THE CONTINUUM OF SPEECH RHYTHM[1]: COMPUTATIONAL TESTING OF SPEECH RHYTHM OF LARGE CORPORA FROM NATURAL CHINESE AND ENGLISH SPEECH[2]

Matthew Benton[*], *Liz Dockendorf*[*], *Wenhua Jin*[*],
*Yang Liu***, *Jerold A. Edmondson*[*]

[*]The University of Texas at Arlington, [**]The University of Texas at Dallas
matt@mjbenton.com, liz1695@yahoo.com, jinwenh@hotmail.com,
yangl@hlt.utdallas.edu, j.edmondson@sbcglobal.net

ABSTRACT

Past research on the dichotomy of language rhythm classes (stress- vs. syllable-timing) has typically been performed on constructed speech data, e.g. "The North Wind and the Sun" text. Our research goes beyond the previously established speech rhythm studies by combining: (1) a data set of 175 minutes of audio from large corpora of natural English and Chinese speech and (2) natural language processing techniques to compute phonetic segment-statistics. Our findings generally agree with the previous result that Chinese and English fall into distinct rhythm categories. However, when individual speaker data were considered in our analysis, an overlapping continuum across both languages was shown to exist. These results indicate that using "ideal" data to measure speech rhythm does not fully explain the division between languages.

Keywords: speech-rhythm, syllable-time, stress-time, corpus, prosody, speech-timing patterns

1. INTRODUCTION

1.1. The Problem of Speech Rhythm

Speech Rhythm has historically been based on the notions of Pike [1], Abercrombie [2], and others. Pike termed the categories *syllable-time*d and *stress-timed languages*. For example, English is commonly considered to be a stress-timed language (emphasizing particular stressed syllables at regular intervals), while Spanish [3] appears to space syllables equally across an utterance.

1.2. Related work

Two modern studies were particularly relevant to this study: Ramus et al. [3] and Grabe and Low [4]. These studies used speech tasks produced by informants reading aloud.

1.2.1. *Ramus et al.: Standard deviation and %V.*

Ramus et al. analyzed five sentences in eight languages (four speakers each). Each sentence was carefully constructed to encompass fifteen to nineteen syllables.

They measured vocalic and consonantal intervals and computed standard deviations within each sentence (ΔV & ΔC), along with the "the sum of vocalic intervals divided by the total duration of the sentence" (%V) [3].

1.2.2. *Grabe & Low: Pairwise Variability Index.*

Grabe and Low employed "The North Wind and the Sun," (a standardized phonetic research text) in an appropriate translation read aloud by a speaker for each of the sixteen languages and recorded in a laboratory setting [4].

The Pairwise Variability Index (PVI) was determined using the procedures proposed in Grabe and Low [4]. The PVI captures the amount of change between durations of successive intervals. It is further broken down to give both a "raw" (rPVI) and a "normalized" (nPVI) value. The rPVI does not take speech rate into account. The nPVI, which includes dividing the pairs by their mean duration, helps to "normalize" for differences in speech rates. (Although speech rate may still be a problematic area of study, we are following previous literature by assuming that nPVI ameliorates speech rate error, though more work is needed).

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

b(m: the number of intervals, *d:* duration)

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

(*m:* the number of items in an utterance, *d:* duration)

Both rPVI and nPVI produced results that confirm the existence of two language types while avoiding difficulties encountered with previous, less refined measurements of syllable and stressed timing. However, two questions remain. Do calculation of rPVI and nPVI from real, naturally occurring data yield similar results to those obtained by [3] and [4] for the same test? Do the traditional theories and assumptions still hold?

2. DATA AND METHODS

2.1. A Computational Approach and Data

To the best of our knowledge, our approach to this problem differs from previous studies in several ways. We studied larger amounts of naturally occurring speech data, considered over 50 speakers of each language, and extracted relevant measurements from raw data with speech recognition software. Each of the previous studies utilized standardized data sets constructed just for the purposes of the experiment. Our data set, however, was compiled from the more naturally occurring speech of broadcast news in Mandarin Chinese (CH) and American English (ENG) and data from English telephone conversations.

The CH broadcast news audio (from CCTV Beijing) and the American ENG data sets (ABC News and CNN broadcast news) were from Linguistic Data Consortium (LDC) audio files [7]. These news files included speech by reporters (Rep – we believe this was in some cases teleprompted) and interviewees (Int). Additional ENG conversation was extracted from the LDC Switchboard data [7]. It consisted of recordings of two bipolar telephone conversations (one of two male speakers and one of two female speakers (Conv)). In total, our data set encompassed 100 minutes of CH speech data and 75 minutes of ENG. Often, interviews conducted during the broadcasts resembled conversational speech more than news-reporting monologues.

The audio files were previously transcribed orthographically and phonetically, and were segmented for "sentence-like" units (a non-trivial task in CH) with all commercials omitted. Transcriptions were also marked for back-channel cues (such as hesitations and pauses) allowing us to disregard overlaps in conversation as well as "non-relevant" speech, such as "uh…" The audio files were then processed by a speech recognizer [6, 8] to align transcripts with audio, to get the segmental units and word boundaries, to measure vowel and consonantal segment durations, and to perform alignments within the larger text. This method gave us "hybrid" transcriptions of characters and segment durations, which allowed extraction of large amounts of consistent data using Python script tools to compute the relevant statistics.

Using these algorithms, we were able to employ naturally occurring data for the same "quantifying" test procedures as Ramus [3] and Grabe and Low [4]. Moreover, we had the added advantage of testing a larger numbers of speakers (119 total individual speakers—62 ENG: 17F; 45M, and 57 CH: 33F; 24M) and longer durations of oral language materials than previous studies used. The contributions of the 119 speakers were narrowed down to eliminate "artifactual" samples (those samples with durations too short to yield conclusive results—those consisting of speakers whose data was less than 0.01% of the total data from each language). After eliminating these samples, there were 92 total speakers (ENG: 15F [8Rep/5Int/2Conv]; 27M [8Rep/17Int/2Conv] and CH: 29F [28Rep/1Int]; 21M [9Rep/12Int]) represented in the data.

The results comprised 569 usable ENG utterances and 527 usable CH utterances. Summing up the durations of the phonological segments, we arrived at approximately 108 minutes (42-ENG, 66-CH) of analyzed speech utterances (disregarding pauses etc.).

2.2. Application of Computation Methods

We were guided in our data processing by the methods of Ramus et al. [3] and Grabe and Low [4] and used Python scripts to output the raw data in spreadsheet format. Our procedures followed [4] in testing normalized and raw Pairwise Variability Indexes [4] and plotting the averages for each speaker, as well as the averages for the data from each language. We plotted averages of the speakers based on gender, genre of language use, and each language as a whole. The averages can be seen in Figure 1. The averages for each language demonstrated agreement with previous studies, showing a clear tendency for the languages to be divided into "rhythm" classes. One finding, keeping the factors genre and gender separate, was that the ENG "male conversational" data and the CH "interview" showed more overlap than we anticipated.

Figure 1: Averages of speakers by language, gender and genre. CH–Chinese, EN–English, F–Female, M–Male, Rep–Reporter, Conv–Conversational, Inv–Interviewee. NOTE: The EN Interviewees are not plotted for clarity (values M: $63.7_{rPVI} \times 59.6_{nPVI}$, F: $63.5_{rPVI} \times 64.8_{nPVI}$).
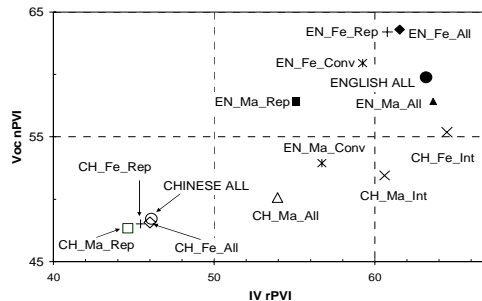
3. RESULTS

3.1. Results based on Ramus et al.

Ramus et al. [3] found that languages could be plotted on a continuum based on %V and ΔC, with stress-timed languages showing higher values on the ΔC axis, and lower %V values than syllable-timed languages. Using this measurement procedure, Grabe and Low [4] found that Mandarin had a higher %V than ENG (British and Singaporean) and lower standard deviations for both vocalic and consonantal intervals. Our own study confirmed their findings for standard deviations, but we found American ENG to have a greater %V value than Mandarin CH.

Independent Samples T-Tests calculations with SPSS were carried out on computed output data (c.f. "sentence" data §2.1), testing for the significance of differences in %V, ΔC, ΔV (all statistical calculations were performed at a 95% confidence level). We found that CH was significantly lower than ENG in these three variables (p<0.001).

3.2. Results based on Grabe & Low

The same data set used above also yielded statistically significant differences between CH and ENG for the values: vocalic (V) nPVI, V-rPVI, Intervocalic (IV) rPVi. Again, CH is significantly lower than ENG (p<0.001).

These results showed mean values of nPVI × rPVI that were generally in line with Grabe and Low, and as expected, considerable diversity was found between individual speakers. We found that the speech rhythm of a language may differ when

generalized but individual speakers may have rhythm values that overlap considerably. This is one result of our study that also appears in [5]. Grabe notes that, when there are multiple speakers of the same language, the observed variation within a language is potentially as great as the variation between languages, yielding the apparent paradox that some speakers of CH or ENG may have values showing "syllable timed" and others "stressed timed" speech. This speaker variation can be seen in Figure 2.

Figure 2: Speaker diversity between EN and CH (circles are English speakers and boxes are Chinese speakers).
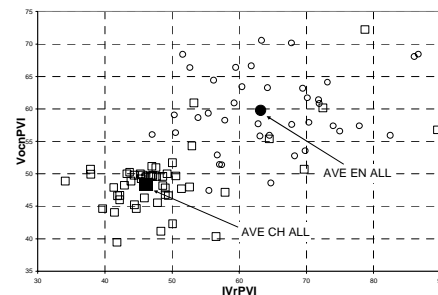
3.2.1. Gender Differences

To test for a significant difference for gender in ENG, we used both male and female news reporters. This restriction eliminated some of the possible genre differences between speakers in news-reporting formats from those in prerecorded interviews. The same type of test was performed on the CH data. (Speakers' gender was determined by native speakers of each language.)

Independent Samples T-Test showed that for gender the only significant differences (p<0.05) we found were in V-nPVI and V-rPVI in ENG speakers and in ΔC in CH speakers. In all these cases, women were significantly higher than men.

3.2.2. Genre Differences

Genres tested were categorized into two types: news anchors-reporters (Rep) associated with the broadcasting station and interviewees (Int), those who took part in a direct interview as well as in the prerecorded sound bites. The genre differences in both CH and ENG were found by comparing the data from the news reporters to the data from the interviewees. Furthermore, we tested the ENG conversational data for differences to the news data. The results showed that there was no significant difference for genre in ENG for any

values (using ANOVA). However, CH news reporters were significantly lower than interviewees in all the values except for %V.

4. DISCUSSION

Grabe [5] says that it may be "premature to establish firm rhythmic typologies" unless they are built "on comparable data from several speakers of several dialects of each number of languages…" Our results strongly confirmed this conclusion. We must also conclude that not only does rhythm vary with individual speakers, but also in some cases with the genre in which it is used. We also note that in our study the female ENG speakers were the speaker group that was most "separated" from the CH news reporters. Even though there was not a statistically significant difference between genders in ENG speakers, (as seen in Figure 1) the female speakers tend to group higher (in regard to V-nPVI, IV-rPVI) in most cases than the male speakers. More study will be needed to confirm this still preliminary result.

Figure 3: Continuum of speakers by language, gender and genre are shown by overlapping layers representing the boundaries of each group. E–English, C–Chinese, F–Female, M–Male, REP–Reporter, INT–Interviewee.
NOTE: only one sample of C_F_INT ($64.5_{rPVI} \times 55.4_{nPVI}$).
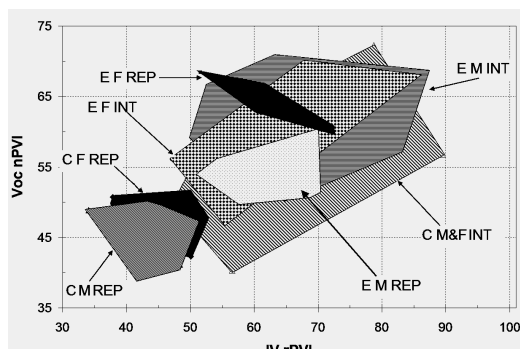
It is also worth mentioning that the news broadcasters in both languages cluster more tightly than the interviewees. Some possible exlanations are that, a. they are using a more regular "normalized" speech form, b. they use less spontaneous oral speech, or c. they have a style more like reading (from teleprompter or speaking rehearsed lines). (The aspect of the input data formation was obviously out of our control.) It is also possible that news reporters are aiming at a language relative "rhythm target" that others speakers strive for in regular speech, but may be impeded by other internal (e.g. cognitive recall) or

external factors (environment, surroundings in which an utterance is made). From the boundary lines on Figure 3 the "linguistic rhythm target" is more convincing for ENG than for CH.

5. CONCLUSION

This study found that phonetic procedures for determining rhythmic categories do extend to naturally occurring CH and ENG and it also demonstrates that natural language processing methods do confirm previous results. Moreover, we have demonstrated that, within the realm of rhythmic distinction genre (and other factors) may also need to be accounted for when using naturally occurring data. Finally, our analysis of naturally occurring data by multiple speakers from each language suggests that an overlapping rhythm continuum exists. This overlap may indicate that, like tones and pitch, rhythm is relative to the speaker and the context in which language is used.

6. REFERENCES

[1] Pike. K. L. 1946. *The intonation of American English.* Ann Arbor: University of Michigan Press.
[2] Abercrombie, D. 1965. *Studies in general phonetics.* Edinburgh: Edinburgh University Press.
[3] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 72,1-28.
[4] Grabe, E., Low, E. L. 2002. Durational variability in speech and the Rhythm Class Hypothesis. In: Gussenhoven, C and Warner, N. (eds), *Papers in Laboratory Phonology 7*. Cambridge: CUP.
[5] Grabe, E. 2002. Variation adds to prosodic typology. In Bel, B and Marlien, I (eds), *Proc. of the Speech Prosody 2002: an International Conference.* Aix-en-Provence, France, 11-13 April 2002.
[6] Hwang, M.-Y. et al. 2006. Investigation on Mandarin broadcast news speech recognition. In: *Proc. of Interspeech* 2006,1233-1236. Pittsburgh, PA, USA.
[7] Linguistic Data Consortium. http://www.ldc.upenn.edu
[8] Venkataraman, A., et al. 2004. SRI's 2004 broadcast new speech to text system. In: *EARS Rich Transcription 2004 Workshop*, Palisades, Nov 2004