Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Capturing Inter-speaker Invariance Using Statistical Measures of Rhythm

Tae-Jin Yoon

Department of Linguistics and Languages, McMaster University, Canada
tjyoon@mcmaster.ca

Abstract

Statistical rhythmic metrics are applied on a Buckeye corpus [1] of spontaneous interview speech in order to investigate the extent of rhythm variability of between-speakers as well as the variability of within-speaker. The corpus consists of speech produced by speakers who share the same regional dialect in North America. The Buckeye corpus is unique in that the speech dataset is obtained from the speakers who have been raised in the same region and hence who share the same dialect from each other. Statistical measures of rhythm metrics are obtained from each of 10 speakers. The results show that the rhythmic measures that capture the least dialectal variance is the normalized pair-wise variability indices calculated based on adjacent consonantal duration and vocalic duration. The finding implies that these statistical measures of rhythm can be used in capturing the dialectal similarities.

Index Terms: speech rhythm, Buckeye corpus, rhythm metrics, rhythmic variability of between-speakers, rhythmic variability of within-speaker

1. Introduction

Current study of speech rhythm has paid more attention to establishing measures that can capture rhythm differences from different rhythm groups, but less attention to the similarities shared by speakers of the same language or dialect. That is, one prime debate in the current study of speech rhythm has focused on universal organization, attempting to find an answer to questions such as ‘are there any distinct types of speech rhythm?’ or is there only a continuum with languages placed along it?’ This focal research question has led researchers to find a better rhythm metrics that can discriminate languages of different rhythm type and that can group together languages with similar rhythmic types ([2][3][4], among others). However, less attention is being paid to the question of what rhythm metrics can capture the rhythmic patterns of those speakers who share the same language or dialect.

The research focus may in part be due to the amounts of data and the labor-intensive methods that have been employed in the typological studies of cross-linguistic rhythm. So far, these rhythm measures require much careful manual marking of the speech, and they are highly dependent on the choice of words. Mostly, they have been limited to carefully designed laboratory experiments. Besides, hypotheses have often been tested on a small chunk of carefully designed spoken utterances in a laboratory setting, or recitations of standard phonetic demonstration texts such as the fable “The North Wind and the Sun” (cf. [5]). The limited data, thus, may suffer from external validity such that the finding may not be generalized to a naturally occurring spontaneous speech.

In this paper, I address the question of what statistical rhythm methods are best suited to capture the rhythmic similarities among speakers sharing the same dialect and present results of the study based on a data-driven approach to the rhythm calculated from speakers who have been raised in the same region.

2. Methods

2.1. Corpus

The Buckeye Corpus of conversational speech contains high-quality recordings from 40 speakers in Columbus, Ohio conversing freely with an interviewer approximately an hour between 1999 and 2000. One characteristic of the corpus is that all forty speakers were recruited from the Columbus, Ohio Community, and they are all native of Central Ohio. That is, they were born in or near Columbus or moved there no later than age 10. The speakers are stratified for age (under thirty (Y) and over forty (O)) and sex (F for female and M for male). I used 10 speakers for the study reported in this study. Presented in Table 1 are gender and age of the speaker, gender of the interviewer, and the total duration (in minutes) during which the speaker spoke in the about 1-hour interview session [1].

Table 1. Speakers in the Buckeye corpus that have been analyzed for the present study.

Speaker	Gender	Age	Interviewer	Duration (in minutes)
S01	f	y	f	16.98m
S02	f	o	m	27.16 m
S03	m	o	m	30.11 m
S04	f	y	f	28.35 m
S05	f	o	f	22.40 m
S06	m	y	f	16.41 m
S07	f	o	f	37.68 m
S08	f	y	f	33.11 m
S09	f	y	f	28.16 m
S10	m	o	f	40.03 m
Total				280.39 m

2.2. Phonetic Segmentation

The current methods of calculating speech rhythm are based on speech data that are annotated based on consonantal and vocalic intervals (see [2][3][4] among other). For their calculation, the acoustic metric require segmentation of the speech signal into vocalic and consonantal intervals, where such intervals are defined as all consecutive segments of the same type (vowel or consonant) irrespective of syllable, morpheme or word boundaries. The speech files in the Buckeye corpus are accompanied by corresponding phonetic and word labels. According to the manual [1], the phonetic labels were obtained by using an automatic phonetic alignment method and then corrected later by manually adjusting incorrectly aligned phones. Thus, it is necessary to convert those individual phonetic labels into consonant and vowel labels depending on their identity and then combining consecutive consonant intervals into a chunk of consonant interval and consecutive vocalic intervals into a chunk of vocalic interval, respectively. A Praat script is written in order to convert the phone labels into a succession of consonantal (C) and vocalic (V) intervals. The sequences of C's and V's were obtained by combining adjacent C's and V's without an intervening pause, end of turn, noise, laughter, a nonspeech sound, an phone marked as incomprehensible by the transcribers, or a segment extraneous to the segment inventory in the corpus. Some previous studies (e.g. [4]) excluded from the sentences the approximants /l/, /w/, /j/, and /r/ to increase the reliability of the segmentation procedure. In this study, however, these approximants were treated as consonants.

Figure 1 is a screen shot illustrating a portion of the Buckeye corpus. The top two windows show waveform and spectrogram displays. Consonantal (C) and vocalic (V) intervals are shown in the tier just below the spectrogram.

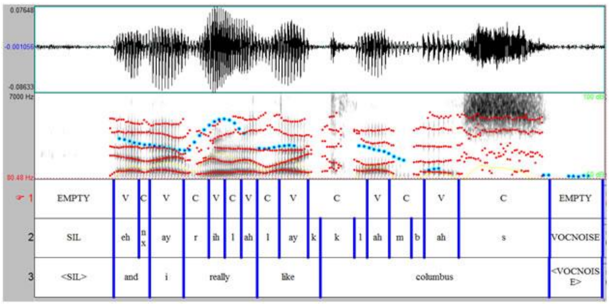

Figure 1: Screenshot of the segmentation of phone-labeled spoken utterances into consonant (C) and vocalic (V) intervals (1st tier), phone and word transcription (2nd and 3rd tiers, respectively). The phone and word transcription are provided. A script is written that converts the phones into consonants and vowels, and generates automatically the CV sequence as shown in the 1st tier.

2.3. Rhythm measurements

In the recent studies on speech rhythm, a series of acoustic metrics based on consonantal and vocalic duration have been designed to distinguish language according to putative stress-, syllable, and mora-timed rhythmic categories. The acoustic measurements suggested include metrics that measure the proportion of vocalic intervals (%V), the duration variability of vocalic intervals (e.g. ΔV, VarcoV, nPVI-V), and duration variability of consonantal intervals (e.g. ΔC, VarcoC, rPVI-C, nPVI-C) (See [4] for a review and evaluation of these metrics). In addition, the speech rate is also calculated by counting the number of syllables realized per second. The metrics used in the present study are described below:

- ΔV: the standard deviation of vocalic interval duration ([2]).
- ΔC: the standard deviation of consonantal interval duration ([2])
- %V: the sum of vocalic interval duration divided by the total duration of vocalic and consonantal intervals and multiplied by 100 ([2]).
- VarcoC: the standard deviation of consonantal interval duration divided by the mean consonantal duration and multiplied by 100 ([6])
- VarcoV: the standard deviation of vocalic interval duration divided by the mean vocalic duration and multiplied by 100 ([4])
- nPVI-V: normalized Pairwise Variability Index (PVI) for vocalic intervals ([3]).
- nPVI-C: normalized Pairwise Variability Index for consonantal intervals
- rPVI-C: raw Pairwise Variability Index for consonantal intervals ([3])
- Speech rate: number of syllables per second

3. Results

In this section, I present the results obtained from applying those rhythm metrics on the ten speakers of the Buckeye corpus.

Figure 2 show the total interval duration for each speaker, with the total duration of the vocalic intervals is stacked on the total duration of the consonantal intervals.

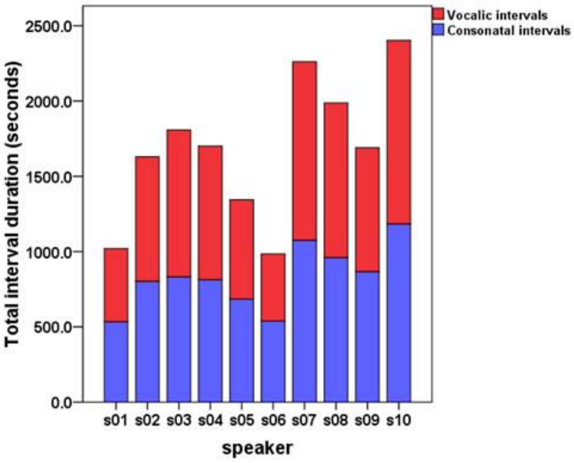

Figure 2: Individual variation of the total interval duration and the ratio of consonantal interval and vocalic intervals.

Figure 3 shows the proportion of the vocalic intervals only. English belongs to the stress-timed language. And typical stress-timed languages such as English exhibits complex syllable structure allowing complex onsets and codas, whereas typical syllable-time languages such as French allow only simple syllable structure. Therefore, it is often hypothesized that the average vocalic duration is relatively short than the average consonantal duration in languages such as stressed-timed English, and the proportion of vowel duration is approximately the same as the proportion of consonant duration in language such as syllable-time French. This hypothesis has been confirmed in a number of controlled studies, including [1]. For example, Ramus et al. (1999) reports the proportion of the vocalic intervals is about 40.1% [1]. However, the controlled studies do not capture other sources of variation for vocalic duration. Vocalic duration is also affected by utterance-level prosodic patterns or disfluency. For example, the naturally occurring utterances will be affected by prosody-induced lengthening effects or disfluecy-induced prolongation [7]. Then the proportion of the vocalic duration may not be as clear-cut as the one reported in [2]. The measure of %V, the portion of the vocalic intervals, as shown in Figure 3, is calculated from the naturally-occurring utterances. The result shows that most speakers except for speakers S06 and S01 must be affected by the other sources of lengthening.

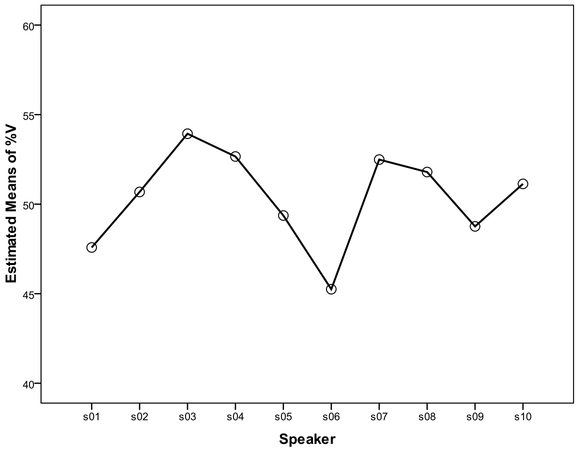

Figure 3: Individual variation of the proportion of the vocalic duration as calculated as %V.

It is often found that the rhythm metric may be influenced by speech rate. Thus, it is possible to imagine that the variation in %V in *Figure 3* may be influenced by speech rate. However, little consistent correlation between %V and speech rate has been found (e.g. [6]). In the present study, the average speech rate for all speakers is 5.12 syl/s (SD 0.5 syl/s). Figure 4 shows the individual speech rate.

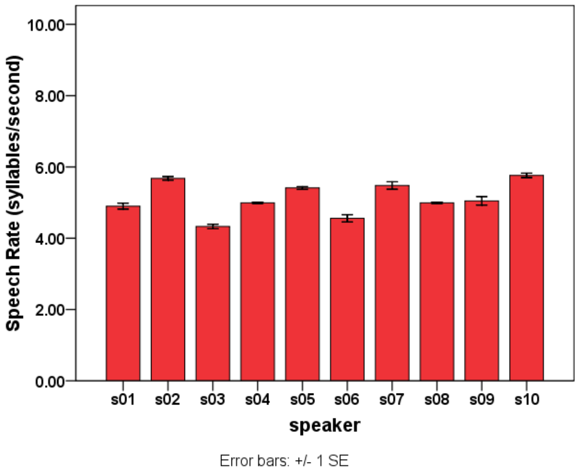

Figure 4: Individual variation of speech rate which is calculated as the number of syllables per second.

Given that the duration of vowel is more likely to be affected by speech rate, metrics of variation duration of vocalic intervals controlling for speech rate have been suggested. White & Mattys (2007) [4], for instance, put forward VarcoV, as a measure of vocalic intervals that is less sensitive to speech rate variation. Figure 5 shows relationship between the mean values of the VarcoV and the mean values of speech rate, for each speaker in the corpus. If the speech rate acts insensitively in the VarcoV metric, we may see less variation on the horizontal axis. To a certain degree, we expect to see the VarcoV value of each speaker may be concentrated around a vertical bar, irrespectively of the differences in their speech rate. In Figure 5, many speakers' VarcoV values are somewhere between 75 and 80. However, there are speakers whose VarcoV values do not fall in this interval. Moreover, one speaker manifests a conspicuously large within-speaker variation in its VarcoV value, in addition to the between-speaker variation. Thus, VarcoV does not seem to be suitable to capture the rhythm similarity.

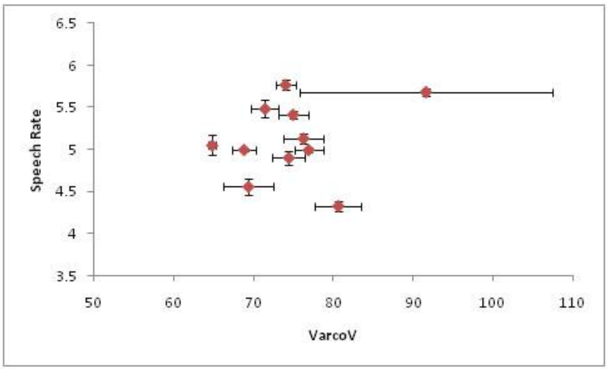

Figure 5: Plot of mean values with standard errors for acoustic parameters of VarcoV and Speech Rate. The whiskers indicate one standard error (i.e. +/-1 SE).

In *Figure 6*, VarcoV is plotted against VarcoC. VarcoC is a rate-normalized metric of consonantal intervals proposed by Dellwo & Wagner (2003) [6]. VarcoC is similar to VarcoV in that the standard deviation of the consonantal interval duration is divided by the mean duration of the consonantal intervals within an utterance. In [4], it is reported that this VarcoC fails to show linguistic discriminative capability between English, Dutch, French, and Spanish, whereas VarcoV appears to be successful at teasing apart these typologically distinct languages. *Figure 6* indicates that VarcoC exhibit less degree of both inter- and intra-speaker variability than VarcoV.

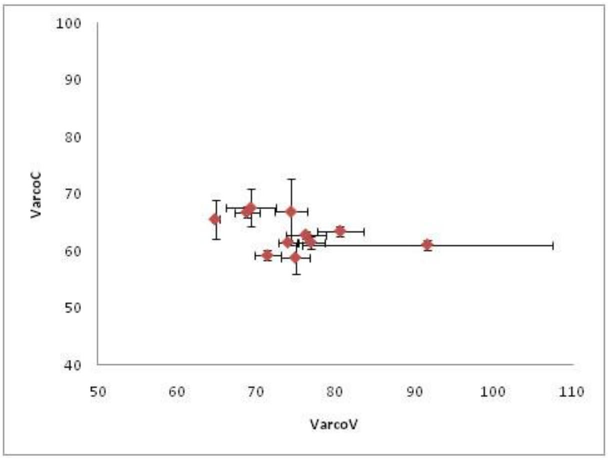

Figure 6: plot of mean values of VarcoV and VarcoC. VarcoC, centring around 62-65 appear to exhibit less degree of both inter- and intra-speaker variability than VarcoV in the Buckeye corpus.

Another metric of duration variability of vocalic intervals that controls for speech rate is the normalized Pairwise Variability Index for Vocalic intervals, henceforth nPVI-V. This nPVI-V metric has been used in a number of previous studies (e.g. [3]). In measuring interval duration variability of adjacent consonantal sequences, Grabe & Low (2002) [3] suggest that normalization for speech rate may not be desirable as it removes linguistically relevant information. Thus, raw Pairwise Variability Index (PVI) for Consonantal interval, henceforth rPVI-C, is instead used instead. Figure 7 illustrates the correlation between rPVI-C and nPVI-V as observed from the 10 speakers in the Buckeye corpus. While nPVI-V is very consistent with respect to both within-speaker and between speakers, rPVI-C reveals quite a large variability in the between-speaker dimension.

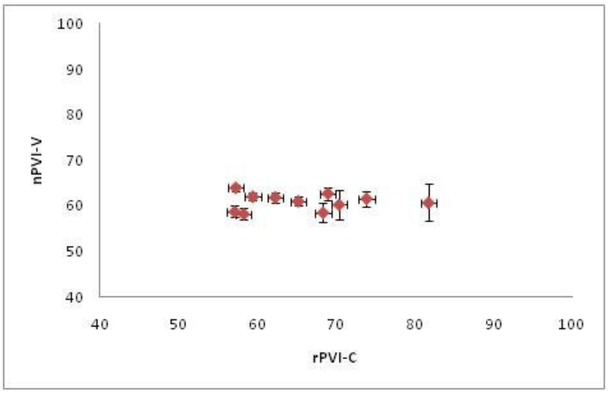

Figure 7: Plot of mean values of rPVI-C and nPVI-V. nPVI-V, proposed in [4], shows quite a consistent pattern across speakers. rPVI-C, proposed in [3], on the other hand, does not function well enough to capture between-speaker invariability.

In Figure 8, rPVI-C instead of nPVI-C is plotted against the nPVI-V. The difference between the rPVI-C and the nPVI-C formula is that the former – the raw Pairwise Variability Index – does not normalize for speech rate, whereas the latter normalizes for speech rate. As shown in Figure 8, nPVI-V and nPVI-C both make a very compact cloud, suggesting that the normalized variability indices are the best rhythmic metrics that capture the speaker's dialect similarity in this study.

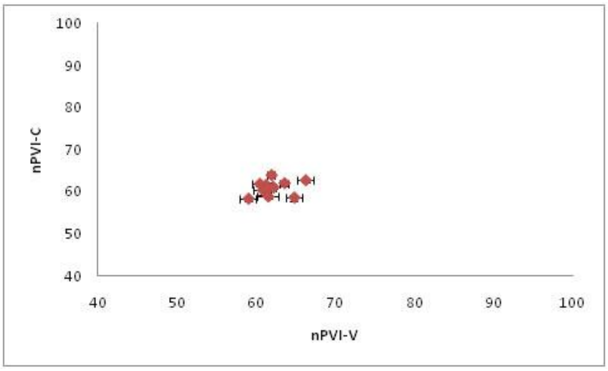

Figure 8: Plot of mean values of nPVI-V and nPVI-C. Both measures show the least inter-speaker and intra-speaker variation.

4. Discussion and Conclusions

In this paper, I demonstrated some of the rhythm metrics proposed in the previous studies may lack the external validity when they are applied to a large corpus of naturally and spontaneously occurring speech sounds. I also illustrated that metrics such as normalized Pair-wise Variability Indices for consonantal and vocalic intervals can capture the within-dialectal similarity better than other metrics. Further study is needed whether these metrics can capture the dialectal similarities and can tease apart different dialects or languages. Given the difference between the controlled laboratory studies and the study reported in this paper based on a corpus of naturally-occurring spontaneous study, it is necessarily to use a large quantity of equivalent naturally occurring linguistic data from other languages or dialects in order to draw a more meaningful and decisive conclusion that the normalized pairwise variability indices are indeed more reliable measures that can capture rhythmic similarity shared by the same dialects and that can discriminate dialectal or language differences.

5. References

[1] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd rel.) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

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

[3] Grabe, E. & E. L. Low (2002). "Durational variability in speech and the rhythm class hypothesis." In C. Gussenhoven & N. Warner (eds.), *Papers in Laboratory Phonology* 7. Berlin: Mouton de Gruyter.

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

[5] International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

[6] Dellwo & Wagner (2003) Relations between speech rate and rhythm. In *Proceedings of the ICPhS*, Barcelona.

[7] Shriberg, E. (1994) *Preliminaries to a Theory of Speech Disfluencies.* Doctoral dissertation, University of California at Berkeley.