Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 37 (2009) 16–28
Journal of Phonetics
www.elsevier.com/locate/phonetics

Rhythm as entrainment: The case of synchronous speech

Fred Cummins*

UCD School of Computer Science and Informatics, University College Dublin, Dublin 4, Ireland

Received 25 January 2008; received in revised form 21 August 2008; accepted 22 August 2008

Abstract

One view of rhythm, not conventionally adopted in speech research, is that it constitutes an affordance for movement. We test this view in an experimental situation in which speakers speak in synchrony with one another. After first establishing that speakers can synchronize with specific recordings, we present two experiments in which the information in the model speech is systematically reduced, allowing an evaluation of the respective roles of the amplitude envelope, the fundamental frequency and intelligibility in synchronization among speakers. Results demonstrate that synchronization is affected by several factors working together. The amplitude envelope, the pitch contour and the spectral qualities of the signal each contribute to synchronization. Intelligibility is not found to be absolutely necessary to support synchronization. This provides initial support for a dynamic account of synchronization among speakers based on the continuous flow of information between them.

1. Introduction

The empirical study of the phonetics of rhythm in speech has conventionally taken measurement of the speech signal as its proper domain. Rhythm has been interpreted as a property of the signal, and great effort has been spent in trying to arrive at signal-derived metrics that serve to index distinctions associated with rhythm. Thus languages of the world have been categorized and sorted based on the presumption that a 'rhythm-type' may be identified with a specific language (Dauer, 1983; Grabe & Low, 2002; Roach, 1982). This rhythm type is presumably deeply embedded in the phonology of the language, e.g. in the definition of prosodic domains such as metrical feet (Hayes, 1995; Liberman & Prince, 1977). Other empirical approaches have sought to distinguish among speech styles (Jassem, Hill, & Witten, 1984; Mixdorff, Pfitzinger, & Grauwinkel, 2005). The entrenched pursuit of a meaningful distinction between syllable-timed speech and stress-timed speech has its origins, not in the categorization of languages, but in differences in speaking style within an individual, that were first described as a 'machine-gun' and a 'morse code' speaking style, respectively (Lloyd James, 1940). Underlying both approaches is the assumption that rhythm is essentially about the regular recurrence of some event delimiting isochronous time intervals (Lehiste, 1977). These intervals are typically assumed to be demarcated by perceptual beats, or P-centers (Scott, 1993), and the acoustic information specifying the P-center is presumed to be the amplitude modulation of the signal (Cummins & Port, 1998). For example, Ramus and Mehler (1999) demonstrated that English and Japanese could be discriminated based on listening to resynthesized stimuli in which all consonants were replaced by /s/, all vowels by /a/, and a flat intonation contour was applied. The authors claim that these stimuli contain information about syllabic rhythm and nothing more, thereby revealing the common assumption that rhythm is best described with reference to the amplitude modulation of the speech signal.

There is, however, a primary sense of the term 'rhythm' that is not captured by these approaches. To most non-specialists, rhythm is intimately connected with music and dance. It is the kind of thing that allows one to move with the music. If there is dancing, there is rhythm. If the music does not support dancing, tapping, nodding, or the like, it is arrhythmic. When used in this sense, rhythm appears as a relationship between an acoustic signal and the potential for movement of an embodied listener. It is an *affordance*. The notion of an affordance is most closely associated with

*Tel.: +353 1 716 2902; fax: +353 1 269 7262.
E-mail address: fred.cummins@ucd.ie

0095-4470/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2008.08.003

the work of Gibson and the school of ecological psychology, though the general recognition of the relevance of the functional significance of environmental features is much older (Gibson, 1979; Heft, 2003). In essence, an affordance is a property of the environment that is of relevance to the movement potential of an organism. A central example of the concept of affordance is the 'climb-ability' of a set of stairs, which does not inhere in the physical properties of the stairs alone, but rather in the relation between the physical property of riser height, and the scale, or leg-size, of an organism. Stairs that are perfectly climb-able for one person may not be so for another if she is of a very different size. In dancing, clapping, toe-tapping, etc., there is a coupling of the actions of the listener to the structural properties of the sound, and it is precisely in those cases in which the signal allows the entrainment of movement of a listener that we call the signal rhythmic. For a fuller account of the concept of affordance, the reader is referred to Chemero (2003) and Stoffregen (2003).

Interestingly, the discussion in the previous paragraph did not need to mention recurrent interval structure, or isochrony, in order to describe this core sense of the word 'rhythm'. It is certainly the case that there are recurring intervals in music, and that listeners are sensitive to this recurrence, such that the movement of a listener may exhibit a similar temporal patterning. But this does not license the inference that the synchronization of movement with sound is based *only* on temporal intervals of equal length. Dramatic tension in a melody may lead to tempo change that an insightful listener can still synchronize with, although temporal intervals are non-constant. Likewise, anyone familiar with musical performance expects a degree of conventional ritardando towards the end of a phrase, without having a sense of the destruction of rhythm. Careful analysis of expressive timing has revealed a great deal of systematicity to deviations from the nominal intervals suggested by written musical notation (Repp, 1996). Rhythm, in other words, is not necessarily about isochrony, but may more accurately describe the relationship between a sound and the movement of a listener.

But does speech exhibit any rhythm whatsoever under this embodied interpretation? There are some cases that appear to constitute the overt entrainment of movement by speech, though they leave open the possibility that speech may be a less effective stimulus for entrainment of movement than music. A speaker gesticulates while speaking, demonstrating a self-entrainment of movement to speech (Port, Tajima, & Cummins, 1996). Effective public speakers carefully time their speech to maximally engage an audience, which often manifests as physical entrainment, as at rallies or more prosaically in nodding ones head along with the speaker (Streeck, 1994).

More subtle evidence of the entrainment of the movement of a listener to the ongoing stream of speech is found in the demonstration of synchrony between new born infants and the speech of their adult caretakers in an early study by Condon and Sander (1974). While suggestive, the methods of this study are problematic in many respects, and there has been a marked absence of more recent follow-on work. A more recent demonstration of the subtle entrainment of the movements of conversational participants is found in work on postural sway by Shockley, Santana, and Fowler (2003).[1] They applied the techniques of embedding to recreate the phase space of sway movement of subjects who were either conversing with each other or with a third party. They found significant coordination of the paired movement traces only when the subjects conversed among themselves. The coordination observed was not directly linked to the speech signal itself in this analysis. While these examples suggest that speech may entrain movement, they also suggest that rhythm, understood as an affordance for movement, may be less effective in speech than in music. Two observations appear apt to draw out the nature of the embodied view of rhythm suggested here.

Firstly, speech movements are relatively small compared to gross limb movements employed in most rhythmic activities. The articulators are small; the energetic constraints on their movement appear to be less important in the determination of the form of their movement than in gross limb movement (Ostry & Munhall, 1985), and the movement is largely hidden from view, precluding a strong cross-modal reinforcement of any entrainment between a speaker and a listener. These considerations might suggest that speech would be relatively ineffective in acting as a stimulus for limb movement, as there is a great deal of disparity between the physical properties of the organs of speech and the limbs.

On the other hand, it has been shown on occasion that rhythmic principles that are operative in the organization of limb movements into rhythmic patterns, may, under suitably controlled circumstances, be shown to be operative in speech as well. The pioneering work of Stetson (1951) demonstrated that there are two potential forms of syllable organization for a continuous stream of alternating vowels and consonants at most rates, e.g. in the sequences /ip.ip/ and /pi.pi/, while at faster rates, there is only one such form of organization.[2] The existence of two forms of stable organization at moderate rates with a transition to a single stable form at fast rates closely parallels results found in studying the forms of rhythmic coordination that are possible when two digits or limbs are made to oscillate at a common frequency (Kelso, 1995). The qualitative similarities extend to such phenomena as hysteresis, critical fluctuation before the bifurcation, etc., and they suggest strongly that similar organizational principles are operative, and that they are best described at a level that is

[1] Thanks to an anonymous reviewer for directing my attention to this work.

[2] The syllable structure observed at fast rates was reported in Stetson (1951) as /pi.pi/, though de Jong has demonstrated that this single faster form differs somewhat from the /pi.pi./ structures found at moderate rates (de Jong, 2001).

sufficiently abstract as to apply to such different effectors as limbs and the vocal tract.

In a similar vein, Cummins and Port (1998) demonstrated that when a short phrase is repeated in time with a metronome, there are a small number of discrete temporal structures that result, and these correspond to a hierarchical nesting of stress feet within the overall phrase repetition cycle. The existence of a small number of discrete coordinative patterns again resembles the limitations on the coordination of the hands or limbs in repetitive tasks.

Both of these examples of 'embodied rhythm' in speech are critically based on repetition and isochrony, and thus they cannot inform us about any putative role for entrainment of action beyond regular recurrence. Spontaneous speech, on the other hand, rarely presents any significant degree of recurrent, periodic, temporal structure. In common with musical timing, however, speech does exhibit the property of a ritardando or local slowing down at the end of major prosodic units (Byrd & Saltzman, 2003). Indeed, the notion that the organizational principles underlying sequencing of gestures in speech and sequencing of movements in the limbs has informed the entire project of articulatory phonology and its task dynamic implementation (Goldstein, Byrd, & Saltzman, 2006; Saltzman & Munhall, 1989).

In the present work, we consider an experimental setting in which the movements of a speaker are coordinated with an external source, but rather than a metronome, or periodic source, we study the synchronization among two speakers reading a text simultaneously, with the instruction to remain in synchrony with one another. In this context, the speech of one speaker acts as the entraining signal for the production of the other in a symmetrical, reciprocal relationship. This is, in many respects, an artificial task. While there are situations in which people speak synchronously, such as prayer repetition, reciting oaths, etc., these are usually highly conventionalized settings and the prosody employed is normally quite stylized. In the synchronous speech setting, we explore the ability of competent speakers to entrain in a somewhat unusual fashion. As native speakers, however, subjects clearly are highly skilled at coordinating their own articulators. We can use the experimental vehicle of synchronous speech to see to what extent these coordinative skills can support a yoking of two distinct production systems.

2. Synchronous speech

In the simplest form of the synchronous speech task, two subjects read a prepared text in synchrony (Cummins, 2003; Krivokapić, 2007). After reviewing the text to be read, the experimenter provides an unambiguous start signal, and the two subjects proceed to read the text, while maintaining synchrony as far as possible. Each speaker can see and hear the other at all times. It has been demonstrated that this task is well within the capabilities of naïve subjects who are competent speakers of a language, and that the asynchrony observed even without practice is smaller than might be expected based on the variability found within and across speakers in other situations. Typical asynchronies reported are about 40 ms, with a slight increase to about 60 ms at phrase onsets (Cummins, 2002). Crystal and House provide estimates of segmental variability in normal read speech that range from 9 ms for a flap at fast rate, to 70 ms for a diphthong at slow rate (Crystal & House, 1982). Given that even a small paragraph as used here will string hundreds of segments together, this sustained synchrony demands some explanation. Even more surprisingly, speakers can perform this task without extensive practice, and it has been shown that practice does not substantially improve their performance (Cummins, 2003). In other words, this appears to be a relatively easy and natural task for subjects, they are very good at it, and they can do it without practice. It is noteworthy that in the very many recordings we have observed to date, it has never once been the case that one speaker was consistently leading the other speaker. Rather, the (very small) lead changes throughout the speaking situation, suggesting that there is no clear leader–follower relationship.

Where two processes are entrained, that entrainment must necessarily be based on some exchange among the processes, allowing the dynamics of one to influence the other. The first recorded example of entrainment was noted by Christian Huygens in the phase-locked oscillation of the pendula of two clocks hung on the same wall (Spoor & Swift, 2000). In this case, the basis for entrainment was clearly a mechanical linkage between the two systems, as the coordination went away when the clocks were hung on different walls. Entrainment among the limbs has been well studied within an individual, as e.g. in the study of gait or finger movement (Kelso, 1995). An experiment by Schmidt, Carello, and Turvey (1990) demonstrated that constraints on the stable coordination of oscillating limbs hold even when each limb belongs to a different person, and the only basis for maintaining an inter-person coordination is visual information. A dynamical account of the coordination observed in a collective task thus requires an understanding of the information exchanged between participants.

In this paper, we examine the relationship between the information present in the speech signal and the resulting synchrony among speakers. From previous work, it is known that two speakers who are physically present in the same room can read a prepared passage of text and maintain a high degree of synchrony. In order to extend the experimental investigation of the basis for such performance, it is first necessary to establish whether speakers can establish a comparable degree of synchrony when speaking along with a recording of another speaker. Using a recording clearly alters the dynamics of the situation somewhat, preventing a mutual adjustment among the speakers. However, if synchrony can be achieved under these circumstances, it will then be possible to examine the

role of specific kinds of information in the speech signal in the process of synchronization. This can be achieved by selectively altering the recorded speech (e.g. by removing pitch information), and quantifying the degree of asynchrony that results in attempting to speak along with the altered signal. By manipulating the recordings to which subjects are attempting to entrain, it may be possible to shed some light on the kind of information that supports the entrainment of movement, and thus contribute to a physical understanding of rhythm in an embodied sense.

The remainder of the paper is structured as follows: A method for quantifying the degree of asynchrony in two parallel recordings is briefly described, and full details are provided in Appendix A. An initial experiment is required in order to ascertain whether subjects can synchronize with recordings. This experiment serves also to identify a subset of recordings that are relatively easy for subjects to synchronize with. These recordings are used as stimuli in the two principal experiments that follow. In Experiment 2, the stimuli are altered in some straightforward ways and synchronization performance is measured. Results suggest further modification to stimuli that might be informative, and these are applied in Experiment 3. The discussion section then integrates the findings of these two experiments and returns to the topic of rhythm as an affordance for the entrainment of movement.

2.1. Measuring asynchrony in parallel recordings

Previous estimates of asynchrony of two parallel speakers were based on the times of clearly identifiable points in the waveforms (vowel onsets, stop releases, etc.). These are necessarily irregularly distributed and sparse in time. While they served to establish estimates of mean asynchrony, they are poorly suited to more rigorous quantitative study as required here. For the present purposes, a measurement technique is required that aligns the two utterances in a continuous fashion, providing a quantification of the degree of stretching or compression required to map or warp one utterance onto the other. Dynamic time warping (DTW) was used to assess the degree to which two utterances were well aligned, and hence to arrive at a quantitative estimate of asynchrony between the two. Full details of the procedure are provided in Appendix A.

In order to provide a sense of scale, we estimated asynchrony for three sets of utterances, as shown in Fig. 1. On the left, are asynchrony estimates for matched utterances where both speakers were live and attempting to synchronize with each other. This represents asynchrony found under optimal conditions for synchronization. On the right are data from a limiting case which forms part of Experiment 3. In this particular condition, speakers heard a sequence of six stimuli corresponding to the six phrases that constitute the Rainbow Text. Each phrase was introduced by three isochronous beeps 0.5 s apart, and after another 0.5 s the phrase started. In this manner, phrase onsets were entirely predictable. The first phrase

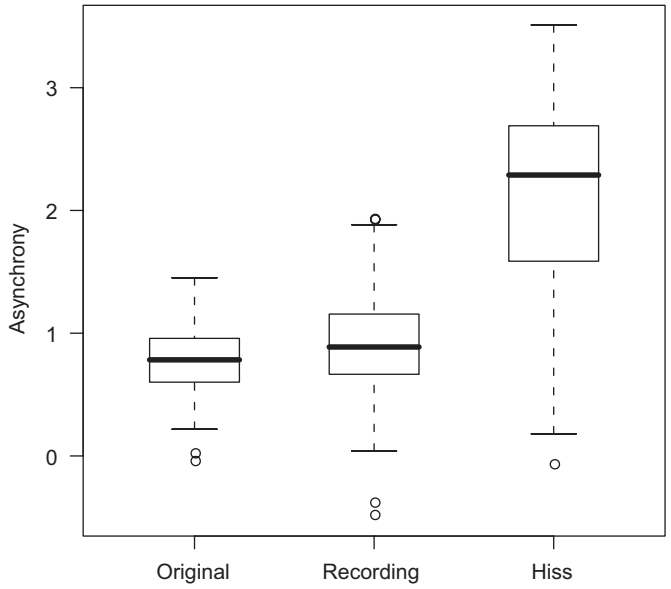

Fig. 1. Distribution of asynchrony estimates for three conditions. Left: live synchronous speech ($n = 72$). Center: synchronization with a recording of synchronous speech ($n = 240$). Right: reading along with an uninformative hiss with duration matched to a model utterance. Asynchrony is based on comparison of the speaker's utterance with the (unheard) model utterance ($n = 160$).

was unmodified speech, so that the speaker could attune to the speaking rate of the model speaker. The remaining five phrases were replaced with an unmodulated hiss, and thus provided no basis for synchronization whatsoever. The duration of the hiss matched that of the original speech phrases they replaced. Subjects were clearly getting no continuous information in this condition. We scored their productions against the five time-aligned original phrases (i.e. speech, and not hiss). The distribution of asynchrony estimates thus represents performance where the approximate speaking rate is known, phrase onsets are perfectly predictable in time, but no other information is available.

In the middle, for comparison, are asynchrony measurements for speakers in Experiment 1 who were attempting to synchronize with a recorded model utterance. It can be seen that their performance is much more similar to the live case than to the control condition.

3. Experiment 1: can speakers synchronize with a recording?

As the present experimental goals require speakers to synchronize with degraded speech, it is necessary to first determine whether speakers can synchronize at all with recordings, and if so, which recordings are best suited to this end. In previous work (Cummins, 2002), it was found that speakers seemed to be able to synchronize with a recording of a text, and that for two of the three speakers, synchrony was improved if the recording itself had originally been obtained in a synchronous condition; that is, if the recording was of synchronous speech. Estimates of asynchrony were approximate, however. The present

method allows a more reliable quantification of asynchrony.

3.1. Experiment 1: methods

Subjects attempted to synchronize with recordings of normal, unconstrained speech, and with recordings of synchronous speech. A corpus of recordings of 36 speakers was available (Cummins, Grimaldi, Leonard, & Simko, 2006). This corpus includes high quality recordings of speakers reading the first paragraph of the Rainbow Text (see Appendix B) both in an unconstrained fashion alone, and in synchrony with a co-speaker. In the latter case, the two speakers were recorded onto separate channels. From this corpus, 12 speakers (6m, 6f) of Hiberno-English were chosen based on an informal appraisal of fluency and naturalness. This provided 12 solo recordings and 12 (single channel) synchronous recordings. The recordings were modified so that a series of three isochronous beeps at 0.5 s intervals preceded the start of each of the six sentences of the text, ensuring that each sentence onset was maximally predictable.

The 24 recordings were played in random order to four subjects (2m, 2f, Hiberno-English speakers), who were instructed to synchronize as well as possible with the recording. Subjects were familiar with the text before recording began. The text was displayed on a computer screen with each of the six individual phrases appearing on a separate line. Subjects listened to the recordings through headphones, and spoke along with the recording into a head-mounted near-field microphone. Their own speech was likewise routed to the headphones, so that subjects actually heard the recording in one ear and their own production in the other. This dichotic presentation had been found to facilitate synchronization in pilot work.

3.2. Experiment 1: results

Asynchrony scores were obtained automatically for each reading using the procedure described in Appendix A.

Fig. 2 (left panel) shows the quantitative estimate of asynchrony, in units derived from the warp path, for each of the four subjects when synchronizing with the model recordings obtained in a synchronous speaking condition. Each box plot contains asynchrony estimates from 5 phrases spoken with 12 models. On the extreme left, for comparison, is the asynchrony obtained in the original recording situation, when the model speakers employed here functioned as target speakers for live co-speakers. Individual data points are for single phrases, and the estimate of asynchrony is normalized using the number of frames in a phrase, to allow comparison of asynchrony measures across phrases of different length.

Mann–Whitney tests comparing asynchrony of each of the four subjects with asynchrony from the original recording session showed that the degree of synchrony is affected by synchronizing with a recording (all $p<0.01$). Although synchrony is somewhat reduced, it will become clear that the effect size is very small compared with that observed in subsequent experiments.

Fig. 2 (right panel) compares synchronization performance when the recording itself is either normal speech or synchronous speech. For three of the four subjects, synchronization was better when the recording itself was recorded in a synchronous condition (Wilcoxon paired signed rank test: all $p<0.01$, except for m1, n.s.).

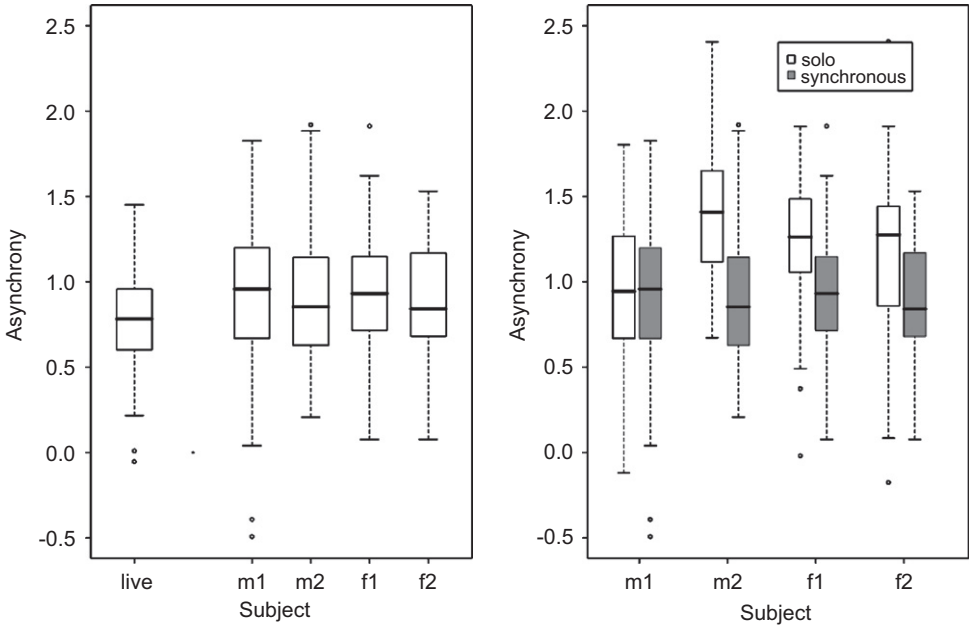

Fig. 2. Left: Asynchrony for four subjects synchronizing with a recording. On the extreme left, the asynchrony originally observed while making the recording (with different co-speakers) is shown. Right: asynchrony for four subjects when the original recording is either normal unconstrained speech ('solo'), or synchronous speech.

(Note that the gray boxes in the right panel show the same data as the group of four boxes in the left panel.)

This first experiment served to establish that synchronization performance is not greatly affected when appropriate recordings are used as models. As a side effect, it allowed the identification of model recordings that were easiest for subjects to synchronize with. Based on these initial results, the four recorded speakers with whom subjects exhibited the best synchronization were selected as models for subsequent experiments. In all following experiments, these synchronous recordings were used as model recorded utterances, thus ensuring that synchronization was facilitated as far as possible.

4. Experiment 2: synchronizing with reduced stimuli

In this and in the subsequent experiment, recordings were altered using a variety of techniques in order to assess the relative importance of the remaining information in the signal in supporting synchronization among speakers. In each case, subjects were recorded as they tried to synchronize with the altered recordings. To assess their performance, their production was aligned with the *original*, unaltered, recording, ensuring that the onset of the altered phrase as heard, and the corresponding phrase in the original recording, were exactly aligned in time. Asynchrony was then computed as described in Appendix A.

Previous work on the perception of rhythmic beats has pointed to the importance of amplitude envelope modulation in perceiving rhythm in speech (Morton, Marcus, & Frankish, 1976; Scott, 1993). In order to evaluate this, we employed three different conditions, each of which degraded the speech signal somewhat, while leaving very low frequency modulation more or less intact.

It is also known that the fundamental frequency contour is critically aligned in time with the segmental and syllabic content of an utterance (Bruce, 1990; Pierrehumbert & Steele, 1989). It may also be the case that pitch perception contributes to the perception of rhythmicity, though perceived rhythmicity is difficult to test. Certainly, stress has been consistently implicated in the perception of rhythmicity (Dauer, 1983) and F0 is known to be a major correlate of perceived stress (Beckman & Edwards, 1994). One obvious component of the speech signal to manipulate is thus the fundamental frequency (condition MONO).

4.1. Experiment 2: methods

Four subjects (3m, 1f, Hiberno-English speakers) listened to modified stimuli in four experimental conditions and in an unmodified control condition. All subjects were new to the synchronization task. Stimuli were presented in random order. They listened to the model recordings through headphones as before, and attempted to synchronize with what they heard. Their instructions asked them to 'stay in time with the other speaker', and noted that this might be difficult, but they should do their best to 'speak in time' with what they heard. Recordings were made in a quiet office environment using near-field head-mounted microphones. Asynchrony was evaluated over the final five sentences of the paragraph, by aligning the subject's recording with the *original* recording and estimating the optimal warp path as before.

Model stimuli were prepared based on the synchronous recordings of the four speakers (2m, 2f) to whom subjects could best synchronize in Experiment 1. As before, each of the four recordings was a reading of the full paragraph reproduced in Appendix B. In each case, the first sentence was left unaltered, so that subjects could attune to the subject's speaking rate.

In a first condition, we resynthesized the utterances with a constant F0 of 100 Hz (condition MONO). This lets us selectively evaluate the relative importance of F0 in synchronization. Although both male and female recordings were resynthesized with a fixed F0 of 100 Hz, the sex of the original speaker was still clearly recognizable, as males and females have systematic differences in formant structure as well as F0.

In a second condition the speech was low pass filtered with a cut-off at 500 Hz (LPF). This speech, although radically altered, remains intelligible.

In a third, signal correlated noise was generated by probabilistically flipping the sign on each sample with a probability of 0.5. This manipulation preserves the amplitude envelope of the original, but renders the speech entirely unintelligible (SCN). This latter condition allows testing of the importance of the amplitude envelope alone in synchronization. Finally, the SCN stimuli were altered to exaggerate the intensity modulation by down sampling to 16 kHz, low pass filtering with a 4 kHz cut off, and using Praat's 'deepen band modulation' function to enhance the modulation of the envelope (Boersma & Weenink, 2005). The resultant stimuli are of course still unintelligible, but it is possible that enhancing the envelope modulation might provide a useful cue for synchronization. This condition is labelled BAND. Samples of all stimuli used in this study are available at http://tinyurl.com/4l5xk2.

4.2. Experiment 2: results

Fig. 3 shows asynchrony produced by each of the four subjects. There is considerable variability across subjects in their ability to synchronize with these recordings. In particular, Subject m3 does not show significant increase in asynchrony, despite the severe modification of the stimulus. In general, ORIG and MONO produced comparable degrees of asynchrony, while LPF was somewhat harder to synchronize with and SCN and BAND were considerably harder. A repeated measures analysis of variance with condition and co-speaker as factors and employing the Geisser–Greenhouse correction to degrees of freedom showed a main effect of condition ($F(1,94) = 51$, $p < 0.001$), while co-speaker and the interaction were not

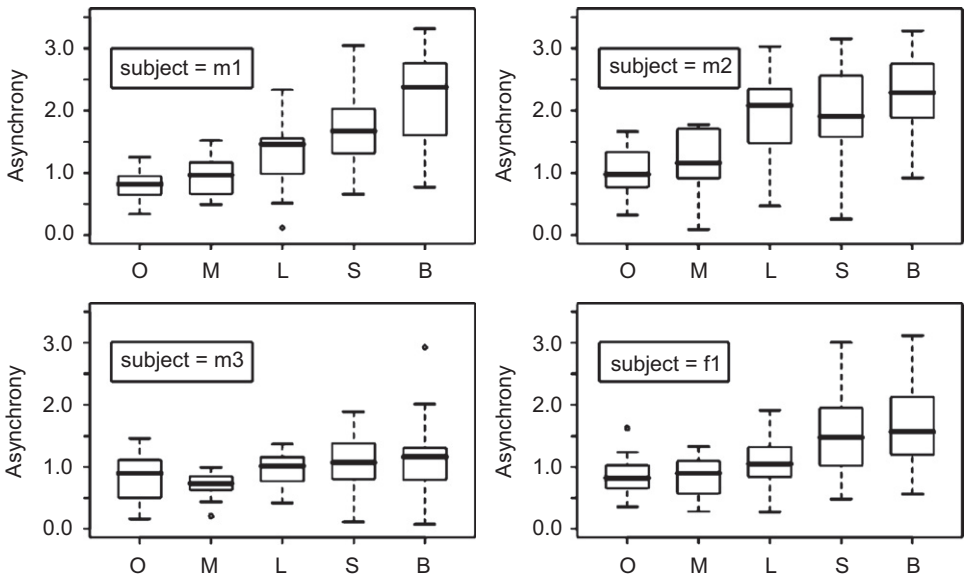

Fig. 3. Asynchrony as a function of condition for four speakers. Conditions: 'O': original, unmodified stimuli (ORIG), 'M': MONO; 'L': LPF; 'S': SCN; 'B': BAND.

significant. Post hoc tests were done using paired *t*-tests, with Bonferroni correction of the alpha level to protect the family-wise error rate. All pairwise comparisons were significantly different from one another with a family-wise alpha of 0.05, except for the comparison between ORIG and MONO, which were not significantly different.

From these results, it appears as if fundamental frequency information is not necessary in order to synchronize with speech. No individual subject was markedly worse in the MONO condition, and the overall ORIG–MONO comparison was not significant. On the other hand, the amplitude envelope alone in either the SCN or the BAND condition does not seem to have been sufficient for synchronization for three of the four subjects. The LPF condition did not generate consistent results across this small subject set.

These initial results pose several questions. Firstly, the naive assumption that F0 information might be crucial for synchronization, and the alternative assumption that synchronization might depend only on the amplitude envelope were both found wanting. This topic is revisited in the next experiment.

One of the four subjects appeared to be able to remain in synchrony with the models irrespective of the amount of signal degradation. It is worth noting that the four model speakers displayed considerable temporal variation among themselves. While three of them has almost identical articulation rates, the fourth (model speaker M_2) spoke at a considerably faster rate (5.96 syll/s compared with 5.06, 5.03 and 5.06 for model speakers M_1, F_1 and F_2, respectively). There is thus some manifest variation in the ability of subjects to exploit the information provided for synchronization. The richest information is present during the first, unaltered, phrase. From this, it may be possible to extract sufficient speaking rate information to ensure reasonable synchronization in subsequent phrases, as subject m3 appears to be doing.

Two of the altered stimulus forms focussed on the macroscopic amplitude envelope variation (SCN and BAND). There are two potential weaknesses in the method used to construct these stimuli. Firstly, the use of signal correlated noise for both of these stimulus types produces signals that are very harsh sounding. It is possible to impose an amplitude envelope on carrier signals with spectral characteristics that are less unpleasant than white noise. It is also notable that many studies of speech rhythm have identified a restricted frequency range as potentially containing the amplitude envelope information responsible for the perception of rhythm in speech. Scott (1993) and Cummins and Port (1998) both have focussed on the amplitude envelope in the approximate range of 500–1500 Hz, that seems to best predict the location of P-centers, or beats, in speech.

With these observations in mind, we conducted a follow-up experiment with some novel forms of signal degradation.

5. Experiment 3: uncovering the roles of amplitude and frequency in synchronization

5.1. Experiment 3: methods

Four new stimulus types were prepared. In a first, control, condition (HISS) phrases 2–6 of the paragraph were replaced by a hiss (white noise, band-pass filtered with cut offs at 500 and 1500 Hz, normalized to 70 dB, constant amplitude). As before, the first phrase was left unaltered, and each phrase was preceded by three introductory tones to ensure that phrase onset was predictable. Subjects thus have available to them a rough measure of the model

speaker's global speaking rate from the first phrase, and the exact time of onset of each of the following phrases. If their productions were then found to be well aligned with the unaltered model phrases, that would strongly suggest that the continuous exchange of information between speakers is not required to support synchronization.

A second stimulus set (BP-SCN) was prepared in similar fashion to the SCN stimuli of the previous experiment, but the speech was first band-pass filtered, excluding frequencies below 500 Hz and above 1500 Hz. The resulting stimuli were low pass filtered with a ceiling of 2000 Hz.

A further set of stimuli was constructed using a modulation signal derived from the amplitude envelope of the band-pass filtered speech signal, but employing a different, vowel-like carrier (VOWEL). For this, a single pitch period from a sustained vowel spoken by a male was excised and repeated to provide a continuous vowel-like carrier. This was then modulated using the amplitude envelope of the band-pass filtered speech signal (500–1500 Hz). The resultant signal had a constant F0 of 111 Hz.

Finally, the stimuli of the VOWEL condition were resynthesized with pitch contours extracted from the original recordings (F0-VOWEL). The resynthesized stimuli thus had the band-pass filtered amplitude envelope and pitch information of the original, but no further information about phonetic content.

Eight subjects (5m, 3f) from Eastern Ireland participated. No subjects had taken part in either of the previous experiments. Each subject first read the paragraph alone, providing an estimate of their preferred, unforced, reading rate. They then listened to each of the model speakers, and attempted to synchronize with them (ORIG1). In a subsequent block, the four altered stimulus types were presented, together with a repeat of the unaltered stimulus (ORIG2), yielding a block of 20 trials, presented in random order. Recording methods and conditions were otherwise exactly as in the previous two experiments.

5.2. Experiment 3: results

Fig. 4 shows asynchrony scores for the two unaltered readings, and the four altered conditions. Clearly, synchronization is adversely affected by the removal of all segmental information, in which speech is rendered completely unintelligible. However, there do appear to be differences between the four conditions of interest. A repeated measures ANOVA with experimental condition and model speaker as factors, with degrees of freedom adjusted using the conservative Geisser and Greenhouse correction for non-sphericity, shows main effects of both condition $[F(1,186) = 126,\ p < 0.001]$ and model $[F(1,310) = 6.8,\ p < 0.01]$, and no interaction. Cursory examination of the difference in performance when synchronizing with the different model speakers shows that speakers were slightly more successful at synchronizing with model speaker M_1, with no apparent differences in performance for the other three models (note: model speaker M_2, not M_1, was the speaker with the relatively fast articulation rate).

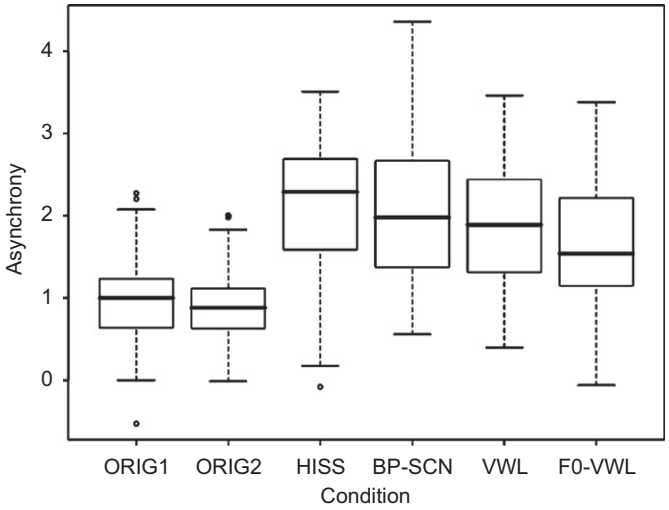

Fig. 4. Experiment 3: Asynchrony as a function of condition for four speakers.

Three planned comparisons were foreseen. Firstly, as HISS served as a control condition, we were interested in whether the least promising of our altered stimuli (BP-SCN) was at all better than the HISS stimulus, that has no information beyond phrase duration. Secondly, VOWEL and BP-SCN differ only in the spectral quality of the amplitude-modulated carrier. A difference in performance here would point to the importance for synchronization of spectral information other than the slow amplitude modulation of the signal. Finally, VOWEL and F0-VOWEL differ only in the addition of a pitch contour. Although the previous experiment had demonstrated that F0 information was not necessary for synchronization, it is possible that it might facilitate synchronization, especially with degraded stimuli as used here. For all three planned comparisons, matched t-tests were conducted. As the number of planned comparisons is small, no correction of alpha was made, but the statistical significance observed holds at the 0.05 level even when the conservative Bonferroni correction is applied.

The comparison of HISS and BP-SCN did not reveal a significant difference $[t(159) = 0.76,$ n.s.$]$. Changing the carrier did make a difference, however, as synchronization with the VOWEL stimuli was better than with the BP-SCN stimuli $[t(159) = 2.66,\ p < 0.01]$. Finally, the addition of the fundamental frequency contour further improved synchronization performance significantly $[t(159) = 3.56,\ p < 0.001]$.

6. Discussion

In this series of experiments, we have demonstrated that speakers can synchronize with a suitable recording (albeit not as well as with a live co-speaker), and that synchronization is dependent in complex fashion on a variety of kinds of information in the speech signal.

It was not self-evident that speakers would be able to synchronize with a recording. The synchronous speaking condition typically demands accommodation by both speakers to the joint task. In Experiment 1 it was found that synchronization with a recording was facilitated when that recording itself bore the hallmarks of accommodation, by virtue of being recorded in a synchronous situation, albeit with a different speaker. Even with the best exemplars, synchronization performance in Experiment 1 was not quite as good as that typically obtained when two speakers are live. The synchrony obtained was, however, sufficiently precise, such that a differentiation among the various conditions in Experiments 2 and 3 was possible, with performance ranging along a continuum from unaltered speech (best) to reading along with an unmodulated hiss (worst).

Two conditions in Experiment 2 employed speech that was still intelligible (MONO and LPF). Of these, monotonous speech was not significantly worse that unmodified speech, while low pass filtered speech was somewhat worse. Intelligibility alone does not appear to be sufficient to explain synchronization performance, as there was marked differences among the conditions in Experiment 3, even though all modified stimuli were unintelligible, as all segmental information was absent.

The various ways in which the stimulus was modified allows for a series of comparisons that collectively rule out an overly simplistic account of the dependence of synchronization performance upon either the fundamental frequency or the amplitude envelope modulation. Fundamental frequency was not essential to good synchronization as shown in Experiment 2. Nonetheless, the restoration of the intonation pattern that differentiates conditions VOWEL and F0-VOWEL in Experiment 3 demonstrates that under some conditions, pitch variation may contribute substantially to synchronization.

Several conditions were relevant to the issue of the role of amplitude envelope modulation. In Experiment 2, the use of signal correlated noise, which maintains the amplitude envelope of the full-spectrum signal, was thoroughly ineffective as a stimulus for synchronization. Although there is limited support in the literature for the importance of amplitude modulation within a restricted frequency range of approximately 500–1500 Hz in the perception of timing (Scott, 1993), no improvement was found in the present experiment by using signal correlated noise derived from a band-pass filtered signal (condition BP-SCN). Synchronization with BP-SCN stimuli was not any better than synchronization observed when the stimulus was reduced to a maximally uninformative hiss. The same amplitude envelope did prove more useful, however, when the signal being modulated was more speech like, in being derived from a sustained vowel (BP-SCN vs. VOWEL). The carrier signal here contains no information whatsoever about timing in the original, and so it must be surmised that processing of the envelope modulation is dependent on the speech-like nature of the carrier. This interpretation receives further support from the finding that additional improvement can be obtained by superimposing the fundamental frequency contour onto the carrier (VOWEL vs. F0-VOWEL).

Together, these results suggest that synchronization among speakers is facilitated both by intelligibility, and by specific information within the signal, some of which may be processed in a speech-specific manner. The speech signal is tremendously rich. The few stimuli employed here do not begin to exhaust the possibilities for information transfer during synchronization. But they do serve to caution that the role of the amplitude envelope, that has frequently been supposed to be a principal carrier of macroscopic, rhythmic, information in the signal may be somewhat overstated. There is a complex interplay between amplitude, fundamental frequency and spectral characteristics that remains to be further clarified.

One further question not addressed by the present study is whether information used in synchronization is continuously distributed throughout the signal, or whether instead some portions of the signal are more important than others. A great deal of attention has been paid to the importance of prominent syllable onsets in rhythmic perception, and syllable onset information was preserved in all conditions except HISS, in which only phrase onsets were predictable. Given the range of performance exhibited across the remainder of the conditions, it seems that syllable onsets alone are not sufficient. F0 peak information is another candidate for punctate information that subjects might exploit for synchronization. It is possible that the combination of F0 and onset information seen in the F0-VOWEL condition allows the identification of phrasal accents. This alone cannot explain why a modulated sustained vowel (VOWEL) should offer a better basis for synchronization than modulated noise. Collectively, then, a combination of envelope modulation, F0 and long-term spectral properties are implicated in facilitating synchronization among speakers.

The utility of continuous physical information, in the absence of segmental information, is consistent with an entrainment model of synchronization, and it raises the question of why this might be. Why would speech, that is typically spoken without a simultaneous counterpart, facilitate entrainment? We noted above that overt cases of entrainment by speech, as at rallies, are relatively rare, and the link between the movement elicited and the speech may be more tenuous than an entrainment account would suggest. Furthermore, although there are impressionistic accounts of tight temporal coordination across speakers in turn-taking (Couper-Kuhlen, 1993), these have not been convincingly backed up by quantitative studies (Bull, 1997).

A first argument, well known from the literature, is the idea that neural processes underlying the production and perception of speech may be similar, or, as it is often stated, perception and production of speech may employ common representations (e.g. Liberman & Mattingly, 1985). This

idea has several considerations in its favor: cognitive economy suggests that employing common representations may obviate the need for developing and maintaining two entirely different, highly complex, systems. More tellingly, this unified account does not require a hypothetical 'translation process' whereby motor and linguistic units of incommensurable composition are mapped onto each other (Fowler, Rubin, Remez, & Turvey, 1981; Goldstein & Fowler, 2003). By recognizing that action and perception are intimately linked, that they are commensurate, and that each process may harness the other, speech production may be studied in a more naturalistic light than heretofore, avoiding the deep metaphysical problems that strictly symbolic accounts of cognition inevitably run foul of Clark (1997) and O'Regan and Noë (2002). The motor theory of speech seems compatible with recent work in neuroscience that has identified mirror neurons that are specific to the form of an action, whether it be carried out by the subject, or seen in a third party (Rizzolatti & Arbib, 1998). Using transcranial magnetic stimulation, Pulvermuller, Hauk, Nikulin, and Ilmoniemi (2005) recently demonstrated direct links between systems for lexical retrieval and limb action. Finally, employing motor structures (gestures) as phonological primitives has proved efficacious in the development of the theory of articulatory phonology, that has provided parsimonious accounts of many phenomena observed in speech (Browman & Goldstein, 1995).

Given the emphasis on the essential intertwining of perception and production both here and in many current approaches to understanding perception and action (O'Regan & Noë, 2002), it is worth noting that a loose analogue of the synchronization task may be found in the perception literature on talker normalization (Goldinger, Pisoni, & Logan, 1991; Wong, Nusbaum, & Small, 2004). Listeners to speech are able to ignore great amounts of variation in speech produced by different speakers, and extract the constant linguistic structure that is the speaker's intended message. In similar fashion, speakers are able to discard inessential variability and produce speech shorn of idiosyncratic and expressive variation.

Beyond such speech-specific considerations, we might reasonably expect that the principles underlying coordination of the articulators are not fundamentally different from those employed in movement generally. This turns out to be the case for the production of periodic units, both manually and vocally (Cummins & Port, 1998; Kelso & Munhall, 1988). While neither arm movement nor speech are typically periodic in any strict sense, limb movement is frequently highly coordinated across individuals, as in passing objects, shaking hands, playing sports, etc. This coordination is supported in part by the continuous flow of visual (and sometimes haptic) information between the individuals concerned (Schmidt & O'Brien, 1997). Our results suggest that acoustic information may play a significant role in coordinating the movements involved in producing speech when two speakers synchronize.

In the present experiments, and in previous work on synchronous speech, it has been demonstrated that the speech signal can be used to entrain the speech production of a co-speaker. Whether one considers this to be an instance of any *rhythmic* phenomenon or not may be largely a matter of personal taste. The term 'rhythm' is certainly rich enough to admit of multiple interpretations. However, by focussing attention on the ability of a signal to entrain the movements of another person, the topic of speech rhythm appears in a new light—one in which it is continuous with our understanding of rhythm from the related domains of music and movement.

Acknowledgments

This work has been supported by the Science Foundation Ireland through Principal Investigator Grant no. 04/IN3/I568 to the author.

Partial results of Experiments 1 and 2 were presented as a poster at ICPhS 2007. The method for quantifying asynchrony presented in Appendix A was introduced at an ISCA workshop in Athens. The author wishes to thank Dani Byrd, Mark Tiede and two anonymous reviewers, whose input significantly improved the present contribution.

Appendix A. Details of measurement of asynchrony in parallel recordings

In order to arrive at an automated quantitative measure of asynchrony between two time-aligned utterances, it will be expedient to first describe the typical implementation of the DTW algorithm, and then describe its modification for the present purposes.

A.1. Dynamic time warping

The DTW is a well-known algorithm, commonplace in speech recognition, that allows one to identify an optimal warping of one sequence onto a referent, with some common-sense constraints such as monotonicity and continuity (Meyers & Rabiner, 1981). Fig. 5 (left) illustrates the path identified by DTW in aligning two short symbolic strings. As one progresses from the bottom left square, one can choose only the square to the North, East or to the North–East as the best match at any given point. In the given example, the *b* in String 2 matches two elements in String 1, while the *ccc* sub-string maps onto a single *c* element in String 1.

In applying DTW to speech, we typically convert the speech to a parametric form, such as Mel frequency-scaled cepstral coefficients (MFCC), calculated for short overlapping frames, and treat the resultant sequence of MFCC vectors as the 'strings' to be warped onto one another, whereby each feature vector acts as a single 'symbol'. Euclidean distance measures provide a similarity metric

with which a decision to advance in a N, E or NE direction can be made.

A.2. Application of DTW to synchronous speech

In order to estimate asynchrony, we start with two time-aligned utterances. In the above experiments, these were typically a subject's production, on one hand, and the underlying model utterance on the other. We arbitrarily take the model utterance as a referent and use DTW to map the novel production (the comparator) onto this referent. Sequences of MFCC vectors are calculated for each utterance using default parameters: Hamming window of 1024 samples, or approximately 23 ms, with an overlap of half a frame across successive frames; filters ranging from 0 Hz to the Nyquist frequency of 22,050 Hz. The first 12 coefficients are retained, without inclusion of delta coefficients or the zero-th coefficient. MFCCs were computed using the voicebox toolbox available through http://mathworks.com. A warp path is then computed using DTW, as described above. We employed a standard implementation, likewise available from Mathworks. This provides the warp path as illustrated in Fig. 5.

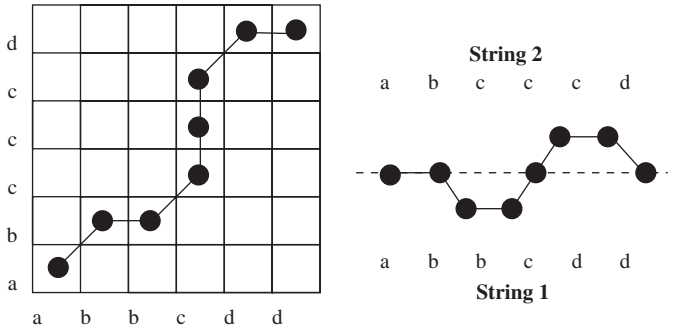

Fig. 5. Illustration of standard dynamic time warping path estimation (left). The comparator is shown along the *y*-axis, the referent along the *x*-axis. Right: conversion of the warp path into a time-aligned function, suitable for estimating asynchrony.

To arrive at an estimate of asynchrony,the warp path is then redrawn, with the SW-NE diagonal as the time axis, as shown in the right hand panel in Fig. 5. Steps in the DTW algorithm that move NE constitute a step of one frame width in the horizontal direction. Steps N or E each constitute deviations towards one or other string, and each such step advances 0.5*frame width along the horizontal time axis. The resulting path is illustrated in Fig. 5 (right panel).

Ranges for which this function is either increasing or decreasing correspond to areas of relative contraction or expansion, respectively, required to warp the comparator onto the referent. The unsigned area under the curve provides an estimate of the degree of asynchrony between the two utterances. Pilot testing was done using a wide variety of speech parameterizations, and with utterances that clearly manifested both good and bad synchronization. It was found that the stability and reproducibility of the algorithm was improved if the summed area under the warp path was confined to those portions of speech estimated to be voiced in the referent. Finally, distributions across a lot of test cases were found to be highly positively skewed, as is common for interval data (Rosen, 2005).

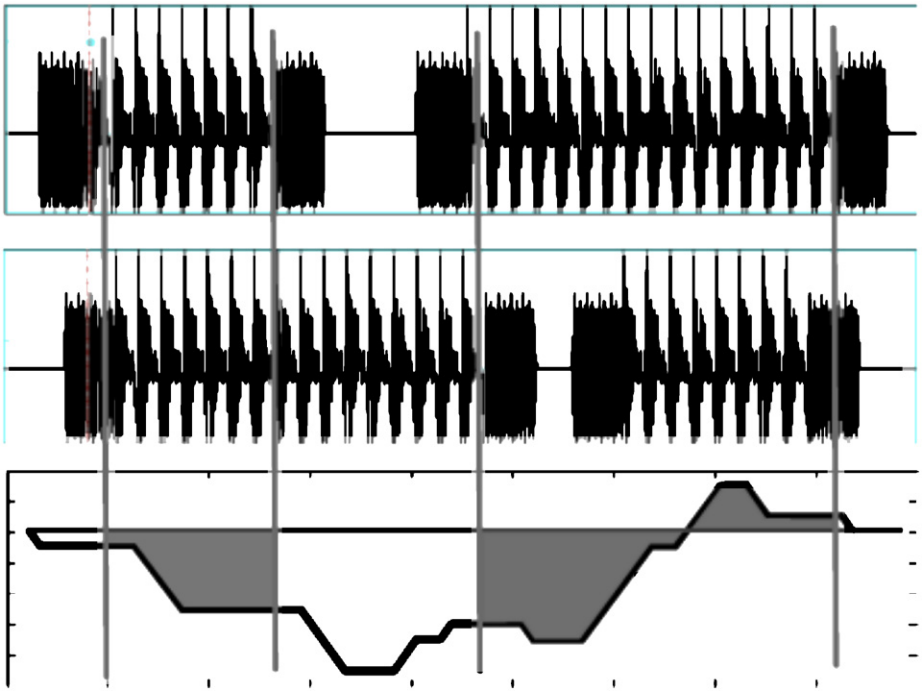

Fig. 6. Two time aligned utterances with associated warp curve. The area contributing to the asynchrony score is shaded. The upper utterance serves as a referent, the lower as a comparator.

The symmetry of the distribution was greatly improved by taking logs of the summed area.

In the context of the present study, all estimates are based on readings of the Rainbow Text. This has six phrases of unequal length. In order to make asynchrony estimates comparable across phrases of different length, the summed area under the warp curve is divided by the number of frames used in the computation.

Fig. 6 illustrates the process on a simple artificial signal. Each signal here comprises a sequence of burst-like and vowel-like parts. The top signal serves as a referent. In this simplified example, the first vowel-like portion of the referent would have to be stretched to align with the comparator, while the second would require compression. This is captured as areas below and above the horizontal line, respectively, in the warp path. Summation of these areas is done only over stretches in which the referent is voiced (shaded in the bottom panel). Finally, the summed area under the curve is divided by the number of frames used in the summation, and is then log transformed.

In order to illustrate the utility of this measure, Fig. 1 shows the distribution of asynchrony estimates for three sets of data. On the left are estimates from 72 paired phrases, taken from 12 dyadic readings of the Rainbow Text in which both speakers were live. In the center is a similar distribution of estimates taken from four speakers attempting to synchronize with 12 recordings of the Rainbow Text, in which the recordings were made while speakers were speaking in synchrony with another speaker. For each reading, the asynchrony estimate is based on the final five of the six phrases of the text, producing 240 data points in all. On the right are estimates from eight speakers reading the text along with a stimulus that is an uninformative hiss. The onset of the hiss is signaled to the speaker, and the duration of the hiss is matched to the duration of a model reading of the text, but the hiss is otherwise completely uninformative. The estimate of asynchrony is made by aligning the speaker's utterance with the model utterance on which the hiss was based. This thus serves as a control condition, and shows what kind of asynchrony might be expected if speakers were, in fact, independent of one another. Although the distributions overlap, it can be seen that the method employed clearly separates the latter control condition from the other two, and that synchrony with the recording is, indeed, comparable to that obtained with a live speaker.

Appendix B. Text used in all experiments

The text read by subjects in all experiments reported herein was the first paragraph of the Rainbow Text.

> When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

References

Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7–33). Cambridge: Cambridge University Press.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer (version 4.6.03) [computer program]. ⟨www.praat.org⟩.

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port, & T. van Gelder (Eds.), *Mind as motion* (pp. 175–193). Cambridge, MA: MIT Press.

Bruce, G. (1990). *Alignment and composition of tonal accents: Comments on Silverman and Pierrehumbert's paper*. Papers in laboratory phonology I: Between the grammar and physics of speech (pp. 107–114).

Bull, M. C. (1997). *The timing and coordination of turn-taking*. Ph.D. thesis, University of Edinburgh.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, *15*(2), 181–195.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Condon, W. S., & Sander, L. W. (1974). Synchrony demonstrated between movements of the neonate and adult speech. *Child Development*, *45*, 456–462.

Couper-Kuhlen, E. (1993). *English speech rhythm*. Philadelphia, PA: John Benjamins.

Crystal, T. H., & House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, *72*(3), 705–716.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, *3*(1), 7–11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, *31*(2), 139–148.

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). The CHAINS corpus: CHAracterizing INdividual Speakers. In *Proceedings of SPECOM'06* (pp. 431–435). St. Petersburg, Russia.

Cummins, F., & Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145–171.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*, 51–62.

de Jong, K. (2001). Rate-induced resyllabification revisited. *Language and Speech*, *44*(2), 197–216.

Fowler, C. A., Rubin, P., Remez, R., & Turvey, M. (1981). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production* (pp. 373–420). San Diego, CA: Academic Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goldinger, S., Pisoni, D., & Logan, J. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152–162.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. A. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 215–249). Cambridge: Cambridge University Press.

Goldstein, L., & Fowler, C., (2003). Articulatory phonology: A phonology for public language use. In *Phonetics and phonology in language*

comprehension and production: Differences and similarities (pp. 159–207).

Grabe, E., & Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven, & N. Warner (Eds.), *Papers in Laboratory Phonology 7* (pp. 515–546). Berlin/NewYork: Mouton de Gruyter.

Hayes, B. (1995). *Metrical stress theory*. Chicago: University of Chicago Press.

Heft, H. (2003). Affordances, dynamic experience, and the challenge of reification. *Ecological Psychology*, *15*(2), 149–180.

Jassem, W., Hill, D. R., & Witten, I. H. (1984). Isochrony in English speech: Its statistical validity and linguistic relevance. In D. Gibbon, & H. Richter (Eds.), *Intonation, accent and rhythm. Research in text theory*, Vol. 8 (pp. 203–225). Berlin: Walter de Gruyter.

Kelso, J. A. S. (1995). *Dynamic patterns*. Cambridge, MA: MIT Press.

Kelso, J. A. S., & Munhall, K. G. (Eds.), 1988. *R. H. Stetson's motor phonetics: A retrospective edition*. College-Hill, San Diego (Originally published 1928).

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162–179.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*, 253–263.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

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

Lloyd James, A., 1940. *Speech signals in telephony*. Cited in Abercrombie (1967; p. 171).

Meyers, C. S., & Rabiner, L. R. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, *60*(7), 1389–1409.

Mixdorff, H., Pfitzinger, H. R., & Grauwinkel, K. (2005). Towards objective measures for comparing speaking styles. In *Proceedings of SPECOM* (pp. 131–134). Patras, Greece.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*, 405–408.

O'Regan, J., & Noë, A. (2002). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, *77*(2), 640–648.

Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, *46*, 181–196.

Port, R. F., Tajima, K., & Cummins, F. (1996). Self-entrainment in animal behavior and human speech. *Online proceedings of the 1996 Midwest artificial intelligence and cognitive science conference*. URL ⟨http://www.cs.indiana.edu/event/maics96/proceedings.html⟩.

Pulvermuller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*, 793–797.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, *105*(1), 512–521.

Repp, B. H. (1996). Patterns of note onset asynchronies in expressive piano performance. *Journal of the Acoustical Society of America*, *100*(6), 3917–3932.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience*, *21*(5), 188–194.

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (Ed.), *Linguistic controversies* (pp. 73–79). London: Edward Arnold.

Rosen, K. M. (2005). Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *Journal of Phonetics*, *33*(4), 411–426.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333–382.

Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 227–247.

Schmidt, R. C., & O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. *Ecological Psychology*, *9*(3), 189–206.

Scott, S. K. (1993). *P-centers in speech: An acoustic analysis*. Ph.D. thesis, University College London.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326–332.

Spoor, P. S., & Swift, G. W. (2000). The Huygens entrainment phenomenon and thermoacoustic engines. *The Journal of the Acoustical Society of America*, *108*(2), 588–599.

Stetson, R. H. (1951). *Motor phonetics* (2nd ed.). Amsterdam: North-Holland.

Stoffregen, T. A. (2003). Affordances as properties of the animal-environment system. *Ecological Psychology*, *15*, 115–134.

Streeck, J. (1994). Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction*, *27*(3), 239–267.

Wong, P., Nusbaum, H., & Small, S. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173–1184.