Prosodic Characteristics of Synchronous Speech

Fred Cummins
Department of Computer Science
University College Dublin
fred.cummins@ucd.ie

September 5, 2000

Abstract

When speakers are constrained to read a text in synchrony with one another, they make marked changes to their prosody. We compare timing- and intonation-based variables obtained from unconstrained speech, speech elicited when attempting to read together with a recording, and speech obtained from two speakers speaking together. Speakers are remarkably successful at synchronizing their speech, and the degree of synchronicity is much higher when both speakers are "live" than when synchronizing with a recording. Pitch range is substantially reduced in the synchronous condition. Speakers appear to agree on the position and duration of silent pauses. It is still unclear whether the synchronous speech is enabled by having speakers produce default durations and phrasing, or whether there is a dynamic entrainment among the speakers.

1 Introduction

As competent speakers, we continuously alter speaking style, changing from a formal to an informal register, adopting an intimate tone one moment only to switch to a brasher style more suited to public address (Beckman, 1996). We also modulate our speech to suit the perceived needs of our listeners, e.g. the non-native speaker, the hearing-impaired or the hungover. Each change of style brings substantial changes to a host of measurable quantities, and prosodic variables are particularly affected. The ability to change style effortlessly suggests that the level at which speakers exert control over their prosody is very high indeed, and the relationship between such high-order variables as "formality" and measurable quantities such as "pitch range" remains to be explored.

Lindblom (1990) has suggested that one complex dimension along which speakers can vary their speech is that of hyper/hypo-articulation. Speech towards the "hyper" end of the continuum is associated with a need for clarity and expressiveness, where speech carries a high informational load, while hypo-articulated speech is possible where the informational load and the communicative demands are less. This has proven to be a useful way to understand some of the variation normally found in speech, and provides a possible link between such high-order information as the "need for clarity" with the lower-order variables of phonetic observation. Variation along a single dimension (the avowedly simplistic H–H dimension) is associated with changes in a host of observables, such as jaw displacement, tongue trajectory, etc.

Given the large gap between putative high-order control variables and our measured quantities, we should actively seek experimental tasks in which global changes to prosody are reliably induced. In the present work, we offer some initial empirical findings about global changes made to prosody when speakers are constrained to speak together—a condition we call Synchronous Speech (SS).

Speaking together is something we are all reasonably familiar with from group settings such as classrooms, churches, assemblies, etc. Typically, the text which is recited by a number of people is very familiar and may be associated with a stereotypical and highly idiosyncratic prosodic form, as for example when American school children recite the Pledge of Allegiance, or Catholics recite a decade of the Rosary. Given any agreed text, however, two or more speakers may attempt to read in synchrony. Success at this endeavour would be of interest for many reasons. It would suggest that *either* speakers are falling back on default phrasing, using unmarked durations, pauses, accents, etc, *or* that speakers enter into a negotiated form of speech in which each takes cues from the other to ensure synchronization of the result. Either result would be of great theoretical interest.

In this work we present an initial experimental investigation of synchronous speech. The central question addressed is whether the condition of speaking along with another speaker does, as we suspect, induce consistent and far-reaching changes to macroscopic timing and intonation (prosody). We then seek to roughly characterize the changes so induced, and to pose questions for further research which will help to illluminate the issue of high-level prosodic control.

2 Methods

Four subjects (2 m, 2 f) participated. All were from the area around Dublin, Ireland. Readings of the familiar Rainbow Text were obtained in three conditions. In the 'solo' condition, subjects first practiced reading the text aloud, then 12 recordings were obtained, without any 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. 12 trials per subject were obtained (4 target recordings taken randomly from each of the 3 other subjects). Finally, in the 'synchronous' condition, each subject-pair read the text 4 times in synchrony. In this latter condition, subjects were seated comfortably next to one another. Each wore a head-mounted near-field microphone, and recordings were made onto the right and left channels of a single stereo file. Subjects were free to look at one another throughout.

3 Results

Overall, subjects were successful in their attempts at speaking in synchrony. Recordings in the 'synchronous' condition showed a high degree of synchrony, as quantified below. A notable feature was the agreement shown among speakers on the placement of pauses (silence of greater than 200 ms duration). The appendix gives the Rainbow Text divided into major phrases. In the 'synchronous' condition, there were only 4 instances of a pause occurring at any location other than the edges of these phrases, and pauses at these edges were uniformly present. In the 'solo' readings, on the other hand, 48 pauses at points other than these phrase edges occurred, and one subject (LK) omitted the pause between the last two phrases on 4 occasions.

For each subject we compared the articulation rate (speech duration less pauses > 200 ms) across conditions, as well as the pause-to-speech ratio. Results are presented in Table 1. There is no consistent effect of condition, as changes in both articulation rate and pause-to-speech ratio are seen in both directions for different speakers. The speech rates produced in the 'synchronous' condition lie between the extrema observed in the 'solo' condition.

To quantify the degree of synchronicity within a trial, we measured the offset between corresponding vowel onsets which were readily identifiable in the two parallel recordings. This gave between 30 and 67 observations per trial (with the exception of one trial which had recording

	Articulation rate		Pause/speech ratio	
HA	faster	$p < 0.001$	larger	n.s.
LK	slower	$p < 0.001$	smaller	$p < 0.01$
TC	slower	n.s.	larger	$p < 0.001$
TG	slower	$p < 0.001$	smaller	n.s.

Table 1: Direction of rate effects ('synchronous' condition compared with 'solo' condition).

	HA	LK	TC	TG
recording	0.085	0.048	0.064	0.059
synchronous	0.026	0.023	0.026	0.026

Table 2: Average lag magnitude (man of trial means). All differences significant ($p < 0.001$).

problems, for which only 14 points were obtained). The mean of the magnitude of the lag between speakers was calculated for each trial in the 'recording' and 'synchronous' conditions, and t-tests were done for each subject to see if the average lag differed across conditions. In all cases, the average lag size was greater ($p < 0.001$) in the 'recording' condition than in the 'synchronous' condition. Table 2 gives the average lag (mean of trial means) for each subject and condition. The 'synchronous' condition clearly elicits speech with a greater synchrony between speakers than does the 'recording' condition, and the average lag is about one third that observed in the 'recording' condition.

Finally, pitch ranges in the three conditions were compared. F_0 traces were obtained using the pitch tracking module of the Wave Surfer software suite (`www.speech.kth.se/wavesurfer/`). To minimize the influence of outliers, an operational definition of pitch range for one trial was adopted as being the smallest range within which 90% of the F_0 values for one trial fell. Table 3 gives the average range (and standard deviation) for each subject and condition. In all cases, the 'synchronous' condition had the smallest mean range of F_0, and for three of the four subjects the 'recording' condition had a smaller mean range than the solo reading.

	solo	recording	synchronous
HA	97 (5.3)	77 (5.8)	68 (7.8)
LK	70 (6.5)	49 (7.8)	45 (7.5)
TC	58 (5.2)	48 (4.9)	32 (3.2)
TG	41 (8.5)	46 (13.3)	30 (5.8)

Table 3: Mean (s.d.) range within which 90% of measured F_0 values fell within a trial

4 Discussion

This pilot study has demonstrated that speakers are capable of establishing a tight degree of synchrony when reading a familiar text together. The close alignment of the two speech channels is considerably stronger when speakers are both "live", i.e., when neither is a static recording. In particular, speakers have no problem in tightly coordinating the placement and timing of pauses.

The basis for this coupling is still unclear. Two hypotheses seem to arise. According to the first, the speakers each have an unconscious knowledge of default timing values for speech. In the synchronous condition, they revert to these defaults, and because they share common values, the resulting speech is highly similar across speakers. The second hypothesis suggests that speakers exchange information continually throughout the reading, so that a dynamic entrainment results. If this is the case, it should be possible to discern the informational basis for this entrainment, but this remains to be done. For example, it is conceivable that speakers are capable of enhancing the degree to which their speech is strictly metrical, which would enhance the predictability of future events. If so, metricality would be a putative high-order control variable for speech which speakers modify directly. Future research will seek to discriminate between these two hypotheses and to examine the possibility that speakers are actually enhancing the predictability of their speech.

5 Acknowledgements

Thanks for insightful discussion with Gérard Bailly and Plinio Barbosa. Also to Robert Battye for measurement and suggestions.

References

Beckman, M. E. (1996). A typology of spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 7–26. Springer Verlag, New York.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic, Dordrecht.

6 Appendix

Canonical division of the "Rainbow" text into phrases. This division was manifest almost without exception in the synchronous condition.

- When the sunlight strikes raindrops in the air they act like a prism and form a rainbow
- The rainbow is a division of white light into many beautiful colors
- 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