Variability of speech rhythm in synchronous speech

Volker Dellwo & Daniel Friedrichs

Phonetics Laboratory, University of Zurich, Switzerland
Volker.dellwo@uzh.ch, Daniel.friedrichs@uzh.ch

Abstract

Speakers are able to speak in synchrony to another speaker or to a recording of another speaker. The present research studied whether and, if yes, speakers change their speech rhythm when synchronizing to another speaker. We developed a measure (SRratio) which monitors on a scale between 0 and 1 whether the durational characteristics of a speaker's synchronous speech are closer to his/her own read speech or closer to the characteristics of the speech of the speaker he/she is synchronizing to. Four speakers (synchronization speakers) synchronizing to twelve recorded sentences produced by four other speakers (target speakers) were studied. The durational characteristics we analyzed were %V and nPVI-v. Results for SRratio suggest that complex processes are going on with main effects for synchronization speakers and target speakers and interaction of the two factors.

Index Terms: speech prosody, rhythm, synchronous speech

1. Introduction

Speech rhythm has been studied widely in terms of the durational characteristics of consonantal (c) and vocalic (v) intervals (Ramus et al., 1999, Grabe & Low, 2002, Loukina et al., 2011). For this reason a large variety of measures (henceforth: rhythm measures) has been developed over the past decade as for example the percentage over which speech is vocalic (%V, Ramus et al., 1999), the Pairwise Variability Index (PVI; Grabe & Low, 2002) which calculates the average durational difference between consecutive v or c intervals in an utterance or the standard deviation of c or v intervals (deltaC and deltaV respectively; Ramus et al., 1999). Numerous variants of these measures based on different rationales were further developed for example by normalizing for speech rate variability (VarcoV, VarcoC; Dellwo, 2006, White and Mattys, 200), by taking into account different combinations of c and v intervals (Barry et al., 2009) or by using different interval types altogether (e.g. voiced and voiceless intervals; Dellwo et al., 2007). For an in depth overview of different rhythm measures see Loukina at al., (2011) .

The original motivation behind the development of these measures was that they were believed to be correlates of auditory language specific rhythmic characteristics (Ramus et al, 1999, Grabe and Low, 2002), as for example whether a language reveals higher perceptual syllabic regularity (syllable-timed languages) or inter-stress interval regularity (stress-timed languages). This process is problematic, however, as it remains a matter of great debate to what degree languages vary in terms of auditory speech rhythm (Dauer, 1983) and if such perceptual categories exist, how they are reflected by rhythm measures (Arvaniti, 2009).

By now there is large body of evidence in the literature that rhythm measures are language specific (rather than rhythm class specific) but that rhythm measures also show considerable within language variability, for example as a factor of accent (White and Mattys, 2007), speaker (Yoon, 2010) or utterance (Wiged et al., 2010) or it can indicate degrees of speech disorders (White et al., 2010). In our lab at Zurich University we are currently particularly interested in the between speaker variability and how knowledge about between speaker durational variability can be applied to forensic speaker identification. In pilot studies we found, for example, that %V remains stable between speakers even when within-speaker rate variability is high (Dellwo & Koreman, 2008).

The present study is in the vein of speaker specific rhythmic characteristics and we wanted to know whether and, if yes to what degree, acoustically measurable rhythmic characteristics change when a speaker imitates the rhythm of another speaker. To make people imitate rhythmic characteristics of other speakers we asked them to speak in synchrony to recordings of other speakers' voices (Cummins, 2002, 2009) similar to close-shadowing reading techniques (Marslen-Wilson, 1973). Even though there may be effects on the rhythmic characteristics of a speaker as a result of synchronization difficulties (in particular when synchronizing to recorded speech; Poor & Ferguson, 2008) we believe that speakers will (a) changer the rhythmic characteristics of their speech and (b) possibly adapt rhythmic features of another speaker under such a condition. Consider, for example, speaker x who typically spends little time on vocalic parts in speech (i.e. has a low %V) and speaker y who does the opposite (i.e. has a high %V). It seems possible that speaker y takes on this feature when synchronizing to x's voice. There may also be speaker specific differences in the way that speakers which are better at synchronizing their speech to another speaker will also adapt more of this speaker's rhythmic characteristics while speakers which have problems synchronizing to other speakers might change their rhythmic characteristic in a more random fashion.

For the present experiment we had speakers speaking in synchrony (sync speakers) to recordings of read sentences from other speakers (target speakers). In addition we obtained read versions of the same sentences by the sync speakers. To introduce more temporal (possibly rhythmic) variability in the speech of the target speakers, we recorded native and non-native speakers of the language under investigation (German). We then analyzed how the synchronous version of the sync speakers compares rhythmically to their own read speech and to the read speech of the target speakers. To perform this comparison we developed a measure that indicates where the rhythm of the synchronous speech version of a sync speaker lies on a scale from 0 (read version of the sync speaker) to 1 (read version of the target speaker).

2. Method

2.1 Subjects

Eight speakers of German (4 f, 4 m, age between 20 and 30) took part in this experiment. This group was subdivided into two equally sized subgroups of 4 target speaks and 4 sync speakers (genders distributed equally across the subgroups).

In the target speaker groups two speakers were native speakers of German while the other two were native speakers of Italian with high competence in German but clearly audible Italian accents. All speakers of the sync group were German native speakers.

2.2. Material

Three German sentences were used as reading material for this experiment:

- *Die Frau des Apothekers weiss immer was sie will*,
- *Das Theater hat viele neue Aufführungen geplant*,
- *Er wollte sich seiner Schwächen einfach nicht bewusst werden*

All speakers were recorded in a sound-treated booth with high-end digital recording equipment. Read versions of all three sentences (above) were recorded by each of the eight speakers in the experiment. In addition the sync speakers recorded the sentences synchronizing to the read versions of each sentence of each of the target speakers. This resulted in 12 synchronous sentences (3 sentences *4 sync speakers) for each sync speaker, making 48 synchronous sentences in total (12 sentences * 4 sync speakers).

2.3 Recording procedure

To record the read versions of the sentences speakers read the sentences from a piece of paper in the recording booth. Sentences for the synchronous speech condition were presented to speakers over headphones as done in Poore & Ferguson (2008). Sentences were preceeded by three 1 kHz sinusoids of 50 ms duration that were spaced by silent intervals of 500 ms. The interval between the last beep and the onset of the sentence was also 500 ms so that listeners could find the point when they would have to start the synchronization. For the synchronous speech productions speakers did not see the sentences in a written form. Instead they heard the sentences (including the beep tones) five times in a row with one second intervals between the offset of the sentence and the onset of the next beep tone. Speakers were asked to listen to the first presentation and speak in synchrony to all other presentations (two to five). Sync speakers were typically very well in synchrony with the target speakers by the fifth presentation of the sentence. All sentences were presented to speakers in one session between the presentation of five repetitions of one sentence to the presentation of the next sentence there was a 3 second interval. The total duration of the synchronization session was under 15 minutes to avoid voice or concentration fatigue of speakers.

2.4 Data editing

The onset and offset of each segment was labeled manually in the read and synchronous sentence recordings using Praat's annotation function (www.praat.org). From the segment labeling we automatically produced a labeling (in a different tier in a Praat TextGrid) of consonantal and vocalic intervals by turning segment labels into their respective sound category labels (consonant or vowel) and combining consecutive consonantal or vocalic segments into consonantal and vocalic intervals respectively.

2.5 Computing results

Based on the durational information from consonantal and vocalic intervals we calculated rhythm measures for each sentence. Amongst the wide availability of different metrics (Loukina et al, 2011) we picked two that have typically been functional in other domains of rhythm (e.g. between-language rhythmic characteristics):

- %V: The percentage over which speech is vocalic (Ramus et al., 1999)
- nPVI-v: The rate normalized average difference between consecutive vocalic intervals in an utterance (Grabe & Low, 1999).

To analyze how rhythmic measurements of the synchronous speech can be quantified between the read version of the sync speaker and the respective read version of the target speaker

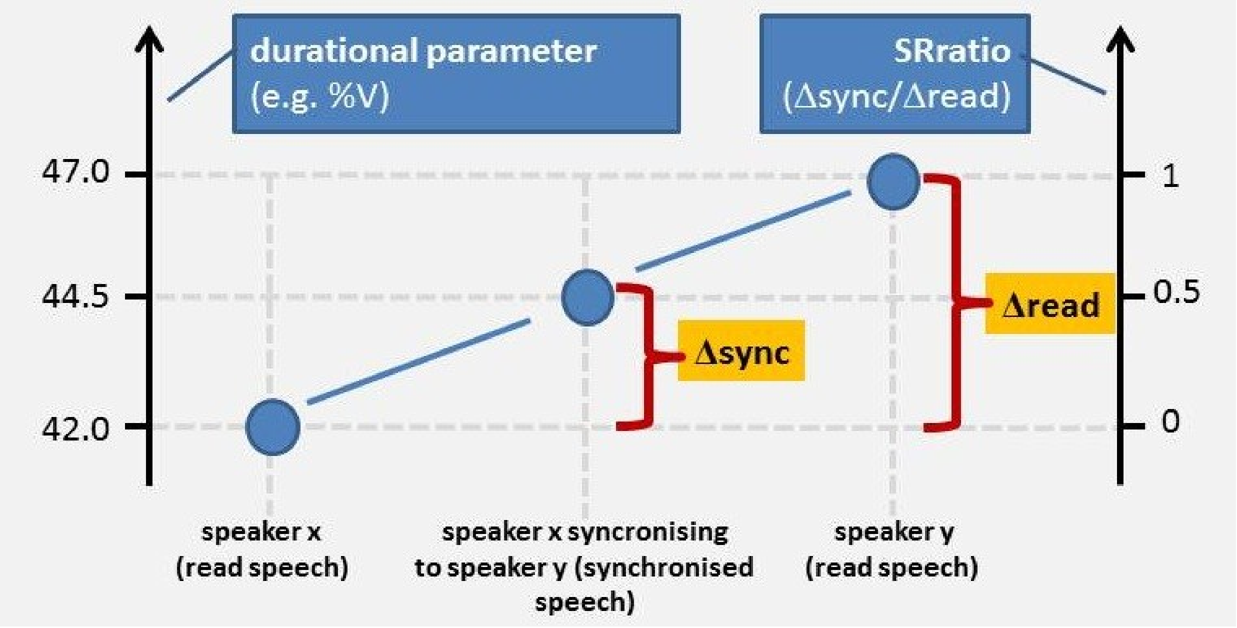

Figure 1: A sketch for the calculation of SRratio (right scale) based on three rhythmic measurements (here: %V, left scale): the read speech of a sync speaker (speaker x; left dot), the read speech of a target speaker (speaker y; right dot) and the synchronous speech of a sync speaker (speaker x; middle dot).

we created a measure (SRratio). Figure 1 contains a sketch of the relevant parameters for this calculation. In the figure the sync speaker is labeled as speaker x and the target speaker is speaker y.

We first computed the difference between the rhythm of the synchronous speech version and the read version for each sentence produced by each sync speaker for each rhythm metric (deltasync). Then we calculated the difference between the read version of the same sentence for the target speaker and the read version of this sentence for the sync speakers (deltaread). Finally we calculated the ratio between deltasync and deltaread (deltasync/deltaread), wich is the SRratio. This computation turns the three values of the absolute scale for a rhythm measures (here %V for the read version of the sync speaker [left dot] the synchronous version of the sync speaker [central dot] and the read version of the target speaker [right dot]) into one single value on the SRratio scale (right scale in the figure). This scale can then be read and interpreted in the following way:

- SRratio = 0: There is no difference between the synchronous and the read version of a sync speaker. This could possibly mean that the rhythm measure reveals maximum speaker idiosyncrasy as there is no change in the particular durational characteristics when synchronizing.
- SRratio = 1: There is no difference between the synchonsized speech of a sync speaker and the read speech of a target speaker. This could possibly mean maximum adaptation of a sync speaker to a target speaker.

We further have two outcomes that are difficult to interpret at the present point but might mean that sync speakers were not well capable of performing the synchronization task:

- SRratio < 0: The sync version of a speaker is lower than both the read version of the sync speaker and the target speaker.
- SRratio > 1: The sync version of a speaker is higher than both read speech of the target speaker and the sync speaker.

2.6 Data exclusion

In cases where the read version of the target speaker was close in measurable rhythm or rate to the read version of the sync speaker deltaread resulted in very small absolute values. In these cases small changes in deltasync resulted in dramatically high or low SRratio values. These cases, however, are also of minor interest: when there is no rhythmic difference for a certain parameter between the sync and the target speaker we would not expect the synchronous version to vary considerably. On the other hand, for an evaluation of how good our measurement procedure works these values might be very interesting as they can show us how much the synchronous speech may vary when both the target and the sync read versions are close to being the same. At the current point we have not yet implemented these results into the model but we are planning to do so in the future. For now we have excluded all values that resulted from this condition. The number of values excluded for each rhythm measure was not higher than 10% (i.e. not more than 5 of 48 values were excluded), which reveals that these cases were not very common.

3. Results & Discussion

Results for SRratio for %V are shown in Figure 1, for nPVI-v in Figure 2. Both figures show the results (a) by sync speakers (top charts) and (b) by target speakers (bottom charts). All charts show that there is considerable variability of SRratio as a factor of both target and sync speaker for both %V and nPVI-v. ANOVA tests revealed that all main effects of either sync speaker or target speaker were significant ($p<0.05$) to highly significant ($p<0.005$). However, we also found significant interactions between sync speaker and target speaker, as can be expected given the high descriptive variability between sync and target speakers. This situation makes it difficult to interpret the main effects but a few interesting features can be detected at this point.

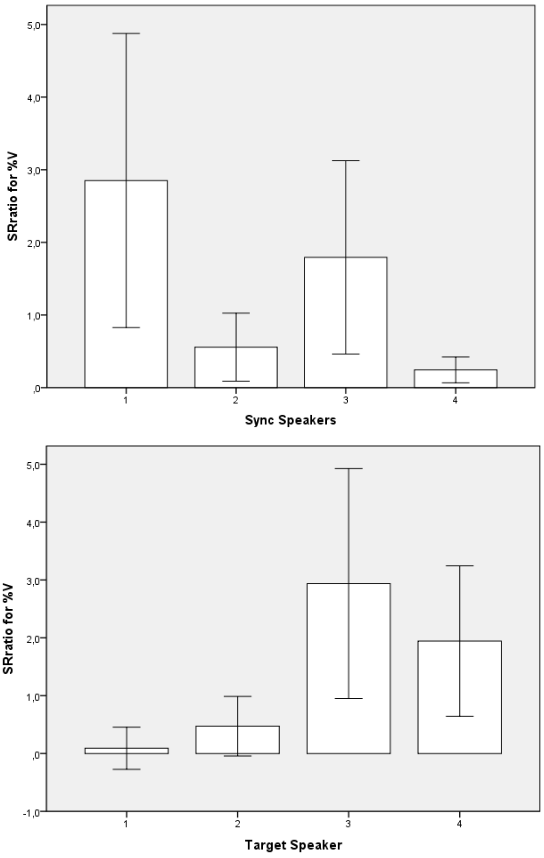

Figure 2: Bar chart of results for SRratio for %V by sync speakers (top chart) and target speakers (bottom chart). Top of bars show mean values, the whiskers show +-1 standard error.

For SRratio in case of %V we see that in case of target speaker 1 there is only very little adaptation. Values remain more around 0, while values are between 0 and 1 for target speaker 2. It is possible that target speaker 2 contains features that result in a different %V to the sync speakers and to which the sync speakers tend to adapt.

It is also very interesting to note the high SRratio values for %V in case of target speaker 3 and 4 (mean values around 3 and 2) which indicates a strong target overshoot. These were the two non-native speakers of German so it appears that the native German sync speakers did not find it as easy to synchronize to German with an Italian accent. This seems plausible as time domain features are possibly much more unpredictable in German L2 speech than in L1 speech (White & Mattys, 2007). The German speakers, incapable of synchronizing to these temporal characteristics, thus produce versions that are neither close to their own speech nor to the speech of the target speakers anymore.

Given this result we must also take into consideration that the sync speakers 1 and 3 were most affected by this (top graph in Figure 2) as their results for SRratio should be most responsible for the high %V values in case of targets 3 and 4 (bottom graph). We can possibly conclude here that target speakers affect the production of %V in different ways in synchronous speech and that synchronization speakers are also affected to very different degrees.

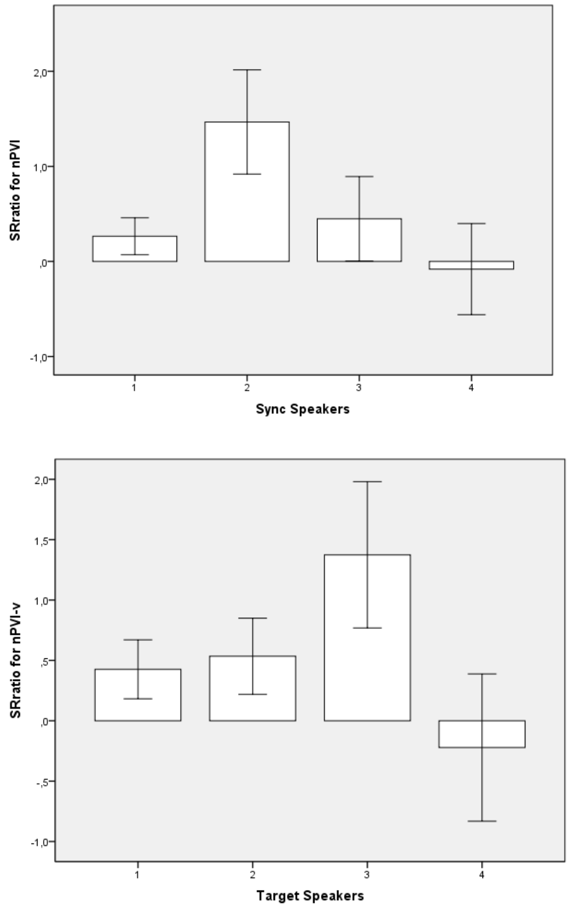

Figure 3: Bar chart of results for SRratio for nPVI-v by sync speakers (top chart) and target speakers (bottom chart). Top of bars show mean values, the whiskers show +-1 standard error.

In Figure 3 we see a very similar situation for SRratio in case of nPVI-v as we saw it in the case of %V. Again there is high variability as a factor of sync and target speakers with sync speakers one and two least affected by the durational characteristics of their target peers (top graph). For the target speakers we find that speaker 4 leads sync speakers to adopt a lower nPVI-v in their synchronous compared to their read speech.

4. Conclusions

In the present paper we have analyzed changes of speech rhythm in synchronous speech. Interpreting the data presented here we need to be aware that we are dealing with a small number of measurements for a rather large number of different conditions at this point (different sync and target speakers, native and non-native speakers amongst target speakers). We are currently collecting more data for smaller number of conditions to gain more descriptive and statistical power but we can probably conclude at the current point that in synchronous speech acoustically measurable rhythm is affected in different ways depending on the speaker synchronizing to someone else and the speaker someone synchronizes to.

So far we have only looked at a small number of 'rhythm measures' based on durational characteristics of c and v intervals. We are currently extending the number of parameters and are also working on including parameters that are based on syllable peaks and prosodic contours rather than interval durations as these measures might capture the actual rhythmic changes between normal and synchronous speech more effectively.

5. References

[1] Arvaniti, A. "Rhythm, timing and the timing of rhythm", Phonetica 66: 46-63, 2009.

[2] Barry, W., Andreeva, B., and Koreman, J. "Do rhythm measures reflect perceived rhythm?" Phonetica 66: 78-94, 2009.

[3] Cummins, F. "On synchronous speech" Acoustic research letters online, 3(1): 7-11, 2002.

[4] Cummins, F. "Rhythm as Entrainment: The Case of Synchronous Speech" Journal of Phonetics, 37(1): 16-28, 2009.

[5] Dauer, R. "Stress-timing and syllable-timing reanalyzed", JPhon 11: 51-62, 1983.

[6] Dellwo, V "Rhythm and speech rate: A variation coefficient for deltaC", Language and Language Processing: Proceedings of the 38th Linguistic Colloquium, 231-241, 2006

[7] Dellwo, V., Fourcin, A. and Abberton, E. "Rhythmical classification of languages based on voice parameters", Proceedings of ICPhS XVI: 1129-1132, 2007

[8] Dellwo, V. and Koreman, J. "How speaker idiosyncratic is acoustically measurable speech rhythm", Abstract in Proceedings of IAFPA meeting, Lausanne/Switzerland, 2011.

[9] Grabe, E. and Low, E. L. "Duration variability in speech and the rhythm class hypothesis", Laboratory Phonology 7: 515-546, 2002.

[10] Loukina, A., Kochanski, G., Rosner, B. and Keane, E. "Rhythm measures and dimensions of durational variation in speech", J. Acoust. Soc. of Am., 129(5): 3258-3270, 2011.

[11] Marslen-Wilson, W. "Linguistic structure and speech shadowing at very short latencies." Nature, 244: 522-523, 1973.

[12] Poor, M. A. and Ferguson, S. H. "Methodological variables in choral reading", Clinical Linguistics & Phonetics, 22(1): 13-24, 2008.

[13] Ramus, F., Nespor, M. and Mehler, J. "Correlates of linguistic rhythm in the speech signal", Cognition 73: 265-292, 1999.

[14] White, L. and Mattys, S. L. "Calibrating Rhythm: First and second language studies", JPhon 35: 501-522, 2007.

[15] Wiget, L., White, L., Schuppler, B., Grenon, L., Rauch, O.l and Mattys, S. "How stable are acoustic metrics of contrastive speech rhythm? " J. Acoust. Soc. Am. 127(3): 1559-1569, 2010.

[16] Yoon, T.J "Capturing inter-speaker invariance using statistical measures of speech rhythm", Electronic Proceedings of Speech Prosody, Chicago, 2010.