Explorations in the prosodic characteristics of synchronous speech, with specific reference to the roles of words and stresses

Fred Cummins[1], *Judit Varga*[2]

[1]UCD School of Computer Science and Informatics, University College Dublin
[2]School of Psychology, Trinity College Dublin
fred.cummins@ucd.ie,vargajudit89@gmail.com

10.21437/SpeechProsody.2014-24

Abstract

We examine the prosodic characteristics of read speech produced alone or in synchrony with a co-speaker in English. Previous work has demonstrated a marked difference between these two speaking conditions in Mandarin, but not English. We employ word lists that are either simple sequences of trochees, or complex lists with regular stress alternation but irregular word boundaries. Inter-onset intervals are examined and no major differences between solo and synchronous interval sequences are found. Viewed from the perspective of two generative models, however, there is weak evidence for some small difference in the dependence of interval duration on serial position.

Index Terms: synchronous speech, stress timing, word lists, joint speech

1. Introduction

Synchronous speech is a laboratory variant of the more general phenomenon of joint speech, where speakers utter the same words in unison [1, 2]. Familiar ethological examples include collective prayer, and the chant of protesters. In a synchronous speaking task, novel texts are typically employed, and subjects do not have difficulty in reading these while keeping in time with one another. The speech so produced is perhaps best characterized as unmarked. To conform to the task demands, speakers must shear their speech of unpredictable temporal variation, stemming, for example, from dramatic expression, or idisyncratic phrasing. This results in speech in which linguistic contrasts are preserved, but many sources of variability that serve to make phonetic analysis both rich and complex are otherwise absent. Synchronization among speakers has since been employed by several researchers as a means for eliciting speech suited to phonetic analysis [3, 4, 5, 6]. Most such work has been done on English, although O'Dell and colleagues (2010) have employed the same methods in studying Finnish speech rhythm and Kim and Nam (2008) examined Mandarin Chinese.

Underlying the use of synchronous speaking to elicit phonetic data is the strong assumption that there is no additional source of alteration to the speech introduced by the device of having speakers synchronize. It is known that synchronous speech tends to be relatively slow in rate, though within the range of rates adopted by speakers when speaking alone [1]. At issue is rather whether there is any form of systematic prosodic alteration to synchronous speech other than a relatively slow speaking rate. To date, this has been tested only informally, by listening to synchronous speech and noting that indeed, it sounds like unremarkable English speech.

In a recent study comparing English and in Mandarin Chinese, we observed that Chinese synchronous speech appeared to exhibit an exaggerated syllable timing compared with speech produced by one person at a time (hereafter, "solo speech") [7, Sample recordings available in Supplementary Materials online]. Sentences were produced almost as if they were lists of unconnected words, and this was evidenced by a slight difference in PVI calculated based on syllable onsets. We hypothesize that there is an interaction between the means that best satisfy the demands of synchronization, and the phonological structures of the language, such that the relatively simple syllable and word forms of Chinese lend themselves to regular, temporally predictable production in a way that the more complex phonological structures of English do not. Additional evidence for this hypothesis was found in that synchronization among Chinese speakers was more resistent to perturbation induced by having slightly mis-matched texts than their English counterparts, for whom such intervention frequently led to complete cessation of speaking [7]. A follow up study including a wider variety of languages is currently underway.

We here return to the question of whether English speech produced synchronously is, in fact, unaltered, compared to solo speech. Several consideration reveal this question to be more complex that it appears at first blush. Solo speech, and joint speech, are each produced with great amounts of variability due to context, purpose, and the identity and concerns of the speakers. Neither variety admits of reduction to a simple unmarked form that can stand for all others. Synchronous speech, more narrowly circumscribed, may, indeed appear as a largely invariant speaking style, as novel texts are used that are divorced from any ongoing behavioral context, and the additional constraint of remaining in synchrony prevents the overlay of any overly dramatic or expressive phrasing. With what should it then be compared? What kind of (solo) speech might serve as a gold standard? Put like this, the question is clearly unanswerable. However if we limit the domain of possible texts radically, we may make some meaningful comparisons between the two styles. This is the approach adopted here, where we use simple word lists of 8 unconnected words.

By using simple word lists, we make use of texts that admit of very little variability when read either alone or together. This serves to constrain the potential variability of the solo speech. But using invariant word lists represents a drastic simplification. In order to then re-admit some potential temporal complexity, we contrast simple word lists (8 trochees) with complex lists in which the relative sequence of stresses and word onsets is non-coincident, such that either word onsets or stress onsets may form the basis of regular timing, but not both. If synchronization is facilitated by enhancing an underlying regularity (as we suspect in Chinese) then we might observe a more

regular sequence of either word onsets or stressed syllable onsets in synchronous productions of complex lists compared to solo productions.

The use of word lists allows us to also evaluate the temporal characteristics of the speech in terms of models of sequential production. Two highly influential models are the Wing and Kristofferson model of interval timing [8] and the hierarchical timing model of Rosenbaum [9]. The Wing & Kristofferson model has found frequent application in teasing apart sources of variability in tapping studies [10, 11, and many others]. It assumes that overt behavior is shaped by a distinct timing process (the central clock) which influences the movement of effectors. Both clock and peripheral physiology are assumed to be potential sources of variability in observed behavior, but the assumptions of the model permit decomposition of that variability into distinct clock and peripheral sources, which may be subject to distinct pathologies, or independent perturbation. The model rests on the assumption that the two sources are independent, and this assumption is warranted only if the lag one autocorrelation of an interval sequence lies between zero and -0.5 (for full justification, see Wing and Kristofferson, 1973). We test this below.

Rosenbaum's model looks for evidence of hierarchical structure in short regular sequences that would suggest constraints on timing that are not strictly sequential (as in Wing & Kristofferson's model). Hierarchical execution of movement plans involves the depth-first traversal of a tree, and would lead to a dependence of interval duration on serial position within sequence. For short sequences of 8 taps, this leads to alternating short-long patterning from binary grouping at the lowest level, with additional short-long alternation at higher levels of composite units of two or four taps. As our word lists consist of short sequences of 8 accents or stresses, we ought to look for any sign of non-sequential, hierarchical influences that would be indicated by a non-monotonic dependence of interval duration on serial position. Such non-sequential effects would be compatible with some form of hierarchical production model, while a specific pattern in which interval lengths are ordered as $\{4\} > \{2,6\} > \{1,3,5,7\}$ would fit the specific form of the model adduced on the basis of sequences of 8 taps in the 1983 paper.

2. Methods

25 dyads (50 speakers) took part in the study. Speakers were either relative strangers (12 dyads) or were highly familiar couples (13 dyads). All familiar dyads were of mixed sex, while among the strangers, 7 were mixed, 2 were male-male and 3 were female-female. Ages ranged from 21 to 56, and all were native speakers of Hiberno-English. Subjects were recruited on the campuses of two Dublin universities, and ethical approval was provided by the School of Psychology at Trinity College Dublin.

Subjects were recorded as part of a larger data gathering exercise. Relevant to the present study, they each read 4 word lists alone ("solo") and 4 word lists together ("sync"). Each group of 4 comprised 2 simple trochaic lists and 2 complex lists (see below). Solo readings were done before synchronous in all cases. Readings were done using head mounted microphones, and recordings were made to parallel audio channels.

Half the wordlists were simple, in which case they consisted of 8 trochees, selected so that stressed syllable onsets were of simple CV form, with C $\in$ {b,d,g}. Sample simple list: *banter, body, dagger, guinness, batty, dancer, bingo, gutter*. The other half were complex. In complex lists, there was a regular alternation of stressed and unstressed syllables (half began with a stressed, half with an unstressed element), but word boundaries were selected to be irregular, due to varying numbers of syllables per word. Sample complex list with a weak initial syllable: *deny, debunking, boot, divide, deduction, bike, barbaric, ban*, with a strong initial syllable: *bad, debugging, body, boot, debacle, banter, bog, degrading*. Both word onset and stressed syllable onsets were constrained to be of CV form with C $\in$ {b,d,g,v,n,m,l}. Each list in each condition was unique and was spoken exactly once. A total of 11 lists were discarded due to speech errors or mispronunciation.

Word onsets and stressed syllable onsets were located in time using a P-centre estimation algorithm, first presented in Cummins and Port (1998). This identifies the time of an onset as the halfway point through a local rise in the amplitude envelope of the bandpass filtered signal (with cut offs at 500 and 2000 Hz). This algorithm works well when syllable onsets are suitably constrained, as here, and in a very few cases where no appropriate local rise was found, manual measurement was made (less than 1% of data points). This provided a sequence of 8 word onsets and 8 stressed syllable onsets for each list. For simple lists, these coincide, providing a single set of 8 onsets, or 7 successive intervals. For complex lists, these provide 2 alternate ways of looking at the list, as 7 intervals demarcated either by word onsets or by stressed syllable onsets.

3. Results

We have three kinds of interval series: trochaic, complex calculated from word onsets, and complex calculated from stressed syllable onsets. We first examine the distribution of interval durations in the solo and synchronous conditions, for each kind of series. Fig. 1 provides an overview of the distribution of interval durations as a function of serial position. The first thing to note is that there is no obvious macroscopic difference in the central values or serial characteristics of the synchronous and the solo data. As has been documented before, the variability across speakers seems to be reduced in the synchronous case [13]. Fig. 2 shows the distribution of the interval durations for each series in each condition. It is clear that the solo distributions are right skewed, while the synchronous distributions are more compact and more nearly symmetric. The reduction in variance is confirmed by one-sided F-tests which verify reduced variability in the synchronous condition for the trochees (F(643,643)=2.7, $p < .001$), the word onsets in complex lists (F(643,587)=1.9, $p < .001$) and stress syllable onsets in complex lists (F(615,587)=2.8, $p < .001$).

The complex lists were designed so that word onsets and stress syllable onsets did not coincide for every word. In order to see whether subjects imposed regularity on the word or stress onsets, we calculated the normalized Pairwise Variability Index for successive interval durations within each series. This provides a measure of variation among successive intervals, and is minimized in isochronous series, and maximized in alternating long-short series. The nPIV is calculated as

$$\mathrm{nPVI} = 100 \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \quad (1)$$

Fig. 3 shows the nPVI scores for each series type and condition. In both solo and synchronous conditions, the intervals formed by successive word onsets in the complex lists are con-

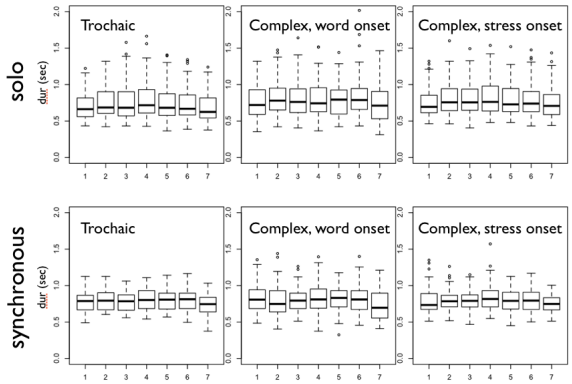

Figure 1: *Interval duration as a function of serial position.*

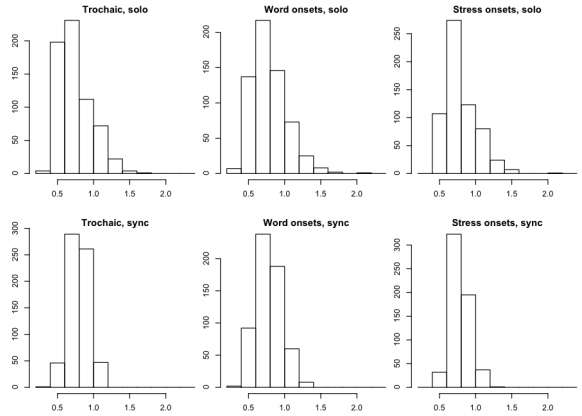

Figure 2: *Histograms of interval durations.*

siderably less regular (higher nPVI) than those formed by the stress syllable onsets. There is no evidence here of any increase in regularity (lower nPVI scores) in the synchronous condition compared with the solo.

We now turn to analyses that examine the assumptions and predictions of the two generative models, the hierarchical production model of Rosenbaum [9], and the clock model of Wing and Kristofferson [8]. The hierarchical production model makes the general prediction that serial position will influence duration, and, for sequences of 8 events, or 7 intervals, it predicts the relative durations based on tree-traversal distance of a hypothetical underlying metrical tree.

In order to investigate the effect of serial position on interval duration in the solo lists, a repeated-measures ANOVA was carried for each of the three series with familiarity (two levels) as a between subjects factor, and serial position and repetition as within subject factors (7 and 2 levels, respectively). For the trochees, there was a main effect of serial position ($F(6,264)=6.6$, $p < .001$) and a main effect of repetition ($F(1,44)=4.2$, $p < .05$). For the word onsets in complex lists there was only a main effect of serial position ($F(6,252)=3.8$, $p < .01$), and similarly for the stressed syllable onsets in complex lists there was only a main effect of serial position ($F(6,252)=3.7$, $p < .01$). A similar analysis for the synchronous lists is complicated by the fact that the utterances of two speakers speaking in unison can not be considered at all independent of one another. We therefore conducted a similar analysis for the first and second speaker separately, and report only those findings that were consistent across both speakers. There was a main effect of serial position only for the trochaic series ($p < .001$), but not for either the word onsets or the stressed syllable onsets.

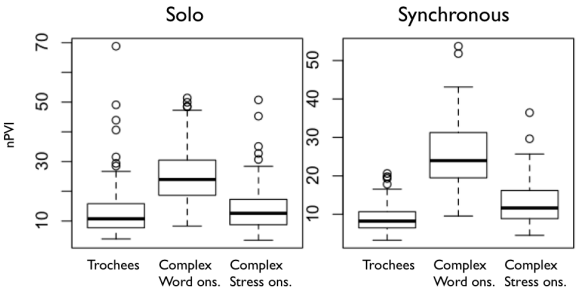

Figure 3: *nPVI scores for each series type and condition.*

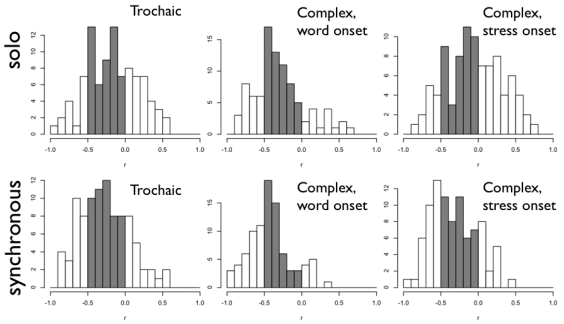

Figure 4: *Histograms of lag 1 autocorrelations.*

There are thus effects of serial position in the solo data, but not in the synchronous data, except for the maximally regular trochaic series. For the solo data, this is in accordance with the general premise of the model of Rosenbaum et al., but does not test the predictions of the specific form of that model described in Rosenbaum et al. (1983), which predicts that interval 4 should be the longest, followed by intervals 2 and 6, with intervals 1, 3, 5 and 7 somewhat shorter. This structure is not evident in the aggregate data, and there is not enough data per subject or dyad to test this on a by-subject/by-dyad basis. However, simple calculation of the index of the longest interval in each series does not support the Rosenbaum predictions, as Interval 4 is the longest interval in no more than 30% of series in the trochaic solo series, and less in all other cases.

Turning now to the Wing and Kristofferson model, it is predicated upon an assumption of the separability of variance in timing due to two independent sources: a central clock or timekeeper, and a peripheral effector system. This hypothetical independence is possible only if the lag 1 autocorrelations within each series fall in the range [-0.5, 0]. Lag 1 autocorrelations outside that range violate the assumptions of the model. In the many studies that have employed this model, the proportion of the data violating this constraint, and hence the validity of the model, has varied greatly form case to case.

Fig. 4 shows histograms of the lag 1 autocorrelations for each series. Only those series falling within the grey bins ac-

cord with the basic assumptions of the Wing and Kristofferson model. The proportion of series that violate those assumptions ranges from 41% to 55%, and violations are found both above and below the bounds of [-0.5, 0]. The serial production model of Wing and Kristofferson is thus not appropriate for interpreting these data. Violations are as common in the solo data as they are in the synchronous case.

4. Discussion

The principal question addressed in this small study is whether systematic differences between speech produced alone or in synchrony with a co-speaker can be found in English. Such differences have been found in Chinese, where synchronous speech has been found to be more list-like, with regularization of inter-syllable onset intervals [7]. In order to constrain the kind of prosodic change that might be observed, lists were employed, both simple trochaic sequences, and complex sequences in which either word onsets or stress syllable onsets might for the basis for any hypothetical regularization during synchronous production.

One difference we expected to find, and did, is that interval timing is less variable between subjects in the synchronous condition. This difference does not imply any systematic change to the prosodic features of any given utterance, however.

At first blush, there appears to be little in the way of variation in the prosodic features of synchronous utterances compared to solo utterances. In both cases, trochees are produced with a low pairwise variability, and the complex lists are produced with similar regularity in the sequence of stressed syllable onsets (but not word onsets). This is itself valuable empirical evidence that stress feet, in the sense of Abercrombie,[1] and not lexical units, constitute temporally extended constituents that may, under specific conditions, form the basis of isochronous sequences in English speech. This is in line with observations from the speech cycling paradigm in English [12], and contrasts with observations made under similar constraints in Korean [14] and Japanese [15].

We then examined the appropriateness of two generative models of sequential interval production that might be called to task in accounting for these data. The simpler model of simple sequential production based on the hypothesis of a modality-independent timer is provided by Wing and Kristofferson (1973). The presuppositions of that model were found to be very inappropriate for the present data set, with almost half of all observations showing non-local serial dependencies that violate the assumptions of the model. This was true for solo and synchronous data in equal measure.

When we viewed the data through the lens of the hierarchical production model of Rosenbaum, however, we observed some possible differences between solo and synchronous interval sequences. A liberal application of this model, that captures the notion of hierarchical production, but remains agnostic about the precise form of control or implementation underlying such production, predicts only that there will be position dependent differences among interval durations that are not a simple function of the immediately preceding neighbor (as in Wing and Kristofferson). This was true of both simple and complex lists in the solo condition, but only true of the simpler trochaic lists in the synchronous condition. The more rigorous prediction that falls out of positing a specific binary tree underlying production, that accounted very well for tapping data in the original study, did not well describe the present data.

Given the failure of the more specific model to capture for form of the non-local dependency of interval duration on serial position, we must be cautious in drawing strong conclusions. The difference betwen solo and synchronous data we observed is slight, and not yet well characterized. Prosodic differences between solo and synchronous speech in English are slight, if they exist at all, and at this stage, it appears that synchronous speech may still be a valuable way of eliminating variability from English speech while leaving linguistic contrasts unaffected. On the basis of our experinece with Mandarin, however, it is clear that this assumption does not generalize to all other languages, and that even in the case of English, further investigation is warranted.

[1] The Abercrombian stress foot is the interval from one stressed syllable onset to the next, including any and all intervening unstressed syllables. It assumes a simple binary stress distinction that may or may not be an appropriate characterization of English stress, and that certainly does not generalize to all languages.

5. References

[1] F. Cummins, "On synchronous speech," *Acoustic Research Letters Online*, vol. 3, no. 1, pp. 7–11, 2002. [Online]. Available: http://ojps.aip.org/ARLO

[2] ——, "Practice and performance in speech produced synchronously," *Journal of Phonetics*, vol. 31, no. 2, pp. 139–148, 2003.

[3] J. Krivokapić, "Prosodic planning: Effects of phrasal length and complexity on pause duration," *Journal of Phonetics*, vol. 35, no. 2, pp. 162–179, 2007.

[4] M. Kim and H. Nam, "Synchronous speech and speech rate," *Journal of the Acoustical Society of America*, vol. 125, no. 5, p. 3736, 2008.

[5] M. A. Poore and S. Hargus-Ferguson, "Methodological variables in choral reading." *Clinical Linguistics and Phonetics*, vol. 22, no. 1, pp. 13–24, January 2008.

[6] M. L. O'Dell, T. Nieminen, and L. Mustanoja, "Assessing rhythmic differences with synchronous speech," in *Speech Prosody 2010 Conference Proceedings*, vol. 100141, 2010, pp. 1–4.

[7] F. Cummins, C. Li, and B. Wang, "Coupling among speakers during synchronous speaking in English and Mandarin," *Journal of Phonetics*, vol. 41, no. 6, pp. 432–441, November 2013.

[8] A. M. Wing and A. B. Kristofferson, "Response delays and the timing of discrete motor responses," *Perception and Psychophysics*, vol. 14, no. 1, pp. 5–12, 1973.

[9] D. A. Rosenbaum, S. B. Kenny, and M. A. Derr, "Hierarchical control of rapid movement sequences," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 1, pp. 86–102, 1983.

[10] J. Gibbon, R. M. Church, and W. H. Meck, "Scalar timing in memory," *Annals of the New York Academy of sciences*, vol. 423, no. 1, pp. 52–77, 1984.

[11] S. M. Rao, D. L. Harrington, K. Y. Haaland, J. A. Bobholz, R. W. Cox, and J. R. Binder, "Distributed neural systems underlying the timing of movements," *The Journal of Neuroscience*, vol. 17, no. 14, pp. 5528–5535, 1997.

[12] F. Cummins and R. F. Port, "Rhythmic constraints on stress timing in English," *Journal of Phonetics*, vol. 26, no. 2, pp. 145–171, 1998.

[13] F. Cummins, "Synchronization among speakers reduces macroscopic temporal variability," in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 2004, pp. 304–309.

[14] Y. Chung and A. Arvaniti, "Speech rhythm in Korean: Experiments in speech cycling," in *Proceedings of Meetings on Acoustics*, vol. 19, 2013, p. 060216.

[15] K. Tajima and R. F. Port, "Speech rhythm in English and Japanese," *Phonetic interpretation: Papers in laboratory phonology VI*, pp. 317–334, 2003.