Assessing Rhythmic Differences with Synchronous Speech

Michael L. O'Dell, Tommi Nieminen, Liisa Mustanoja

University of Tampere, University of Jyväskylä

michael.odell@uta.fi, tommi.k.nieminen@jyu.fi, liisa.mustanoja@uta.fi

Abstract

In this study, we looked at speech rhythm in Finnish using the technique of synchronous speech developed by Cummins [5]. As predicted, synchronous reading resulted in a reduction of timing variation. The homogeneity achieved, however, did not necessarily represent average behavior, but often an extreme outside the range of performance in the solo reading. While the synchronous speech task was easy for speakers in general, there were clear differences in how difficult it was for speakers to synchronize their speech. These differences were not however related in an obvious way to differences between speakers themselves. It would appear that most of the work of synchrony was achieved at approximately the level of pause group; at finer levels speakers did not consistently adjust their timing to improve synchronization.

1. Introduction

Synchronous speech, in its simplest form, means a situation where two subjects simultaneously read aloud the same written text while trying to keep in synchrony. The experimenter gives a signal to begin, after which the task is completely up to the subjects. Previous experiments have demonstrated that the task is well within the capabilities of naïve subjects, and that they do well at it without any previous training. Typical asynchronies (time lags) reported during the task are on the order of 40 ms, with a 50% increase after pauses. [5]

Our initial motivation for this experiment was the fact that it has never been carried out with Finnish subjects. Finnish has been claimed to be different than English rhythmically, although Cummins has raised doubts about the whole rhythm dichotomy [5]. Furthermore, since we know that Finnish rhythm varies dialectally, we thought that synchronizing might be more difficult for subjects with different dialect backgrounds.

In addition we were interested in the following questions. Cummins reported that synchronous speech diminishes variation [3]. Would this be the case for our Finnish subjects as well? What kinds of individual speaker differences (such as dialect related timing differences) are reduced? At what level does synchronization occur? What factors influence difficulty of synchronizing?

2. Methods

We had a total of 20 subjects, 16 female and 4 male, in 10 pairs (A–J); however, because of a mistake in the experiment we were forced to leave out one of the pairs (I). We used a Sanako 300 language studio system (Tandberg Educational) where the subject(s) used one or two of the students' computers (recording stations) and the experimenter used the teacher's control station to allow subjects to hear each other via headset earphones and to start the recordings centrally. Students' recording stations were chosen to be equidistant from the teacher's station and as far apart from each other as possible.

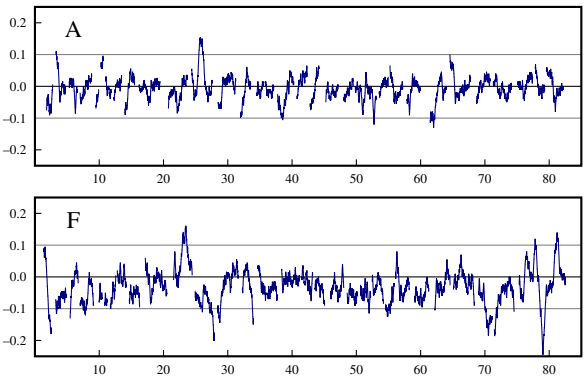
In our test setup, the subjects read the same text twice: first alone as if to practice it, then synchronously in pairs. The text used was a short excerpt from a children's story. Both readings were recorded in order to study what actually happens when speakers need to keep up with one another during the reading. In what follows we refer to these two readings as solo and synchronous recordings.

In order to assess the synchronization process we computed a Dynamic Time Warp (DTW) for each speaker pair using Praat [1]. All time warps were computed using MFCCs with a 20 ms window and time step of 10 ms, and restricting slope to a range of $[1/3, 3]$ (i.e. each 10 ms frame of one signal was matched to one, two or three 10 ms frames in the other signal). The DTW during pauses is essentially random and does not reflect the time relationship for the two speakers. Therefore pauses were marked by hand and portions of the DTW corresponding to pause for both speakers were deleted.

A DTW can be expressed in many forms; for our purposes the most useful was as a time series of average time vs. time difference measurements (also used in e.g. [6, 4]). Two examples of DTW in this format are shown in Fig. 1, with average time $(t_1+t_2)/2$ on the x-axis and time difference $(t_1-t_2)/2$ on the y-axis.

It should be noted that there are potential problems with DTW. We would like to follow speakers' phase differences at corresponding times. Instead, the characterization used here substitutes time differences at corresponding phases (phases corresponding to each other can be found with DTW or traditional segmentation procedures). This procedure will tend to underestimate the phase difference when either speaker is speaking relatively fast and overestimate the difference when either speaker is speaking relatively slow. The hope is that on average these errors will cancel out.

Figure 1: Examples of DTW in average time vs. time difference format (pairs A and F; time in seconds).



Another difficulty is that at present our procedure makes no provision for deletion, insertion or substitution of material, partly for ease of computation and partly because it is not clear how such discrepancies between speakers should be scored.

As a crude statistic of overall asynchrony we used the root mean square (RMS) of the time differences in the DTW. For a finer analysis of the synchronization process we modeled the DTW itself as a random walk[1] including an inward *bias* factor (attraction to zero difference) and a *persistence* factor (tendency to continue in the same direction) in addition to a general *diffusion* factor, cf. e.g. [8]. Diffusion is a general measure of noise in the process, inward bias is roughly analogous to the restoring force of harmonic motion and persistence to inertia. We used the WinBUGS program [10] to perform Bayesian estimation of these three model parameters (diffusion, bias, persistence) for each speaker pair. The parameter estimates for persistence were highly correlated with the diffusion estimates (cf. Fig. 5). This may be an artefact of restricting the slope of the DTW.

3. Results and discussion

Overall performance. It was obvious that all subjects did in fact achieve synchronization, for instance from the fact that total durations varied widely in the solo reading task but were nearly identical for the synchronous task (see Fig. 2). Another obvious feature which stands out clearly in Fig. 2 is the fact that subjects spoke more slowly in the synchronous task. The same result was observed by Cummins [2]. We take this to mean that synchronizing placed an extra burden on the speakers.

Results for the synchronous task in terms of RMS can be seen in Table 1, with speaker pairs ordered from "best" (lowest RMS) to "worst". These are similar to results from previous studies on English, but with slightly greater asynchrony (the RMS time difference used here should be doubled to compare with the time lags reported by Cummins [2]). The difference may be due to the slightly different experimental setup, e.g. the fact that our subjects could not see each other (Cummins [2] reported that synchrony was slightly worse in this condition) and heard each other only over earphones.

A finer comparison of subject pairs' performance is possible using the parameter estimates for random walk diffusion, persistence and inward bias shown in Fig. 5. It is obvious from these parameters that there were clear differences in the performance of different subject pairs. It is not so obvious, however, what caused these differences, other than just individual variation.

We examined various differences between speakers in a pair to see what factors might influence synchronization difficulty. Comparing our worst pair (F) with our best pair (A) one observation is that A had the lowest mean age (25 yrs) while F had the highest mean age (52) so age may be a factor here, although it is far from consistent (cf. Table 1). In addition pair F consisted of our slowest speaker paired with our fastest speaker (judging from the solo task), which may have made synchronizing more difficult. However pair D, which was fairly close in solo reading rate (and also had the second lowest mean age, 26½), had the second to worst RMS.

One of our hypotheses was that dialect differences might increase synchronization difficulty. In addition to their rate difference, the speakers of pair F also came from different dialect backgrounds, which may have hindered their synchronization. However, pair A had a roughly equivalent dialect difference and managed to get the best RMS score.

There was no consistent effect on RMS associated with whether the two subjects in a pair knew each other previous to the experiment, or to whether speakers differed in sex.

[1] There is an intimate connection between random walk analysis and dynamic systems analysis. For noisy oscillators Pikovsky *et al.* noted that in general, "... during the synchronous epochs the phase difference resembles a random walk motion ..." [9]. The random walk approach seems promising and we hope to develop it much further in the future.

Figure 2: Total duration (right) and total pause duration (left) of utterances (in seconds) for solo task (upper open circles) and synchronous task (lower filled circles).

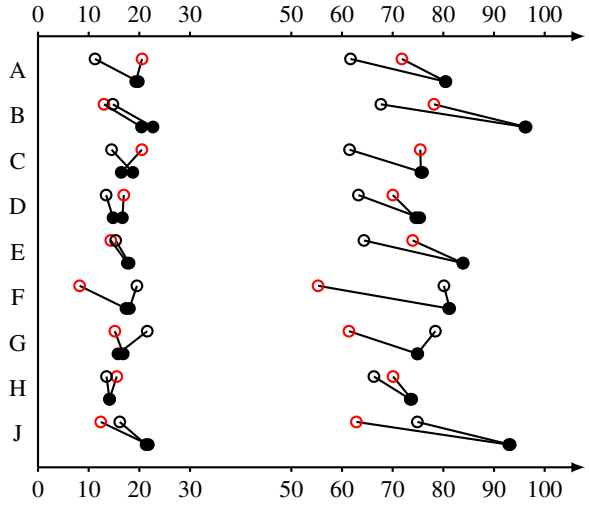


Table 1: Synchrony scores and various subject pair attributes.

subj. pair	RMS	acquainted	ages (mean)	sex
A	36.9	no	23, 27 (25)	♀♀
C	37.4	yes	42, 48 (45)	♂♀
E	40.6	yes	34, 48 (41)	♀♀
H	41.3	no	31, 33 (32)	♀♀
J	44.1	yes	34, 40 (37)	♂♀
G	45.1	yes	39, 47 (43)	♀♀
B	52.9	yes	35, 64 (49½)	♀♀
D	54.7	no	25, 28 (26½)	♀♀
F	60.7	no	43, 61 (52)	♂♀

Pausing behavior. As can be seen in Fig. 2, total pause time for pairs was also closer in the synchronous task. Interestingly, total pause time did not consistently increase, indicating that the slowing mainly affects the running speech between pauses.

In addition to slowing down in the synchronous task, all speakers increased the number of pauses (with one exception who maintained the same number). For speakers with relatively few pauses in the solo task the number increased dramatically in the synchronous task (Table 2). In one sense this is paradoxical, since it would seem that there is very little information about mutual phase difference available to speakers during pauses as compared to running speech. (There is perhaps more than might be expected, e.g. sounds of breathing and click-like sounds as speakers open their mouths to resume speaking.) On the other hand, more frequent pausing, *provided that pauses are coordinated*, reduces the amount of synchronizing necessary between pauses. In any case, the *difference* in number of pauses for each pair was smaller in the synchronous task compared to the solo task, with the notable exception of pair J, as can be seen in Table 2. This means that pausing behavior was more regular in the synchronous task, but it was by no means an average of the be-

havior in the solo task. It might be more accurate to say speakers executed a maximal number of pauses in the synchronous task.

It may be that overall slowing is related more directly to increased pausing. Because there are more pause groups in the synchronous task, pause group duration decreases on average while total duration increases. The situation thus calls to mind the phenomenon of rhythmic gradation (*rhythmische Abstufung*), which has been one motivation for our Coupled Oscillator Model (COM) [7]. We are currently expanding our model to handle pausing and interspeaker coupling.

Pair J was obviously a special case in regard to pausing. Their strategy seemed to consist of very frequent pauses in the synchronous task, resulting in many more pauses than any other pair. While their reading thus does not sound fluent, they did achieve a medium range RMS score for synchrony with this technique.

Table 2: Number of pauses.

subject pair	solo 1	synchronous	solo 2
A	31	33 — 32	20
B	26	35 — 34	32
C	26	30 — 31	29
D	28	31 — 32	26
E	27	31 — 33	30
F	21	32 — 34	31
G	22	32 — 32	30
H	31	31 — 29	23
J	30	50 — 44	27

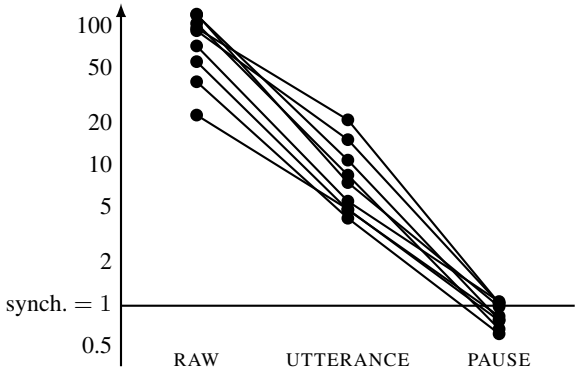
At what level does synchronization occur? One indication of different behavior at different levels is the effect of synchronizing on pauses. We used the solo recordings of speaker pairs to approach this question in another way. Using the Manipulation feature of Praat we artificially stretched the solo recordings for each speaker pair to match the average durations of their synchronized recording at various levels based on hand marked units. We then applied DTW to these artificially synchronized pairs and compared the result with the actual performance in the synchronous task.

The results are shown in Table 3 and Fig. 3. When no adjustment is made (RAW) the RMS is very large because the speakers have typically spoken at different rates. When rate is adjusted by uniformly stretching to make beginning and end match (UTTERANCE) the RMS is considerably less, but still several times larger than the reference level of the synchronous task. However, when pause groups are also matched in duration (PAUSE), the RMS falls to the reference level or below. This would seem to indicate that most of the work of synchronization was accomplished at the pause group level, that is, by coordinating pausing and speaking rate between pauses.

Table 3: RMS of DTW (ms) for synchronous task and solo utterances matched at various levels.

subject pair	synch.	RAW	UTTERANCE	PAUSE
A	36.9	3478.4	574.2	38.3
B	52.9	2989.3	258.0	44.5
C	37.4	4569.9	415.0	29.7
D	54.7	2217.3	232.3	34.4
E	40.6	2975.2	228.6	43.5
F	60.7	7521.2	464.7	59.8
G	45.1	4791.4	390.5	30.7
H	41.3	964.9	203.4	32.3
J	44.1	4334.6	952.3	46.0

Figure 3: RMS of DTW relative to synchronous task (= 1) for solo utterances artificially synchronized at various levels.



Half-long vowel. A well known feature of many Finnish dialects (but notably not in the Tampere region where most of our subjects were from) is the so called "half-long" vowel, referring to the phonetic lengthening of a short CV syllable after another CV syllable. The ratio of V_2 duration to V_1 duration in CVCV structures is traditionally used to quantify this dialect difference [11, 12].

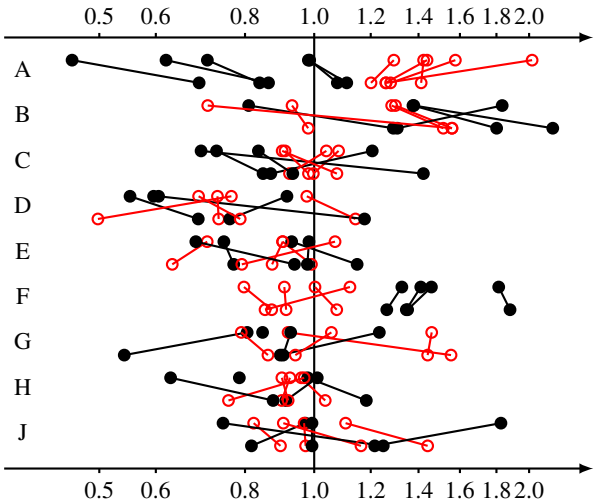
Measurements for the subjects in the various conditions are shown in Fig. 4. Two pairs (A and F) exhibited a consistent difference in the half-long vowel feature in their solo readings. The differences observed were compatible with the speakers' background information.

For pair A, the difference in V_2/V_1-ratio diminished consistently in the synchronous reading condition but did not disappear. It should be noted, however, that this diminished difference did *not* consistently improve local synchronization, but was sometimes achieved at its expense.

One possible explanation of this behavior is that both speakers spoke more "expressively" in the solo task, and due to their dialect difference this extra expressiveness often stretched V_2 for the first speaker but V_1 for the second speaker in CVCV-structures.[2] Whatever the mechanism involved, the syn-

[2]The COM [7] does in fact predict that V_2/V_1-ratio should vary as a function of prosodic organization.

Figure 4: V_2/V_1-ratio for words in the solo task (upper points) and synchronous task (lower points).



chronous task does appear to have reduced variation for this pair of speakers.

For all other pairs, including pair F with a consistent difference in V_2/V_1-ratio, neither speaker showed a significant reduction in variation for the synchronous reading. In both cases it would appear that synchronization did not occur at a sub-word level.

4. Conclusions

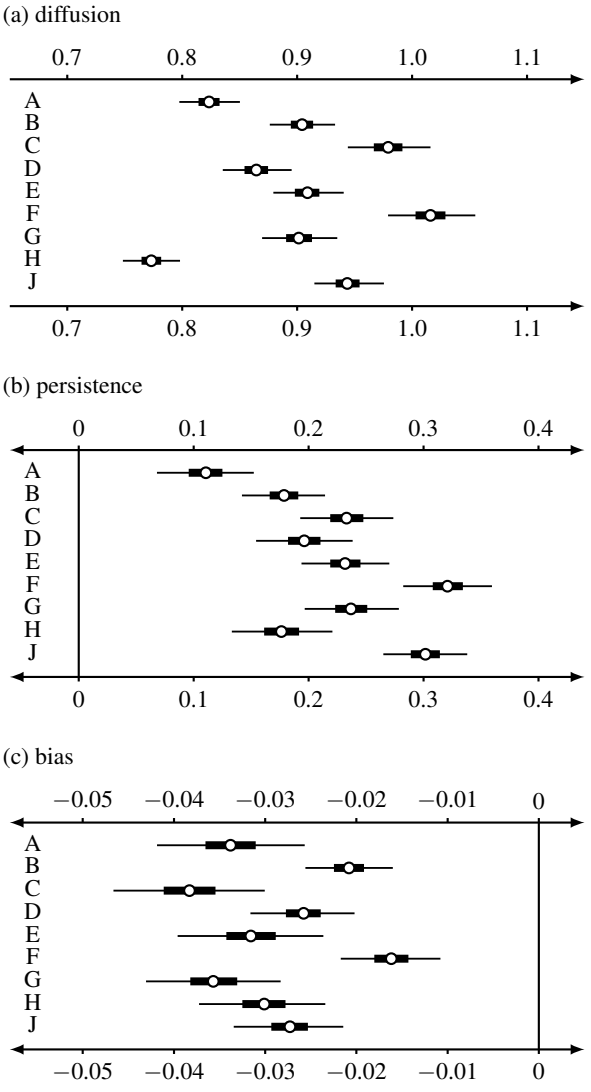
The synchronous speech task was easily and successfully performed. Synchronizing obviously placed an extra burden on subjects because it slowed their average speaking rate without exception, sometimes quite dramatically.

As predicted, there was a reduction of variation in the synchronous task. The homogeneity achieved, however, did not necessarily represent average behavior, but often an extreme outside the range of performance in the solo task.

There were clear differences in the performance of different speakers (and perhaps in the strategies they employed). These differences were not however related in an obvious way to speaker differences, with the possible exceptions of age and natural speaking rate.

It would appear that most of the work of synchrony was achieved at approximately the level of pause group, whereas for smaller units speakers did not adjust their timing directly as a result of synchronization. In particular, speaker differences in word level timing due for instance to dialect differences were not adjusted to achieve better synchronization.

Figure 5: Credible intervals for random walk parameters. The dot indicates the median of the posterior distribution, the thick line indicates the 50% CI and the thin line indicates the 95% CI.



5. References

[1] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.1.20) [Computer program], 2009. Last retrieved Oct. 31, 2009, from http://www.praat.org/.

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

[3] F. Cummins. Synchronization among speakers reduces macroscopic temporal variability. In *Proc. 26th Annl. Meeting of the Cognitive Science Society*, pages 304–309, 2004.

[4] F. Cummins. Measuring synchronization among speakers reading together. In *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece*, 2006.

[5] F. Cummins. Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, 37(1):16–28, 2009.

[6] M. O'Dell. *Intrinsic Timing and Quantity in Finnish*. Acta Universitatis Tamperensis 979. Tampere University Press, 2003.

[7] M. O'Dell and T. Nieminen. Coupled oscillator model for speech timing: Overview and examples. In M. Vainio, R. Aulanko, and O. Aaltonen, editors, *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*. Peter Lang, 2009.

[8] C. Patlak. Random walk with persistence and external bias. *Bulletin of Mathematical Biology*, 15(3):311–338, 1953.

[9] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2001.

[10] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, Version 2.10*. Cambridge: Medical Research Council Biostatistics Unit, 2005.

[11] K. Wiik. Suomen murteiden vokaalien kestoista. In O. Aaltonen and T. Hulkko, editors, *XIII Fonetiikan päivät Turku 1985 / XIII Meeting of Finnish Phoneticians — Turku 1985*, number 26 in Publications of the Department of Finnish and General Linguistics of the University of Turku, pages 253–317, 1985.

[12] R. Ylitalo. *The Realisation of Prominence in Three Varieties of Standard Spoken Finnish*. Acta Universitatis Ouluensis, Series B, Humaniora 88. University of Oulu, 2009.