Coupling Among Speakers During Synchronous Speaking in English and Mandarin

Fred Cummins[a,*], Chenxia Li[b], Bei Wang[b]

[a] *UCD School of Computer Science and Informatics, University College Dublin*
[b] *College of Chinese Minority Languages and Literature, Minzu University of China*

Abstract

The laboratory task of synchronous speech is considered as an experimental analog of the ubiquitous phenomenon of choral speaking. We here consider some implications that arise if we regard two synchronous speakers as mutually entrained systems. Firstly, the degree of synchrony should be a function of the strength of coupling between speakers. Secondly, the entrainment would necessarily be vulnerable to perturbation. We test both these predictions, first in English and then in Mandarin Chinese. We demonstrate that modulation of the auditory link between speakers strongly affects synchronization in both languages. We also find that mismatched texts are an effective way of inducing speech errors in English, but not in Mandarin. The errors found in English frequently involve the complete breakdown of the act of speaking. An unexpected finding is that Mandarin may be pronounced with a distinct syllabic regularity in the synchronous condition. A post hoc analysis attests that the syllable is more regularly timed in synchronous Mandarin than when spoken by one person, but this effect is absent in English. We hypothesize that the strongly articulated syllable provides synchronous Mandarin with a stability in the face of perturbation.

Keywords: Synchronous speech, syllable timing, speech errors, entrainment, Mandarin Chinese

*Corresponding author
Email address: fred.cummins@ucd.ie (Fred Cummins)

1. Introduction

Choral, or joint, speaking is a common practice in schools, houses of worship, sports events, and demonstrations worldwide. Despite the ubiquity of this mode of collective speaking, it has hardly been subjected to empirical study at all, with the exception of a small specialized literature that examines the role of co-speaking in ameliorating stuttering (Kalinowski and Saltuklaroglu, 2003). Earlier studies focussed exclusively on the potential role joint speaking might play in pronunciation training (Jones, 1950). In many situations in which choral speaking is conventionally employed, the texts used are highly over-practiced, for example in classroom recitation of the Pledge of Allegiance in the United States, or recitation of the Our Father or Hail Mary in Catholic rites. These conventional recitations typically display an exaggerated and highly stylized prosody, with phrases broken into short chunks, prominent pauses, and unconventional intonation contours. A contrasting example is provided by the recent employment of the "human microphone" at protests in New York at which electronic amplification was prohibited by municipal authorities. In order to circumvent this limitation, a practice emerged of reading a text to be announced in short phrases, each of which was repeated in unison by large numbers of bystanders, thereby amplifying the sound while bypassing the municipal prohibition. Here too, informal listening suggests characteristic changes to the prosody of each utterance arising from the demands of choral speaking, but no opportunity for over-practice, giving rise to conventionalized prosodic patterns arises.

A highly constrained experimental analog of choral speech has been presented in the laboratory study of Synchronous Speech (Cummins, 2002, 2003, 2009, 2011). Within this paradigm, two speakers are presented with a novel text, usually short, and, after a single silent reading, they are asked to read in synchrony with one another, starting after a signal from the experimenter. This differs from choral speech more generally both in restricting the number of speakers to two, and in

the use of unfamiliar texts. It has repeatedly been found that subjects can typically synchronize with one another without undue difficulty. The degree of synchrony attained was initially surprising. Mean asynchrony was found to be approximately 40 ms, which is equivalent to a single frame in a conventional video sequence. This rises to ca. 60 ms at the start of phrases after a pause. The speech so produced is necessarily shorn of unpredictable expressive variability, but it is not typically perceived as sounding odd or prosodically atypical. Synchronous speaking has been used as an elicitation methodology in several studies, as it has been found to greatly reduce temporal variability, especially pause variability, while generating speech in which linguistic contrasts are still fully expressed (Cummins and Roy, 2001; Cummins, 2004; Krivokapic, 2007; Kim and Nam, 2008; O'Dell et al., 2010).

As a form of synchronized joint action, synchronous speech exhibits some intriguing and idiosyncratic properties. If we restrict our definition of synchronization to "doing the same thing at the same time", there are relatively few human activities in which people display tight temporal alignment. These include activities such as dancing, playing music in unison, and synchronized sports such as diving, rowing, trampolining, and swimming. In each of these cases, there is a strong external constraint or constraints that seems to facilitate mutual synchronization. In many (music, dancing, rowing), there is a clearly perceptible pulse or beat which allows constant registration among actors. In others (trampolining, diving), the action is strongly constrained by the mechanical coupling of organism and environment, and especially by the force of gravity. Synchronized speaking is unique in that there is no regular beat (Dauer, 1983), and the activities of the speech articulators are relatively shielded from the physical environment and not strongly constrained by gravity.

Elsewhere, we have argued that the phenomenon of synchronization among speakers provides a

theoretically valuable case study of coordination among skilled actors that can serve in the development of dynamical accounts of embodied action Cummins (2011, 2013). We will not recapitulate these arguments here except to point out that our understanding of the phenomenon is that two speakers talking in unison are usefully regards as coupled, through the medium of sound, so that we can ask questions of the properties of the dyad that are distinct from questions we can ask about the individuals within that dyad.

We here present some new experimental data that serve to flesh out this point of view. Three hypotheses are tested. *The first hypothesis is that the degree of synchrony will be parametrically modifiable by manipulating the strength of the coupling between subjects.* We can illustrate this best by analogy with a familiar case of coupling: the three legged race. In a three-legged race, two components (the runners) are physically coupled by tying their medial legs together. The coupling here is uncontroversial and physically evident. A relatively weak coupling, as would arise from looser ties between the legs, would allow each runner a modicum of relative freedom during the race, whereas a tighter coupling would ensure that the movements of the two runners are more tightly synchronized. In an analogous fashion, we here manipulate the relative strength of the coupling between the speakers by modulating the relative intensity of the sound of the speaker's and co-speaker's voice. The hypothesis we are examining here is that the bond between the speakers is expected to be stronger as the contribution of the co-speaker's voice tends to dominate. By manipulating the degree of feedback, we seek to directly influence the degree of synchrony shown by the speakers. In previous work, we have developed a computational method for quantifying the degree of asynchrony between two matched and time-aligned utterances (Cummins, 2009). This can be employed to compare synchronization under different coupling conditions.

A second elaboration of the three-legged race metaphor also arises, and illustrates our second

hypothesis. Successful runners in such a race exhibit a high degree of coordination. But this achievement is perilous, and the very existence of a successful running pair is threatened by any misstep or error. Abrupt tumbles and the inevitable cessation of any running whatsoever are common sights at student race days. If we are justified in describing two synchronous speakers as a coupled system, it should be possible to say when the system exists, and when there are merely two people talking, without mutual coupling. This line of thought was motivated by the observation in previous synchronous speech studies that a speech/reading error by one participant frequently led to a complete breakdown of speech in both speakers at the same time. Not all errors have an effect as strong as this, but it had been observed sufficiently often to warrant further investigation. In unaccompanied reading, it is essentially never the case that speech catastrophically falls apart. To return to the three-legged race analogy, if the ties are relatively loose, each runner can compensate for a perturbation with a small degree of relative independence. If tied more tightly, the system may exhibit greater synchronization, but this comes at the price of brittleness, and a tumble is more likely to ensue. *The second hypothesis we explore is that the coupled system will be vulnerable, to the point of dissolution, if it is perturbed.* We perturb the system by having a small number of mismatched texts among a large list of sentences to be read. Any awareness, on the part of the speakers, that they are not saying the same thing at the same time, will have the potential to induce a catastrophic error, resulting in the cessation of speech. This is explored in tandem with the manipulation of feedback strength.

Finally, we decided to examine Mandarin Chinese as well as English. Prior to this study, we are aware of only a few studies that employed the synchronization paradigm in a language other than English (Kim and Nam, 2008, Mandarin Chinese; O'Dell et al., 2010, Finnish). *We here test a third hypothesis that our initial dynamical characterisation of synchronization is independent*

of the details of the prosodic and linguistic characteristics of the language being spoken. Strictly speaking, this is not an experimental hypothesis, as we proceed in the expectation that no difference will be found. Although we are not aware of any reports of differences in the basic phenomenon of synchronization, it seems judicious to be open to the possibility that languages with rather different rhythmic characteristics might, perhaps, differ in the manner in which speakers couple. Because there were some methodological differences between the procedures employed in each language, we report results from each language separately, and consider the overall pattern of results thereafter.

2. Experiment 1: Coupling and Errors in English Synchronous Speech

2.1. Methods

Twelve subject pairs were recruited. For English, there were 8 male-male dyads and 4 female-female. Mean age was 32 yrs., s.d. 12. Inter-dyad familiarity was not controlled and ranged from largely familiar (classmates, work colleagues) to complete strangers. Subjects were recruited on a university campus in Dublin, Ireland. All subjects spoke Eastern Hiberno-English as their native language. None had any known pre-existing speech or hearing problems. Informed consent was obtained in accordance with the guidelines for research on human subjects at University College Dublin.

For each dyad, a list of 9 blocks of 6 sentences each was prepared. English sentences were taken from those used in the CSLU Speaker Identification corpus (Cole et al., 1998) and the TIMIT corpus (Garofolo et al., 1993). Lengths ranged from 7 to 23 syllables. Sentences used to induce errors, and for measurement of asynchrony are provided in Appendix 1.

Subjects were given formal instruction, and each dyad read six practice sentences at the start of the recording session. Practice sentences were not reused in the experimental blocks. For each sentence, the experimenter counted "two... one..." and provided a baton stroke of the hand to

signal the point at which speaking should start. Subjects then read the sentence, attempting to remain in synchrony with one another. Two sentences were repeated in each block, and asynchrony measurements as a function of recording condition are restricted to these two sentences. Most of the remaining sentences were non-repeated fillers which are not further analyzed, except that there were 6 sentences, one each in blocks 2,3,4 and 6,7,8, in which subjects were given mismatching sentences, differing in medial position by one lexical item (e.g. "It's been about two years since Davy kept shotguns" and "It's been about two months since Davy kept shotguns"). Subjects were not told that there were some mismatched sentences, although most noticed at some point. When errors occurred, the experimenter dictated what would be the response. If the error occurred in one of the two sentences that were included in each block for the measurement of asynchrony, subjects were asked to repeat the reading. If an error occurred in either a mismatched pair, or in a filler sentence, no attempt to remedy the situation was made and subjects moved on to the next sentence.

The hypothesis to be tested was that the degree of synchrony would be parametrically variable by modifying the strength of coupling among speakers. The strength of coupling was operationalised by regulating the relative volume of each speaker. Blocks of 6 sentences were distinguished by coupling conditions. Subjects were seated beside one another. Each wore a head mounted microphone (Shure SM10A), and full-cup earphones (Beyer Dynamic DT 100 or 150). The signal routed to the earphones differed across the three coupling conditions. The three conditions were selected to provide qualitatively different amounts of auditory linkage, and, by hypothesis, coupling, among subjects. Sound pressure levels were not rigorously controlled, but headphone levels were set such that speaking under all three conditions was comfortable. In the SELF condition, the earphones relayed the speaker's own voice. This did not completely exclude hearing the co-speaker, as volume

levels through headphones were moderate, ambient sound was still clearly perceptible, and speakers sat in close proximity to each other. In the BOTH condition, headphones relayed both speakers' voices at equal level, while in the OTHER condition, the headphones relayed only the co-speaker's voice. Of course here the speaker also receives self-generated sound, through both airborne sound and bone conductance, as well as proprioceptive feedback. Feedback routing was set at the start of each block, providing a brief interval between blocks. Each recording session lasted approximately twenty minutes.

Two distinct analyses were conducted, each based on the three different coupling conditions. In the first analysis, a quantitative estimate of asynchrony was made using the method introduced in Cummins (2009). The method works by representing each of the waveforms, which are aligned in real time, as a sequence of MFCC vectors, and then using dynamic time warping to find the optimal deformation that would warp one utterance onto the other. The area under the warping curve, restricted to voiced portions of the signal, provides the quantitative estimate of asynchrony. We use the 12 MFCC coefficients, omitting the zero-th coefficient and we do not employ delta coefficients. Dynamic time warping is an asymmetrical procedure in which one signal acts as a referent onto which the other is warped. As the method is restricted to voiced parts of the signal, the voiced intervals are determined with respect to the referent. For this reason, each analysis was conducted twice, with each of the two signals acting in turn as the referent, and the average of the two estimates was taken as the asynchrony score. This procedure provides a robust estimate, i.e. it is not sensitive to microphone characteristics or speaker identity, and it has previously been shown to allow the qualitative discrimination among different synchronization conditions (Cummins, 2009).

The second analysis was based on the 6 sets of mismatched sentences. For each paired recording, two pairs of raters judged the severity of errors induced by the deliberate mismatch between

sentences. Each pair of raters could listen to the speech, and interact with a visual display of the waveform and spectrograph. Scoring was done blind, i.e. raters did not know what condition each sentence was from. Each single-channel recording was listened to as often as required, and each scorer decided independently how to score any resulting error using the following criteria:

- Score 0: no noticeable dysfluency
- Score 1: noticeable dysfluency, but no pause or cessation of speech above 500 ms and no missing speech elements (phonemes or syllables)
- Score 2: as for 1, but with a pause equal or greater than 500 ms, or with evidence of omitted speech material (phonemes or syllables)
- Score 3: complete breakdown in speaking. This may be recorded even if there is a subsequent attempt to restart speaking

The two scorers within the rating pair then conferred to agree on a single score to be recorded. Table 1 shows the inter-rater agreement (unweighted Cohen's Kappa = 0.72) obtaining among the two pairs of raters (not among individuals within a rating pair). Critically, agreement for the principle category of interest, Category 3 (complete breakdown) is very high.

	0	1	2	3
0	66	5	0	0
1	7	19	11	0
2	0	1	6	1
3	0	0	1	25

Table 1: Confusion matrix showing agreement among two pairs of raters for errors by English speakers.

2.2. Results

We first examine asynchrony as a function of the three coupling conditions. Fig. 1 shows the calculated asynchrony (arbitrary units, derived from area under the warping curve) as a function of coupling condition for the two sentences separately. Three outliers for the data from Sentence 1 have been removed to better show the data. These are SELF: 27.5, 18.8 and OTHER 14.7. The SELF condition clearly produces utterances that are less well synchronized than the other two conditions, and there is no obvious difference between the other two conditions. For inferential tests, the asynchrony data were log transformed to make their distribution more approximately normal. The above observation was confirmed by a repeated measures ANOVA with condition and sentence as within-dyad factors (Condition: $F(2,22)=92.1$, $p < .001$; sentence and interaction n.s.). Pairwise t-tests, with a Holm adjustment of p-values to control family-wise error rate, show a significant difference between asynchrony in the self condition and each of the other two conditions ($p < .001$), while they in turn do not differ from one another. Similar findings appear when the ANOVA is done for each sentence separately.

Speech errors were scored that arose in cases where the texts provided were mismatched on one medial lexical item. Each dyad encountered six such mismatches. Fig. 2 (left) shows the distribution of error severity scores obtained by averaging the error ratings of the two pairs of scorers. A first observation is that the experimental manipulation was successful at inducing speech errors, although they were not inevitably occasioned. Overall, there were 25 (of 142 total) "catastrophic" errors, in which both pairs of raters agreed that there had been a complete breakdown in speaking. (One paired reading is omitted, as a serious speech error occurred before the mismatched lexical item.) From the marked bimodality in the distribution of error scores, it seems plausible that these Category 3 errors are distinct from the more minor errors, and thus perhaps of distinct origin.

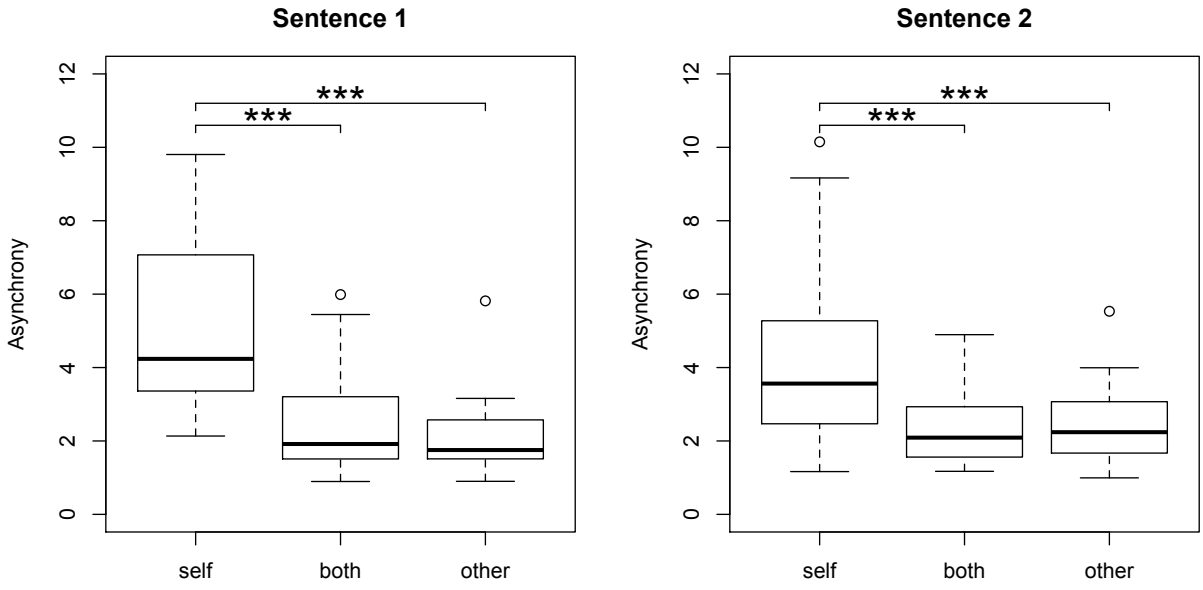

Figure 1: Asynchrony as a function of coupling condition for the two English sentences.

The right hand plot shows the distribution of error severity scores as a function of coupling condition. Mann-Whitney tests reveal significant differences between SELF and the other two conditions ($p < .001$), which in turn do not differ from each other.

2.3. Discussion of Experiment 1

Viewing two synchronized speakers as a single dyadic domain within which two individual components (the speakers) are mutually coupled led to the hypothesis that coupling would be facilitated by increasing the relative amplitude of the co-speaker's voice. This was borne out by the observed results, although the means employed to regulate the coupling—gross modification of the volume of the co-speaker as heard through the headphones—is relatively crude. The results obtained did not distinguish between the BOTH and OTHER conditions. It is not possible, therefore, to infer quantitative aspects of the emergent coupling among speakers. It may be that coupling is an all-or-nothing phenomenon, or it may be that the degree of non-independence among speakers

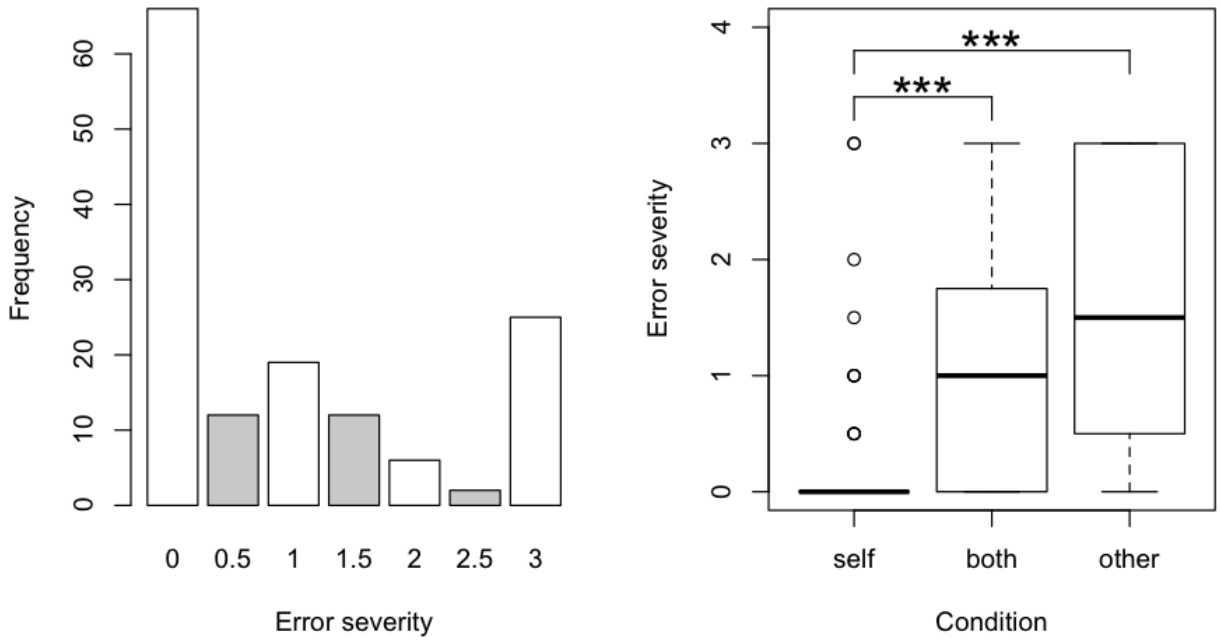

Figure 2: Left: Boxplot of average error scores. Right: Error severity as a function of coupling condition.

is continuously and monotonically related to the relative amplitude of the co-speaker. Further investigation of this will require more carefully controlled regulation of the relative amplitude of the two voices. This may necessitate confronting the difficulty of properly assessing the relative contribution of bone conductance and proprioceptive feedback.

The associated prediction from our second hypothesis that complete breakdown in speaking would occur with greater frequency as coupling increased was also supported, although again the methods employed did not reveal a graded difference between the BOTH and OTHER conditions. The error distribution speaks strongly for the existence of a distinct class of error not observed under more normal reading conditions. While there was some disagreement among judges with respect to the two intermediate degrees of error that were operationalized using somewhat ad hoc criteria, there was almost complete agreement on the significant presence and seriousness of Category 3 errors, "Complete breakdown in speaking".

3. Experiment 2: Coupling and Errors in Chinese Synchronous Speech

3.1. Methods

For Mandarin Chinese, there was one male-male dyad, two mixed dyads, and 9 female-female dyads, for a total of 12 speaking pairs. Mean age was 22.9 yrs, s.d. 2.7. Subjects were recruited on campus at Minzu University of China, Beijing. Within dyad familiarity was, again, not controlled for. All participants were native speakers of standard Mandarin Chinese.

Free translation of the sentences used for Experiment 1 were made. These had syllable counts ranging from 8 to 23 syllables. Where translation suffered, priority was accorded to phonological composition, and not sentence meaning, which was allowed to vary freely. Sentences used for estimating asynchrony and for inducing errors are again listed in Appendix 1. Subjects were given formal instruction as before, and they completed six practice sentences that were not reused prior to data collection. Block structure and experimental procedures were otherwise the same as for Experiment 1.

Because of the paucity of prior experience in applying the synchronous speech method beyond English, it was decided to obtain comparable data from speakers reading alone ("solo speech"). Subjects were therefore invited back in to re-read all sentences as before, but without a speaking partner. For 9 of the dyads, this took place on the day after the synchronous recording. The three remaining dyads were recorded 20 days after the first, synchronous, session, due to logistical difficulties.

In recording the Chinese data, handheld Shure SM58 microphones and Sony MDR-7506 full-cup earphones were employed.

3.2. Results

Asynchrony scores as a function of condition are shown in Fig. 3. Two outliers have been removed from the SELF condition in each sentence. They are: Sentence 1: 26.5, 49.7; Sentence 2: 33.3, 36.5. Notice that the y-axis here has a different scale than that in Fig. 1. This is because the asynchrony values in the SELF condition exhibit much greater variability and range than was the case with the English speakers. The range and variability in the other two conditions is comparable to the English case.

For inferential statistics, we again log transformed the asynchrony data. A repeated measures analysis of variance with sentence and condition as within-dyad factors revealed a main effect of condition ($F(2,22)=77.5$, $p < .001$), but no effect of sentence or interaction. Posthoc pairwise t-tests, with a Holm adjustment of p-values to control family-wise error rate, yielded significant differences between each pair. All p values are $p < .001$ unless noted otherwise. For Sentence 1, SELF–BOTH: $t(35)=9.6$; SELF–OTHER: $t(35)=10.6$; BOTH–OTHER: $t(35)=3.7$ ($p < .05$). For Sentence 2, SELF–BOTH: $t(35)=5.6$; SELF–OTHER: $t(35)=6.0$; BOTH–OTHER: $t(35)=4.2$.

Errors were scored in the same fashion, and using the same explicit procedure, as for the English data, but raters were native Mandarin speakers. Table 2 provides the confusions among scorers, and the unweighted Cohen's Kappa score is 0.70, which is very slightly less than the value obtained for the English data.

	0	1	2	3
0	72	1	0	0
1	13	33	8	0
2	0	1	10	0
3	0	0	3	3

Table 2: Confusion matrix showing agreement among two pairs of raters for errors by Chinese speakers.

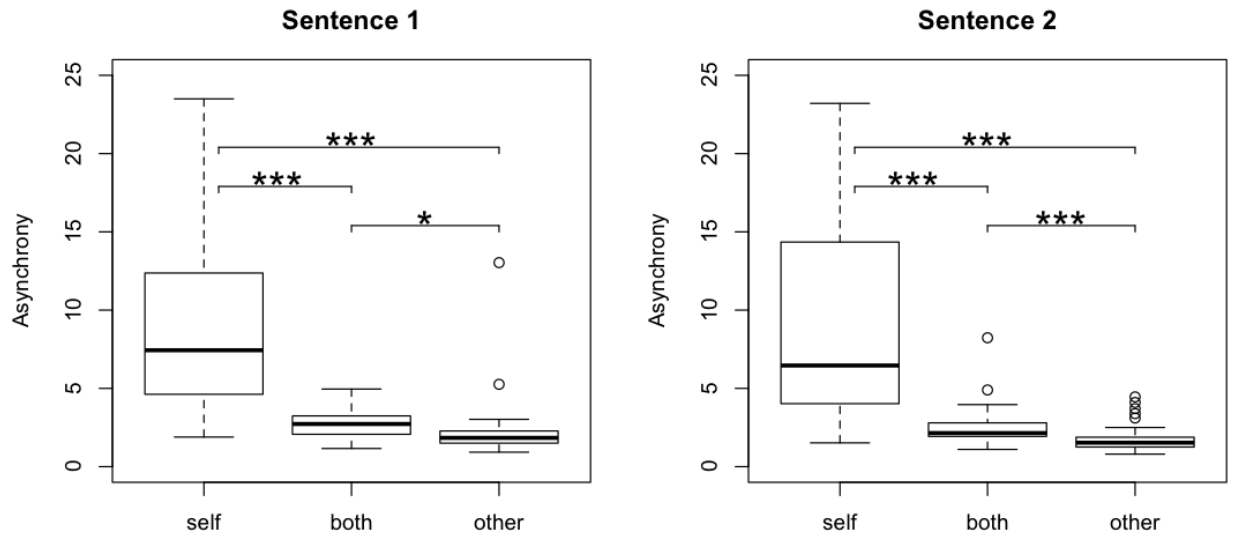

Figure 3: Asynchrony as a function of coupling condition for the two Chinese sentences.

An immediate, and somewhat surprising, observation is that there were far less significant errors among the Chinese speakers than observed in Experiment 1: 3 unambiguously severe errors in Chinese against 25 in English. Fig. 4 shows the distribution of errors overall and as a function of condition. Mann-Whitney tests revealed significant differences for each pair of conditions ($p < .01$ for BOTH–OTHER, $p < .001$ otherwise).

3.3. Discussion of Experiment 2

The role of coupling in modulating the degree of synchronization among speakers was much as seen with the English data, with the exception that there was a quantitative difference in the expected direction between the BOTH and OTHER conditions for one of the two sentences. Again, it was found that increasing the relative amplitude of the co-speaker's voice led to greater synchronization within a dyad. The principal difference between the two languages lies in the domain of errors induced by having mismatched lexical items in occasional sentences. Whereas the English data clearly showed a distinct class of severe errors, this is not evident in the Chinese data, even though the texts read, and the kind of mismatch employed, were directly comparable across

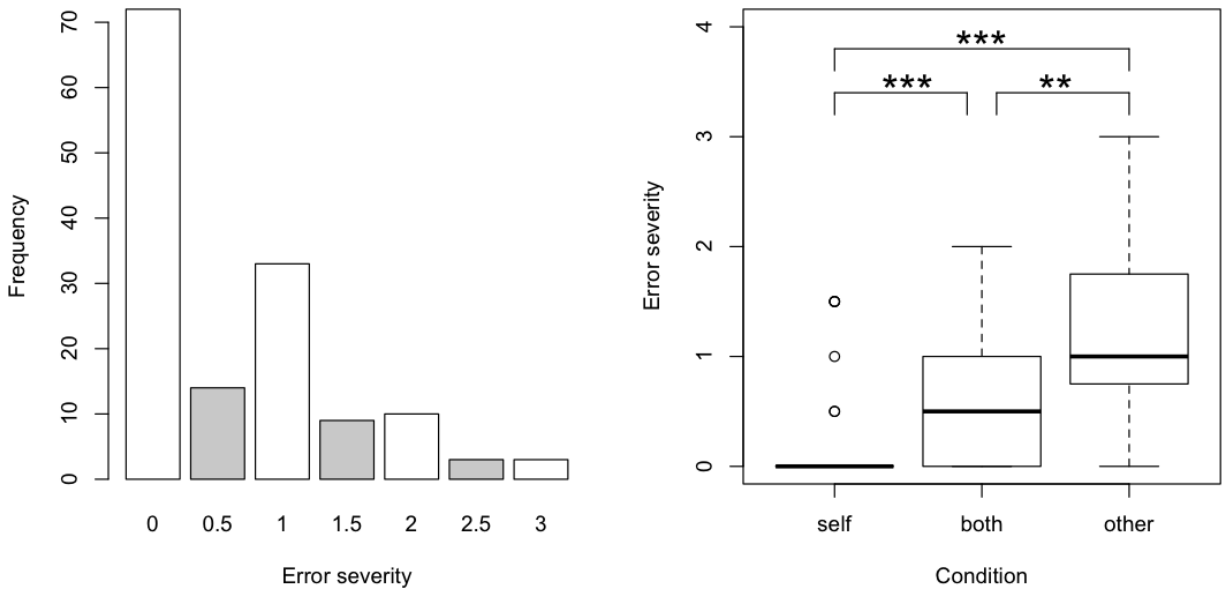

Figure 4: Left: Boxplot of average error scores. Right: Error severity as a function of coupling condition.

languages.

A further observation was made about the Chinese recordings that served to differentiate them from the English set. For several of the 12 dyads (most especially Dyads 1, 6, 7, and 12), speech produced synchronously was perceived by the authors to be markedly more syllable-timed[1] than speech produced when reading alone, or by the other dyads. This observation is strongly at odds with our third hypothesis, which was that the dynamic characterisation of synchronization would be independent of the prosodic and linguistic characteristics of the language being spoken. Representative examples of contrasting solo and synchronous utterances from Dyads 1 and 12 are provided in the supplementary materials accompanying this article online.

To date, no systematic alteration of prosody has been observed in synchronous speech, but the

[1]The term "syllable timing" is used with severe reservations in this instance. The term has been popular in the discussion of rhythmic differences obtaining between languages, where it is conventionally teamed with the contrasting classes of "stress timed" and often also "mora timed". As argued at length elsewhere, this rhythm typology is not empirically supported, despite its widespread discussion (Cummins, 2012).

bulk of data collected under such conditions has been in English[2]. Further investigation of this apparent phenomenon seems warranted on two counts. Firstly, it might alert experimenters to prosodic alteration due solely to the condition of synchronization, which might have consequences for the employment of synchronous speech as an elicitation tool to obtain expressively neutral, but otherwise unaltered, speech (Zvonik and Cummins, 2002; Krivokapic, 2007). Secondly, it might shed some light on the manifestation of the syllable in Chinese as speaking conditions are varied.

A recent study in German did find some alterations to the normalized Pairwise Variability Index (nPVI, vowel based, Grabe and Low, 2002) and %V (the proportion of an utterance that is voiced, see Ramus et al., 1999) in recordings made when subjects read in synchrony with recorded speakers (Dellwo and Friedrichs, 2012). Subject numbers were very small (4 synchronizers and 4 targets, 3 sentences), the biggest changes were observed when speakers read along with non-natively produced targets, and synchronization was not among live speakers, so that inferences pertinent to the present case are hard to draw.

Given the observation that the Chinese data appeared to have more regular syllabic timing in the synchronous condition than in the solo, we decided to conduct a post-hoc analysis, comparing the two conditions. A matching data set was constructed for English, making use of recordings from a pervious experiment. We then employed a variant of the well-known Pairwise Variability Index (PVI) to compare speech across the two speaking conditions within the two languages. Full details of the post-hoc comparison are provided in Appendix 2. The PVI index used was based on syllable duration, with duration measurements based on P-center estimates for each syllable (Cummins and Port, 1998).

Figure 5 provides an overview of the normalized PVI scores. A repeated measures ANOVA

[2]Michael O'Dell has communicated to us that some differences in Finnish prosody have been found between synchronous and solo speech. Data presented at RPPW Leipzig, 2010. Slides available at `http://bit.ly/WxR16H`

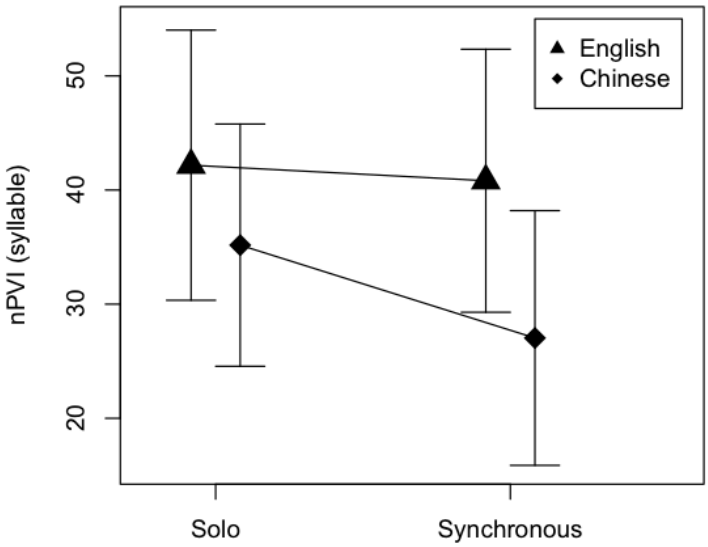

Figure 5: Normalized Pairwise Variability Index based on syllable durations. Error bars show one standard deviation.

with language as a between subjects factor, and condition as a within subjects factor shows main effects of both language ($F(1,22)=21.3$, $p<.001$), and condition ($F(1,22)=29.0$, $p<.001$), as well as an interaction ($F(1,22)=14.8$, $p<.001$). Post-hoc t-tests with Bonferroni protection for multiple comparisons show significant differences between solo and synchronous in Chinese, but not in English.

4. General Discussion

The overarching goal of the present study was to further explore the nature of the bond between speakers when they speak in synchrony. We employed the synchronous speaking task developed in many of our previous studies, in the understanding that this is a laboratory-specific manipulation that bears some similarities to joint or choral speech, but does not capture all the properties thereof. In particular, we know *a priori* that choral speech exhibits characteristic prosodic alterations that have not previously been observed in work on English synchronous speech. We now first summarize the empirical results obtained, and then use them as the basis for further consideration of the effects

of synchronization among speakers.

Experiment 1 employed the regulation of auditory feedback among English speakers to quantitatively influence the coupling among them. In keeping with our first hypothesis, a rather coarse link between the degree of synchrony that speakers manifest and the type of feedback that is provided was observed. When the auditory feedback from the co-speaker was substantially reduced, synchrony decreased. The method of adjustment used did not allow a fine-grade of control over the total feedback available to subjects, and so further work remains to be done to see if synchrony can indeed be manipulated in a gradient fashion through feedback manipulation. In Experiment 2, where the language employed was Mandarin Chinese, the basic finding that the SELF condition led to less synchrony than the other two conditions was replicated, and in addition, all three coupling conditions were found to be significantly different from one another for both sentences. There is thus evidence to shore up the claim that synchronous speakers are coupled, and the coupling is enabled through the auditory feedback available to speakers.

Both experiments also included mismatched texts, designed to induce speech errors. If speakers are well-described as coupled, then the coupling itself will be potentially vulnerable to disruption by a perturbation, such as an induced speech error. In English, we saw that such errors frequently lead to a complete cessation of speaking on the part of both speakers. This had been observed anecdotally before, but we here obtain empirical support for the claim that there exists a common type of abrupt cessation error when speaking synchronously that does not, or not frequently, occur in normal speech. In Chinese, the count of such errors was much lower than in English. Chinese speakers appear to be less vulnerable to the kind of catastrophic collapse of joint speaking found in English. We conducted a Kolmogorov-Smirnov test for goodness of fit, to see if the distribution of errors in English differed significantly from Chinese. The resulting p-value was 0.06, thus not reaching

significance. Our second hypothesis is thus supported in English only, and an unexpected interaction between our second and third hypotheses was found, so that the stability of the synchronized dyad appears to be dependent on the prosodic characteristics of the language employed.

Taken together, our findings suggest that the dynamical account of synchronized speaking, in which two speakers are mutually coupled, can capture some of the qualitative characteristics of joint speaking under these circumstances. This suggests that the further development of dynamical systems models of coupling among speakers might be a useful strategy for the still undeveloped scientific study of joint speech more generally.

We turn now to the unexpected finding of Experiment 2, that catastrophic speech errors were far less common in Chinese than in English, and link it to the equally unexpected finding that the syllabic prosody of Chinese speakers was greatly exaggerated in the synchronous condition—an effect of speaking synchronously that has not been hitherto observed in English, and that was predicted *not* to occur by our third hypothesis. We suspect that these two observations are not independent.

Firstly, we might note that although prosodic alteration has not been documented in the highly constrained laboratory task of synchronous speaking, it is commonplace in choral speech more generally. The highly stylized prosodic forms found in repetitions of prayers, or in recitals of the oaths of allegiance are familiar examples. It could be argued that these prosodic characteristics arise because of over-practice, rather than synchronization. A recent illustrative counter-example is provided by the "human microphone" employed during the Occupy Wall Street protests. Here, texts were entirely novel, and a cursory listening will assure the interested party that the prosody is not that which would be expected if the text were read alone. Other factors contribute here too, not least the fact that these texts are shouted aloud, rather than read, and a full account of

the prosodic effects of synchronization will be outstanding until considerable further work has been done, both in the laboratory and in more ethologically relevant situations. In the meantime, it would be wise to caution researchers who employ synchronous speech as an elicitation tool, that the absence of reports of prosodic alteration should not be taken as support for a claim that there are no such effects.

The nature of the prosodic change found in Chinese is of particular interest. The impression obtained on informal listening was one of exaggerated syllable timing. The reader can listen to the examples provided in the supporting materials to verify this themselves. Based on our analysis presented in the appendix, we can reasonably claim that the syllables in the synchronous condition are more regularly timed (less syllable-to-syllable durational variability) than in the solo readings. The nPVI metric that showed this clearly was based, unsurprisingly, on syllable durations, rather than vowel or consonantal intervals. The characterization of prosodic change provided by an altered nPVI index may be descriptively accurate, but it is somewhat unilluminating.

When we also take into account the almost complete absence of the kind of catastrophic speech error found in English, a possible explanatory account becomes available. The syllable has many faces. To the phonologist it may be a unit of structure, to the poet, a unit of quantity. To the phonetician, the syllable may be more or less well defined in any given utterance, and when it is present, it can, on some accounts, be interpreted as a unit of coordination. This has found expression within Articulatory Phonology, where the syllable is taken as a domain within which individual gestures are mutually linked: onset consonants with the nuclear vowel, and coda consonants each with the preceding segment, in a chain (Browman and Goldstein, 1988). Within the recent Embodied Task Dynamic model, the relative timing of an onset consonant and the nuclear vowel was found to be the single-most invariant feature of all possible pairwise gestural timing

relations, as speech rate and degree of articulatory precision were varied over wide ranges (Simko and Cummins, 2011). Here, again, the syllable (in this case the simplest CV organization) appears as a domain of coordination among gestures. The notion of a syllable as a coordinative domain has been repeatedly advanced in quite different contexts in the past (Fowler, 1983; Cho and Keating, 2001; Jong, 2001; Byrd et al., 2005).

The suggestion that arises from our unexpected observations is that the exaggerated syllable may provide a coordinative stability to speech production that is not available in the case of English, and that thus provides the synchronous speakers with a degree of relative stability in the face of an experimentally induced perturbation. If the (phonetic) essence of the syllable is a coordinative relation among its component gestures, then we might observe speech in which this relation is very highly constrained, and speech in which the temporal relations among gestures are more variable. The former, we suggest, would present us with phonetically well-defined syllables. In the context of a synchronization task, a well-defined syllable is highly constrained in time and space. If speakers share the capacity to speak with such well defined syllables, then the possibility of tight synchronization would be enhanced, as they would share the temporal constraint imposed by the strongly defined syllable.

An exaggerated "syllable-timing" could thus be interpreted as a task-specific functional response, helping to ensure that the task constraint of synchronized speaking is met despite possible external perturbation. This interpretation is, of course, somewhat speculative. It arises in the post hoc consideration of two unexpected empirical observations. However, it may point the way towards the future study of prosodic alteration to speech when spoken jointly. In this way, the laboratory study of synchronous speech may help to guide the ethological study of joint speech. It also suggests that one might usefully extend the current work to explore the degree of temporal

constraint exhibited by the Mandarin syllable as speaking context varies.

The work we have presented herein adds to our understanding of what is going on as speakers demonstrate their remarkable ability to speak in time with one another. The dynamical perspective we employ has been presented as a framework within which one can look for the hallmarks of coupling, emergent stability, and shared constraint. This might be a useful perspective to employ as the work is extended in at least two potentially rich directions—to the study of joint speech and joint speaking more generally, and to the study of coordinative structures that emerge as a function of speaking context or task.

Acknowledgements

This work was funded in part by Project 111 of the Minzu University of China. We are grateful to the editor, to two anonymous reviewers, and to Michael O'Dell for very substantial feedback that has greatly improved the manuscript.

Appendix 1: Sentences used

We here list the non-filler sentences that were used in Experiments 1 and 2. For each language, two sentences were repeated within each block, and these formed the basis for the asynchrony measurements.

Repeated sentences

- She had your dark suit in greasy wash water all year.
- Alice's ability to work without supervision is noteworthy.
- Mā mā bǎ nǐ de hēi xī zhuāng rēng dào dài yóu de shuǐ lǐ le.

  (Mother threw your black suit in greasy water.)

- Sūn dá chāo qiáng de dú lì gōng zuò néng lì líng rén chēng zàn.

  (Sunda's extremely good ability to work independently is worthy of praise.)

Mismatched sentences

In each language, there were six pairs of sentences, differing in a single lexical item. These were designed to induce speech errors.

1. Assume for example a situation where a {farm/house} has a packing shed and fields
2. Will you please {describe/confirm} government policy regarding waste removal?
3. Brush fires are common in the {dry/bare} underbrush of Nevada.
4. The fish began to {leap/jump} frantically on the surface of the small lake.
5. It's been about two {years/months} since Davy kept shotguns.
6. How much will it {take/cost} to do any necessary modernizing and redecorating?

1. Nà tiáo hóng sè de yú {jīng1 huāng/xùn sù} de tiào chū le shuǐ miàn.

   (That red fish jumped out of the water {in a panic/rapidly}.)

2. Yóu xī yān yǐn qǐ de huǒ zāi zài {mù chǎng/lín qū} shí yǒu fā1 shēng.

   (Fires due to smoking are frequent in the {farm/forest}.)

3. Dà wèi yǐ jīng kāi le {liǎng nián/sān tiān} mó tuō chē shàng bān le.

   (It's been {two years/three days} since Dawei went to work by motorcycle.)

4. Kě yǐ xiǎng xiǎng yí xià {jīng yíng/yōng yǒu} yí gè nǒng chǎng de chǎng jǐng.

   (It is possible to imagine the situation of {running/owning} a farm.)

5. Zhè gè fáng zi chóng x$\bar{\text{i}}$n {fěn shuā/zhuāng xiū} dà gài xū yào duō1 shaǒ qián?

   (How much money does it approximately cost to {re-paint/redecorate} the house?)

6. Nǐ kě yǐ xiàng jiē dào {chéng qīng/què rèn} yí xià zhè jiàn shì ma?

   (Can you {describe/confirm} this matter to the government office on the street?)

Appendix 2: Post hoc comparison of English and Chinese Syllable Timing

We conducted a post-hoc analysis of the unexpected alteration to syllabic prosody in Chinese speakers when speaking synchronously. The Chinese material examined consisted of 6 of the filler sentences from each of 6 dyads in both solo and synchronous conditions. Filler sentences were chosen to ensure that text materials were sufficiently varied. Recordings were taken from the BOTH coupling condition, as this is the standard set up previously employed in synchronous speech studies (equal auditory feedback from oneself and one's co-speaker. The dyads were chosen so as to include three dyads for whom the effect seems to be particularly unambiguous (Dyads 1, 7 & 12), and three for whom it was not as obvious (Dyads 3, 8 & 11). A corresponding English data set was retrospectively assembled from the published CHAINS speech corpus (Cummins et al., 2006). Therein, it was possible to select six dyads that matched the Chinese dyads in sex. Sentences matching the syllable count of the Chinese sentences were extracted form the larger corpus. These sentences were, as before, selected from the CSLU Speaker Identification corpus (Cole et al., 1998) and the TIMIT corpus (Garofolo et al., 1993). Within the CHAINS corpus, both solo and synchronous recordings were available. In this case, the solo and synchronous recordings had been obtained on the same occasion, embedded within a larger set of materials to be recorded.

4.1. Methods

For each paired reading (solo and synchronous), we were interested to see if we could find an empirical correlate that accorded with our perception of prosodic change from solo to synchronous in the Chinese data. To the ear, it certainly seemed as if the traditional description of "syllable-timing" might be of use in characterizing the perceived shift: in the synchronous condition, syllables sounded more evenly timed, to the extent that some readings sounded as if they were lists of syllables, rather than continuous sentences (illustrated in the Supplementary Materials online).

We therefore decided to employ a variant of the familiar normalized Pairwise Variability Index (nPVI) based on syllable durations.

The use of nPVI for syllable durations is not widespread, but seemed to be appropriate given the perceptual impression of syllabic regularity. Syllable onsets were taken as the best estimate of the P-center for a given syllable, as computed using a variant of the algorithm introduced in Cummins and Port (1998). For this, the speech is first bandpass filtered with cut offs at 500 and 2,500 Hz. The smoothed amplitude envelope was then used to identify local rises in amplitude at syllable onset, and a P-center, or beat, estimate was placed midway through such rises. The algorithmic estimation of P-centers is inherently noisy, and no perfectly satisfactory method has yet been developed. All P-center estimates were reviewed and corrections made manually where necessary. A protocol was once again employed whereby measurement issues were resolved under consideration of both the solo and its matched synchronous utterance. In this way, syllable duration estimates are employed consistently in matched utterances, so that the comparison of solo and synchronous speech tokens is as reliable as possible.

The nPVI is calculated as:

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

For the sake of completeness, we decided to also apply two conventional rhythm metrics to the data. The conventional metrics employed are the Normalized Pairwise Variability Index, nPVI (vowel), introduced by Grabe and Low (2002) and the %V measurement used by Ramus et al. (1999). For nPVI (vowel) measurements, vocalic intervals were identified by following the loose criteria provided in Grabe and Low (2002), complemented by the requirement that interval boundary criteria were applied in identical fashion for matched solo and synchronous utterances. Thus,

for example, if one utterance did not allow a partition of a glide-vowel sequence, no attempt was made to identify such a partition in the matched utterance. In this way, despite the inherent nosiness of the measurement procedure, directly comparable segmentation criteria were employed for all utterances to be compared. For %V measurements, the same segmentation into vowel and consonants was made, and the proportion of the entire utterance, multiplied by 100, provided the required metric.

As reported above, a repeated measures ANOVA based on the syllabic PVI metric with language as a between subjects factor, and condition as a within subjects factor shows main effects of both language ($F(1,22)=21.3$, $p<.001$), and condition ($F(1,22)=29.0$, $p<.001$), as well as an interaction ($F(1,22)=14.8$, $p<.001$). Post-hoc t-tests with Bonferroni protection for multiple comparisons show significant differences between solo and synchronous in Chinese, but not in English.

As expected, nPVI (vowel) was in general lower for Chinese than English, but this measure did not appear to be sensitive to the manifest change in prosody that was observed between the solo and synchronous conditions in Chinese. This was confirmed by a repeated measures analysis that identifies a main effect of language ($F(1,22)=115$, $p<.001$), but no main effect of condition, nor an interaction.

For %V, larger values were associated with Chinese than with English (as expected from the typological literature), and larger values in each language in the synchronous condition than in the solo condition (unexpected). A similar RM ANOVA shows main effects of both language ($F(1,22)=80.2$, $p<.001$) and condition ($F(1,22)=87.4$, $p<.001$), and the interaction is not significant. Post-hoc t-tests with Bonferroni protection for multiple comparisons shows significant differences between solo and synchronous in both Chinese and English.

In summary then, the syllabic PVI did a good job of capturing the perceived alteration to

syllable timing, and confirmed that syllable durations were more nearly equal in the synchronous Chinese productions than in the corresponding solo ones. The more conventional rhythm metrics did not perform well.

References

Browman, C., Goldstein, L., 1988. Some notes on syllable structure in articulatory phonology. Phonetica 45 (2-4), 140–155.

Byrd, D., Lee, S., Riggs, D., Adams, J., 2005. Interacting effects of syllable and phrase position on consonant articulation. Journal of the Acoustical Society of America 118 (6), 3860–3873.

Cho, T., Keating, P. A., 2001. Articulatory and acoustic studies of domain-initial strengthening in Korean. Journal of Phonetics 29.

Cole, R., Noel, M., Noel, V., 1998. The CSLU speaker recognition corpus. In: Proc. ICSLP. Sydney, Australia, pp. 3167–3170.

Cummins, F., 2002. On synchronous speech. Acoustic Research Letters Online 3 (1), 7–11.
URL `http://ojps.aip.org/ARLO`

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. In: Proceedings of the 26th Annual Meeting of the Cognitive Science Society. pp. 304–309.

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

Cummins, F., 2011. Periodic and aperiodic synchronization in skilled action. Frontiers in Human Neuroscience 5, published online, December 31, 2011.

Cummins, F., 2012. Looking for rhythm in speech. Empirical Musicology Review 7 (1–2).

Cummins, F., 2013. Towards an enactive account of action: Speaking and joint speaking as exemplary domains. Adaptive BehaviorIn Press.

Cummins, F., Grimaldi, M., Leonard, T., Simko, J., 2006. The CHAINS corpus: CHAracterizing INdividual Speakers. In: Proc of SPECOM'06. St Petersburg, RU, pp. 431–435.

Cummins, F., Port, R. F., 1998. Rhythmic constraints on stress timing in English. Journal of Phonetics 26 (2), 145–171.

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

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

Dellwo, V., Friedrichs, D., 2012. Variability of speech rhythm in synchronous speech. In: Proceedings of Speech Prosody 2012, V. 2. pp. 539–542.

Fowler, C., 1980. Coarticulation and theories of extrinsic timing. Journal of Phonetics 8 (1), 13–133.

Fowler, C. A., 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. Journal of Experimental Psychology: General 112 (3), 386–412.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V., 1993. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium 10 (5), 0.

Grabe, E., Low, E., 2002. Durational variability in speech and the rhythm class hypothesis. Laboratory Phonology 7.

Jones, M., 1950. Choral reading and speech improvement. Western Journal of Communication (includes Communication Reports) 14 (1), 19–21.

Jong, K., 2001. Effects of syllable affiliation and consonant voicing on temporal adjustment in a repetitive speech-production task. Journal of speech, language, and hearing research 44 (4), 826.

Kalinowski, J., Saltuklaroglu, T., 2003. Choral speech: the amelioration of stuttering via imitation and the mirror neuronal system. Neuroscience & Biobehavioral Reviews 27 (4), 339–347.

Keller, E., 1990. Speech motor timing. In: Hardcastle, W. J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic, Dordrecht, pp. 343–364.

Kim, M., Nam, H., 2008. Synchronous speech and speech rate. Journal of the Acoustical Society of America 123 (5), 3736.

Krivokapic, J., 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. Journal of Phonetics 35 (2), 162–179.

O'Dell, M., Nieminen, T., Mustanoja, L., 2010. Assessing rhythmic differences with synchronous speech. In: Speech Prosody 2010-Fifth International Conference. Vol. 100141. pp. 1–4.

Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73 (3), 265–292.

Simko, J., Cummins, F., 2011. Sequencing and optimization within an embodied task dynamic model. Cognitive Science 35 (3), 527–562.

Zvonik, E., Cummins, F., 2002. Pause duration and variability in read texts. In: Proc. ICSLP. Denver, CO, pp. 1109–1112.