A Preliminary Analysis of the Relationship of Speech Rate to Speech-Timing Metrics as applied to Large Corpora of Non-Laboratory Speech in English and Chinese Broadcast News

Matthew Benton

Department of Linguistics and TESOL, The University of Texas at Arlington, USA

matthew.benton@mavs.uta.edu

Abstract

A renewed interest, in recent years, has occurred in the area of speech rhythm (traditionally defined by categories of speech timing patterns based on perceptual or acoustic durations of stresses, syllables, or moras). Since accurate categorization seems to be a three dimensional problem (durations of vocalic intervals, intervocalic intervals, and speech rate/tempo), some studies have made provision for differences in speech tempo by providing metrics with rate normalizing parameters based on the intervocalic intervals or the vocalic intervals (VarcoC & nPVI-V respectively). This study applies these different metrics on larger corpora of many speakers and more naturally occurring speech.

Index Terms: speech rhythm, pairwise variability, speech tempo, syllable timing, stress timing

1. Introduction

Traditionally, the research problem of speech timing (or speech rhythm) was done primarily based on the listener's perception of the timing patterns or possibly by the perceptual beats of a particular language. The idea was popularized early on by Pike [1], Abercrombie [2], and (later by) Dauer [3]. Due to the vast improvement of computing and speech processing over the last few decades, more recent work building on [3], has been done based on phonetic science by using acoustic correlates of the speech signal in an attempt to give empirical evidence to the perceptual qualities languages have that cause them to be perceived as either of two categories (syllable-timed or stress-timed) (Note: a third category of mora-timing has also been proposed [as in 4], however it is not discussed at length here since neither of the two languages in the current data set seem to be included in that category).

The most common recent methods categorizing languages into rhythm (or timing) categories are those proposed by Ramus et al. [5] (hereafter RNM) and Grabe and Low [6] (hereafter GL). Even as popular as these metrics have become in the area of language categorization, there is still the possibility that the two dimensional aspects of plotting differences between the vocalic and intervocalic (or consonantal) intervals has been skewed by a third dimension of time. Several recent studies have pointed this out [7], [8], [9], and [10].

Being that GL [6] had also introduced the normalized Pairwise Variability Index (nPVI) metric for vocalic sections in 2002, it seems that the issue of rate has continually been thought to interfere with categorization. Furthermore, [7] & [8] have also proposed a variation of the consonantal intervals metric based on the RNM's ΔC (the standard deviation of the consonantal intervals in a unit of speech) called VarcoC (which is the "variation coefficient" of the ΔC or ΔC*100/meanC where meanC is the mean of the C intervals in a particular language – see [7]).

2. Data and Methods

2.1. Data Background

The collection of data used for this study and previous studies on speech timing (by the author and colleagues) [11], [12], and [13] originated as the training data for separate speech research projects investigating, among other things, automatic tagging of sentence units in English (EN) and Chinese (CH) and were using speech recognizer systems to provide forced alignment of an orthographic transcription and the raw audio data [14], [15], and [16]. During these studies, the data sets of time measured corpora of non-laboratory speech for Mandarin CH [17] and American EN [18] were produced as output.

The new form of the data consisted of text files marked for sentence, word, and phone level segmentation with durations which were then processed by Python scripts to compute the speech rhythm metrics used for the statistical analysis (hereafter "data" is meant to be this set of audio processed into a tagged text format). The three corpora used in the study included 569 EN sentences from broadcast news (40 speakers), and 527 CH sentences (49 speakers), and 235 sentences of conversational EN (from 4 speakers of the LDC switchboard collections [14]). Both the EN and CH News corpora included male and female reporters and interviewees. (Note: an additional break down of the data for speakers' gender, genre, and language has been explained in more detail in previous papers – this preliminary study was more focused on the metrics and the statistical relationships of the groups of speakers as a whole rather than the individual speakers' performance.)

2.2. Data analysis method

The larger corpora used for analysis had previously been separated into smaller files based on speaker, gender, and genre by native speakers of each EN and CH [11]. This data was then used to mark speaker by a unique ID tag in order to track different genders or genres types. The text files were then run through a Python script (provided by the first author of [13]) to calculate speaker independent as well as language and genre specific files. For example, there were separate files for EN-Female-Journalist-01, CH-Male-Interviewee-01, EN-Male-Conversational-01, etc. The following speech rhythm metrics were then calculated for each sentence division based on the phonetic segmentation tagging (V & C stand for intervals unless followed by "-ind" – meaning independent phones, and "unit" being V, C-ind, C respectively):

- Sent#: generates a tracking number for each sentence
- #V, #C-ind, #C: number of unit per sentence
- %V: the percent of vocalic time per sentence

- SumV, SumC-ind: total length of unit per sentence.
- meanVLen, meanC-indLen, meanCLen: the mean length of duration of unit per sentence
- ΔV, ΔC-ind, ΔC : standard deviation of unit/sentence
- VpS, C-indpS, CpS: rate in unit per second
- rawPVI_V, rawPVI_C: pairwise variability index
- nPVI_V: the normalized PVI for V
- Sent-Len: the length of the sentence from the beginning of the first word's initial phonetic segment to the end time of the last word's final segment.

Values for ΔC, %V, etc are based on RNM [5] and PVI based on GL [6]. From this output, variables like VarcoC [7] were calculated in excel using the mean of ΔC and meanCLen columns. (VarcoC was actually calculated in two ways: $VarcoC_1$ – using the global mean ΔC for a language/style, and $VarcoC_2$ – using the meanCLen of each individual sentence). Statistics were then calculated using SPSS. Other calculations (i.e. VpS, CpS, etc.) were likewise based on sentence level divisions. Phones per Second (PsS) was calculated as the sum of V and C-ind numbers for each sentence.

2.3. Assumptions

There are a few assumptions that were made about this data and the use of it in this study. First, it was previously assumed that this data is representative of typical speech (primarily speech of the news reporters as being a "prototype" or "target" of normal speech by any given speaker of this particular language). While this may not be the case of some less common dialects, it seemed to the native speakers who divided the data that these samples were "typical" at least perceptually.

A second assumption is the accuracy of the data processing. We have to assume that the data was tagged accurately for phonemes and durations. Based on the results of previous studies [11], [12], and [13], and some test sample spot testing, the data seems to be within acceptable bounds for this type of research. More detailed sampling and hand checking may be necessary to verify this assumption.

The second assumption is particularly an issue when dealing with the sampling of consonant intervals. For instance, since the data was calculated in accordance with GL [6 specifically § 2.3], there may be instances where the C intervals (maybe some V intervals too) should be considered separate sound groups since they stretch across word boundaries (possibly ignoring important contextual inter-lexical pauses). However, in defense of this method, at least in EN, there appear to be many articulations that do slur sounds across word boundaries in typical naturally occurring speech ([19] has many examples of "linking" segments in EN pronunciation – particularly vowels, stops, and continuants).

The final assumption is that we can take accurate "syllable per second" measures from the VpS due to the way this particular data was tagged at the phoneme level. For example in EN all potentially syllabic consonants (e.g [n] in the word 'button") were tagged as if the recognizer "heard" a very short schwa with them. Additionally, it appears (according to [20], [21], and [22]) that a syllable nucleus must be a vowel in CH. All diphthongs in both languages were also recorded with one "phonetic" symbol similar to The CMU Pronunciation Dictionary [23]. Thus, since all syllables must have a nucleus, we are assuming that by counting the "peaks", even though we do not necessarily look at where the "valleys" (i.e. onsets and codas) span, we can still get an accurate count of syllables. Like the first two assumptions, this will likely need to be investigated more empirically in the future. (Note: it is possible that this assumption will not account for "peaks" that occur in multiple V intervals that appear across a word boundary in which the two phonemes would both be their own syllable nucleus rather than being combined in regular speech).

2.4. Aims of this research

There are three main aims or purposes for this study.

2.4.1. Continue testing more natural speech

First, in previous studies done by the author and colleagues, we found that the results of the analysis allowed for the argument that a speech rhythm division did exist between these two languages [11], [12], & [13] (although not always between individual speakers [11]). Those preliminary results also tested the languages only using the typical metrics, %V x ΔC, and nPVI_V x rPVI_C. Additionally, we were able to show the benefit of using a large number of speakers and a large set of data to keep any individual's idiosyncrasies from skewing the results.

2.4.2. Extend the researched data set

Additionally, as was noted in [10], it is important to have enough "representative" data, that which is "large enough, has enough different speakers, enough different utterances spoken in enough different styles of speech" for each language before it can be categorized as belonging to a specific rhythm type. This analysis addresses the issues of having many speakers and many utterances. It also attempts to add to the "styles of speech" qualification by adding more natural speech than was done in [5] and [6] and many studies that have followed (which this author has defined as "laboratory speech").

2.4.3. To test proposed normalization metrics (on a larger speaker base and potentially larger corpora)

As stated above, one of the aims of this research was to expand the speaker base to which these metrics have been applied. The final aim is then to test the "normalization" calculations against the initial calculations using basic statistical test (such as linear regressions) to show the relationship between the calculated metric values and speech tempo.

The first normalization value that was studied was the relationship between VarcoC and ΔC. According to [7], using VarcoC rather than ΔC should help counter any speech tempo effects on the rhythm parameter of C (the intervocalic intervals). This was done by calculating the ΔC of each sentence, taking the mean of those values to use as "meanC" for the variation coefficient. An additional reason this value was chosen was due to a comment in [7] that "a wider range of speakers" may be needed as a test sample.

The second normalization value that was studied was the nPVI_V as set out in [6]. This value was devised to be able to counter the vocalic changes due to speech rate and yet still capture the rhythm values that appear within the acoustic properties of any given speaker. This metric was compared to the rPVI_V (the raw or "un-normalized" value of the PVI for each pair of vocalic segments). During this study, it was not possible to do any tests on the validity of %V compared to VpS as might have been possible with lab-data, since all the sentences in "natural" speech were different, a general comparison was unable to be made in relation to speech rate. Likewise, it didn't seem necessary to work on an rPVI_C to

nPVI_C relationship as a first priority; however, this is something that may be in the works for the future.

3. Results and Discussion

3.1. Results

The results are divided into three basic sections: one for the general results of speech rate calculations, one for the C interval, and then one for the V interval.

3.1.1. Basic results

The first test was done was to verify that the language samples were indeed different in rate parameters. Using SPSS Independent-Samples T-test showed that the two sets of news broadcasts were indeed different from each other ($p<0.001$) in all values. However, when individual speakers' averages were compared, VpS was not significantly different ($p=0.145$).
Similar to the results given in [10], linear regression of EN broadcast news comparing the values of VpS (which as stated above is being used as a measurement for syllables per second) gave, $R^2=0.763$, with a slope of 2.128 ($p < 0.001$). CH on the other hand had an even higher correlation between PpS and VpS ($R^2=0.932$, slope of 2.773 ($p<0.001$). This seems to indicate that at least in this data set the consonantal phonetic segments have a higher likelihood of varying in EN than in CH (particularly since VpS is one component of PpS).

3.1.2. Results for Intervocalic or C intervals

As noted, VarcoC was computed in two ways. The first was to take the global value for meanC (or the mean of the meanCLen) of all sentences. As shown in figures 1 below, this wasn't particularly useful due to the fact that the "global meanC" for each group was very near 100 (the same amount multiplied by ΔC in the "VarcoC" formula (EN News Global meanC = 114.00 ms, CH News =100.68, and EN Conv = 101.91). However, as shown in Figure 1 the "Local" value of VarcoC (that computed for each sentence) was more effective at reducing the correlation with VpS.

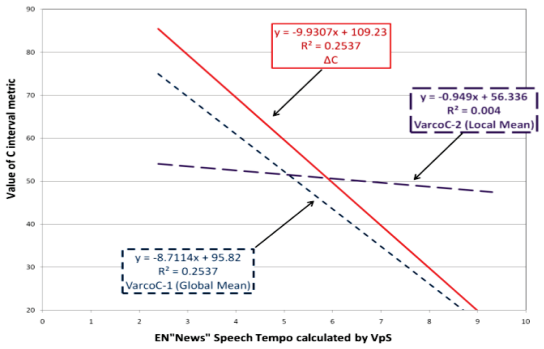

Figure 1: *Trend lines of EN News showing ΔC and VarcoC-Global having a significant relation to VpS (p<0.001) while VarcoC-Local does not (p=0.132)*

Figure 1 shows typical (though the most extreme) results of VarcoC-Local applied to the data. In both other cases, CH News and EN Conv, the slope and R^2 results were as follows: CH News → ΔC -5.34, 0.325; VarcoC-Local -2.43, 0.119; EN Conv → ΔC -3.49, 0.112; VarcoC-Local -2.24, 0.058.

Thus using a global meanC for the denominator was not very effective in this data although it may be if the meanC is quite different from 100. It is also probable that had the data been separated into "speech tempo" categories (as indicated in [10]) and the meanC of each group been used as the denominator the results would have been different.

3.1.3. Results based on normalizing PVI

Comparing the rPVI_V values with the nPVI_V values showed that some normalization had taken place. As shown in Figure 2, the nPVI value did reduce the slopes; however, like VarcoC-Local, all but one of the values were still showing a significant relationship with VpS over the vocalic intervals.

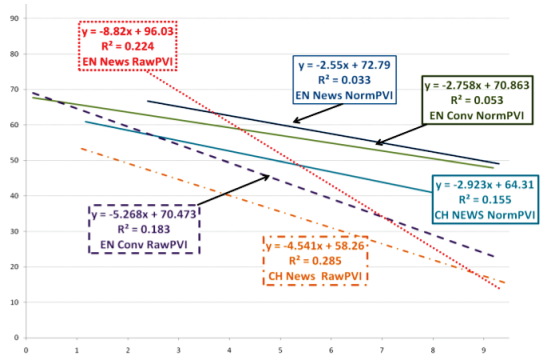

Figure 2: *(y-axis = PVI value, x-axis = VpS) Similar to Figure 1, the normalizing factor of nPVI seems to be somewhat effective over the rPVI when compared to VpS. (For all regression lines p<0.001 except for EN Conv nPVI p=0.006)*

3.2. Discussion

It appears from this data and the statistical results that there is some normalizing of speech rate that happens when the traditional metric of ΔC is changed to VarcoC. Additionally, there seems to be reduction of the relationship when using the nPVI over rPVI for the vocalic intervals.

3.2.1. Proposal for the next step

When looking at the fact that the nPVI does seem to account for some of the tempo variation, I am proposing that possibly the normalization can be widened such that it encompasses the previous spoken segments of a particular utterance. The general PVI formulas (1) and (2) have been used in the past with the assumption that normalization is occurring, but the results of this study seem to suggest that it does not rule out variation due to speed entirely.

$$rPVI = \left[\sum_{k=1}^{m-1} \left|d_k - d_{k+1}\right| / (m-1)\right] \quad (1)$$

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

In a discussion with the last author of [11] about how to widen the scope of normalization, it was proposed that if the normalizing denominator was to be expanded it would be more logical to work backwards to the beginning of the sentence / utterance. Thus in formula (3) below, I have shown what the first iteration of this would look like. In this case, the rate would be normalized over the first two segments as before, but as the third one became available it would be included as well.

$$\mathrm{Re}v\_nPVI_1 = 100 \times \left[\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{(d_{k-1} + d_k + d_{k+1})/3} /(m-1) \right] \quad (3)$$

When taken to the next step, the formula for what I'm calling Reverse-normalized PVI would then attempt to normalize the utterance over all past durations from that particular utterance as shown in formula (4) (where $n \leq$ (m-1) or the number of potential segments over which the normalization could occur and i would be a normalizing index value that could be changed by the researcher as needed).

$$\mathrm{Re}v\_nPVI_i = 100 \times \left[\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{\left(\sum_{i=0}^{n} d_{k-i} + d_{k+1} \right)/(2+i)} /(m-1) \right] \quad (4)$$

It is presumed at this point (as this formula has not been adequately tested on large amounts of data) that this type of normalization will give an even more accurate measure of V with the respect to factoring out the changes in speech rate while still allowing an accurate representation of speakers' potential dialectal or emotional variation. Using a variable index for normalizing may allow for the researcher to find the best value for i for a particular language or for gender/genre groups without over normalizing (which was probably the issue with using a Global meanC value above). The normalization index may be able to be used as a classification metric as well (particularly if it is found that many languages indeed can be classified based on rate of speech alone [9]).

3.3. Potential issues

As mentioned above in assumption about inter-lexical pauses, it is probably necessary to get a more detailed printout of the original data set to check accuracy of durations of both segment time and pauses as well as to verify any possible rounding errors that may have occurred for the reported segment durations. Also, preliminary results indicate that there may indeed be difference in speech tempo within languages based on gender and genre, which need further study. Finally, these results above may also need to be analyzed by more sophisticated statistical methods and on more data samples for a more definitive generalization of their normalization ability across languages.

4. Conclusions

In conclusion, the aims of this research were to expand the measured data for EN and CH to include more speakers and more sentences, to further investigate speech-timing patterns/ speech rhythm of this particular data set with newer metrics than had been done previously, and at the same time to investigate the relationship between speech rhythm and speech rate statistically. The results show that, while there does seem to be some neutralization of speech rate by VarcoC and nPVI, there may be another way to further normalize without over generalizing. The next steps are to investigate the new metric proposed above, the differences between individual speakers in the data, and use more advanced statistics for analysis of normalization.

5. References

[1] K. L. Pike, The intonation of American English. Ann Arbor: University of Michigan Press, 1945.

[2] D. Abercrombie, Studies in general phonetics. Edinburgh: Edinburgh University Press. 1965.

[3] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed", Journal of Phonetis, 1983, pp 51-69.

[4] Port, R, Dalby, J. and O'Dell, M. (1987). "Evidence for mora timing in Japanese." Journal of Acoustical Socitey of America, vol. 81, pp. 1574-1585.

[5] Ramus, F., Nespor, M., Mehler, J. Correlates of linguistic rhythm in the speech signal. Cognition 72,1-28. 1999.

[6] Grabe, E., Low, E. L. "Durational variability in speech and the Rhythm Class Hypothesis", In: Gussenhoven, C and Warner, N. (eds), Papers in Laboratory Phonology 7. Cambridge: CUP. 2002.

[7] Dellwo, V. "Rhythm and Speech Rate: A variation coefficient for deltaC,". in Language and Language-processing. edited by P. Karnowski and I. Szigeti, Frankfurt am Main: Peter Lang, pp. 231–241. 2006.

[8] Dellwo, V. and Wagner, P. "Relationships between language rhythm and speech rate" in Proceedings of the 15th ICPhS, Barcelona, 2003, pp. 3471-474.

[9] Dellwo, V. "The role of speech rate in perceiving speech rhythm", in Speech Prosody 2008, Campinas, Brazil, 2008, pp. 375-378.

[10] M. Russo and W. J. Barry, "Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo," in Speech Prosody 2008, Campinas, Brazil, 2008, pp. 419-422.

[11] M. Benton, et al., "The continuum of speech rhythm: computational testing of speech rhythm of large corpora from natural Chinese and English speech", in *Proceedings of the 16th ICPhS*. Saarbrücken, pp.1269-1272. 2007.

[12] M. Benton and L. Dockendorf, "A Comparison of Two Acoustic Measurement Approaches to the Rhythm Continuum of Natural Chinese and English Speech," in Interspeech 2008, Brisbane, Australia, 2008, pp. 772-775.

[13] L. Dockendorf, et al., "Testing a Large Corpus of Natural Standard Arabic for Rhythm Class," in Interspeech 2008, Brisbane Australia, 2008, p. 771.Strassel, S, Walker, C, and Lee, H. "RT-03 MDE Training Data Speech". Linguistic Data Consortium, Philadelphia. 2004.

[14] Linguistic Data Consortium, http://www.ldc.upenn.edu http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62

[15] Strassel, S., Kong, J. and Graff, D. "TDT4 Multilingual Text and Annotations". Linguistic Data Consortium, Philadelphia. 2005.

[16] Strassel, S, Walker, C, and Lee, H. "RT-03 MDE Training Data Speech". Linguistic Data Consortium, Philadelphia. 2004.

[17] Hwang, M.-Y. et al. "Investigation on Mandarin broadcast news speech recognition." In: Proc. of Interspeech 2006, 1233-1236. Pittsburgh, PA, USA. 2006.

[18] Venkataraman, A., et al. "SRI's 2004 broadcast news speech to text system." In: EARS Rich Transcription 2004 Workshop, Palisades, Nov. 2004.

[19] J. B. Gilbert, Clear speech : pronunciation and listening comprehension in North American English: student's book, 3rd ed. New York: Cambridge University Press, 2005.

[20] M. I. N. Wang and C. Cheng, "Subsyllabic unit preference in young Chinese children," Applied Psycholinguistics, vol. 29, pp. 291-.314, 2008.

[21] M. Ashby and J. A. Maidment, Introducing phonetic science. Cambridge; New York: Cambridge University Press, 2005.

[22] S. Tseng, "Contracted syllables in Mandarin: Evidence from spontaneous conversations," Language And Linguistics; Taipei, vol. 6, p. 153, 2005.

[23] The CMU Pronunciation Dictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[Special thanks to previous co-authors from [11 & 12] who helped with talking through proposed metrics and who made the analysis software and data available for this research]