# Synchronization Among Speakers Reduces Macroscopic Temporal Variability

**Fred Cummins (fred.cummins@ucd.ie)**
Department of Computer Science
University College Dublin
Belfield
Dublin 4
Ireland

**Abstract**

A recent method for restricting inessential variation in speech is presented. Synchronous Speech is obtained by having two subjects read a prepared text in synchrony. Past results demonstrate that this is easy for subjects to do, and that some prosodic variability is greatly reduced when reading synchronously. Particular advantage has been found in the analysis of pauses and fundamental frequency variation, where synchronous speech has been demonstrated to exhibit markedly less inessential variability, thus furthering analysis and modeling. Here, duration ratios within a phrase are compared across synchronous and solo conditions. Variables associated with global timing and with the relationships between phrases are shown to be more consistent in the synchronous condition, while smaller units are not noticeably affected by the speaking condition. No systematic artifacts are found to be introduced by asking subjects to read in synchrony.

## A Method for Restricting Variability

Synchronous Speech is obtained with the simple expedient of having two subjects read a prepared text together, with the minimal instruction to attempt to maintain synchrony (Cummins, 2002). The reason for constraining subjects in this manner is perhaps best appreciated by analogy with the difficult task of attempting to reconstruct a musical score, based only on a recording of a specific musician (Heijink et al., 2000). This task is interestingly similar to the work of the theoretically minded phonetician who attempts to uncover control and timing information, along with combinatorial units, from the continuous stream of speech.

If one were faced with this task, it is worth considering which musician would give one more tractable data: the soloist, or the 14th violin player in the string section. Neither will reproduce the durations (or pitches) specified in their score exactly, of course, due to the inherent underspecification of the score. Both players will overlay some inherent biophysical noise, along with conventional timing variability, such as the predictable decellerando at the end of a phrase. The soloist will add additional complexity, however, in keeping with her role as the expressive focus in performance, making the inverse mapping from the recording to the score considerably more difficult.

Now return to the position of the laboratory phonologist (or theoretical phonetician). An overarching goal is to deduce the units of control which relate to the linguistic message being uttered, and to uncover their mutual relations. This is not so different in kind from the above musical analog, though additional levels of complexity undoubtedly arise. Signal variability which is related to the linguistic content is relevant, while (for many purposes) one might like to find a way to reduce or exclude variability of para- or non-linguistic origin. The approach which I and colleagues have recently been following is to constrain the speaker to speak in time with another co-speaker. For this purpose, speakers read through a given text silently to familiarize themselves with it, and then commence reading together on a signal from the investigator. For many purposes, recording using near field head-mounted microphones onto the left and right channels of a single stereo file is sufficient to separate the two speakers while preserving the relative temporal alignment of speech events.

We call speech collected in this manner Synchronous Speech, and both the task and the product have provided us with much food for thought (Cummins, 2001; Cummins, 2002; Cummins and Roy, 2001; Cummins, 2003). In this paper, I will summarize those findings which have best revealed the advantages of this novel method, then provide some new results which examine the variability of intervals below the whole phrase, and finally provide pointers to areas I believe might benefit from adoption of the method.

## Properties of Synchronous Speech

Synchronizing with a co-speaker, without extensive practice, turns out to be simple for subjects to do (Cummins, 2002; Cummins, 2003). After reading through a simple text once, and being given a start signal, subjects typically manage to keep inter-speaker lags to average values of around 60 ms at phrase onsets, and 40 ms or less after the first syllable or so. Rather surprisingly, extensive practice at the task does not improve the degree of synchrony significantly (Cummins, 2003), although with repeated readings of the same text, and with the same co-speaker, a slight improvement may be detected. Visual contact with the co-speaker does seem to have a small beneficial effect on synchrony, even though subjects are typically attending to a read text in front of them (Cummins, 2003).

In experiments done to date, speakers have not been carefully matched for familiarity, intrinsic speaking rate

or volume. Among the heterogeneous pairs of speakers we have studied to date, most appear to be collaborating, producing speech at a relatively slow rate (but faster than some of the slowest speakers' natural reading rate). We have not yet (in over 60 pairs of speakers) found a speaking pair in which one speaker consistently lagged behind the other. Rather, they seem to genuinely speak together, with a high degree of synchrony.

## Phrasing and Pauses

One of the first properties of Synchronous Speech we noticed, was that phrasing, i.e. the division of a long stretch of speech into intonational units separated by pauses, appeared to be much more consistent in Synchronous Speech than in control readings done alone. In an initial pilot with 4 speakers, we found that in 48 'solo' readings, pauses occurred at points other than major expected phrase breaks 48 times (Cummins, 2002). By contrast, in Synchronous Speech, there were only 4 such idiosyncratic pauses in 24 paired readings.

These findings have been extended in the studies of pauses in Synchronous Speech (Zvonik and Cummins, 2002; Zvonik and Cummins, 2003), who found that inter-speaker variability in pause duration was greatly reduced in Synchronous Speech, compared with 'solo' speech. The reduced variability allowed the identification of a quantitative relationship between pause duration and the length (in syllables) of the preceding phrase—a relationship which was obscured in the rather more variable solo data. Specifically, we found a restricted distribution of clauses of less than 300 ms length. These pauses were far more likely to occur when the preceding phrase was relatively short (less than 11 syllables long)[1]. We examined pause duration in readings by 6 speakers (3 pairs) of 19 short texts (13 distinct authors). Table 1 shows the proportion of pauses in each environment (preceding phrase long or short, following phrase long or short) which were below 300 ms. The preponderance of short pauses in an environment following a short phrase is clear.

Table 1: Number of pauses of duration less than 300 ms as a function of the length of the surrounding Intonational Phrases. For IPs, 'short' is here taken to mean less than or equal to 10 syllables. Reproduced from Zvonik (2004, unpublished PhD thesis).

| Preceding IP | Following IP | Proportion of Short pauses |
|---|---|---|
| short | long | 0.32 |
| short | short | 0.39 |
| long | short | 0.11 |
| long | long | 0.06 |

[1]An earlier observation in Zvonick and Cummins (2003) that a similar relationship obtained between pauses and *following* phrases is probably an artifact of the idiosyncratic text used in that study

## Fundamental Frequency Variability

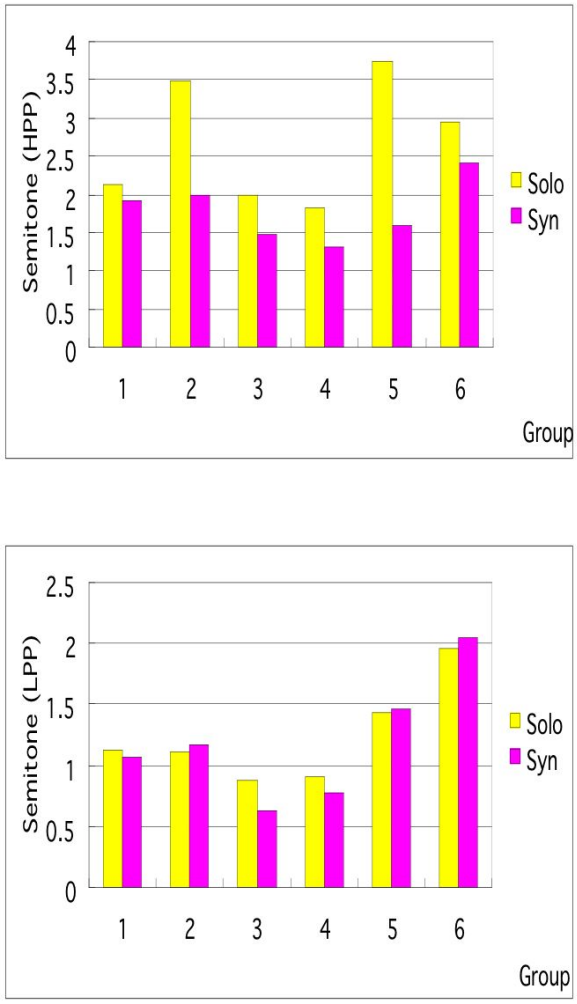


Figure 1: Difference in F0 peaks (HPP, top) and valleys (LPP, bottom) between speakers in a pair in solo and synchronous readings. All F0 values converted to semitones before analysis.

Given the above findings on phrasing and pauses, it seemed natural to examine the effect of speaking synchronously on other prosodic variables. To this end, we recorded six pairs of female speakers reading simple fairy tales (Wang and Cummins, 2003). We identified peaks and valleys in the intonation contour (HPP: High Point of Pitch, LPP: Low Point of Pitch), and looked to see whether these variables were affected by speaking synchronously. We found that the peaks were considerably more highly correlated across speakers within a pair in the synchronous condition (mean $r = 0.72$, s.d = 0.08) than in the solo condition (mean $r = 0.59$, s.d.=0.07). This did not hold for valleys (Synchronous: mean $r = 0.43$, s.d.=0.08; Solo: mean $r = 0.30$, s.d.=0.17). Figure 1 shows the differences in pitch between speakers of a

pair for both conditions. From the peaks (top panel), it can clearly be seen that there is a substantial reduction in inter-speaker differences in the synchronous condition, while the difference is slight or nonexistent for the valleys. This suggests that there might be a difference in the amount of free variability which speakers may employ in the absolute placement of H and L tones, a possibility which has been suggested independently elsewhere (Liberman and Pierrehumbert, 1984).

## On Synchronization

How do speakers manage to synchronize so efficiently? One possibility which can be discounted already is that one speaker provides the lead, and the other follows. As mentioned above, we have yet to find a dyad in which a single leader could be identified. The very short lags between speakers also seem to preclude an explanation along these lines, as the typical lag of 40 ms is simply too short to allow perceptually guided correction while speaking.

Most mathematical models we have of synchronization are based on populations of oscillators (Glass and Mackey, 1988; Strogatz and Stewart, 1993). The mathematics of coupling among periodic sources is complex, but by and large tractable. Powerful predictive models have been constructed of such phenomena as juggling (Beek and Lewbel, 1995), heart cells (Mirollo and Strogatz, 1990), etc. Some have chosen to restrict the term 'synchronization' to the "adjustment of rhythms of oscillating objects due to their weak interaction" (Pikovsky et al., 2001, p. 8). Certainly, speech production is not oscillatory or periodic in anything but the loosest sense, and so the known mechanisms of entrainment among periodic sources can not be invoked here.

The answer, it seems to us, must lie in the shared knowledge speakers have of what is essential and what is redundant, or optional, in the modulation of the speech organs. Speakers of the same dialect must have control structures in common that govern the production of, and temporal relations among, the discrete units of speech. Little has yet been ascertained about the degree to which these putative control structures must coincide among such speakers. Certainly, many of the processes of diachronic change in language which have been described suggest that differences among speakers are not particularly rare, e.g. the age-related differences observed among speakers of Brazilian Portuguese by Major (1981). Nonetheless, the efficiency of communication dictates that most such structures must be shared among speakers. Although we do not have privileged access to the units and processes of speech production, speakers do seem to be able to modify their speech in direct response to the task demands, suggesting that the method of synchronous speech elicitation is a promising technique for tapping speakers' unconscious knowledge of the process of speaking.

An acknowledged limitation of the present is that much of the prosodic richness of spontaneous speech, specifically that associated with information management, speaker's attitudes, etc, is clearly not present in Synchronous Speech. The method requires the reading of a prepared text, and the additional constraint of synchrony places strict limits on the degree of personal interpretation and expression which a speaker can employ. Some of what is shorn away can correctly be considered to be meaningful prosodic structure. This limitation has an upside, however, as the timing which remains is still an immensely rich object of study, and those aspects of speech timing which are preserved (very many!) can be seen more clearly in the absence of the other additional sources of variation in speech.

The closest parallels to the demands of the Synchronous Speech task appear to be met in studies of synchronization among ensemble musicians (Rasch, 1979; Rasch, 1988), and the largely unstudied process of synchronization among dancers. In studying ensemble playing, Rasch (1979; 1988) used the standard deviation of differences in onset time of simultaneous notes in many voices as an index of asynchrony and noted typical values of 30 to 50 ms. The more direct measure of mean lag used in our studies of two voices at a time have provided values of approximately 40 ms.

## Effect of Synchronization on Proportional Durations

An important question about the process of synchronization is whether it introduces artifacts into the temporal structure of speech, or conversely, whether it merely serves to reduce variability and reveal a shared understanding of temporal structure among speakers. Artifacts might be revealed in the systematic alteration of proportional durations, as would be the case if, e.g., unstressed syllables were found to be less reduced, and hence longer compared with stressed syllables. Any such systematic alteration of the durational properties of speech would severely limit the potential of Synchronous Speech to inform researchers about the properties of speech in the more general case.

As one way to investigate this, we here examine means and variances of a variety of interval ratios. By looking at ratios rather than durations, we better capture the relational properties of speech, and simultaneously avoid the difficult issue of rate normalization.

### Methods

Readings of the first paragraph of the Rainbow Text were obtained from 27 pairs of speakers, as described in Cummins (2003). Each subject provided one reading alone and one with a co-speaker, obtained during a larger corpus collection exercise. The second sentence of the passage was chosen for detailed analysis. It reads "The rainbow is a division of white light into many beautiful colors". Reliably identifiable points in the waveform were chosen for measurement (stop releases, V-nasal transitions, etc). Figure 2 illustrates a representative set of measurement points for one recording.

Each variable studied was a ratio of two intervals, and comparisons were made of both mean values (using t-tests) and variability (F-test, one-sided, with the hypothesis of reduced variability in Synchronous Speech). Each

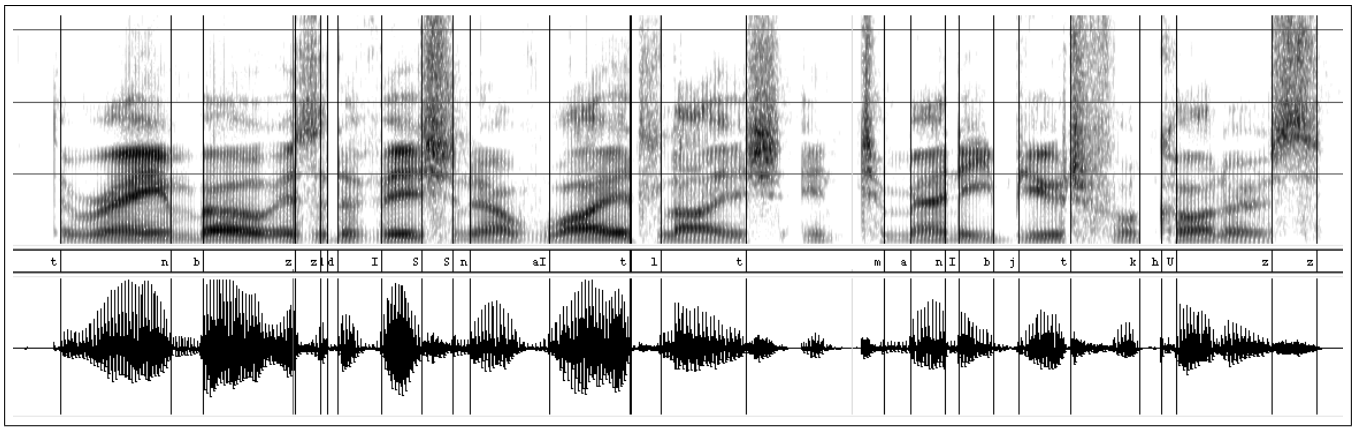


Figure 2: Measurement points for a single recording.

Table 2: Comparison of interval ratios in Synchronous Speech and Solo Speech. Intervals are taken from the sentence "The rainbow is a division of white light into many beautiful colors". Segments at measurement points are capitalized and intervals used are in bold face.

| No. | Variable Type | Interval 1 | Interval 2 | $t$ (df) | $p(t)$ | $F$ (df) | $p(F)$ |
|---|---|---|---|---|---|---|---|
| 1 | juncture/phrase | ligh**T into M**any | rai**Nbow...ighT** | 0.15 (93) | n.s. | 1.67 (50,45) | $< 0.05$ |
| 2 | juncture/phrase[1] | ligh**T into M**any | **Many...colorS** | 0.36 (90) | n.s. | 2.0 (50,45) | $< 0.01$ |
| 3 | phrase/phrase | **Many...colorS** | rai**Nbow...ighT** | -0.27 (89) | n.s. | 2.16 (50,45) | $< 0.01$ |
| 4 | unstressed/stressed syllables | **NY** | **MA** | 1.66 (92) | n.s. | 1.06 (47,45) | n.s. |
| 5 | onset segment/word | **C**olors | **ColorS** | 0.7 (94) | n.s. | 1.04 (50,45) | n.s. |
| 6 | stressed vowel/word | div**I**sion | **DivisioN** | -0.56 (79) | n.s. | 0.56 (43,42) | n.s. |

ratio was expressed as the smaller value divided by the larger, and distributions were checked visually for approximate normality. Adjustment to degrees of freedom as appropriate for distributions of unequal variance was made using the Welch approximation.

Previous results had demonstrated that macroscopic phrasing (the division of an utterance into intonation phrases, the placement and duration of pauses) was significantly less variable in Synchronous Speech. No analysis of the durations of shorter intervals had yet been done. The possibility that duration ratios might be significantly different in Synchronous Speech was of interest, as this would suggest that the process of synchronization introduces artifacts into speech timing, and speech so obtained cannot be considered as unproblematically related to conventional speech. On the other hand, it was of interest to see whether the previous indications of reduced inter-speaker variability would be found with shorter, intra-phrasal, units also.

## Results

**Major Syntactic Juncture:** The sentence studied contains one major syntactic juncture, between "white light" and "into many". The most reliably measurable interval spanning this juncture was delimited by the obstruent occlusion at the end of "light" and the first nasal onset of "many". We examined the ratio of this interval to the duration of each of the surrounding phrases (From the onset of "many" to the fricative onset in "colors" and from the nasal of the initial "rainbow" to the obstruent closure in "light"). Rows 1 and 2 of Table 2 show that the relative duration of the interval spanning the juncture is similar across conditions, whether one takes the preceding or the following phrase as a referent, but the variability of this ratio is substantially reduced in Synchronous Speech.

**Phrase Length:** The durations of the two major phrases ("The rainbow is a division of white light" and "into many beautiful colors") were compared. No difference in the ratios was discernible, but the variability of the ratio was greatly reduced in Synchronous Speech (Row 3, Table 2). This result may be attributable to a greater constancy of speaking rate in the synchronous condition.

**Stressed and Unstressed Syllables:** The word "many" provides unambiguous measurement points which make a comparison of the duration ratio of an unstressed to a stressed syllable within the same word possible. Row 4 of Table 2 provides the results of the analysis in which no significant differences in either ratios or ratio variability was found across conditions.

**Segment 1: Onset.** The length of the initial consonant (closure to voicing onset) in "colors" as a propor-

tion of the word length (/k/ closure to /z/ onset) was examined. Row 5 in Table 2 shows that mean ratio was not different across conditions, nor was ratio variability different in Synchronous Speech.

**Segment 2: Vowel:** The length of the stressed vowel /I/ in "division", expressed as a proportion of the word duration (stop closure to nasal release) was also studied. Again, neither ratio means nor variability was significantly different across conditions.

## Discussion

The variables examined in the present study spanned a range of temporal scales and phonological structures. Those variables which were most directly related to macroscopic temporal structure (i.e. phrasing) all showed significant reduction in Synchronous Speech without any discernible change in mean values. The variables which describe smaller intervals showed no effects. In none of these cases was the proportional duration indexed by the ratio found to differ in its mean value between Synchronous Speech and solo speech, nor was the variability affected by speaking condition.

These results accord well with previous findings on Synchronous Speech and suggest areas for further study. No evidence has yet been found that speaking synchronously produced durational artifacts. The only properties of Synchronous Speech which have been reliably identified to date are a demonstrable increase in the consistency of global timing and phrasing (including intonation) across speakers. Those variables which exhibit substantially reduced variability in the Synchronous Speech condition are those most closely tied to timing at a global level, in which whole phrases are coordinated with respect to one another. Neither the unstressed/stressed syllable comparison, nor the segmental variables exhibited any difference in mean value or variability, suggesting that at a finer timescale there is little if any change to speakers' timing when speaking in synchrony with another person.

Some of the reduction in variability which is observed may be due to the forced maintenance of a constant speech rate. The indexing of speech rate is a notoriously difficult problem. Crude indices such as articulation rate, measured in number of syllables or segments per second, do little to match speakers' intuitions of a continuous abstract 'rate' of speaking. The constraint of speaking together with another speaker places severe limits on the freedom of the speaker to continuously modulate this abstract speaking rate, as any modification must be predictable for the co-speaker also.

[2]Alone among the distributions used herein, the ratios of the juncture to the second phrase were not normally distributed in the solo condition, but were skewed right. A Wilcox rank sum test substantiated the findings of the parametric test.

## Further Exploitation of Synchronous Speech

The earlier examples of studies of pauses and intonation illustrate two different ways in which Synchronous Speech offers a novel approach to the analysis of variability in speech. In the former case, Synchronous Speech provided cleaner data than solo read speech, allowing the identification of temporal regularities which would otherwise be obscured. Synchronization among speakers is a simple and effective way of obtaining high-quality spoken data which is stripped of inessential sources of variability. This interpretation of the character of Synchronous Speech is supported by the durational measurements reported here for the first time. No difference in the fine structure of speech was observed, but variability associated with macroscopic timing was reduced. Future work should also examine finer gradations in the prosodic hierarchy: are there changes at levels between the intonational phrase and the syllable?

In the intonation study, the difference between solo speech and Synchronous Speech was itself a source of information about essential variability. By comparing Synchronous Speech with solo speech, we obtain a partition of variability into essential and inessential parts. It is tempting to associate the essential variability, preserved in Synchronous Speech, with linguistic sources, and inessential variability, absent in Synchronous Speech, with para- and non-linguistic sources, but this step is probably premature at this stage. To gauge the reliability of this attribution of the source of variation to linguistic or nonlinguistic origins will require further targeted research. However, the prospect of obtaining this partition potentially opens up new avenues of exploration for both kinds of variation.

Important information about the quality of Synchronous Speech will come from testing to see if subjects can perceive artifacts in Synchronous Speech, or indeed distinguish it from normal speech in perception tests. This work is ongoing.

One tantalizing possibility is the identification of parameters of free variability which might be exploited in the synthesis of expressive or characterful voices. Synthetic voices are bland, while carefully tailored voices which convey some sense of personality (emotion, expression) are laborious to construct. Adding random variation to synthesis parameters does nothing but reduce intelligibility. However an analysis of the properties of Synchronous Speech and a comparison with solo speech may inform voice designers about those parameters which they are relatively free to vary for expressive purposes. For example, the above study on intonation strongly suggested that an excitable voice might result from an expanded dynamic range of intonation in which the high targets are modified, but not the low targets.

Obtaining synchronous speech is not difficult. All studies of the properties of Synchronous Speech to date have suggested that the principal effect of the constraint of speaking together is to reduce idiosyncratic variability, leaving the essential quality of the speech untouched. This seems to offer two things to the experimental pho-

netician. Firstly, it provides an easy route to cleaner (less variable) data, for studies in which non-linguistic variability is unwanted. Secondly, it may provide a principled manner of partitioning variability, so that intrinsic variability which cannot be voluntarily avoided is retained, while superfluous variability is removed, thus allowing the differentiation of two kinds of variability in speech.

## Acknowledgements

This work has been done with the help of Elena Zvonik and Bei Wang. Funding was provided by an Irish Higher Education Authority grant for collaborative research between Irish Universities and Colleges and Media lab Europe.

## References

Beek, P. J. and Lewbel, A. (1995). The science of juggling. *Scientific American*, pages 92–97.

Cummins, F. (2001). Prosodic characteristics of synchronous speech. In Puppel, S. and Demenko, G., editors, *Prosody 2000: Speech Recognition and Synthesis*, pages 45–49, Krakow, Poland. Adam Mickiewicz University.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, 3(1):7–11.

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

Cummins, F. and Roy, D. (2001). Using synchronous speech to minimize variability in pause placement. In *Proceedings of the Institute of Acoustics*, volume 23 (3), pages 201–206, Stratford-upon-Avon.

Glass, L. and Mackey, M. C. (1988). *From Clocks to Chaos*. Princeton University Press, Princeton, NJ.

Heijink, H., Desain, P., Honing, H., and Windsor, W. (2000). Music representation–make me a match: An evaluation of different approaches to score-performance matching. *Computer Music Journal*, 24(1):43–56.

Liberman, M. Y. and Pierrehumbert, J. B. (1984). Intonational invariance under changes in pitch range and length. In Aronoff, M. and Oehrle, R. T., editors, *Language sound structure: studies in phonology presented to Morris Halle*, pages 157–233. MIT Press.

Major, R. C. (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics*, 9:343–351.

Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal of Applied Mathematics*, 50(6):1645–1662.

Pikovsky, A., Rosenblum, M., and Kurths, J. (2001). *Synchronization: A universal concept in nonlinear sciences*. Number 12 in Cambridge Nonlinear Science Series. CUP.

Rasch, R. A. (1979). Synchronization in performed ensemble music. *Acustica*, 43:121–131.

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In Sloboda, J. A., editor, *Generative Processes in Music*, pages 70–90. Clarendon Press, Oxford.

Strogatz, S. H. and Stewart, I. (1993). Coupled oscillators and biological synchronization. *Scientific American*, pages 102–109.

Wang, B. and Cummins, F. (2003). Intonation contour in synchronous speech. *Journal of the Acoustical Society of America*, 114(4(2)):2397.

Zvonik, E. and Cummins, F. (2002). Pause duration and variability in read texts. In *Proc. ICSLP*, pages 1109–1112, Denver, CO.

Zvonik, E. and Cummins, F. (2003). The effect of surrounding phrase lengths on pause duration. In *Proceedings of EUROSPEECH*, Geneva, CH. to appear.