to be engaged. The influence of rhythmicity on attention is acknowledged, and it is suggested that local increases in rhythmicity occur at times when attention is required to coordinate joint action, whether in talking or music-making. Evidence is presented that suggests that while these short phases of heightened rhythmical behaviour are crucial to the success of transitions in communicative interaction, their modality is immaterial: they all function to enhance precise temporal prediction and hence tightly-coordinated joint action.

1. Introduction

This paper has its origin in work presented at the conference, but not represented in the finished volume. To set the stage, Large provided a summary overview of the application of Dynamic Attending Theory to music, while Scott presented a high level overview of the neural pathways implicated in the processing of speech. Taking the juxtaposition of the processing of speech and music as a challenge that promises to illuminate what studying each alone might not, this paper seeks to draw out the similarities and differences we must confront as we consider how the brain processes speech and music. The emphasis is on the lessons learned that can then inform us about human behaviour, especially with respect to that critical common aspect of musical and speech behaviour: communicative interaction.

The work underlying Large's presentation is Dynamic Attending Theory, first set out in general form by Jones and Boltz [1] (see also [2]) who pointed out its relevance to complexly-structured temporal events, particularly music and speech. DAT's basic tenets are that interactive attention entrains to rhythmic environmental events and that this interactive attunement between the organism and its environment sets up expectations for future events of a related nature. Our sense of timing and rhythm stems from this, and so is tied to intervals defined by events external to ourselves, rather than to intervals defined by anything like an independent internal 'clock'. Complex or hierarchically-structured rhythms are responded to as a number of different tempi, the faster ones nested within, or coupled to, the slower ones. The theory was ground-breaking in its connectedness with established biological processes in non-temporal domains, in how it accorded attention a participatory quality by highlighting how it is guided by interaction with environmental events, and in using attention within this framework to guide analytic attending to local events as well as long-term expectancies. Large and Jones developed DAT [3], introducing concepts of nonlinear dynamics and phase attractors to the model, these having been used for some years in modelling movement, including in the production and to some extent perception of speech e.g. [4, 5]. More recently, Large has taken advantage of advances in neuroscience to enhance the model's biological roots while developing and testing the

model further, especially with respect to rhythm and timing in music. See [6] and Large's paper in this volume for comprehensive reviews, and [7] for a clear yet brief technical description of the mathematical principles of nonlinear resonance.

Consistent with the primary tenet of DAT, Large suggests that the sense of musical pulse and metre arises when neural oscillations in sensory and motor cortices are driven by an external rhythm. Musical rhythm produces a neural resonance with a particular frequency that arises and is perceived as a pulse or tactus. Felt metre arises from the interaction of neural resonances of different frequencies. Such higher-order resonances respond at harmonic (nf) or subharmonic (f/n) frequencies, or integer ratios (e.g. xf/n) of the basic resonance frequency f, where n, x are integers; so different oscillatory frequencies represent different levels of metrical structure. Crucially, Large and colleagues model these interactions as nonlinear. An important property of nonlinearity is that there is a huge response to certain parameter values, or their combinations, and very little response to others that may be quite similar acoustically. This is the principle of quantal theory [8] which, though incomplete as a model of speech perception, nonetheless is both rooted in acoustic theory and provides a valuable basis upon which to build other cognitive processes that together allow speech to be understood. Thus a nonlinear oscillator network does not just transduce the signal, but adds frequency information which allows, amongst other things, pattern recognition and hence pattern completion, thus providing a basis for perception of pulse and metre in simple and syncopated musical rhythm.

Pattern completion is also fundamental to crucial properties of speech perception e.g. (references just representative): understanding speech in noise from 'intelligent interpretation' of wide-ranging spectro-temporal attributes of the signal [9-12], Gestalt-type processes of auditory scene analysis [13-17], on-line use of fine phonetic detail to facilitate access to meaning [18, 19], listeners' temporary adaptation to accents and ambient conditions [20], perhaps partly via tuning of the outer hair cells [7], and the influence of context on how very casual speech is understood [21-23] including influences of speech rate and rhythm early in an utterance on the interpretation of words in later portions of the speech [24, 25]. In both speech and music, pattern completion may explain the subjective experience of communicatively significant pulse when there is no event in the physical signal e.g. [26-28]. In conversation, such silent but felt beats may influence successful turn-taking, and can be crucial in determining, for example, whether a response to someone else's statement is an agreement or a disagreement: longer delays tend to signal disagreement. In Velasco and Large's [7] words, "nonlinear oscillators can track tempo fluctuations of complex rhythms, and deviations from temporal expectancies may provide a means of perceiving structural intentions of performers." Resonance, then, does not just aid memory by simplifying and streamlining, but might underpin an active, multi-level structuring process that allows performers and listeners to "make sense".

In sum, Large's DAT is one of a class of nonlinear dynamical systems models. It allows for the active creation of metrical components that may not be in the signal. Natural neural frequencies, of different periods, work simultaneously and entrain to external stimuli, probably via phase-locking of oscillations to onset rise times [29]. Beta waves mediate auditory-motor networks and allow prediction of the time of rhythmic events, while the faster gamma waves induce the percept of metre. The model appeals for many reasons. These include that it accounts for learning and other forms of flexible response in ways compatible with Hebbian learning [30] and its more recent formulation as functional synergies, and that it offers an explanation of how we feel pulse and metre, and can predict future events, even with irregular external stimuli and missing beats. As developed below, entrainment and prediction seem to be fundamental requirements for successful interactive communication.

The details may be debated. For example, Giraud and colleagues implicate delta, theta and gamma frequencies rather than beta and gamma (reviewed by [31]). Other models propose much the same principles via different processes and for different purposes e.g. [5, 32-46]. This is a reason to take the model seriously—when there is so much we do not know, different beginnings that converge on similar conclusions may be more valuable than unnecessarily polarised viewpoints [47]. However, there are also fashions in scientific thinking, so convergence alone is not enough to merit gravitas.

Oscillatory and dynamical systems models give a greater role to memory and prediction than many earlier ones, and in so doing have arguably opened up ways to investigate and model interaction. They are also general enough to apply to any medium, and in de-emphasizing a particular modality, they help develop understanding of general principles of how the brain controls and coordinates complex behaviour, be it motoric or social e.g. [48, 49]. Both contributions represent a significant step forward, though lacunae remain: for example, Vesper *et al.* [50] argue that dynamical systems model synchronization of immediate behaviour well, but not prediction and planning, for which we need more explicit modelling of shared attitudes and intentionality, as well as representation of other interactants' tasks.

Scott and colleagues propound a similar basic philosophy from a different perspective. They examine response patterns and connectivity in the cerebral cortex in speech production and perception tasks, and additionally compare across species, usually with monkeys. In a series of papers e.g. [51] (summarised for phoneticians in [52]) and [53], Scott suggests that the dorsal 'where/how' pathway, which extends from the auditory cortex to the frontal lobes, via the parietal lobe and sensory and motor cortices, includes amongst its functions tracking rhythm and rates of speech, and anticipating the end of a talker's turn. She points out that, while the behavioural and neuroscientific literature demonstrates that motor areas are clearly involved in speech perception and language comprehension, there is no evidence that activation of motor areas entails phonetic decoding, or is speech-specific. The responses could be any of embodied semantics, syntax (sequential dependencies), an enhanced response to adverse conditions, and interpretation of the 'how' of the signal's provenance. This is a refreshing interpretation, and almost certainly correct. Examples of independent corroboration include fMRI reports of contextual meaning modulating neural responses to isolated verbs [54] and of auditory objects forming (a type of pattern completion) [55], in this pathway.

Parallel statements can be made for phoneme perception: there is no direct, unambiguous evidence that phonemes are a primary unit of sensory perception, or necessary in speech perception, because abstract, context-free phonemes are physically realised as context-dependent allophones, which are themselves composed of auditory properties that are not necessarily discretely bundled with respect to discrete phonemes. Furthermore, and crucial to the current argument, every utterance, even if just one phoneme, is necessarily part of a larger structure. For example, [ɑ] (*ah*) spoken on a falling f0 to indicate understanding, can be described as a phoneme /ɑ/, but also as having various vocalic features and temporal properties characteristic of its utterance in isolation (immediate phonetic context = silence), and it is simultaneously a syllable, a foot, an intonational phrase, a conversational turn with a particular pragmatic and functional purpose, and so on. Change one cluster of properties, such as its attack time, duration and timbre, and it acquires a completely different function, such as the [ʔaa̰] indicative of sudden recall of something which, though possibly relevant, heralds interruption of the first speaker; keep the original properties but change the f0 trajectory from falling to rising, and it will still be the phoneme /ɑ/, but its function changes to one of encouraging the speaker to explain further.

Extrapolating, although we can identify many units of analysis in both music and speech, their properties, which scholars tend to represent as hierarchically structured (though they may not be in terms of neural processing), are such that we do not know what the units of speech perception are. It may not even be sensible to seek them at the neurophysiological level, given current technology. We can guess that there must be some correspondence between units of theoretical linguistics and neural memory circuits, but it is by no means clear whether, and if so how, a given unit is obligatorily activated during the understanding of a speech signal, far less during understanding of and participation in a conversation. Likewise for music: exposure to a particular musical system shapes perceptual responses to an external stimulus both in rhythm and pitch domains [56, 57]. In sum, in both domains, listeners find what they already know about in the signal that they happen to hear.

Both Scott's and Large's work proposes, to different degrees, active creation of percepts, with pattern completion, and with the neuropsychological underpinnings of social interaction—the timing of synchronous activity, and turn taking. The present paper accepts many of their premises to examine data on rhythm perception in speech and music from a single point of view: that the listening brain

constructs musical and speech rhythm/metre; that their construction depends on experience and bears a poorly-understood relationship to the physical signal; that the brain readily interpolates elements into the percept that are physically absent in the signal, and that although there is little evidence for physical isochrony in either medium, both tend to be experienced and hence described as largely isochronous, this misconception being of interest in itself.

While the subjective sense of isochrony merits investigation, it seems unrealistic to assume that its source must lie solely or even mainly in the physical signal. That type of identity between signal and percept is not found in vision, olfaction, or touch. What reaches the retina bears little obvious resemblance to what we experience as a table, or a rabbit, and much learning about an object's functional significance is required for all varieties of table and rabbit to be recognised as such. While, in any modality, certain physical signal properties, and their combinations, are more likely to trigger a particular percept than others, local context and general expectations are enormously influential, and it seems reasonable to assume that all non-reflexive perception results from complex, context-conditioned interaction in distributed neural circuits [58-60].

Extrapolating, four points are argued here. First, that structural and functional properties of music and language systems constrain the extent to which it is possible to achieve 'true' rhythmicity. Second, that listeners familiar with the structural properties of the communicative system compensate perceptually for structurally-imposed departures from rhythmicity, so that their knowledge of the system, and hence of its constraints, influences them to hear more rhythmicity than need be in the signal. Third, that these principles may be sufficient to explain the greater rhythmicity of music than of speech. And fourth, that differences in function produce variation in the degree of rhythmicity achieved in different speech styles. In short, comparing structures and functions of speech and music may help elucidate relationships between physical signal and auditory percepts of rhythm. Let us apply this reasoning to P-centres, an important influence on much of Scott's thinking [61].

2. P-centres in music and speech

A P-centre [62] is the psychological moment of occurrence of each note, or word, in a sequence. Differences in location of P centres for speech and music illustrate why we should expect experience-driven differences between the two. The P-centre in spoken words bears a complex relationship with syllable structure, whether acoustically or linguistically defined. Morton, Marcus, and Frankish [62], and subsequent research [61, 63-68] showed that the most important factor is probably onset rise time or rate of change of the envelope amplitude near the beginning of the syllable, although with a complex relationship between aperiodicity and periodicity. Moreover, other influential factors include segment or syllable duration and the properties of the end of the syllable. In consequence, as Figure 1 shows, considerable temporal variability in the relative alignment of acoustic word onsets is needed to produce the percept of a regular rhythm. For music, there can also be considerable variability in the relationship between the acoustic onset of a note and its P-centre, but the influencing factors seem to be fewer and more closely related to onset properties, presumably because the acoustic offset properties of many musical notes are less variable, and hence more predictable, than those towards the ends of words or syllables in most languages. Musical P-centres depend mainly on the onset rise time and (in a type of threshold effect) the overall amplitude of the notes. Duration and spectral timbre play a much smaller role, and offset properties seem unimportant [69, 70]. Thus high-amplitude notes with short rise times (rapid attacks) have P-centres close to their physical onsets, the precise location depending mainly on the note's amplitude rather than the rise time; whereas low-amplitude notes with slow rise times have later P-centres. In the latter case, duration and timbre can be more influential, and P-centre location is less precise and more easily masked by other sounds.

Insert Figure 1 about here

All these points are broadly compatible with oscillatory models. From a different perspective, Gordon [70], citing Pickles [71], speculated that P-centres for music may originate in firing patterns of octopus, pauser and chopper cells in the cochlear nucleus. However, Pickles [72] notes that mechanisms of temporal coding in the cochlear nucleus and above are still poorly understood; and that although there is solid physiological evidence for sharpening of responses to acoustic transients as the signal passes between the auditory nerve and the inferior colliculus, the vastly multiplied degrees of freedom introduced by the more recent recognition of the descending neurons comprising the corticofugal system mean that neurophysiological understanding of temporal coding is little better than it was 30 years ago. Furthermore, temporal resolution declines in regions higher than the inferior colliculus [73].What we can be sure of is that the corticofugal system allows ample opportunity for prior experience, and hence expectation, to affect the most basic aspects of rhythm perception, right down to the cochlear hair cells [74], including in humans [75]. Abundant evidence now shows that temporally-regulated behaviour in speech, music and other performance activities involves both cerebellum and basal ganglia, possibly with different roles regarding prediction [e.g. 76].

Furthermore, observations about P-centres apply to very restricted circumstances: isolated words, spoken more regularly than is normally the case even in natural-sounding lists, and alternating sequences of two tones [69] or orchestral instruments [70]. The relevance to rhythms of music, and especially of normal conversational speech, is debateable but worth exploring. For example, [77] reported that when accented monosyllabic words in simple read sentences are p-centre aligned, rather than misaligned by 100 ms, listeners tapped more consistently and accurately, especially later in the sentence; interestingly, listeners tapped equally consistently to unmanipulated natural speech. Questions for speech thus include whether the location of P-centres for digits is unchanged when the digits occur in natural connected speech, with and without prosodic focus; whether the location is comparable in the same word placed in short and long feet, and whether the observations made in the literature for ‘vowels’ and ‘consonants’ would be more appropriately couched in terms of periodic and aperiodic onsets or rate of change of amplitude envelopes respectively—those few experiments that explicitly compare non-obstruent with obstruent consonants (i.e. nasals and approximants with stops and fricatives) suggest that these details do matter [66, 67, 78, 79]. Work preceding the discovery of P-centres [80-82] as well as Marcus’ own [63] suggests strong but probably imperfect correspondences between periodic onsets and P-centres, but much remains to be clarified. . The recently-introduced concept of P-centre clarity [78] seems particularly promising. Clarity reflects the breadth of the time window within which a P-centre may be subjectively experienced (P-centre subjective ‘precision’), and is related to both the abruptness of the initial event onset and the offset between the onset and the P-centre. So, for example, syllables beginning with simple onsets like /b/ have clearer P-centres than those beginning with complex onsets like /sp, spl/, with onsets like /n, l/ intermediate in clarity. The importance of these observations is developed below. For music, one obvious question is whether the cited facts hold even for well-known sequences of notes i.e. familiar tunes, because presumably some of the end-of word influences in speech are due to listeners knowing what the end of the word is: are musical P-centres affected by the ends of notes when the melody is familiar?

Why are P-centres different in speech and music? Can we exploit this difference to find a common explanatory principle for perception of auditory rhythm? Scott and McGettigan [64:2] comment that “perceptual centers are associated with increases in mid range spectral energy (around 500–1500 Hz; Marcus, 1981), i.e., with the onsets of the first formants in speech…thus linked to the onsets of vowel sounds.” The regularity here seems not especially connected with F1, which is lower than 500 Hz in many vowels, and rarely as high as 1500 Hz even in children’s speech; further, voiced consonants (including approximants) can have a considerable amount of low-frequency energy. The influencing factor seems more likely to be the onset of higher-amplitude energy (relative to ambient levels) across a wide range of the frequencies characteristic of the sound source(s). This might account for the relatively simpler relationship between P-centre and physical onsets in music compared with speech, since source characteristics change less within a musical instrument than a human talker. Local-contrast and/or threshold effects might account for the strong influence of high- but not low-amplitude fricatives (/s/ but not /f/ in English digits). However, no formula has been found. A more nuanced

explanation seems to be required, one that includes acoustic properties, but also prior knowledge about how the communicative medium is structured and functions.

The hypothesis explored here is that, given that the function of a communication heavily constrains the form it takes, then if speech is being used to convey referential meaning, the words (phonology and morphology), and their sequencing (grammar) are largely dictated by the particular language. There is often more than one way to express the same meaning, lexically, grammatically and pragmatically, but the talker is nonetheless severely constrained by linguistic and pragmatic factors, and some forms are more natural and hence frequent/acceptable than others. Music, in contrast, though subject to a wide variety of practical and cultural constraints, and at least as able to vary rhythm for expressive purposes, nonetheless lacks both the variety of obligatory phonological influences on syllabic complexity, and explicit reference to real-word objects and concepts, that characterise language. Thus there may be a closer relationship between physical signal and perceived rhythm in music than in speech because music is subject to fewer intrinsic, structural, constraints.

Put simply, when speech is used to convey referential meaning, then the words, and their sequential order, are largely dictated by the language. When words differ in number of syllables, and syllables differ in structural complexity, stress placement, and duration, then strict rhythmic cycles may have to take second place to intelligibility. Prosodic structure, pragmatic function and emotion add further constraints. Resultant differences can be large. For example, in an utterance of *everyone was happier after Geoff re-strained the fruit* (nuclear syllable *strained*) the metrically weak syllable *re-* was 108 ms, whereas it was only 33 ms in *everyone was happier after Geoff restrained the brute* (nuclear syllable *brute*) spoken by the same speaker in the same speech style and rate. The spectro-temporal difference is central to the meanings and therefore must be made, yet the durational difference is about the same as that producing poor predictive tapping [77]. Durational differences in weak CV syllables such as English *re-* will be amongst the smallest. Much greater differences occur between stressed syllables.

The cumulative effect of a series of these small adjustments can be large. Native listeners can be expected to know about and to compensate for them, just as they compensate for many other contextual effects, ranging from intrinsic vowel f0 [83] and nasalization of English vowels [84], to phoneme restoration [85, 86] and fricative distortions due to holding a pen in the mouth [20]. Music, being less constrained by such factors, can accord the beat relatively greater importance.

However, speech styles do exist in which the perceived beat is paramount: nursery rhymes provide a prime example, though many forms of adult poetry also have a clear rhythmic structure, as does Shakespearean blank verse. Other forms of speech maintain a clear rhythm for short periods: lists [87], parts of persuasive oratory [88] and parts of infant-directed speech as one of many ways to facilitate the infant's prediction and hence language learning [89]. Speech registers can be intuitively ranked from most to least rhythmic thus: nursery rhymes and playground games/chants, metrical poetry, infant-directed speech (especially if non-referential), persuasive oratory, and any kind of talk, whether didactic or conversational, in which conveying a precise meaning is paramount, including, presumably, when cognitive load is high. Knight [88] has corroborated the validity of this intuitive ordering by showing greater variability in 'tapping to the beat', and increasingly low subjective ratings of rhythmicity, as one progresses through metred poetry, persuasive oratory, didactic speech, and conversation. She suggests that periodicity can serve both attentional and persuasive functions, presumably at least partly via entrainment.

Importantly, styles with clear rhythm should still be subject to the structural phonological-syllabic constraints noted above. Presumably languages which possess little variation in both syllable structure and rhythmic properties (Korean, Tamil, Spanish, Italian) should be relatively more isochronous than English in each particular style. Evidence from PVI studies suggests that they are, although such generalisations must be used cautiously [90, 91] and probably require more composite measures than those relying purely on duration [90, 92]. Laboratory studies of English and French reiterant speech show that repeated CV syllables like /ma/ can be quite isochronous yet reflect rhythmic patterns

appropriate to particular real utterances when the task is to substitute such syllables for the actual syllables of specific sentences [93-96]. Interestingly, [95] found durational modifications for reiterant speech phrases of 3-5 syllables that reflected word stress, final (but not nonfinal) position, and foot length, with feet of the same size being of constant duration—a metrical structure similar to that of Western music.

Sufficient control of language, talker, register and structure, then, with measures of rhythm that incorporate acoustic influences other than duration, may go a long way towards validating subjective judgments of rhythmicity in speech, and so offer closer similarity to perception of musical rhythm. McGowan and Levitt [97] took a significant step in this direction by showing, for three rhythmically-contrasting English dialects, correspondences between PVIs for individuals' spontaneous speech and the way those individuals played Scottish reels.

But differences seem likely to remain. These may be accounted for by distinguishing differences in function, possibly over quite short time periods. The degree of rhythmicity in speech may reflect the extent to which the speaker wishes to evoke prosocial and affiliative responses in the listener via rhythmic entrainment [88, 98]. Poetry is designed to move the listener, political speech to persuade, to inspire and induce loyalty, and so on. The argument is appealing, though modification for the widely varying functions of conversational speech needs to be introduced, and is returned to below.

Such functional differences in speech can be loosely described as contrasting in the relative balance accorded to phatic/emotional vs referential meaning. The emphasis must be on *relative* balance, since speech with a major phatic function can be referential, and highly referential speech often includes phatic or emotional force. Indeed, referential meaning is presumably most effective when it is highly intelligible yet conveys the commitment of the speaker and arouses engagement in listeners. Further, even the most mundane communications are likely to arouse memories, and wherever memory is implicated, so too will emotions be [cf. 99]. Even individual words may excite a complex network of associations. It would thus be surprising if the limbic system were not habitually active during speaking and listening. The literature confirms that it is e.g. [99]. The hippocampus, parahippocampus and cingulate cortex are regularly reported as active during speech/language processing, but these arguably have multiple functions, not all of which need be classed as emotional. More tellingly, the amygdala, commonly accepted as mediating more purely emotional and vigilance responses, is activated during listening to narrative speech [99-101] but not to their decontextualized sentences or words [101]. Although amygdala activation has been reported when listening to lists of independent, descriptive spoken sentences read in an unengaging unemotional style, those sentences were, presented in noise, reactions to which may have produced their own emotional response [102]. Interestingly, despite the large literature showing limbic activation while listening to music, Alluri *et al.*[103], comparing Beatles songs (with lyrics) against a medley of non-vocal instrumental music (Vivaldi, Miles Davis, Booker T and MGs, and The Shadows) found limbic system involvement only for the music with lyrics. While acknowledging that part of the reason may be methodological (less homogeneity in the non-vocal medley set of music), they attributed this difference to greater limbic system involvement when music includes lyrics. Presumably, the lyrics must be intelligible or known beforehand.

In summary, languages will differ in the degree to which regular physical beats are achievable, as will musics, since they too differ in structure. Within a given language, different speech styles will emphasize rhythm more or less, depending partly on the intention to persuade or to gain the audience's positive emotional involvement with the speaker. However, it would be a mistake to equate degree of rhythmicity with location of the speech style along a continuum of phatic/emotional to referential speech, for both logic and neuroscience tell us that whenever there are memories and stories, there are emotions, and even strongly emotional language still has phonological and syntactic constraints. Nonetheless, these differences in balance amongst the various communicative functions of language may in large part dictate the greater predictability of rhythm in music than in speech. A hypothesis yet to be tested is that everyday conversational talk varies in rhythmicity depending on (a)

its current particular function, (b) the availability of alternative lexical and syntactic formulations, and (c) the particular phonological constraints that the chosen words impose.

3. Common substrates of speech and music

An obvious question is whether speech and music share a common substrate that enables or facilitates successful joint action and communication. Musical and linguistic skills can be mutually enhancing [87, 104, 105], although there are many counter-examples, and much domain-specificity in the application of skills [106]. The challenges to well-controlled comparisons are significant, including a tendency for linguists to overestimate the homogeneity of musics and musicians, and vice versa. However, the burgeoning literature on evolution of speech and music e.g. [98, 107] suggests that, if there is a shared substrate, then rhythmic entrainment seems likely to be implicated.

Temporal properties that music and speech share include frequent use of a wide range of basic rates or tempi—for various purposes; local rate changes (rubato) which also serve various purposes (emotional expression, asides, floor-holding); phrase-final lengthening; predictable tonal endings (cadences, nuclear tones), and, amongst suitably enculturated listeners, the percept of higher-order structure (metre, feet, intonational phrases). In both media, deviation from rhythmicity indicates emotion, and phrasing for greater intelligibility. As noted above, in speech the demands of the specific words and grammar used also cause deviations from rhythmicity, though in certain circumstances words and structures are chosen with rhythm accorded higher-than-usual priority. In music, skilled composers and improvisers write/play to accommodate the constraints imposed by the instrument: with less-skilled players, deviation from rhythmicity can indicate where such constraints operate.

Consistent with most other perceptual approaches, it is now commonly accepted that listeners construct regular rhythms. In Large's model, the brain's way of dealing with structural demands of the medium allows considerable flexibility: the build-up of resonances has windows of allowable variation, phase correction, and other forms of re-adjustment so that perceived rhythm is more regular than the regularity of physical events giving rise to that perception. Constraints may be such that this sense of regularity extends over longer durations and hence is more powerful in music than in speech, but is not qualitatively different. Consistent with Large's general orientation, when listeners understand what is being said, they also appear to construct patterns of brain activation that are highly correlated with those of the speaker. For example, using fMRI, [108] showed spatial and temporal coupling between listeners' brain activity and that of a speaker telling a story, but only when listeners understood the speaker: story-telling in a foreign language that the listeners did not understand elicited no such coupling. Most correlated patterns in listeners were delayed relative to the speaker's; but importantly, some were anticipatory, and the greater the anticipatory speaker–listener coupling, the greater the degree of understanding as assessed in an independent measure. For broader discussion, see [42] (language) and [109] (music).

Most literature discussed so far has examined either music or speech, but not both. If coupling of neural oscillations between individuals underpins successful communicative interaction, then we should find similar processes in music and speech, differing only in terms of the demands of the medium, and the function of the particular interaction. We should find evidence for shared attention, and for mutual entrainment making possible the required alignment, or coordination, of actions. Further, if hypotheses from music are correct, that participatory music-making of all types facilitates a sense of shared intentionality and enhances social bonding [98], then we should find not only that individuals entrain to co-produce improvised music, but that they enjoy doing so—i.e. that doing so is itself rewarding. Finally, if rhythmic entrainment underpins coordinated (and hence successful) communication, then we would expect to find not only similar processes affecting both music and speech, but also relatively seamless transitions between the two. The following section summarizes a preliminary investigation of this last hypothesis.

4. New data on rhythmic entrainment in spontaneous talk and joint improvisational music-making

Jointly-improvised music and spontaneous conversation were compared because both are natural activities that require interactants to manage moment-by-moment unscripted sequenced actions. Our emphasis was on mutual cooperation in achieving joint goals. The data, from a study described in detail by [110], involve analysis of temporal properties of joint improvisational music-making between five same-sex pairs of friends aged 18-31, three musician pairs (two pairs male) and two non-musician pairs (one pair male). Musicianship is not distinguished here as it had no discernible effect on the parameters discussed. They were recorded (5 audio microphones, 4 video cameras) for about one hour talking, doing simple non-musical cooperative 'construction' tasks and improvising together on percussion instruments (a circular one-octave xylophone, a kalimba, various drums, and clapsticks).

Episodes from the 10 minutes of improvisation were extracted when there was talking either simultaneously, or in close proximity, with obvious attempts to make music together that lasted for several seconds. Episodes which were ambiguous in any way were excluded, as were sections where the players explicitly counted in or otherwise discussed in detail what to play—the music was spontaneous. Most main extracts lasted between 30-60 s, range 25-91 s; they were further subdivided if tempo or rhythm changed appreciably. Music within these episodes was classed subjectively as 1) 'successful' (27 cases); 2) 'breakdown': the disorderly cessation of a bout (5 cases); or 3) 'unsuccessful': a new bout was attempted, but was subjectively experienced as relatively incoherent, though still 'music' (15 cases). The validity of this subjective classification of success was confirmed by differences in variability of the tactus, operationally defined as the mean inter-onset interval (IOI) of the 'main beats in the bar', each being coincident with the onset of a performed note. Although the mean pulse rates (mean IOIs) of successful and unsuccessful bouts were close (731 ms and 664 ms respectively) and within preferred/spontaneous tempo [111], the mean standard deviation of the IOIs within each successful bout (64 ms) was almost half that of those in unsuccessful bouts (112 ms). Thus there was far more tempo variability in unsuccessful bouts. Since much of the music was essentially treated as unpitched percussive, especially by some of the non-musicians, melodic or harmonic measures of musical success were not applicable.

Next, the hypothesis described at the end of the preceding section was tested by comparing the temporal locations of musical pulse onsets and of f0 maxima or minima in stressed syllables (pitch-accents in the ToBI system) in each individual's speech. In anticipation of future work, these f0 maxima/minima are called *pikes*, after Loehr [112]. In Loehr's definition, pikes are often multimodal, with gesture, for example, being an important contributor, but the modalities that contribute to any given instance of a pike can vary across instances. For speech, the only requirement for the event to be called a pike is the presence of a pitch-accented syllable, from which the time of occurrence of its local f0 maximum (or (more rarely) minimum, if more prominent) is measured. Figure 2 shows waveforms recorded from two talkers' close mouth microphones, combined spectrogram with superimposed f0 contour, and various labelled tiers for a short extract from a successful bout. For the two pikes that occur in this extract, one spoken by the Right hand talker and the other by the Left talker, the deviation between the pike and the nearest pulse onset is indicated, expressed as a percentage of the IOI within which the pike falls. This is the pulse-to-pike (p2p) deviation. For this analysis, the five cases of breakdown (plus speech), were included in the unsuccessful set to simplify and increase the power of the statistical analyses, since there were too few for a separate analysis and their mean p2p deviations were spread evenly across the range of successful and unsuccessful bouts.

Figure 3 shows that the majority of p2p deviations was around 20-30% in both successful and unsuccessful+breakdown bouts, but when the music was successful the distribution was significantly skewed towards smaller deviations, whereas it was more normally distributed in unsuccessful bouts (Mann-Whitney U (27,20) = 159, p = 0.009, one-tailed test). A number of alternative parametric and non-parametric analyses explored the effects of using shorter or faster IOIs in the case of ambiguous pulses. For example, the same analyses were re-run with p2p values of 50% ±5% removed, as if the pulse rate were doubled. All gave the same pattern of results. Thus the hypothesis is supported: talkers

seemed to entrain with one another such that spoken pikes occurred significantly closer to musical pulse onsets in and around successful compared with unsuccessful joint improvisation.

Insert Figure 2 about here: Music and speech \ data for demo figs \ tribal.pptx

Insert Figure 3 about here

The entrained speech was not always simultaneous with the musical bout but preceded or followed it in 23% of cases (7 before and 4 after the music). For these, the p2p deviation was calculated by adding 'virtual pulses' of uniform duration backwards or forwards in time until the first or last pike occurred within a pulse IOI. The IOI of these added pulses was usually that of the first IOI they preceded or of the last IOI they followed, for talk preceding and following music respectively. In a small minority of cases, when that initial or final IOI was not representative of the bout's average, either the next IOI or the average of the entire bout was taken, depending on which was more representative of the local IOIs. Results for these data were not measurably different from those in which the speech occurred within the bout.

In other words, the synchronization of speech and music seems not due solely to the pulse of the music; pulse may emerge in the speech prior to the musical pulse being produced. When such temporally-aligned speech precedes a successful musical bout, it seems to seed the musical pulse so that players can start to play synchronously at the same tempo without overt negotiation. This is a tentative finding but is found across all participants, musician and non-musician alike as well as in pilot data with slightly different tasks. If it proves robust, it supports the interpretation that talking together and joint music-making share a common resource of mutual temporal alignment that allows successful initiation and continuation of spontaneous interaction; and that neural substrates of this alignment may include coupled neural oscillation between interlocutors, focussed within particular temporal windows whose duration is dictated by the function of the communication.

Analyses by Richard Ogden [113] of the same participants' question-answer turns during purely conversational parts of our recordings lead to the tentative conclusion that, when an answer follows a question in the next turn, at least its initial pike, or the initial sound that the answerer makes (such as a click, in-breath, *um* etc), is usually rhythmically synchronised with the last two or three pikes of the question. That is, the last 2-3 pikes of a question tend to have fairly equal IOIs, signalling (with other factors) that the end of the turn is imminent. These final pikes set up an underlying pulse which seems to provide the next speaker with a periodicity with which to coordinate the beginning of the answer. Sufficiently primed, two beats is all it takes to establish rhythmicity in music [114] so 2-3 beats across two turns may be all it takes to achieve and confirm successful alignment and mutual affiliation in conversation. Similar conclusions are arrived at by Widdess [115], in an analysis of how an Indian *dhrupad* singer coordinates with his drum accompanist during improvisation to achieve simultaneous endings. The players improvise relatively independently, but the drummer provides understood anchor points from time to time throughout, and, crucially, the singer's small (apparently imperceptible) timing adjustments to make the last three or so beats of a section periodic allow the successful achievement of the required simultaneous ending, *sam,* while accommodating all and only the necessary syllables. Perhaps it is within these relatively short but interactionally critical domains that we should be looking for mutual alignment of neural oscillatory activity, and modelling the locus of

predictive behaviour. And perhaps, to a human brain, it does not matter whether the interactional medium is speech, music or dance, i.e. sound, vision, or movement.

5. Concluding remarks

Cortical oscillatory rhythms may serve to focus attention on crucial parts of a sensory signal, one consequence being to enable interactants to coordinate and align their behaviour at different metrical levels, but without the need for predicting rhythm over long temporal domains. That such attentional focus may only be needed for relatively localised parts of a signal is compatible with evidence for the ubiquity of rapid, repeated, apparently unconscious, and (in the case of interaction) mutual adjustments of phase rather than period, for cerebellar and (dorsal) thalamo-cortical oscillatory cycles, tapping behaviour, and discrimination of 'out of time' events e.g. [116-118]. An appealing aspect of this interpretation is that it emphasises the importance of attended-to detail in the physical signal while allowing much latitude in the extent to which higher-order resonances, or metrical structure, need be physically present in the signal, as demonstrated by Large's work. The emphasis on the brain's creation of structure as an integral part of perceptual processing helps to show the conceptual difficulties in postulating a stark difference between 'exemplar' and 'abstractionist' processing of words, which has been a focus of psycholinguistic research in the past two decades or so. While theorists may regard the incoming signal as an exemplar, it is at no time functionally independent of stored, high-level, abstractions in the brain. Both the afferent exemplar and the stored knowledge are physically instantiated in the brain, and thus both are presumably affected by memory and expectation because the incoming signal is modulated by the corticofugal system. That is, the corticofugal system's rich projections from the cerebral cortex down through the auditory pathway to the cochlea appear to tune and otherwise modulate neuronal responses to incoming sensory signals. Given this, the commonly-drawn distinction between 'top-down' and 'bottom-up' processing would seem to have little basis in reality, for once a neural signal has started on its route towards the cortex, it is already subject to 'interpretation' (modulation of various types) from cortical activity, and such modulations seem to continue throughout the complex synapses of the auditory pathway. Modality and (speech or music) domain may be irrelevant to such processes, and, in the case of timing, based on rhythmic entrainment that requires only the possibility of synchronising one type of behaviour with another. However, variation in processing speed of visual, auditory and tactile signals, and in the degree to which different types of speech can achieve physical rhythmicity, suggest that the cognitive system must account in some way for the source of entrainment. Further, it is suggested that communicative function typically overrides maintenance of rhythmicity when the two are in conflict. When they are not in conflict, however, as in much music and dance, and some forms of spoken communication, influences leading to rhythmic entrainment can profoundly influence the experience of the participants.

6. Figure captions

Figure 1: Relative temporal offsets required to achieve perceptual isochrony in a particular set of spoken digits (from Morton *et al.* 1976, [62]).

Figure 2. Waveforms, spectrogram with f0 contour, and praat tiers demonstrating the principles of the p2p analysis. See text for explanation.

Figure 3. Distribution of mean p2p intervals for successful and unsuccessful (+ breakdown) bouts of music.

7. Short title

Situational influences on rhythm

8. References

[1] Jones, M. R. & Boltz, M. 1989 Dynamic attending and responses to time. *Psychological Review* **96**, 459-491.

[2] Jones, M. R. 1976 Time, our lost dimension: Toward a new theory of perception, attention and memory. *Psychological Review* **83**, 323-355.

[3] Large, E. W. & Jones, M. R. 1999 The dynamics of attending: How people track time-varying events. *Psychological Review* **106**, 119-159.

[4] Saltzman, E. & Kelso, J. A. S. 1987 Skilled actions: A task-dynamic approach. *Psychological Review* **94**, 84-106.

[5] Tuller, B. Case, P. Ding, M. & Kelso, J. A. S. 1994 The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology-Human Perception and Performance* **20**, 3-16.

[6] Large, E. W. 2008 Resonating to musical rhythm: Theory and experiment. In *Psychology of Time*. Grondin S (Ed.) pp. 191-231. West Yorkshire: Emerald.

[7] Velasco, M. J. & Large, E. W. 2011 Pulse detection in syncopated rhythms using neural oscillators. *Proceedings of the 12th Annual Conference of the International Society for Music Information Retrieval*, 186-190.

[8] Stevens, K. N. 1989 On the quantal nature of speech. *Journal of Phonetics* **17**, 3-45.

[9] Greenberg, S. 2006 A multi-tier framework for understanding spoken language. In *Listening to speech: An auditory perspective*. Greenberg S, Ainsworth W (Eds.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[10] Cooke, M. P. 2006 A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America* **119**, 1562-1573.

[11] Heinrich, A. Flory, Y. & Hawkins, S. 2010 Influence of English r-resonances on intelligibility of speech in noise for native English and German listeners. *Speech Communication* **52**, 1038-1055.

[12] van Engen, K. J. & Bradlow, A. R. 2007 Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America* **121**, 519-526.

[13] Bregman, A. S. 1990 *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

[14] Warren, R. M. 1984 Perceptual restoration of obliterated sounds. *Psychological Bulletin* **96**, 371-383.

[15] Samuel, A. G. 1996 Phoneme restoration. *Language and Cognitive Processes* **11**, 647-653.

[16] Hawkins, S. 2010 Phonological features, auditory objects, and illusions. *Journal of Phonetics* **38**, 60-89.

[17] deWitt, L. & Samuel, A. G. 1990 The role of knowledge-based expectations in music perception: Evidence from musical restoration. *Journal of Experimental Psychology: General* **119**, 123-144.

[18] Smith, R. & Hawkins, S. 2012 Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics* **40**, 213-233.

[19] Hawkins, S. Clayards, M. & Gaskell, G. in prep Parallel processing of morphological and phonemic information: An eye-tracking experiment.

[20] Kraljic, T. Samuel, A. G. & Brennan, S. E. 2008 First impressions and last resorts. *Psychological Science* **19**, 332-338.

[21] Pickett, J. M. & Pollack, I. 1963 Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech* **6**, 151-164.

[22] Ernestus, M. Baayen, H. & Schreuder, R. 2002 The recognition of reduced word forms. *Brain and Language* **81**, 162-173.

[23] Ernestus, M. 2014 Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua* **142**, 27-41. (10.1016/j.lingua.2012.12.006)

[24] Heffner, C. C. Dilley, L. C. McAuley, J. D. & Pitt, M. A. 2013 When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes* **28**, 1275-1302. (10.1080/01690965.2012.672229)

[25] Morrill, T. H. Dilley, L. C. McAuley, J. D. & Pitt, M. A. 2014 Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. *Cognition* **131**, 69-74. (10.1016/j.cognition.2013.12.006)

[26] London, J. 2012 *Hearing in Time*. 2nd ed. Oxford: Oxford University Press.
[27] Stivers, T. 2008 Stance, alignment, and affiliation during storytelling: When nodding Is a token of affiliation. *Research on Language & Social Interaction* **41**, 31-57. (10.1080/08351810701691123)
[28] Pomerantz, A. 1984 Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In *Structures of Social Action Studies in conversation analysis*. Atkinson JM, Heritage J (Eds.) pp. 57-101. Cambridge: Cambridge University Press.
[29] Snyder, J. S. & Large, E. W. 2005 Gamma-band activity reflects the mteric structure of rhythmic tone sequences. *Cognitive Brain Research* **24**, 117-126. (10.1016/j.cogbrainres.2004.12.014)
[30] Hebb, D. O. 1949 *The Organization of Behavior: A Neurophysiological Theory*. New York: Wiley.
[31] Giraud, A.-L. & Poeppel, D. 2012 Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience* **15**, 511-517.
[32] Carpenter, G. A. & Grossberg, S. 2003 Adaptive resonance theory. In *The Handbook of Brain Theory and Neural Networks*. Arbib MA (Ed.) pp. 87-90. Cambridge, MA: MIT Press.
[33] Grossberg, S. 2005 Linking attention to learning, expectation, competition, and consciousness. In *Neurobiology of Attention*. Itti L, Rees G, Tsotsos J (Eds.) pp. 652-662. San Diego: Elsevier.
[34] Pearce, M. T. & Wiggins, G. A. 2006 Expectation in melody: The influence of context and learning. *Music Perception* **23**, 377-405.
[35] Roy, D. 2005 Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence* **167**, 170-205.
[36] Roy, D. 2005 Grounding words in perception and action: Computational insights. *TRENDS in Cognitive Sciences* **9**, 389-396.
[37] Todd, N. P. M. 1992 The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America* **91**, 3540-3550.
[38] Todd, N. P. M. O'Boyle, D. J. & Lee, C. S. 1999 A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research* **28**, 5-28.
[39] Tuller, B. 2004 Categorization and learning in speech perception as dynamical processes. In *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. Riley MA, Van Orden GC (Eds.) pp. Ch. 8 http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp.
[40] Wilson, M. & Wilson, T., P. 2005 An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review* **12**, 957-968.
[41] Tily, H. Gahl, S. Inbal, A. Snider, N. Kothari, A. & Bresnan, J. 2009 Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* **1-2**, 147-165.
[42] Hasson, U. Ghazanafar, A. A. Galantucci, B. Garrod, S. & Keysers, C. 2012 Brain-to-brain coupling: a mechanism for creating and sharing a social world. *TRENDS in Cognitive Sciences* **16**, 114-121. (10.1016/j.tics.2011.12.007)
[43] 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.
[44] Schroeder, C. E. Wilson, D. A. Radman, T. Scharfman, H. & Lakatos, P. 2010 Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology* **20**, 1-5.
[45] Garrod, S. & Pickering, M., J. 2004 Why is conversation so easy? *TRENDS in Cognitive Sciences* **8**, 8-11.
[46] van Noorden, L. & Moelants, D. 1999 Resonance in the perception of musical pulse. *Journal of New Music Research* **28**, 43-66.
[47] Zatorre, R. J. & Gandour, J. T. 2008 Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 1087-1104. (10.1098/rstb.2007.2161)
[48] Wolpert, D. M. Doya, K. & Kawato, M. 2003 A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society London B* **358**, 593-602. (10.1098/rstb.2002.1238)

[49] Buszáki, G. & Draguhn, A. 2004 Neuronal oscillations in cortical networks. *Science* **304**, 1926-1929.
[50] Vesper, C. Butterfill, S. Knoblich, G. & Sebanz, N. 2010 A minimal architecture for joint action. *Neural Networks* **23**, 998-1003.
[51] Scott, S. K. & Johnsrude, I. S. 2003 The neuroanatomical and functional organization of speech perception. *TRENDS in Neurosciences* **26**, 100-107.
[52] Scott, S. K. 2003 How might we conceptualize speech perception? The view from neurobiology. *Journal of Phonetics* **31**, 417-422.
[53] Scott, S. K. McGettigan, C. & Eisner, F. 2009 A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience* **10**, 295-302.
[54] Raposo, A. Moss, H. E. Stamatakis, E. A. & Tyler, L. K. 2009 Modulation of motor and premotor cortices by actions, action words, and action sentences. *Neuropsychologia* **47**, 388-396.
[55] Shahin, A. J. Bishop, C. W. & Miller, L. M. 2009 Neural mechanisms for illusory filling-in of degraded speech. *NeuroImage* **44**, 1133-1143.
[56] Hannon, E. E. & Trehub, S. E. 2005 Metrical categories in infancy and adulthood. *Psychological Science* **16**, 48-55.
[57] Endress, A. D. 2010 Learning melodies from non-adjacent tones. *Acta Psychologica* **135**, 182-190. (http://dx.doi.org/10.1016/j.actpsy.2010.06.005)
[58] Barsalou, L. W. 2008 Grounded Cognition. *Annual Review of Psychology* **59**, 617-645.
[59] Bar, M. 2007 The proactive brain: Using analogies and associations to generate predictions. *TRENDS in Cognitive Sciences* **11**, 280-289. (10.1016/j.tics.2007.05.005)
[60] Bar, M. 2009 The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 1235-1243. (10.1098/rstb.2008.0310)
[61] Scott, S. K. 1998 The point of P-centres. *Psychological Research* **61**, 4-11.
[62] Morton, J. Marcus, S. M. & Frankish, C. 1976 Perceptual centres (P-centres). *Psychological Review* **83**, 405-408.
[63] Marcus, S. M. 1981 Acoustic determinants of perceptual centre (P center) location. *Perception and Psychophysics* **30**, 247-256.
[64] Scott, S. K. & McGettigan, C. 2012 Amplitude onsets and spectral energy in perceptual experience. *Frontiers in Psychology* **3**, 1-2.
[65] Cooper, A. M. Whalen, D. H. & Fowler, C. A. 1988 *Haskins Laboratories: Status Report on Speech Research* **SR-93/94**, 23-32.
[66] Janker, P. 1996 Evidence for the p-center syllable-nucleus-onset correspondence hypothesis. *ZAS Papers in Linguistics (ZASPIL)* **7**, 94-124.
[67] Harsin, C. A. 1997 Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics* **59**, 243-251.
[68] Howell, P. 1988 Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics* **43**, 90-93.
[69] Vos, J. & Rasch, R. 1981 The perceptual onset of musical tones. *Perception and Psychophysics* **29**, 323-335.
[70] Gordon, J. W. 1987 The perceptual attack time of musical tones. *Journal of the Acoustical Society of America* **82**, 88-105.
[71] Pickles, J. O. 1982 *An Introduction to the Physiology of Hearing*. London: Academic Press.
[72] Pickles, J. O. 2012 *An Introduction to the Physiology of Hearing*. 4th ed. Bingley: Emerald.
[73] Scott, S. K. & Wise, R. J. S. 2004 The functional neuroanatomy of prelexical processing in speech perception. *Cognition* **92**, 13-45.
[74] Winer, J. A. 2006 Decoding the auditory corticofugal systems. *Hearing Research* **212**, 1-8.
[75] Khalfa, S. Bougeard, R. Morand, N. Veuillet, E. Isnard, J. Guenot, M. Ryvlin, P. Fischer, C. & Collet, L. 2001 Evidence of peripheral auditory activity modulation by the auditory cortex in humans. *Neuroscience* **104**, 347-358.
[76] Teki, S. Grube, M. Kumar, S. & Griffiths, T., D. 2011 Distinct neural substrates of duration-based and beat-based auditory timing. *The Journal of Neuroscience* **31**, 3805-3812.
[77] Buxton, H. 1983 Temporal predictability in the perception of English speech. In *Prosody: Models and Measurements*. Cutler A, Ladd DR (Eds.) Ch. 9, pp. 111-121. Berlin / Heidelberg: Springer-Verlag.

[78] Villing, R. C. Repp, B. H. Ward, T. E. & Timoney, J. M. 2011 Measuring perceptual centers using the phase correction response. *Attention, Perception and Psychophysics* **73**, 1614-1629. (10.3758/s13414-011-0110-1)
[79] Pompino-Marschall, B. 1989 On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics* **17**, 175-192.
[80] Allen, G. D. 1972 The location of rhythmic stress beats in English: An experimental study I. *Language and Speech* **15**, 72-100.
[81] Allen, G. D. 1972 The location of rhythmic stress beats in English: An experimental study II. *Language and Speech* **155**, 179-194.
[82] Huggins, A. W. F. 1972 On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America* **51**, 1279-1290.
[83] Whalen, D. H. & Levitt, A. G. 1995 The universality of intrinsic F0 of vowels. *Journal of Phonetics* **23**, 349-366.
[84] Krakow, R. A. Beddor, P. S. Goldstein, L. M. & Fowler, C. A. 1988 Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America* **83**, 1146-1158.
[85] Warren, R. M. 1970 Perceptual restoration of missing speech sounds. *Science* **167**, 392-393.
[86] Shinn-Cunningham, B. G. & Wang, D. 2008 Influences of auditory object formation on phonemic restoration. *Journal of the Acoustical Society of America* **123**, 295-301.
[87] Jefferson, G. 1990 List-construction as a task and resource. *Studies in Ethnomusicology and Conversation Analysis* **1**, 63-92.
[88] Knight, S. 2014 *An investigation of passive entrainment, prosociality and their potential roles in persuasive oratory*. PhD thesis, University of Cambridge.
[89] Fernald, A. 2000 Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica* **57**, 242-254.
[90] Nolan, F. J. & Jeon, H.-S. 2014 Speech rhythm: A metaphor? In *Communicative Rhythms in Brain and Behaviour*. Smith R, Rathcke T, Cummins F, Overy K, Scott S (Eds.). London: Philosophical Transactions of the Royal Society B.
[91] Arvaniti, A. 2012 The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* **40**, 351-373.
[92] Cumming, R. E. 2011 Perceptually informed quantification of speech rhythm in pairwise variability indices. *Phonetica* **68**, 1-25.
[93] Larkey, L. S. 1983 reiterant speech: An acoustic and perceptual validation *Journal of the Acoustical Society of America* **73**, 1337.
[94] Nakatani, L. H. & Dukes, K. D. 1977 Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America* **62**, 715-719.
[95] Nakatani, L. H. O'Connor, K. D. & Aston, C. H. 1981 Prosodic aspects of American English speech rhytthm. *Phonetica* **38**, 84-106.
[96] Levitt, A. G. 1991 Reiterant speech as a test of non-native speakers' mastery of the timing of French. *Journal of the Acoustical Society of America* **90**, 3008-3018.
[97] McGowan, R. W. & Levitt, A. G. 2011 A comparison of rhythm in English dialects and music. *Music Perception* **28**, 307-314.
[98] Cross, I. 2012 Music and biocultural evolution. In *The Cultural Study of Music: A Critical Introduction*. Clayton M, Herbert T, Middleton R (Eds.), 2nd ed, pp. 17-27. London: Routledge.
[99] Awad, M. Warren, J. E. Scott, S. K. Turkheimer, F., E & Wise, R., J.S. . 2007 A common system for the comprehension and production of narrative speech. *Journal of Neuroscience* **27**, 11455-11464.
[100] Tzourio, N. Crivello, F. Mellet, E. Nkanga-Ngila, B. & Mazoyer, B. 1998 Functional anatomy of dominance for speech comprehension in left handers vs right handers. *NeuroImage* **8**, 1-16.
[101] Xu, J. Kemeny, S. Park, G. Frattali, C. & Braun, A. 2005 Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage* **25**, 1002-1015.
[102] Davis, M. H. Ford, M. A. Kherif, F. & Johnsrude, I. S. 2011 Does semantic context benefit speech understanding through "top-down" processes? Evidence from time-resolved sparse fMRI. *Journal of Cognitive Neuroscience* **23**, 3914-3932.

[103] Alluri, V. Toiviainen, P. Lund, T. E. Wallentin, M. Vuust, P. Nandi, A. K. Ristaniemi, T. & Brattico, E. 2013 From Vivaldi to Beatles and back: Predicting lateralized brain responses to music. *NeuroImage* **83**, 627-636.
[104] Merrett, D. L. Peretz, I. & Wilson, S. J. 2013 Moderating variables of music training-induced neuroplasticity: a review and discussion. *Frontiers in Psychology* **4**, 1-8. (doi: 10.3389/fpsyg.2013.00606)
[105] Patel, A. D. 2010 *Music, Language and The Brain*. Oxford: Oxford University Press.
[106] Peretz, I. 2012 Music, language, and modularity in action. In *Language and Music as Cognitive Systems*. Rebuschat P, Rohrmeier M, Hawkins J, Cross I (Eds.) pp. 254-268. Oxford: Oxford University Press.
[107] Clayton, M. Sager, R. & Will, U. 2005 In time with the music: The concept of entrainment and its significance for ethnomusicology. *ESEM counterpoint* **1**, 1-82.
[108] Stephens, G. J. Silbert, L. J. & Hasson, U. 2010 Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the USA* **107**, 14425-14430.
[109] Loehr, J. D. & Palmer, C. 2011 Temporal coordination between performing musicians. *The Quarterly Journal of Experimental Psychology* **64**, 2153-2167. (10.1080/17470218.2011.603427)
[110] Hawkins, S. Cross, I. & Ogden, R. 2013 Communicative interaction in spontaneous music and speech. In *Language, Music and Interaction*. Orwin M, Howes C, Kempson R (Eds.) pp. 285-329. London: College Publications.
[111] Fraisse, P. 1982 Rhythm and tempo. In *The Psychology of Music*. Deutsch D (Ed.) pp. 149-180. London: Academic Press.
[112] Loehr, D. 2007 Aspects of rhythm in gesture and speech. *Gesture* **7**, 179-214. (10.1075/gest.7.2.04loe)
[113] Ogden, R. & Hawkins, S. 2014 Temporal co-ordination in conversation. Meeting of the British Association of Academic Phoneticians; Oxford.
[114] Hasty, C. F. 1997 *Meter as Rhythm*. Oxford: Oxford University Press.
[115] Widdess, R. 1994 Involving the performers in transcription and analysis: A collaborative approach to dhrupad. *Ethnomusicology* **38**, 59-79.
[116] Repp, B. H. 2005 Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin and Review* **12**, 969-992.
[117] Kotz, S. A. 2014 Timing and prediction in audition: From sound to speech. In *Communicative Rhythms in Brain and Behaviour*. Smith R, Rathcke T, Cummins F, Overy K, Scott S (Eds.). London: Philosophical Transactions of the Royal Society B.
[118] Cope, T. E. Grube, M. & Griffiths, T. D. 2012 Temporal predictions based on a gradual change in tempo. *Journal of the Acoustical Society of America* **131**, 4013-4022. (10.1121/1.3699266)

Figure 1

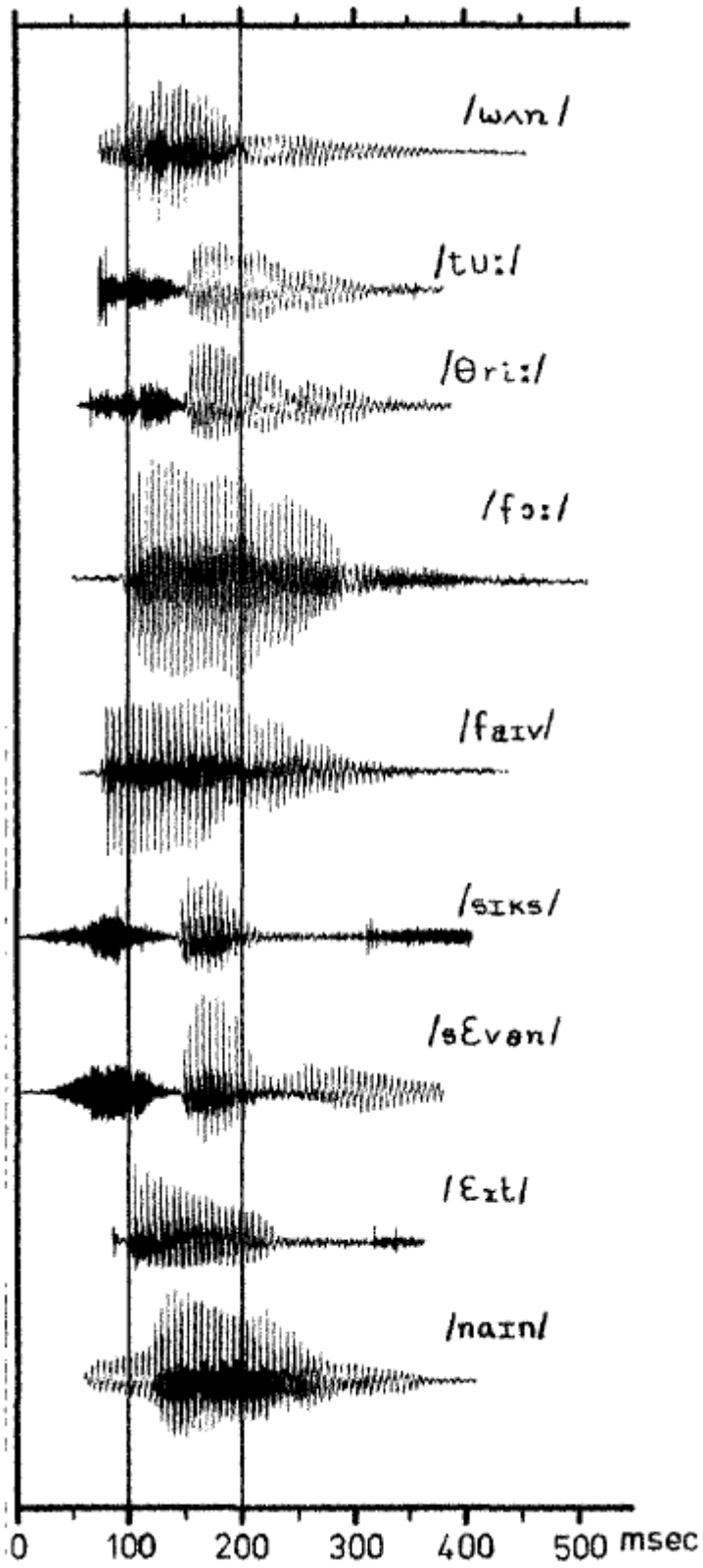

/wʌn/
/tu:/
/θri:/
/fɔ:/
/faɪv/
/sɪks/
/sɛvən/
/ɛɪt/
/naɪn/
0
100
200
300
400
500 msec

Figure 2.

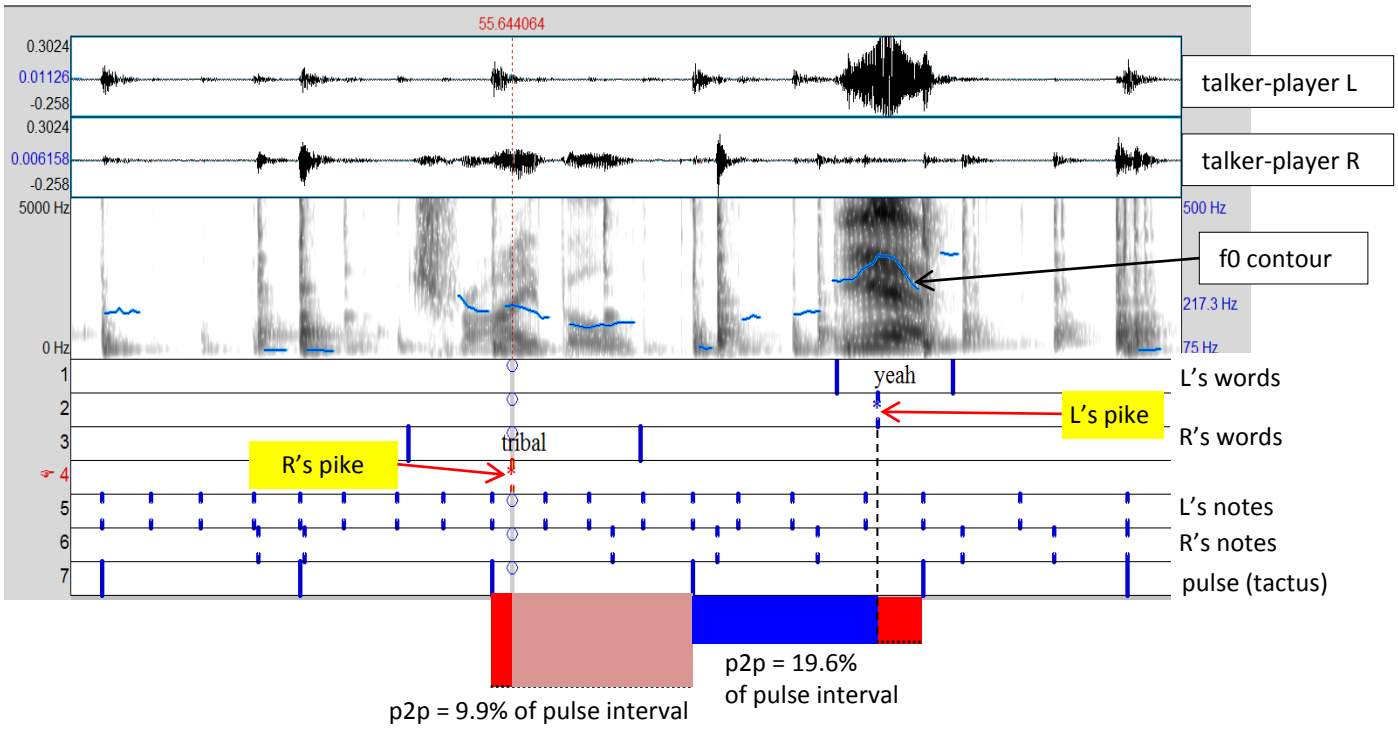
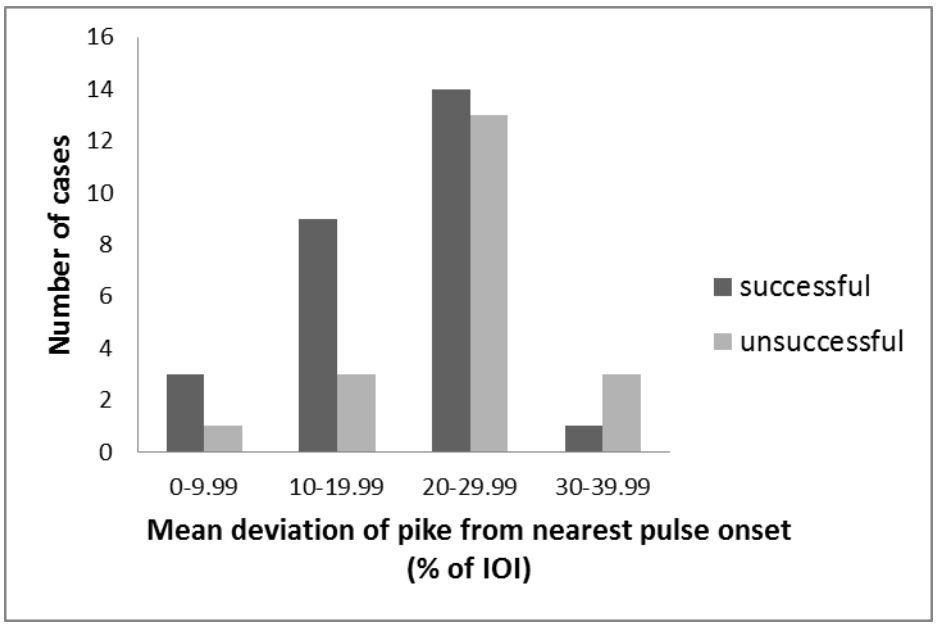

55.644064
0.3024
0.01126
-0.258
0.3024
0.006158
-0.258
5000 Hz
0 Hz
500 Hz
217.3 Hz
75 Hz
talker-player L
talker-player R
f0 contour
1
2
3
4
5
6
7
yeah
tribal
L's pike
R's pike
L's words
R's words
L's notes
R's notes
pulse (tactus)
p2p = 9.9% of pulse interval
p2p = 19.6%
of pulse interval

Figure 3


16
14
12
10
8
6
4
2
0
Number of cases
successful
unsuccessful
0-9.99
10-19.99
20-29.99
30-39.99
Mean deviation of pike from nearest pulse onset
(% of IOI)