SPEECH SYNCHRONIZATION: INVESTIGATING THE LINKS BETWEEN PERCEPTION AND ACTION IN SPEECH PRODUCTION

Fred Cummins

University College Dublin
fred.cummins@ucd.ie

ABSTRACT

Speakers can achieve a high degree of synchrony when reading a prepared text together. Under these constraints, there is necessarily a very tight coupling of production and perception. In a first experiment, we demonstrate that speakers can successfully synchronize with selected recordings of others obtained in a synchronous speaking condition. We then have speakers attempt to synchronize with modified recordings, in which the original recording is replaced with altered speech. The goal is to find out the physical properties of the speech signal which permit the coupling required for synchronization. It is demonstrated that the energy envelope itself is not sufficient to support coupling, while pitch information is essentially unimportant.

Keywords: Synchronization, perception-action coupling, rhythm

1. INTRODUCTION

It has been shown that speakers can achieve a very high degree of synchrony over protracted sequences of utterances when asked to read a prepared text together [5]. Typical median asynchrony values are 40 ms, with a slightly greater typical asynchrony at phrase onsets (ca. 60 ms). Given the great degree of variability in speech timing within and across individuals, this demonstrates a remarkably tight coupling of the speech production systems of the two speakers involved. Several accounts of this feat are possible.

Firstly, speakers could be monitoring the speech of their co-speaker, and adjusting their timing in response. This seems to be a plausible account of occasional dysfluencies encountered under these conditions, but does not provide a satisfactory account of such tight responsive coupling, due to the time lags involved [13]. Secondly, subjects could be responding to the task demands by stripping their speech of its idiosyncratic timing elements, and falling back on neutral, unmarked temporal specifications which are presumed to be held in common by competent speakers of a language. Again, this account lacks plausibility, due to the variability known to be present in even relatively unmarked, laboratory, speech [8].

A third possibility is suggested by recent findings about direct linkages between motor action and perception, with the discovery of mirror neurons in monkey, and by inference, in humans [12, 11]. Neurons have been identified in several brain regions which exhibit sensitivity to the perception of a specific intentional action, such as grasping or reaching, and these same neurons are found to fire when the subject performs the same action. Identification of mirror neurons has provided the first neurophysiologically plausible account of direct linkages between motor action and high level perception. A possible role for such a common representation has long been suggested within speech theory, in the motor theory of speech perception [11]. As we learn more about the intimate linkages of perception and action, it seems appropriate to consider a model in which the two speakers are viewed as mutually entrained systems, and to ask what the basis for this entrainment might be.

Entrainment between movements need by no means be restricted to multiple effectors within a single individual, as in a gait [4] or periodic finger waving [9]. In an elegant experiment, Schmidt, Carrello and Turvey demonstrated that two people constrained to wave their lower legs in synchrony with one another exhibit the same signatures of a single coupled system as found in bimanual oscillation [14]. In this case, the basis for the entrainment is clearly *information*, specifically visual information provides the necessary linkage between the two systems, allowing their respective movements to exhibit reciprocal influence.

With this in mind, the present series of experiments is designed to investigate whether it is possible to identify an *informational* basis for entrainment among speakers speaking in synchrony. In a preliminary experiment, we ask whether people can synchronize with recorded speech as well they do with speech of a live co-speaker. Those recordings to which subjects can best synchronize are then used in a second experiment. Here, speech is systematic-

ally degraded, and we measure the effects of each form of degradation upon synchronization performance.

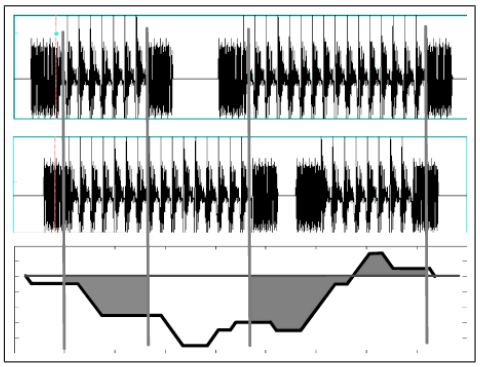
2. EXPERIMENT 1: SYNCHRONIZING WITH RECORDED SPEECH

Recordings of 12 speakers (6m, 6f) of Hiberno-English reading the first paragraph of the Rainbow Text were made both when speakers read alone, and when they read in synchrony with a co-speaker. The 12 speakers were chosen from a larger corpus of 12 female and 16 male speakers, and selection was based on an informal appraisal of fluency and naturalness. The recordings were modified so that a series of three isochronous beeps at 0.5 sec intervals preceded the start of each of the 6 sentences of the text, ensuring that each sentence onset was maximally predictable.

The 24 paragraphs were played in random order to 4 subjects (2m, 2f, Hiberno-English speakers), who were instructed to synchronize as well as possible with the recording. At the same time, the text of the paragraph was displayed, with clear separation among the 6 sentences. Synchronization was measured automatically using a procedure described in full in [6]. In brief, the speech is parameterized as standard MFCC feature vectors, and Dynamic Time Warping is used to find the optimal warp path from one sequence of feature vectors to the other. The amount of warping required is a function of the temporal alignment of the two utterances, and this is quantified as the unsigned area under the warp function. The method is illustrated in Figure 1, in which two utterances and their associated warp path are shown. The shaded area under the curve is summed to arrive at an estimate of asynchrony.

At issue in this preliminary experiment was the question of whether subjects could synchronize at all with recordings of speech, and if so, whether they were facilitated in the task if the recording was of speech which was, in turn, recorded under synchronous speaking constraints. To answer the first question, we estimated the degree of synchrony achieved by the subjects speaking in time with speech which was originally recorded in a synchronous speech setting, and compared this to the synchrony achieved by the co-speakers in the original setting, where both speakers were 'live'. Results are shown in Figure 2, in which the y-axis plots the quantitative estimate of synchrony in units derived from the warp path. Perfect synchrony results in a score of 0. Although synchrony is somewhat reduced when synchronizing with the recordings ($t(71)=-4.1, p<001$), performance appears roughly comparable, and will

Figure 1: Quantification of asynchrony between two stylized utterances. The warp path obtained by dynamic time warping has been replotted as a function of (referent) time at the bottom. Portions of the area under the curve corresponding to voiced portions of the referent (top waveform) are summed.



serve as a baseline comparison when evaluating synchrony scores in subsequent conditions. Individual data points are for single sentences, but because the quantification of asynchrony is a normalized measure which is insensitive to the length of phrase used, the estimates are numerically comparable across the two experiments reported here.

Figure 2: Synchrony estimates for speakers in a live setting and synchronizing with recording. The y-axis shows measured asynchrony.

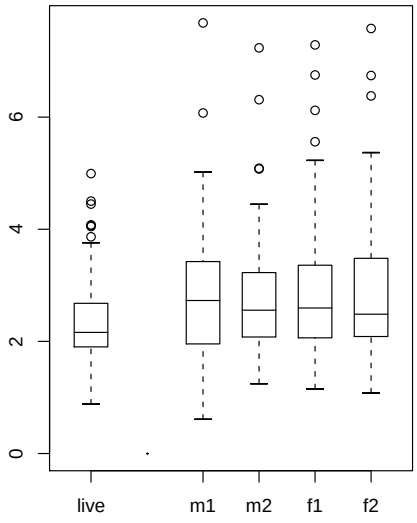
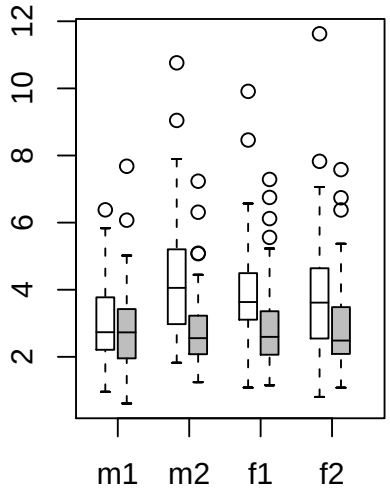


Figure 3 compares synchronization performance where the recording itself is either normal speech or synchronous speech. For three of the four subjects, synchronization is greater when the recording itself was recorded in a synchronous condition (all

$p<0.01$, except for m1, n.s.).

Figure 3: Performance when synchronizing with a recording which is normal speech (white bars) or synchronous speech (grey bars).



Based on these results, the four recorded speakers with whom subjects exhibited best synchronization were selected, and their synchronous recordings were used as stimuli in Experiment 2. In this manner, we ensure that the synchronization task is as easy as possible.

3. EXPERIMENT 2: SYNCHRONIZING WITH DEGRADED SPEECH

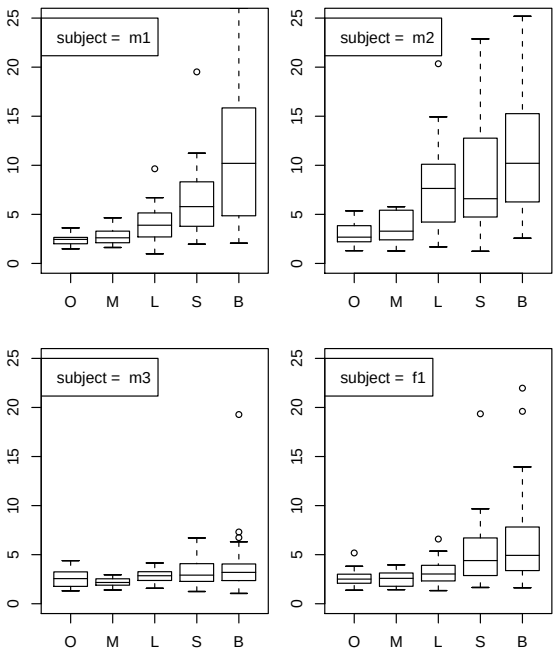
Four modifications were made to the original recordings. For one set, the fundamental frequency was resynthesized at a constant value of 110 Hz, producing monotone (MONO) utterances. In another, the speech was low pass filtered with a cut off frequency of 500 Hz (LPF). In a third, each sample was randomly (p=0.5) flipped, producing signal correlated noise which preserves the amplitude envelope of the original but renders speech entirely unintelligible (SCN), 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 [2] to enhance the modulation of the envelope. The resultant stimuli are of course still unintelligible, but we reasoned that enhancing the envelope modulation might provide a useful cue for synchronization. We label this condition BAND.

Four subjects (3m,1f, Hiberno-English speakers) listened to to the modified stimuli and an unmodified control presented in random order. In each case, the first sentence was left unaltered, so that subjects could attune to the subjects voice quality and speaking rate. Synchrony was evaluated over the remaining five sentences, by aligning the subject's recording with the *original* recording and estimating the warp path for voiced sections, as before.

3.1. Results

Figure 4 shows synchrony achieved 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 disimprovement, despite the severe modification of the stimulus. In general, ORIG and MONO produced comparable degrees of synchrony, with LPF somewhat worse and SCN and BAND considerably worse. A repeated measures analysis of variance with condition and co-speaker as factors showed a main effect of condition ($F(4,377)=36, p<.001$), while co-speaker and the interaction were not significant. Tukey HSD post hoc tests revealed that synchrony estimates for BAND and SCN were significantly worse than ORIG for three of the four speakers (m1,m2,f1, $p<.01$), and LPF was worse then ORIG and MONO for m2 ($p<.01$).

Figure 4: Performance when synchronizing with original unmodified stimuli ('O'), MONO ('M'), LPF ('L'), SCN ('S') or BAND ('B') stimuli. Each panel shows data from one subject.



4. DISCUSSION

We set out with the goal of looking for the informational bases for entrainment among speakers. As a first step, it was necessary to ensure that synchronization with a static recording was possible, and we found that this is indeed possible, and is facilitated when the recorded model was originally obtained under matched speaking conditions (i.e. when speaking in synchrony with another speaker). We then made some fairly obvious modifications to the recordings, to selectively remove some possible bases for entrainment. Pitch information (F0) was found to be entirely dispensable, as no subject showed worse performance for the monotone stimuli than for the originals. The low pass filtered speech was somewhat worse for one subject, and the trend towards a performance decrement was observed in two others. With a cut off at 500 Hz, this speech remains largely intelligible under these optimally predictable circumstances. The remaining two conditions, BAND and SCN, were markedly worse for three of the four subjects. In both cases, the speech is completely unintelligible, but the slow envelope modulation remains, and is accentuated in BAND. Our interest in these conditions came from the naive belief that macroscopic timing information (or 'rhythm' broadly construed) is largely specified by the envelope modulation. This was clearly not sufficient to support synchronization here.

Further experiments with systematically altered speech will be necessary to pursue this question further. Of special interest is the finding that the envelope modulation alone does not provide sufficient information for synchronization. This adds to findings from other studies that perceived 'rhythm' in speech is not merely a function of syllable onset timing, as these onsets are preserved in SCN, and were even accentuated in BAND to no effect [7, 1]. This calls into question those approaches to the study of rhythm in speech which assume that macroscopic timing is conveyed primarily via the amplitude envelope [3, 10]. The relative unimportance of fundamental frequency in synchronization is somewhat surprising. It is well known that some points in the fundamental frequency contour are tightly tied to the segmental content, suggesting an intimate link between timing and pitch variation. This link was entirely removed from the MONO utterances, resulting in clearly distorted speech, yet synchronization was unaffected.

5. REFERENCES

[1] A.-P. Benguerel and J. D'Arcy. Time-warping and the perception of rhythm in speech. *Journal of Phonetics*, 14:231–246, 1986.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program]. www.praat.org, 2005.

[3] L. Bosch and Sebastián-Gallés. Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, 65:33–69, 1997.

[4] J. J. Collins and I. Stewart. Hexapodal gaits and coupled nonlinear oscillator models. *Biological Cybernetics*, 68:287–298, 1993.

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

[6] F. Cummins. The quantitative estimation of asynchrony among concurrent speakers. Technical Report UCD-CSI-2007-2, School of Computer Science and Informatics, University College Dublin, 2007.

[7] R. M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62, 1983.

[8] E. Keller. The variation of absolute and relative measures of speech activity. *Journal of Phonetics*, 15:335–347, 1987.

[9] J. A. S. Kelso. *Dynamic Patterns*. MIT Press, Cambridge, MA, 1995.

[10] T. Nazzi, J. Bertoncini, and J. Mehler. Language discrimination by newborns: towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24:756–766, 1998.

[11] G. Rizzolatti and M. A. Arbib. Language within our grasp. *Trends in Neuroscience*, 21(5):188–194, 1998.

[12] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[13] E. Saltzman and J. A. S. Kelso. Skilled actiions: A task dynamic approach. *Psychological Review*, 94:84–106, 1987.

[14] R. C. Schmidt, C. Carello, and M. T. Turvey. 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, 1990.