Measuring synchronization among speakers reading together

Fred Cummins
School of Computer Science & Informatics, University College Dublin, Ireland
https://doi.org/10.36505/ExLing-2006/01/0020/000020

Abstract

It has been demonstrated that speakers are readily able to synchronize with a co-speaker when reading a prepared text together. The means by which a high degree of synchronization is attained are still unknown. We here present a novel measure of synchrony which allows us to follow the time course of synchronization among two speakers, based on the parallel acoustic signals. The method uses traditional frame-based cepstral features and a slight variant on standard dynamic time-warping. We develop the method based on a novel corpus of synchronous speech, comparing its estimates of synchronicity with hand estimates. The method out-performs laborious manual estimation, and allows us to now begin to study the dynamics of synchroni-zation among speakers.

Synchrony Among Speakers

It has been demonstrated that speakers are readily able to synchronize when reading prepared texts together (Cummins, 2003). The degree of synchrony achieved is remarkable (typically with lags of about 40 ms), and does not improve much with practice. Synchronization in joint activity is, of course, quite common, but typically such activities are periodic in nature. Collaboration in working a two-man saw, in juggling, dancing or playing music, all rest on a periodic basis. Despite the naive impression of speech as 'rhythmic', it is practically never regularly periodic (Dauer, 1983). This poses the question, then, of how speakers manage to maintain such tight synchrony without extensive practice. In order to study the process and timecourse of synchronization, a method is required to continuously assess the degree of synchrony obtaining among speakers.

Assessing Synchrony

Previous work on synchrony among speakers has relied on a point-wise estimate of synchrony. In so doing, a few points are chosen which are clearly identifiable in both speech waveforms. The lag between speakers at each of these points serves as an instantaneous estimate of synchrony.

ExLing 2006: Proceedings of 1st Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece

Using this method, Cummins (2003) reported that asynchrony at phrase onsets was slightly greater than medially (62 ms, vs 40-44 ms medially). The point-wise method does not allow a continuous estimate of synchrony, and, in particular, it is difficult to see when, and how often, the lead changes between the two speakers.

Continuous assessment using Dynamic Time Warping

Dynamic Time Warping (DTW) is a well known algorithm, commonplace in speech recognition, which allows one to identify an optimal warping of one sequence onto a referent, with some common-sense constraints such as monotonicity and continuity (Meyers and Rabiner, 1981). Figure 1 (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.

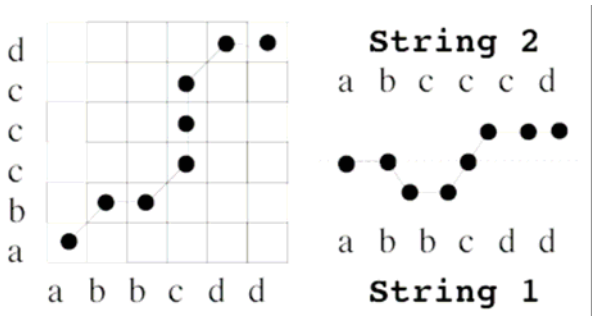

Figure 1: Illustration of standard Dynamic Time Warping path estimation (left) and conversion to an estimate of 'synchrony' (right).

The path identifies by DTW can be used to derive a continuous assessment of synchrony among speakers as follows:

We first represent the speech of each of two speakers as a sequence of equally spaced vectors. Conventional MFCCs appear to do just fine for this, though one could consider using any other parametric representation such as PLP, Rasta, or even LPC coefficients. One simplification we can make is to require that each speech sample be of the same length as the other (equal number of frames). This may involve including silence at the start of one or other sample.

We then generate the optimal warping path, similar to that in Figure 2. Notice that this path veers above and below the main SW-NE diagonal.

When it is above that diagonal, the second string ('abbcdd') is ahead of the first, and when it is below the diagonal, the first leads the second.

We now redraw the path, with the SW-NE diagonal as our time axis. Steps in the DTW algorithm which 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 Figure 1 (right panel). It can be seen directly that String 2 leads String 1 initially, and that the lead changes, just after the mid point of the two strings.

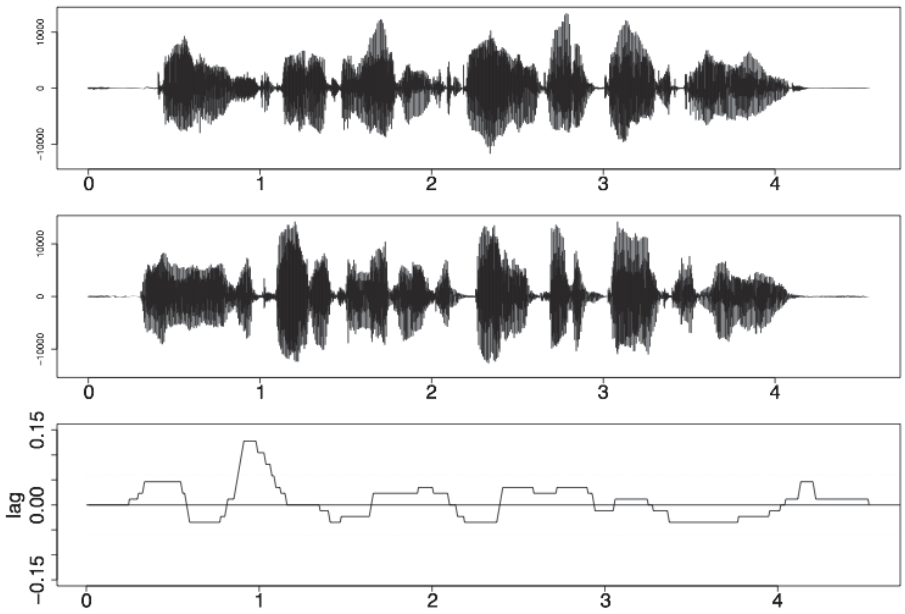

Figure 2: Two synchronous speech waveforms and associated synchrony estimate.

Figure 2 illustrates the result obtained for two phrases spoken in approximate synchrony by two male Irish speakers. When the synchrony estimate lies below the mid-line, the top speaker leads and vice versa. Dotted lines have been added at lags of +/- 60 ms. It can clearly be seen from the lower synchrony estimate that the two speakers are quite closely coupled, and that the lead is exchanged several times throughout this one short phrase (The phrase was "There is, according to legend, a boiling pot of gold at one end".)

Preliminary results

We have developed and tested this method of estimating synchrony on a subset of a database of synchronous speech we have recently collected (Cummins et al, 2006). Synchrony was estimated for 12 female and 12 male speakers reading the first paragraph of the 'Rainbow Text' in dyads. Our initial question was simply whether it was generally possible to identify a leader-follower relationship in a synchronous dyad, or whether the lead repeatedly changed from one to the other. In all 12 dyads, we found regular changes of the lead, and such changes typically happened several times within each individual phrase, as shown also in Figure 2. We were thus able to rule out the hypothesis that synchronous speaking is based on a simple leading-following relationship. Given that the vast majority of lags observed were of a magnitude less than 60 ms, and the shortest speech shadowing lags typically reported are of the order of 200 ms (Marlsen-Wilson, 1973), this is reassuring that synchronous speech does, in fact, require entrainment or coupling among speakers. However, it raises the question of how synchrony is maintained in the absence of a strictly periodic structure. Future investigations will address the following key issues:

[1] What timing information is extracted by one speaker in synchronizing with another, and

[2] To what extent does synchronous speech reveal unmarked, default speech timing, reflecting shared knowledge of what is marked and unmarked within a language/dialect.

Acknowledgements

The present work was funded by a grant from Science Foundation Ireland to the author.

References

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

Cummins, F. 2004. Synchronization among speakers reduces macroscopic temporal variability. Proc. 26th Annl. Meet. Cognitive Science Society, 304-9.

Cummins, F. and Grimaldi, M. and Leonard, T. and Simko, J. 2006. The CHAINS corpus: CHAracterizing INdividual Speakers. Proc. SPECOM'06. To appear.

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

Marslen-Wilson, W. 1973. Linguistic structure and speech shadowing at short latencies. Nature 244, 522-523.

Myers, C. S. and 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.