On synchronous speech

Fred Cummins

Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
fred.cummins@ucd.ie

Abstract: Synchronous speech is speech elicited by asking speakers to read a text in synchrony. The present study investigates the timing characteristics of speech obtained under such circumstances. In a main experiment, subjects read a text alone, with a recording of another speaker, or with another live speaker. The last condition produces a much higher degree of synchrony, even at the left edges of phrases following a pause. Subjects display a high level of agreement in pause placement in the synchronous condition, but add pauses idiosyncratically when reading alone. A small second experiment fails to uncover the informational basis of this synchrony, because some subjects can achieve similar synchrony with a recording of synchronous speech, whereas others appear to require a live speaker. Speech that has been modified in this manner is of immediate interest because it seems to express speaker's attempts to produce maximally predictable speech.

PACS numbers: 43.70.Jt, 43.70.Fq
Date Received: Dec 7, 2000 **Date Accepted:**

1. Introduction

When two people recite a common text in synchrony, they make marked alterations to their speech. If synchrony is to to be achieved across speakers, it is apparent that they must somehow eliminate unpredictable flourishes that would make speech timing unpredictable for their co-speakers. This study investigates the timing characteristics of speech obtained under such circumstances. Its main goal is to see whether speech elicited in synchrony with another can be differentiated from other forms of speech (unconstrained speech, speech in synchrony with a recording). The results should serve as an existence proof and provide an initial baseline for a range of further investigations.

Speaking in approximate synchrony with other speakers is familiar from tasks such as praying, chanting, reciting oaths, etc. A feature of these situations is that the texts are well practiced and usually have a highly stylized prosody. For example, the pledge of allegiance, as recited by American schoolchildren, differs markedly from a reading by one unfamiliar with the text. Nonetheless, a single informal trial will suffice to convince the reader that if two speakers are presented with a text with which they are reasonably familiar, they can read the text in synchrony with what appears to be reasonable success. More generally, speaking in synchrony with another requires rapid attunement of each speaker to the other, and is thus a generalized form of speaker-listener accommodation, as is familiar from situations like infant-directed speech, speech directed at nonnative speakers, etc.

Success at this task is of immediate theoretical and practical interest. If synchrony is possible without elaborate training (as we demonstrate below), then speakers must be able to strip their speech of idiosyncratic features, which would render it unpredictable (and which remain the bane of speech recognition systems). How this could be achieved is not immediately obvious: would speakers revert to some presumed default values for segment and syllable durations? Or would they dynamically exchange cues that allow the upcoming portion of speech to be predicted?

The issues are quite akin to those examined in the study of expressive timing in music performance, where ensemble playing, with its restricted opportunity for stylistic embellishment, provides an analogy to synchronous speech [1,2]. In this case, the musical score provides a representation of idealized timing against which we can measure expressive variation. With

Table 1. Canonical division of the first paragraph of the "rainbow" text into 6 phrases. Vowel onsets in italicized syllables formed the basis of measurements of synchrony reported below.

[When the sunlight strikes raindrops in the air they act like a prism and form a rainbow**]**
[The *rain*bow is a division of white light into many beautiful *col*ors**]**
[*These* take the shape of a long round arch with its path high above, and its two ends apparently *be*yond the horizon**]**
[*There* is, according to legend, a boiling pot of gold at one *end***]**
[*Peo*ple 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 rain*bow***]**

speech, no such referent exists. The literature on factors affecting speech timing is vast, but does not in general aspire to making a sharp distinction between inherent timing factors that must affect speech and factors that a speaker may optionally apply as the communicative situation warrants. Indeed, the difficulties in generalizing results obtained from laboratory experiments to spontaneous (expressive) speech may be due in part to the failure of the laboratory situation to elicit much in the normal range of expressive timing [3,4].

Many experimental tasks attempt to reduce variability in one measure or another, e.g. the near elimination of segmental variation in reiterant speech [5]. In a synchronous speech task, however, we should be able to exploit the tacit knowledge of speakers about both necessary and superfluous aspects to their speech. If this holds up, the experimental condition may be of potential use in a variety of situations. For example, it may be used to identify preferred or unmarked choices from a paradigm, as when speakers appear to choose freely from several possible tunes for a single utterance [6].

Beyond the theoretical significance, an experimental condition in which speakers produced something like "default" speech would be of immediate interest to those attempting to synthesize expressive speech, in that a partition between neutral (predictable) and expressive factors could form the basis for a principled decomposition of the task. Concatenative approaches could economize by distinguishing between the essential and the expressive.

The present work seeks to establish a baseline result and prepare the ground for a more thorough investigation of synchronous speech. The question we ask is whether subjects can, in fact, achieve the goal of speaking in synchrony with one another, when the text does not have a highly stylized prosody, as in prayer or recitation.

2. Experiment I: baseline results with synchronous speech

2.1. Methods

Four subjects (2 males, 2 females, age 20–35) participated. All were from the area around Dublin, Ireland. Readings of the first part of the "rainbow" text (see Table 1) were obtained in three conditions. In the solo condition, subjects first practiced reading the text aloud, after which 12 recordings were obtained with no further constraints on speaking style or rate. In the recording condition, each speaker attempted to read the text in synchrony with a recording (from the first session) of one of the other speakers. Twelve trials per subject were obtained (4 target recordings taken randomly from each of the 3 other subjects). Finally, in the synchronous condition, each possible subject-pair (6 in all) read the text 4 times in synchrony. In this last condition, subjects were seated comfortably next to each other. Each wore a head-mounted near-field microphone (Shure WH20), and recordings were made onto the right and left channels of a single stereo file. Subjects were free to look at each other throughout.

2.2. Results

We first compare the solo condition with the synchronous condition.

Table 2. Mean (and standard deviation) in Hz of the range within which 90% of measured F_0 values fell within a trial.

subject	solo	recording	synchronous
F1	97 (5.3)	77 (5.8)	68 (7.8)
F2	70 (6.5)	49 (7.8)	45 (7.5)
M1	58 (5.2)	48 (4.9)	32 (3.2)
M2	41 (8.5)	46 (13.3)	30 (5.8)

2.2.1. Phrasing

Table 1 divides the "rainbow" text into 6 distinct phrases. In the synchronous condition, pauses (silence of more than 200 ms) are present at these phrase edges without exception. Pauses at other points (such as major syntactic edges within these phrases) occur 4 times in 24 paired readings. By contrast, pauses occur at other points 48 times in the 48 readings in the solo condition. Pauses are absent at these phrase edges 4 times (one speaker only). Thus, speakers display almost complete agreement on pause placement in the synchronous condition but often add additional pauses when reading alone.

2.2.2. Rate

Three speakers show a longer utterance duration in the synchronous condition, and one a shorter. The ratio of pause (silence longer than 200 ms) to speech is larger in the synchronous condition for two of the four speakers, and smaller for the other two. Thus, there is no consistent effect of condition on either speech rate or articulation rate.

2.2.3. Pitch range

Fundamental frequency was estimated using the AMDF pitch estimation provided with the Snack Sound Toolkit[7]. Pitch range (operationally defined as the range within which 90% of measured pitch values lie) is reduced in both the recording and the synchronous condition. Table 2 gives means and standard deviations for each speaker and condition. One way ANOVAs done for each subject individually showed a significant effect of condition for each subject ($p < 0.01$). Post-hoc Tukey HSD tests for differences between means ($\alpha = 0.01$) showed a significant difference between each pair of means for subjects F1 and M1, between solo and recording, and between solo and synchronous for F2 and between recording and synchronous for M2, all in the expected direction (i.e., the range is smaller in the recording condition and smaller again in the synchronous condition). A similar analysis of mean pitch values produced no consistent results across subjects.

We now turn to measurement of synchrony in the recording and synchronous conditions.

2.2.4. Synchrony

We assessed synchrony by looking at the temporal lag between corresponding events in the waveforms of paired speakers. For each phrase but the first, the magnitude of the lag between corresponding vowel onsets at the start and at the end of the phrase was measured. The syllables used are given in italics in Table 1. The data are plotted in Figure 1. In each case, synchrony is greater in the synchronous condition than in the recording condition. A Wilcoxon signed rank test confirms the effect of position within each condition, and a Wilcoxon rank sum test compares corresponding positions across conditions (all p-values < 0.001). Synchrony at the beginning of phrases in the synchronous condition appears to be as good as at the end of phrases in the recording condition. Subjects in the synchronous condition thus have a much easier time predicting when speech will resume after a pause at a major phrase edge.

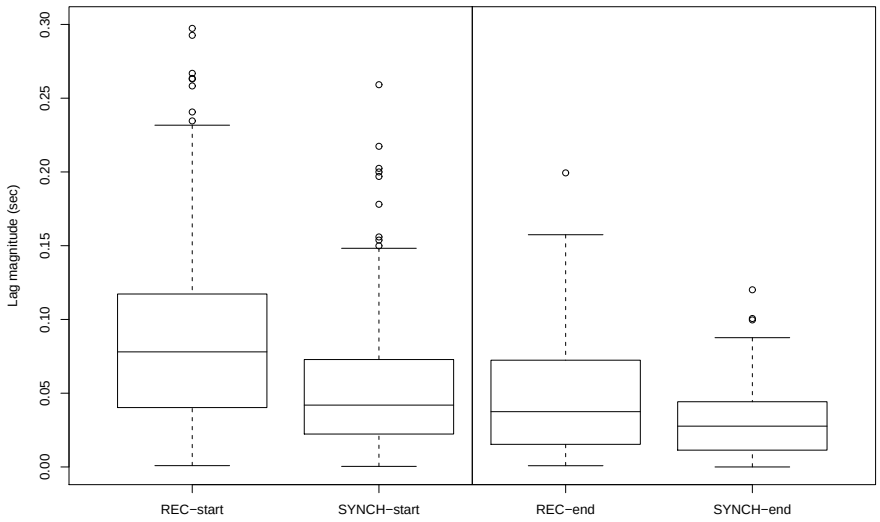

Fig. 1. Absolute value of measured asynchrony at phrase start and ends.

Summarizing, subjects demonstrate clearly that they can achieve a high degree of synchrony. Median lag magnitude in the synchronous condition was 30 ms, (upper quartile 55 ms), but 56 ms (upper quartile: 95 ms) in the recording condition. Remarkably, synchrony is maintained well across pauses when both speakers are "live."

3. Experiment II: Dynamic cues or simplification?

How is synchrony maintained across speakers? In an effort to uncover the informational basis for the ability of speakers to predict the time of phrase onset of their co-speakers, we compared speakers' performance when reading in synchrony with two kinds of recording. In one condition, ("rec-solo"), the recording was obtained in an unconstrained reading (these recordings were taken from the solo condition of Experiment I, o this condition essentially replicates the recording' condition of the previous experiment). In the other ("rec-synch"), recordings were taken from one channel of the synchronous condition of Experiment I. Our hypothesis was that it might be easier to synchronize with a recording of synchronous speech than a recording of unconstrained speech. Three of the original four subjects participated. Each listened to a randomized set consisting of 24 recordings from each conditio, for a total of 48. They were asked to synchronize their speech to the recordings. Asynchrony at phrase start and end was again measured and compared across phrase position and condition.

3.1. Results

We again used nonparametric statistics to evaluate the effect of position within each condition, and of condition for each position separately. Using the Wilcoxon signed rank test, there was a significant effect of position within each condition (asynchrony at phrase starts was routinely larger than that at phrase ends, $p < 0.001$), and this result held when separate tests for each individual were done (all p-values < 0.05). However, when we compare phrase initial position across conditions (Wilcoxon rank sum), we find two subjects for whom synchrony is substantially improved when the recording is a recording of synchronous speech, and one subject for whom there is no such effect. The same holds for a comparison of phrase-final position across conditions, with the same one subject demonstrating no advantage in the "rec-synch" condition.

Clearly, our small pool of subjects are exhibiting different behaviors. Our hypothesis is not unequivocally supported by this pilot experiment. Further work is called for, which more carefully controls the degree of information available to each subject.

4. Discussion

This small study has sought to demonstrate that synchronous speech is both possible and interestingly different from other forms of speech. Subjects were able to maintain a high degree of synchrony with little or no practice. They are thus clearly capable of making their speech very predictable for a co-speaker. The fact that this is possible opens up several immediate avenues of inquiry.

- What is the informational basis that allows subjects to start phrases together after a pause? Pilot results from Experiment II suggest that different subjects may make different use of available cues.
- Do subjects match intonational contours too? The study of intonation has been bedeviled by the difficulty of separating the categorical/linguistic from the continuous/paralinguistic[8]. Synchronous speech may help to settle long contested claims about the linguistic nature of intonational contrasts.
- Is segmental variation similarly reduced in synchronous speech? If so, this may provide an appealing alternative to the well-worn phonetic vice clamps of "Say X again" frames.
- Speech that has been modified in this manner is of immediate interest because it seems to express speaker's attempts to produce maximally predictable speech.

For the above reasons, synchronous speech appears to provide a promising new object of study and, perhaps, also a novel tool in the arsenal of the experimental phonetician.

References and links

[1]R. A. Rasch. "Timing and synchronization in ensemble performance," in *Generative Processes in Music*, edited by J. A. Sloboda (Clarendon Press, Oxford, 1988) pp. 70–90.

[2]B. H. Repp. "Patterns of note onset asynchronies in expressive piano performance," J. Acoust. Soc. Am., **100**, 3917–3932 (1996)

[3]M. E. Beckman. "A typology of spontaneous speech," in *Computing Prosody: Computational Models for Processing Spontaneous Speech*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi (Springer Verlag, New York, 1996) pp. 7–26.

[4]T. H. Crystal and A. S. House. "Segmental durations in connected-speech signals: Current results," J. Acoust. Soc. Am., **83**, 1553–1573 (1988).

[5]M. Y. Liberman and L. A. Streeter. "Use of nonsense-syllable mimicry in the study of prosodic phenomena," J. Acoust. Soc. Am., **63**, 231–233 (1978).

[6]J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, reprinted by the Indiana University Linguistics Club (1980).

[7]K. Sjölander, "The Snack Sound Toolkit," http://www.speech.kth.se/snack/ (2001).

[8]J. Pierrehumbert. "Tonal elements and their alignment," in *Prosody: Theory and Experiment*, edited by M. Horne, vol. 14 of *Text, Speech, and Language Technology* (Kluwer Academic, Dordrecht, 2000) chap. 1.