Rhythmic variability between some Asian Languages: Results from an automatic analysis of temporal characteristics

Volker Dellwo[1], *Peggy Mok*[2], *Mathias Jenny*[1]

[1]Department of Comparative Linguistics, University of Zurich
[2]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
{volker.dellwo, mathias.jenny}@uzh.ch, peggymok@cuhk.edu.hk

Abstract

The rhythmic organization of speech can vary between languages. In the present research we studied rhythmic variability between Mandarin, Cantonese and Thai using automatically retrieved prosodic temporal characteristics from read speech. We measured the variability of intervals between amplitude peaks in the amplitude envelope (<10 Hz) and the durational characteristics of intervals with and without glottal activity (voiced and unvoiced intervals) in speech. Results for between language comparisons revealed significant differences between languages in both amplitude peak interval variability and voiced-voiceless interval durational characteristics. Results are discussed in connection with language specific phonotactic/phonological properties and hypotheses about the perceptual significance of the acoustic measurements in terms of speech rhythm.

Index Terms: Speech rhythm; Southeast Asian Languages

1. Introduction

In the present study we (a) demonstrated two methods to measure speech rhythmic characteristics automatically, (b) applied these methods to the Asian languages Cantonese, Mandarin and Thai and (c) discussed whether structural differences in the linguistic organization of these languages might have an influence on the obtained acoustics rhythmic differences. In the following we motivate the methods and the choice of languages and formulate hypothesis about the acoustically measurable variability we expected in the three languages.

Speech rhythm is a complex phenomenon that arises through the temporal organization of speech on multiple levels. While it had been assumed for a long time that the durations of syllables and/or feet are to a large degree responsible for our percept of rhythmic variability between languages ([1], [15]) the assumption lacked acoustic evidence (see discussion in [17]). This does not seem surprising because syllabic durations are based on the on- and offsets of syllables, which are often highly arbitrary in the acoustic signal. So acoustic measurements of syllable durations must inevitably be noisy and it is questionable whether listeners may judge rhythmic characteristics of speech in terms of syllable durations.

There are probably markers that carry acoustically and perceptually more salient information about the durational structure of speech. More recent approaches argued that durational variability of consonantal and vocalic intervals are to a high degree responsible for rhythmic differences between languages ([8], [17]). While these approaches were shown to be more promising in terms of perception ([17]), listeners still rely on a significant amount of phonological knowledge to identify numerous vocalic and consonantal boundaries (e.g. vowel-approximant-vowel boundaries). For this reason [6] studied durational characteristics of voiced and unvoiced parts of a signal, i.e. parts of the signal with and without glottal activity. They found that the proportional duration over which speech is voiced or the durational variability of voiced intervals can vary significantly between languages. They assumed that such acoustic variability should be highly salient in terms of perceptual rhythmic impressions of speech. Further, there are approaches that study durational characteristics of the speech amplitude envelope ([10], [18]). Since the temporal organization of amplitude peaks in speech might considerably contribute to perceptual beats in speech, such approaches are again likely to be close to the perceptual phenomena of speech rhythm. Both the identification of voiced and voiceless intervals in speech ([6]) as well as amplitude peak information ([12]) can be retrieved from the signal with little arbitrariness and can thus be carried out automatically with appropriate precision and in 2 we demonstrate two methods for doing this.

Why did we choose Cantonese, Mandarin and Thai? These languages reveal a number of structural and prosodic differences that are likely to influence the acoustic durational characteristics under investigation. The following overview is based on [7], [9] and [14]: Historically, Mandarin is drastically reduced in syllable structure compared to Cantonese, in particular in respect to vowel duration and syllable final consonants. Cantonese maintained a higher proportion of final consonants in syllables (there are only two final consonants, -n and -ng. in Mandarin while there are six in Cantonese, -m, -n, -ng, -p, -t, -k), which lead to many historically different words to develop into homophones in Mandarin, resulting in a larger number of disyllabic words. Concerning stress patterns, numerous of these disyllabic words do not reveal word stress, however, words with neutral tone on the second syllable are typically trochaic (i.e. accentuated on the first syllable). The highly frequent monosyllabic words in Cantonese are typically very equally stressed. When disyllabic words are built in Cantonese, both syllables maintain their tone with no reduction und typically no stress difference between the syllables. In summary, because of their structure and stress patterns, Mandarin has more alternating full and reduced syllables while Cantonese does not. This should inevitably lead to a higher variability between syllable peak points in Mandarin compared to Cantonese. We therefore hypothesized that a measure of the durational variability of the intervals between the peaks should be higher in Mandarin than in Cantonese.

The point that makes Thai interesting is that in terms of stress patterns it is somewhere between Mandarin and Cantonese. Both Thai and Mandarin have syllable reductions but in Thai they are on the first syllable (iambic) and in Mandarin on the second (trochaic). This should in principle result in a very similar overall organization of the amplitude envelope. However, the reductions are stronger in Mandarin than in Thai. In Thai, for example, the tones are not lost on the unstressed syllable while in Mandarin they are. Reduction in Thai is likely to be a reduction in syllable duration, possibly with some vocalic centralization but the reduction processes are typically rather weak. For this reason we hypothesized that variability measures of inter-peak-intervals in Thai should be stronger than in regularly stressed Cantonese but not as strong as in variably stressed Mandarin.

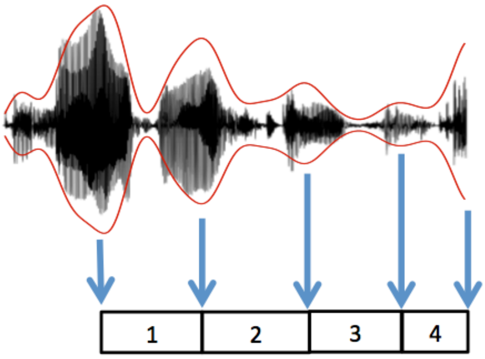

Figure 1: *The extraction process of inter-peak-intervals. An amplitude envelope (red line) is calculated from the waveform. Each peak point was extracted automatically with a peak filter and resulted in inter-peak-intervals (1-4).*

For the durational organization of voiced and voiceless intervals between the languages there are other language specific factors that might be influential. Concerning the consonantal structure Mandarin has no voiceless codas (only /n/ and /ng/ in syllable final position). Cantonese possesses all plosives and nasals in coda position but no fricatives. If plosives occur in final position they are not released. Because of the lack of voiceless codas there are no voiceless consonant cluster in Mandarin while in Cantonese they may exist across syllable boundary. Because of the resulting larger number of voiceless consonants in Cantonese we hypothesized that the proportional overall time during which this language is voiced (i.e. glottal activity is present) is lower compared to Mandarin. As the voiced intervals might more often be interrupted by the insertion of voiceless consonants in Cantonese, we further hypothesized that in Cantonese there is a higher variability of voiced intervals. In respect to Thai, it needs to be noted that it has a three-way distinction for initial stops (fully voiced, voiceless, and voiceless aspirated), and its final consonants are very similar to Cantonese (with an extra glottal stop). Also, Thai syllable structure is very similar to Cantonese. The only difference is that in Thai there are some initial clusters with /l/ and /r/ as the 2nd consonants (e.g. /kr/), but these clusters are often pronounced as single consonants in conversational speech (e.g. /k/). For these reasons it should be plausible to assume that Thai might reveal some overall durational characteristics that lie between Cantonese and Mandarin.

2. Data and methods

2.1. Subjects

7 native Cantonese, 7 native Mandarin and 4 native Thai speakers were recorded by the second author. All speakers, apart from 3 Thai speakers, were undergraduate students at the Chinese University of Hong Kong. None of them reported any speech or hearing related health disorders.

2.2. Recordings

Each speaker was recorded reading the North Wind and the Sun passage in their respective native language. For the Cantonese, Mandarin and 1 Thai speaker the recordings took place in a sound-treated room at The Chinese University of Hong Kong. The remaining 3 Thai speakers were recorded in a sound-treated room at the University of Hong Kong. Recordings were made directly on disk (sampling rate 22050 Hz, quantization: 16 bit). The speakers practiced reading the story as many times as they liked before the recording.

2.3. Data editing

Silences longer than 50 ms were automatically identified in the signal using Praat signal processing software (Boersma & Weenink, 2012; function: 'To TextGrid (silences)...'). The filled speech intervals between pauses (henceforth: IPI for inter-pause-intervals) were extracted from the recordings resulting in 189 IPIs in total (between 8 and 14 IPIs for each of the speakers; 77 for Mandarin, 73 for Cantonese and 39 for Thai).

The syllabic peaks were identified in the amplitude envelope of the signal. The amplitude envelope was extracted by (a) full wave rectifying the signal and (b) low-pass filtering the signal below 10 Hz. The resulting sound file is a wave (red line in Figure 1) where each peak is roughly the amplitude peak of a syllable. Peaks were identified automatically in the low-pass filtered wave using a peak filter written by the first author. A threshold was implemented to avoid the detection of very small peaks.

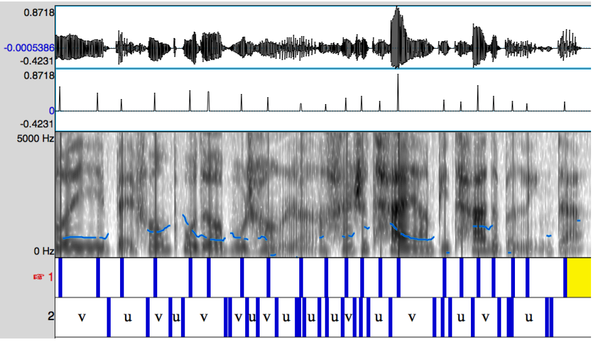

Figure 2: *Speech waveform (top) and waveform containing the extracted amplitude peaks (below). Spectrogram (middle) and TextGrid (bottom) containing the inter-syllabic-peak intervals on the first tier and the voiced-unvoiced intervals on the second tier.*

At each peak point an impulse was placed in the time-domain with the amplitude of the peak in the signal. The resulting waveform can be viewed in Figure 2 (second channel). For the ease of calculations a Praat TextGrid was created and an interval boundary was placed under each sample in the wavform that contained a peak.

Voiced and unvoiced intervals were identified using Praat's function 'To TextGrid (vuv)...' which is available for

PointProcess objects and creates a TextGrid object which contains the boundaries between voiced and unvoiced intervals. See example in Figure 2 (tier 2).

2.4. Measurement procedures & statistics

For each IPI we calculated the following variables:

- The standard deviation of inter-syllable-peak intervals (deltaPeak).
- The rate of inter-syllable-peak intervals (ratePeak).
- The standard deviation of voiced intervals (deltaVoiced)
- The ratio between the overall duration speech is voiced as opposed to unvoiced (%VO)

Previous research showed that the standard deviation of syllables, or consonantal and vocalic intervals is correlated positively with speech rate ([3], [19]) and a rate normalization method was suggested by [4]. This is based on calculating the logarithm of each interval duration to the base *e*. An additional effect of this method is that it turns typically positively skewed data distributions of interval durations into normally distributed distributions and thus makes the data suitable for statistical models like ANOVA. To avoid speech rate influence on our measures and to create normally distributed duration data we applied the log-transform method to our data. ANOVAs were carried out for each dependent measure with the three languages as a factor using the R software package. In case of significant main effects, Bonferroni corrected post-hoc paired t-test were calculated to identify effects between individual languages.

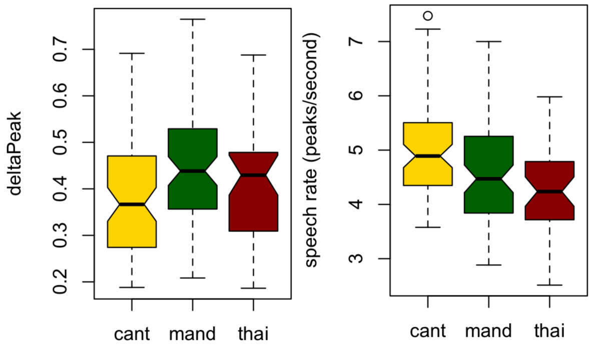

Figure 3: *Inter-peak interval variability (deltaPeak), top, and rate of inter-peak intervals, bottom, for Cantonese (cant), Mandarin (mand) and Thai (thai) natives (L1) and speakers of English (L2).*

3. Results

3.1. Inter-syllable-peak interval variability

The descriptive results for deltaPeak and the ratePeak can be viewed in Figure 3. For the inter-syllable-peak variability it can be seen that it is highest in Mandarin and lowest in Cantonese. Thai is more similar to Mandarin with a tendency of the second quartile to reveal lower variability. The effect was significant (F[2,186]=7.54, p<0.001). Post-hoc analysis showed that there was a significant difference between Cantonese and Mandarin (p=0.00044) but not between Thai and Mandarin (p=0.23) nor between Thai and Cantonese (p=0.47). For the rate data the main effect was also significant (F[2,186]=10.6; p<0.001). There were significant differences between Mandarin and Cantonese (p=0.014) and Cantonese and Thai (p=0.000038) but not between Mandarin and Thai (p=0.098).

3.2. Voiced-unvoiced variability

Figure 4 contains the durational variability of voiced intervals (deltaVoiced), left, and the percentage over which speech is voiced (%VO), right. It is visible that for both durational variables the two languages Cantonese and Mandarin differ most. For the variability of voiced intervals, Mandarin has the highest values and Cantonese the lowest. Thai is between the two languages, however, it appears that it is more similar to Mandarin. The ANOVA showed that the effect of language is again significant (F[2,186]=13.45, p<0.001). The descriptive similarities of the languages are confirmed by the post-hoc test. Both Cantonese and Mandarin (p=0.0000027) and Cantonese and Thai (p=0.0074) differ significantly but Mandarin and Thai do not (p=0.78).

For the proportional voicing it appeared that Mandarin was least voiced and Cantonese most, again with Thai between the two. However, in this case Thai appeared more similar to Cantonese. The main effect of language was again significant here (F[2,186]=23.36, p<0.001) and the post-hoc test was in line with the descriptive picture: Mandarin and Cantonese differed significantly (p=0.0000000012) and so did Mandarin and Thai (p=0.000092). There was no significant effect for Cantonese and Thai (p=0.67).

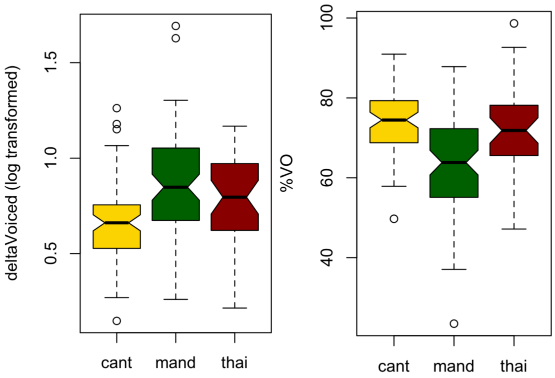

Figure 4: *Voiced-voiceless ratio (%VO), right, and durational variability of voiced intervals, left, for Cantonese (cant), Mandarin (mand) and Thai (thai).*

4. Discussion

We found language specific differences in automatically retrieved durational characteristics between Cantonese, Mandarin and Thai in the intervals between amplitude peaks of a low-frequency amplitude envelope as well as between voiced and voiceless interval durations. All between language effects were significant.

Concerning our predictions, the observed results for the inter-amplitude-peak variability fit well. We did find a higher inter-peak interval variability for Mandarin compared to Cantonese. It seems plausible that this pretty clear effect is driven to a high degree by the reduction patterns present in Mandarin as opposed to Cantonese. The Thai variability revealed more similarities with Mandarin. This is also plausible in terms of expected variability. There was, however, some tendency visible for a possible overlap with Cantonese. Given the weak reductions in Thai, a similarity with both Cantonese and Mandarin was therefore expected. From the shape of the box-plot in Figure 3 it can be inferred that the data distribution of the Thai inter-peak-interval variability was slightly positively skewed (the second quartile has a much larger range compared to the third). Because of the rather small amount of data for Thai (N=39) it was difficult to interpret this situation. It might be that it resulted from a bimodal distribution in which one peak is closer to the Cantonese mode, the other to the Mandarin mode. It is also possible that there was just more of a natural spread in the data. It would be interesting to explore this situation further with a larger dataset in the future.

The rate data in Figure 3 (right) revealed that Cantonese has the highest number of inter-amplitude-peak intervals, followed by Mandarin and Thai at the lowest end. One would thus assume that Cantonese is perceived as fastest, Mandarin and Thai might be perceived as slower. Interestingly, this was supported by a number of informants who are familiar with these languages. It will be interesting to see in the future whether the inter-amplitude-peak intervals are good correlates of perceptual speech tempo differences between languages and how they compare as predictors of speech tempo with commonly applied measures like the number of syllables per second. We would expect that the number of syllables and the number of inter-peak-intervals are possibly correlated to a high degree ([12]).

For voiced-voiceless interval characteristics the results did not conform with our expectations, in fact, they showed the exact opposite picture. The higher number of voiceless consonants in Mandarin did not lead to an overall higher %VO and the more frequent interruptions did not lead to durationally more variable voiced intervals. Comparing the inter-peak variability graph (Figure 3, left) with the voiced interval variability graph (Figure 4, left) we can see pretty much the same picture, namely that in both cases Mandarin revealed the highest variability, Cantonese the lowest and Thai had the tendency to show more overlap with Mandarin. This comparison suggests that maybe the segmental phonotactic differences between the languages have less of an influence than the lexical compositions and stress arrangements. The result is also in line with our previous research ([13]), which showed that Mandarin reveals higher variability in terms of consonantal interval durations compared to Cantonese. It might also be that other phenomena were responsible for the result that we have not taken into account yet. Cantonese and Thai, for example, have quantitative vowel contrasts but not Mandarin. It is possible that this phenomenon lead to the overall higher time that Cantonese and Thai speakers spent on vowels as opposed to consonants. It would again be interesting to explore these phenomena more and tease apart various factors that influence voicing patterns to better understand what durational measurements of voicing patterns tell us. Like with the tempo differences it would further be interesting to see whether the obtained acoustic differences in variability have a perceptual correlate. Given the results it seems reasonable to assume that Mandarin should be perceived as rhythmically more irregular than Cantonese with Thai being closer to Mandarin.

One drawback of the study was that the data sample was not particularly large (in particular for the Thai group). Given that the measurements can be carried out fully automatically on the data it seems feasible to analyze larger datasets in the future to test whether the results would replicate. It would also be interesting to investigate whether the between language variability is consistent when other sources of within language variability come into play (e.g. speaking style). In addition, we found in other studies that the measures under investigation vary within language as a function of speaker ([5], [14]). While it seems perfectly feasible that within and between language effects can be maintained it will be inevitable to tease apart the effects for a more in depth understanding of these variables.

5. References

[1] Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburg: University Press.

[2] Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer. Online www.praat.org, 05-15-2012.

[3] Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for ΔC. In: P. Karnowski, & I. Szigeti (Eds.), Language and language-processing. In: Proceedings of the 38th linguistics colloquium 2003, Piliscsaba, Hungary (pp. 231–241). Lang: Frankfurt am Main etc.

[4] Dellwo, V. (2009) Choosing the right rate normalization methods for measurements of speech rhythm. In: Proceedings of AISV, 13-32.

[5] Dellwo, V., Leemann, A., Kolly, M.-J. (2012) Speaker idiosyncratic rhythmic features in the speech signal. In: Electronic Proceedings of Interspeech, Portland/Oregon/USA.

[6] Dellwo, V., Fourcin, A. and Abberton, E. (2007) Rhythmical classification of languages based on voice parameters. In: Proceedings of the 16th international congress of phonetic sciences (ICPhS) 2007, Saarbrücken, Germany (pp. 1129–1132).

[7] Duanmu, S. (2000) *The Phonology of Standard Chinese*. Oxford: OUP.

[8] Grabe, E. & Low E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), Laboratory Phonology 7 (pp. 515-545). Berlin/New York: Mouton de Gruyter.

[9] Iwasaki, S. and Ingkaphirom, P (2005*) A Reference Grammar of Thai*. Cambridge: CUP.

[10] Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. In: *Cognition*, 93, 225–254.

[11] Leemann, A., Kolly, M.-J., Dellwo, V. (2014) Speaker-specificity in the time domain: implications for forensic voice comparison. In: *Forensic Science International* 238, 59-67.

[12] Mermelstein, P. (1975) Automatic segmentation of speech into syllabic units. In: J. Acoust. Soc. Am., 58, 880-883.

[13] Mok, P., Dellwo, V. (2008) Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. in Proceedings of Speech Prosody, Campinas, Brazil, 423-426.

[14] Matthews, S. and Yip, V. (1994) *Cantonese: a comprehensive Grammar.* London, New York: Routledge.

[15] Pike, K. (1945) *The intonation of American English.* University Press: Ann Arbor.

[16] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Version 3.0.0. http://www.R-project.org, 2013.

[17] Ramus, F., Nespor, M. & Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. In: *Cognition*, 73, 265-292.

[18] Tilsen, S. and Johnson, K. (2008) Low-frequency Fourier analysis of speech rhythm. In: *J. Acoust Soc. Am*. 124, 34-39.

[19] White, L. & Mattys, S. L. (2007) Calibrating rhythm: First language and second language studies. In: *Journal of Phonetics*, 35, 501-522.