euronoise

Synchronous speech and speech rate

Miran Kim[a] and Hosung Nam[b]

[a]Suny, Dept. Linguistics, S201, SBS building, Stony Brook, NY 11794-4376, USA
[b]Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511, USA
mrkim@ic.sunysb.edu

Synchronously read speech has been shown to reduce a high degree of variability exhibited by speakers in laboratory recording: e.g., pause duration and placement, and speech rate. However, quantitative analysis of speech rate variation per se has rarely been reported in studies on synchronous speech. This study examines global and local patterns of speech rate variation in Mandarin Chinese, which is expected to show a relatively stable speech rate variation as measured in the number of syllables per second. The speech data were analyzed and compared in terms of mean speech rate and the variations within a subject, across subjects, and across dialects. Our findings show that speakers exhibit lower and less variable speech rates when they read together than when they read alone. This global pattern is consistently observed across dialects maintaining the unique local variation patterns of speech rate for each dialect. We conclude that simultaneous speakers lower their speech rates when reading together such that the variability of speech rates found in simultaneous speakers is ensured to decrease by lowering speech rate in both global and local patterns. This characteristic is a hallmark of synchronous speech.

1 Introduction

Synchronously read speech, in which speakers are asked to read a text together, has been shown to reduce a high degree of variability: e.g., pause duration and placement [1, 2, and 3]. However, quantitative analysis of speech rate variation per se has rarely been reported in studies on synchronous speech.

This study focuses on the following two questions. First, how do speakers with different speech rates compromise when they read a text together? The second question is how the variability of speech rates differs between read-alone and read-together speech across speakers and across repetitions.

Synchronous speech in English, for example, has been shown to exhibit less variation in tempo and more regular pauses in duration and placement [3]. On one hand, speech rate variation per se has not been reported in detail in synchronous speech studies. On the other hand, the syllable may not be a reliable unit to describe tempo variation for English. This is partly due to the fact that variability of syllables highly depends on stress involvement: in particular, a schwa does not vary by tempo as much as a stressed vowel does [4, 5]. In other words, in stress-timed languages, changes in speech rate would mostly influence stressed syllables, but not unstressed syllables to the same degree.

We examine global and local patterns of speech rate variation in Mandarin Chinese, where no segmental reduction is expected. We expect this language to show a relatively stable speech rate when measured in "the number of syllables produced per second [6 as cited in 4]".

2 Experiments

A total of 8 Mandarin Chinese speakers were recruited in this study: 6 from Taiwan (4 females and 2 males) who are bilingual in the Southern Min dialect and Mandarin Chinese, and 2 female speakers from Shanghai, Mainland China, whose native language is Mandarin Chinese. Ages ranged from 24 to 35. Subjects were paired keeping gender and dialect homogeneous: P2 is excluded in the pair-wise comparison because S4 turned out to speak too little Min to be grouped with S3. Dialectal grouping is taken into consideration under the assumption that speech tempo is also a part of the speech style that belongs to a dialect. Table 1 summarizes the linguistic backgrounds of the subjects grouped by dialect and gender.

Pair	P1		P 2		P 3		P 4	
Region	**T**aiwan						**C**hina	
Nat. language(s)	**M**andarin Chinese							
	S. Min				**Q**uan Zhou		(Shanghai)	
	S1	S2	S3	S4	S5	S6	S7	S8
Gender	**F**		**M**		**F**		**F**	
Labeled as	TMS-F		TMS-M/ TM-M		TMQ-F		CM-F	

Table 1 Linguistic backgrounds of subjects for grouping

An anecdotal fable was selected as reading material: the text consisted of 7 sentences with 15 potential intermediate phrases (ips), which were indicated by a visual space between two phrases in the written text. The first two ips are excluded in the analysis because the beginning parts can vary widely depending on the speakers' initiation of reading. Each ip varies in terms of segmental composition and number of syllables, as shown in Table 2:

ip #	1	2	3	4	5	6	7	8	9	10	11	12	13	**Total**
# of syll.	12	8	10	5	6	7	6	6	7	7	10	7	7	**98**

Table 2 Number of syllables in the reading material

Two different transcriptions were used for the given text, making them suitable for the two major regional groups of speakers: one is transcribed using the archaic Chinese letters that are used in Taiwan, and the other is transcribed as in Mainland China. A copy of the text was mounted at the speaker's eye level, allowing speakers to look at each other, whenever necessary, by simply moving their eyes. Data were collected in a noise-proof recording room, using a digital recorder (Marantz PMD 660). Stereo channels were set for RT in order to collect two speech signals from two speakers in pairs simultaneously.

2.1 Read-Alone (RA) speech

RA recordings were obtained under the presence of a partner who will participate in the RT session afterward. This setting was intended to provide the speakers in pairs with chances to experience or monitor each other's speech tempo. For each recording, two speakers were seated facing each other with approximately 1 meter between them. After a 2 minute silent reading period, one speaker was instructed to read the entire text alone at a normal tempo signaled by a series of three isochronous beeps at 1 second intervals. After one completion of the first speaker, the second speaker took the turn also signaled by the beeps. In this way, 5 repetitions of Read-Alone were obtained for each speaker.

2.2 Read-Together (RT) speech

RA subjects also participated in a subsequent RT session, where speakers were instructed to read the text together with their partner followed by three isochronous beeps at 1 second intervals: an analogy of singing in a chorus was given as a part of the instruction. The speakers were asked to read at a normal tempo and as synchronously as possible. The speakers were cautioned not to discuss any kinds of strategies to improve their performance. No information on the read together task was provided in advance of the RT session. A total of 5 repetitions of RT reading were obtained for each speaker.

After each completion of RT repetitions, speakers were asked to evaluate their performance on a scale of 1-10, where 10 means "it could not be better". This was devised to ensure that the RT readings were reasonably well-performed and representative of RT characteristics. All speakers evaluated at least one reading at 9 or 10, and the average of evaluation amounts to 7.95, based on which we accept our RT data as well-performed ones.

3 Results

This section shows the results of speech rate (SR) variation when the speakers read together in a pair. Global and local aspects of speech rate variation will describe the temporal characteristics of RT reading. The global pattern refers to the general mean differences in speech rate across repetitions and subjects. The local variation looks at the speech rate variation at ip levels.

Speech data collected from 4 pairs were analyzed at ip levels within each utterance. A total of 80 utterances are analyzed: 2 speech types (RA/RT) × 5 (repetitions) × 8 (subjects).

3.1 Global variation of speech rate in RA/RT

Two speakers in pairs were given the chances to monitor the tempo characteristics of their partner, set by the given experiment environment. Therefore, when they were asked to read the text together, the experience of the partner's speech tempo in RA session is supposedly used in the RT task. Keeping this condition in mind, let us look at Fig.1, where the mean speech rate for each subject is specified with the standard deviation values across repetitions. RA and RT results are shown in the figure in comparison.

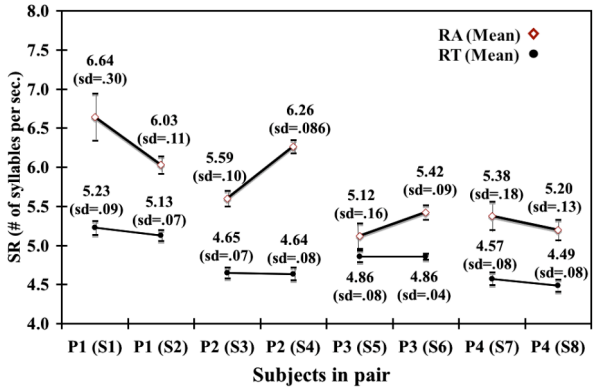

Fig. 1 Mean SR of RA and RT across repetitions

The slope of each line illustrates the degree of differences in terms of averaged speech rate between two subjects in each pair: the greater the slope of the line, the bigger the difference between the two.

First, we describe the speech rate variation referring to the mean speech rate values (MSR) within a subject and across subjects. One of the most consistent patterns found in Fig.1 is that speakers lower their speech rates when they read the text with a partner. This pattern is consistent across pairs. The mean SR across subjects is 5.5 (sd=.54) for RA and 4.8 (sd=0.26) for RT. The mean difference between RA and RT describes that speakers slowed down their speech in RT by 0.90 SR in average.

A pair-wise comparison is shown in Table 3, where the Mean speech rate of each pair is quantified. The speech rates by the two reading types are also compared within a pair.

Pair	P1	P2	P3	P4
Gender	F	M	F	F
Mean SR in RT	5.18	4.65	4.86	4.53
Mean SR in RA	6.33	5.93	5.27	5.29
SR diff. bt. Pairs: RT/RA	|P1-P2|=0.53/0.40		|P2-P3|=0.21/0.66	
	|P1-P3|=0.32/1.06		|P2-P4|=0.12/0.64	
	|P1-P4|=0.65/1.04		|P3-P4|=0.33/0.02	
Averaged MSR diff. in RT/RA	MSR (RT) = 0.36 MSR (RA) = 0.64			

Table 3 Pair-wise Mean SR comparison in RA/RT

Each pair shows a decreased speech rate in RT, and this tendency is consistent across pairs. The averaged Mean SR difference by reading types describes that every pair in RT shows a slower speech rate than in RA and the pair-wise speech rate variation is reduced in RT by 0.28 SR.

Now we look at the speech rate variability within a subject and across subjects, referring to standard deviation values. In RT, as shown in Fig.1, the within-subject standard deviation varies ranging from the smallest .086 SR (S4) to biggest .30 SR (S1). This variation range become reduced in RT resulting in the range between 0.04 SR (S6) and 0.09 SR (S1). This reduction in standard deviation is consistent across subjects, which means that every speaker shows a more consistent speech rate in RT than in RA: in other words, the RT speech rate is less variable across repetitions within a speaker. This is also seen in the standard deviation values averaged across subjects by the two reading types: RA (Avg., 5.5; sd=0.54) and RT (Avg., 4.8; sd=0.26) across subjects. Here we see that the SD value across subjects in RT is reduced to the half of the SD in RA.

One interesting question arises from Fig.1 when looking at the speech rates between the subjects in P4. These two speakers had very close speech rates to each other: 5.38 and 5.20 for S7 and S8 respectively. Therefore, one may question that it may be unnecessary for these two speakers to adjust their speech rates because the speech rates are already similar to each other. However, both of the speakers lowered their speech rates in RT by 0.81(S7) and 0.71 (S8) respectively. The question is why speakers should change their speech rates even though their normal speech tempi are close to each other. This question leads us to look inside the speech rate variation at ip levels within a subject.

3.2 Local variation of speech rate in RA/RT

For the local variation analysis, we look at speech rate changes at ip units, within and across reading types. Since each ip has different segment composition, local speech rates supposedly vary across ips. In addition, there may exist dialectal differences in segment quality and duration as well as tone differences, all of which are involved in temporal organization. Therefore, we look at local variation of speech rate by subject and dialect: the comparison was made between P1 (TMS-F) and P4 (CM-F), which are homogeneous in gender but most distant in terms of dialectal variation.

Fig. 2 illustrates the speech rate variation over 13 ips for P1 (S1 and S2) and P4 (S7 and S8). Two different repetitions (the first and the fifth ones) for each speaker are chosen to examine their local variation across repetitions. Note that by looking at local variation, we are interested in how local speech rate varies in the two reading conditions (dotted lines for RA and solid ones for RT).

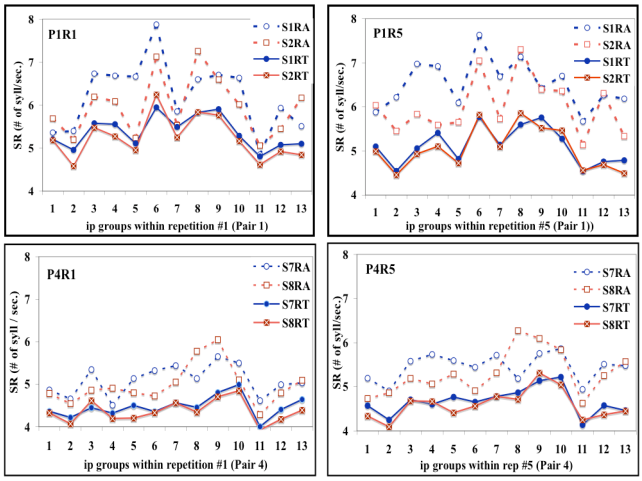

Fig. 2 Local SR variation in RA and RT from P1 (Top) and P4 (Bottom): the first (L) and the 5th (R) repetitions.

As in Fig 2, the RA variation between two subjects within a pair is more variable than that for RT: in the upper two figures taken from P1 (S1 and S2), the local speech rates in RA (dotted lines) exhibit substantially variable difference for each ip whereas those in RT (solid lines) are very close to each other. This pattern is consistent across pairs (P1 and P4: top to bottom) and across repetitions (R1 and R5: left to right).

In summary, the speakers decrease their speech rates in order for them to ensure less variability in RT. This is supported by the speech rate pattern found in Pair 4, where the two speakers reduce the mean speech rates regardless of their similar speech rates even in RA. Accordingly, less variability in speech rate is found to be a main characteristic of RT.

Before we conclude our discussions, we will explore further implications behind the slow-down strategy and the reduced variability, which are described by the distribution of the mean SR and by the standard deviation, respectively. The next section introduces a perceptual notion of "Just Noticeable Difference," based on which further discussions are made beyond the statistical descriptions.

3.3 Interpretation of SR differences

There has been a study on "Just noticeable difference for tempo (JND) [7]", which describes how much difference in speech rate is needed to make two utterances perceived as different. It has been reported that a 5% tempo change from the original recording could be perceived as being faster or slower by Dutch speakers [7].

This notion is interesting to see what the SR differences found in RA and RT can tell us more about speech rate characteristics in synchronous speech. Before we apply this notion to our results, we verify whether the 5% differences between readings are applicable to our results. In order to do so, we look at the speech rate variation obtained from one speaker (S4: TM-M) who provided additional readings at various tempi with the same text material: Normal (N1≈N2≈N3), Slow (SL1≈SL2) < Slower (SL3), and Fast (F1) < Faster (F2) << Fastest (F3). This instruction was made in order to see how much SR differences are produced in changing speech tempi. Figure 3 illustrates various speech rates by the instructed tempo categories.

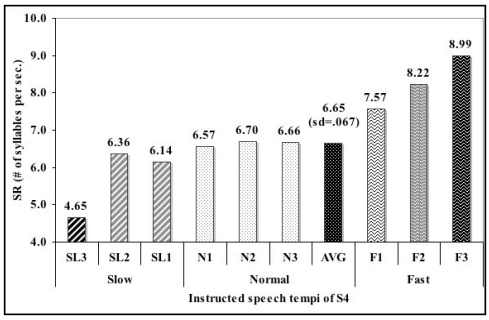

Fig. 3 RA readings of S4 at varied tempi.

The speech rate differences among normal speech (N1, N2, and N3) are smaller than JND values of each reading, which are 0.33, 0.34, and 0.33 respectively. On the other hand, the SR differences between two tempo categories are all greater than the corresponding JND values: for example, SR (|N3-SL1|) > JNDs (N3 and SL1). The results from SR differences between readings in S4 can be summarized as follows: speech rate differences are bigger than JND across categories and smaller within each category. This result is consistent with the instructions given for a speech tempo (SL3 < SL2, SL1 <N1, N2, N3 < F1 < F2 <F3).

Based on these results, 5% of JND reference can be said to properly work within and across tempo categories. In addition, we can understand the SR differences in an explanatory way: for instance, the value of 0.52 SR is enough to shift speech tempi (from N3 to SL1).

Now let us revisit the results found in RA/RT readings. If a mean speech rate difference between two subjects/readings exceeds the corresponding JND value(s), we will interpret that the two subjects/readings show perceptually different speech rates. Table 4 shows our JND analysis by comparing the mean SR with JND values, and describes the results by the reading types (RA/RT).

The JND analysis can be summarized as follows: the speakers within a pair (P1-P3) show bigger SR differences than 5% JND values in RA, and smaller differences in RT. This analysis enables us to say that the RA speech rates within a pair were not perceived as being identical, but as identical in RT.

Pair ID		P1		P2		P3		P4	
Subject ID		S1	S2	S3	S4	S5	S6	S7	S8
Mean SR	RA	6.63	6.03	6.26	5.59	5.12	5.42	5.38	5.20
	RT	5.23	5.13	4.65	4.64	4.86	4.86	4.57	4.49
5% of SR (JND)	RA	0.33	0.30	0.31	0.28	0.26	0.27	0.27	0.26
	RT	0.26	0.26	0.23	0.23	0.24	0.24	0.23	0.22
JND analysis (within pairs)	RA	0.61> JND(S1,S2)		0.67> JND(S2,S3)		0.30> JND(S5,S6)		0.18< JND(S7,S8)	
	RT	0.1< JND(S1,S2)		0.01< JND(S2,S3)		0.0< JND(S5,S6)		0.08< JND(S7,S8)	

Table 5 JND analysis within pairs

Another point, though not shown in the table, has to be made: the mean distribution across subjects/pairs in RT shows a quasi-convergence ranging from 4.49 to 5.23 SR. Less variation found in RT between subjects may not be a surprising result because the synchronization task itself entrains the speakers within a pair. However, it is surprising and interesting to find that the RT variation across subjects/pairs is also reduced, knowing that there is no explicit dynamic entrainment across pairs. This might be because speakers reduce the variability of speech rate to the degree that their speech rates are perceived identical. Results show that the JND values in RT ranges from 3.23% to 5.7%, which converges to around 5% JND, while it ranges from 3.9% to 11% in RA. The JND value might be language-specific and requires further studies on Chinese. Alternatively, less variation across subjects in RT can be viewed as an optimal effort for speakers to reduce their variation in speech rate. In other words, RT speech between speakers is slowed down and converging to a point that the variability of speech rates is effectively suppressed. We leave these questions for further studies.

4 Conclusion

The Read-Together context is interesting to show how speakers adjust their speech tempo in the synchronization task. The results show that all the speakers take a slow-down strategy at a converging point exhibiting less variability.

Considering that RT context is an entrainment situation between two dynamic speakers of different frequencies in terms of speech rate, one possible way to resolve the different speech tempi can be to adjust the rates between the SRs of the two speakers in RA. As Fig.1 showed, no such case was found. More importantly, P4 does show the slow-down effect even though the speakers in P4 are not different from each other in speech rates. The local variation of speech rates described in section 3.2, demonstrates that speakers in pairs can effectively decrease variability of speech rates, which is a hallmark of synchronous speech, by slowing down both their speech in RT condition.

Finally, our JND analysis shows that speech rates in RA can be perceptually different but identical in RT between pair-wise subjects. This implies that speech rates in RT converge within the range that speakers perceive them identical, which is not the case in RA.

Acknowledgement

Preparation of this paper was supported in part by the National Science Foundation under Grant No. 0325188.

References

[1] Fred Cummins, "On synchronous speech", *Acoustic Research Letters Online* 3(1), 7-11 (2002)

[2] Fred Cummins, "Speech synchronization: investigating the links between perception and action in speech production", *ICPhs XVI*, Saarbrucken, 529-532 (2007).

[3] Elena Zvonik, Fred Cummins, "Pause duration and variability in read texts", *Eurospeech*, Geneva, CH, 777-780 (2003)

[4] Brigitte Zellner, "Fast and slow speech rate: a characterization for French", *ICSLP, 5th International Conference on Spoken Language Processing*, Vol. 7, 3159-3163 (1998)

[5] Thomas H. Crystal, Arthur S. House, "Articulation rate and the duration of syllables and stress groups in connected speech", *J. Acoust. Soc. Am.* 88(1), 101-112, (1990)

[6] F. Grosjean, A. Deschamps, "Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes", phénomèmes d'hésitation. *Phonetica,* 31, 144-184 (1975)

[7] Hugo Quene, "On the just noticeable difference for tempo in speech", *Journal of Phonetics* 35, 353-362 (2007)