Rhythm measures with language-independent segmentation

Anastassia Loukina[1], *Greg Kochanski*[1], *Chilin Shih*[2], *Elinor Keane*[1], *Ian Watson*[1]

[1] Phonetics laboratory, University of Oxford, United Kingdom
[2] EALC/Linguistics, University of Illinois, Urbana-Champaign, USA
anastassia.loukina@phon.ox.ac.uk

Abstract

We compare 15 measures of speech rhythm based on an automatic segmentation of speech into vowel-like and consonant-like regions. This allows us to apply identical segmentation criteria to all languages and compute rhythm measures over a large corpus. It may also approximate more closely the segmentation available to pre-lexical infants, who have been claimed to discriminate between languages. We find that within-language variation is large and comparable to the language-to-language differences we observed. We evaluate the success of different measures in separating languages and show that the efficiency of measures depends on languages included in the corpus. Rhythm appears to be described by two dimensions and different published rhythm measures capture different aspects of it.

Index Terms: linear discriminant typology acoustic phonetics speech segmentation experimental

1. Introduction

The rhythm of speech is a subjective impression which is presumably derived from acoustic properties. Recently, a group of quantitative statistical indices have been proposed to capture the rhythmic properties of languages. We will follow Barry et al. [1] and collectively call these indices rhythm measures (RMs).

To date, observed differences have generally been interpreted as differences between languages or groups, but more recent studies have revealed substantial variability between speakers and texts. For example, Keane [2] showed that differences between Tamil speakers exceeded those separating different languages.

Furthermore, most current measures rely on manual segmentation, which can be a very subjective process. Previous studies emphasized potential ambiguities and discrepancies in manual segmentation [cf. 3]. As Ramus [4] has pointed out, discrepancies between labelling principles make it 'virtually impossible' to ensure consistent segmentation between different studies.

In this paper we apply published rhythm measures to a large corpus of data in order to test whether rhythm measures can reliably separate languages. To avoid inconsistencies introduced by human segmentation, we use a simple automatic segmentation into consonant-like and vowel-like regions. Such segmentation offers consistent language-independent treatment of the acoustic signal. It also makes it possible to apply rhythm measures to a larger corpus of data than previously used in similar studies.

2. Data and methodology

Our corpus consisted of 1843 short texts recorded from 41 speakers of Southern British English, Standard Greek, Standard Russian, Standard French and Taiwanese Mandarin. The texts included extracts from "Harry Potter" in the original or translation, fables and the fairytale Cinderella.

Speakers were 20-28 years old; all were born to monolingual parents and had grown up in their respective countries. At the time of the recording all speakers were living in Oxford. Speakers of languages other than English had lived outside their home country for less than 4 years. The recordings were made in the soundproof room of the Oxford University Phonetics laboratory, using a condenser microphone, and recorded direct to disc at a 16 kHz sampling rate. The texts were presented on the screen in standard orthography for each language. The speakers had a chance to repeat any text if they were not satisfied with their reading. The recordings took place on two or three sessions on separate days.

2.1 Automatic and manual segmentation

Many rhythm measures are calculated on the basis of vocalic and intervocalic intervals. Numerous perceptual studies using a processed signal have shown that both adults and infants can identify the language without access to segmental information [for references see 5, 6]. This has led to an increasing interest in segmentation which is based on acoustic basis and not phonological units. For example, Ramus [6] note that rhythm measures should ultimately be computed 'in more general terms, e.g. in terms of highs and lows in a universal sonority curve'. Potential outcomes of such computation have been demonstrated by [7].

In this paper we have segmented speech based on loudness and irregularity. The process yields three types of segments: silences, vowel-like segments with a nearly periodic waveform (1), and segments where the waveform is not periodic (2). Category (2) can include frication and/or regions with rapid changes in the waveform. Our algorithm[1] computes time series of specific loudness and aperiodicity [8, 9]. These values are smoothed and then compared against the thresholds in Figure 1 to generate transitions from one discrete state to another. The segmentation is controlled by 5 parameters: [a] a smoothing time constant for the loudness and irregularity time series (this has the effect of suppressing very short segments); [b] the normalised loudness of the silence-to-nonsilence transition; [c] the normalised[2] loudness of the nonsilence-to-silence transition (i.e. the transitions have hysteriesis); [d and e], the irregularity for the 2→1 and 1→2 transitions respectively.

The parameters are set by an optimization procedure and apply to the entire corpus. They are adjusted to minimize the

[1] Source code is available at http://sourceforge.net in the "speechresearch" project under z2009aesopRM.
[2] Normalization involves subtracting an estimated noise floor, and then scaling so that the average loudness is unity.

mean-squared difference between the number of regions generated by the segmentation and the number predicted from the phoneme-level transcription of each utterance[1]. Based on expected transcription of the text, the number of occurrences of state (2) is matched to the number of sequences of vowels and sonorants; state (1) to the remaining phonemes; and silences are weakly constrained to appear at about 10% as often as the other regions. The resulting parameters for our corpus are: 0.022 (seconds), 0.622, 0.0001, 0.382, 0.479, respectively.

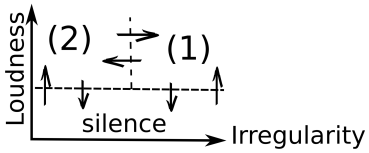

Figure 1: *Transitions between the three states.*

A comparison of the result against human segmentation by three professional phoneticians showed that state (2) corresponded to vowels and sonorants, state (1) corresponded to obstruents and the silence corresponded to pauses.

While pauses were consistently matched with silences in all languages, there were certain differences in the distribution of consonants between (1) and (2). Most notably, automatic segmentation reflected differences in the realization of stop consonants: in Mandarin /t/ was often lenited with sustained voicing and thus classified as state (2). There was also a noticeable difference in the segmentation of voiced plosives in French and English reflecting the difference in acoustic correlates of phonological voicing in these two languages.

Some acoustic classifications of segments differed from standard phonetic classification. For example, English [h] was consistently classified as 'vowel-like' (state 2) or part of silence. This agrees with the view that in English and possibly in other languages [h] is acoustically closer to approximants than to other fricatives [11: 326]. Similarly, devoiced vowels and sonorants in phrase-final position were consistently classified as 'consonants'.

2.2 Rhythm measures

Based on the segmentation described above, we computed the following rhythm measures as described in previous studies: %V [6], ΔV [6], ΔC [6], VI [12], CrPVI [13], VnPVI [13], CnPVI [13], Vdur/Cdur [14], PVI-CV [1], med_CrPVI [15], med_VnPVI [15], YARD [16], nCVPVI [17], VarcoΔC [18], VarcoΔV [18]. Although we follow the literature in using V and C in our labels, these really refer to states (2) and (1) respectively.

Comparison between expected transcription and segmentation showed that on average vocalic segments correspond to 1.4 syllables. This number was higher for Russian and Greek (1.57 and 1.62 respectively) and lower for Mandarin and English (1.27 and 1.30). One vocalic region generally corresponds to one syllable, but adjacent syllables are frequently fused together, e.g. if vowels were separated by sonorants.

Previous studies of RMs differed in their treatment of pauses and pre-pausal syllables. To estimate the potential effect of such differences, we have computed several values for each measure. For the first, we calculated the scores for each interpause stretch (IPS) and then computed an average for each text weighted by the duration of each IPS. For a second variant, we also applied the same algorithm, but did not include the final consonantal and vocalic intervals of each IPS. The third variant was computed across the whole text including intervals spanning a pause.

2.3 Classifier

In order to compare the intra-group variation in RMs to the inter-group variation, we apply classifier techniques as used in [8]. The classifier[2] draws boundaries between different languages, and we measure how often it can correctly predict the language, based on the RMs. A high success rate means that RMs from different groups are well separated; poor performance means that the intra-group variation is large so the probability distributions of the RMs of the languages overlap. Assuming that the RMs capture the rhythmic differences between the languages, success or failure of a classification corresponds roughly to whether a listener could identify the languages based on rhythm after listening to a single paragraph.

We used a classifier that assumes that the log likelihood ratio between the probabilities of any languages is a linear function of the rhythm measures fed into the classifier. The classifier's prediction is the estimate of which language was most likely to have produced the observed RMs. Classifiers were built with 16 different non-overlapping combinations of training and test sets. We report averages.

We used z-tests to test the significance of difference between success rate and chance for each classifier and also differences between classifiers.

3. Results

3.1 Classifiers based on single measures.

We tested all 45 variants of the 15 RMs described above, building a classifier for each variant (i.e. attempting to predict the language from a single RM).

Our results showed that different ways of computation had little effect on the overall efficiency of measures in separating languages. Overall, classifiers based on RMs computed without pre-pausal intervals performed slightly better than classifiers based on RMs computed using two other algorithms. However these differences were not large, nor did they affect the overall ranking of measures.

Table 1. *Results for classifiers based on one measure.*

RM	PCorrect	RM	PCorrect
PVI-CV	31%	VI	37%
VarcoΔC	33%	VarcoΔV	37%
ΔV	34%	nCVPVI	38%
YARD	34%	Vdur/Cdur	39%
ΔC	34%	%V	40%
CrPVI	35%	med_VnPVI	41%
CnPVI	35%	VnPVI	43%
med_CrPVI	36%		

[1] We used [10] to transcribe French texts.

[2] Source code is available at http://sourceforge.net in the "speechresearch" project under "g_classifiers-0.28.0".

The success rates for each measure are shown in Table 1. The table gives values for the variant computed across inter-pause stretches without final syllable (chance performance=30%). The success rate of classifiers based on many single measures was only slightly above the chance level. Measures based on vocalic intervals were generally more successful in separating languages than measures based on the variability of sonorant regions. Differences between the classifiers that are larger than 3% are significant at $P<0.01$.

There is no evidence that classifiers based on any single measure could better distinguish between languages traditionally assigned to different rhythm classes (e.g. English and French) than between languages from the same rhythm class (e.g. English and Russian). However these classifiers performed better at distinguishing Mandarin from other languages.

We also built and tested classifiers which separate each pair of languages (45 1-dimensional classifiers for 10 pairs of languages). This showed that some measures are better than others in separating specific pairs of languages. VnPVI consistently separated Mandarin from other languages. At the same time, CnPVI and CrPVI were more successful in separating French and Greek or French and Russian, while Greek and Russian were best separated by YARD.

3.2 Classifiers based on two or three measures

We then tested all 120 pairs of RMs with three variants for each pair, building 360 2-dimensional classifiers. While pairs of RMs were more effective than singletons, no pair correctly classified more than 50% of the data. The combinations which proved most efficient were %V-medVnPVI (49%) and %V-VnPVI (48%), medCrPVi-medVnPVI (48%) Vdur/Cdur-VnPVI (48%), and CrPVI-VnPVI (47%). The combination of %V and ΔC correctly classified 44%, while VarcoΔV-VarcoΔC achieved 40%. Differences between the classifiers that are larger than 3% are significant at $P<0.01$.

Although the most efficient pairs of measures achieved a similar success rate, they differed in how well they identified specific languages. Table 2 shows the percentage of correct identification achieved by these pairs for each language[1].

Table 2. *% of correctly classified data.*

RM	E	F	R	G	M
%V-med_VnPVi	72	0	33	19	69
Vdur/Cdur – VnPVI	73	0	34	15	70
CrPVI-VnPVI	70	2	2	50	70
%V-ΔC	75	16	34	25	64

We also ran 6 three-dimensional classifiers which combined most successful pairs of measures and singletons. The success rate of most efficient of these classifiers did not exceed the success rate of most efficient two-dimensional classifiers.

3.3 Multidimensional classifiers

As the next step we explored higher-dimensional classifiers based on more than three RMs. We built 3 15-dimensional classifiers (1 for each of the RM variants) and one 45-dimensional classifier, using all the measures. The overall success rate of these classifiers was not significantly better than the success rate of the most efficient two-dimensional classifiers (50% for 15-dimensional classifers and 52% for 45-dimensional classifier).

At the same time, the classifier based on all measures showed less differences than pairs of measures in percentages of correctly identified texts for each language (E: 60%, F: 26%, R: 37%, G: 47% and M: 71%).

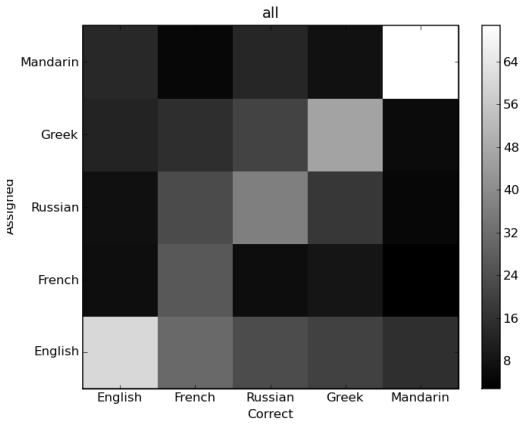

Figure 2: *Confusion matrix for 45-dimensional classifier.*

Figure 2 shows the confusion matrix for this classifier. The grey scale corresponds to percentage of data from language X (horizontal axis) classified as Y (vertical axis). Higher percentage is shown in greater brightness. The squares on the diagonal are correct classifications. Bright areas off the diagonal indicate classification mistakes.

3.4 Speech rate

Several studies have pointed out that speech rate may be an important factor when computing rhythm measures [cf. 4]. A classifier based on the average duration of underlying syllables was not significantly better than chance, which proves that differences in speech rate alone do not account for perceived differences between languages.

To gain a better understanding of the effect of speech rate, we then built 45 2-dimensional classifiers based on speech rate combined with one or another RM. The most efficient combinations are presented in Table 3. As in Table 1, the reported values are computed for inter-pause stretches without final syllables. The number in parentheses indicate the success rate for the measure without including the speech rate information. Significant differences are marked with asterisk[2].

Table 3. *Results for classifiers based on speech rate*

RM	Pcorrect
CrPVI	44% (35%) *
med_VnPVI	47% (41%)
Vdur/Cdur	48% (39%) *
%V	48% (40%) *
VnPVI	48% (43%)

The RMs that were most efficient in combination with speech rate were also the ones which were most efficient on their own. Adding speech rate led to a substantial increase in the success rate of measures that are not normalized by speech rate. Therefore we can conclude that even though speech rate

[1] E=English, F=French, R=Russian, G=Greek, M=Mandarin.

[2] The difference needs to be greater than 8% to be significant. The threshold is determined by how much the classifier performance varies from one choice of training set to another.

cannot separate languages on its own, it is definitely one of the variables in the 'rhythm equation' and needs to be included in any model of rhythm.

4. Discussion

Rhythm measurements based on automatic segmentation reveal rhythmic differences between languages. At the same time, there exists substantial variation within languages, which makes it impossible to reliably separate languages based on the rhythm of a single paragraph. These results agree with studies on human language identification. It has been repeatedly shown that when presented with a processed signal without segmental information, people are not able to correctly identify the language of all samples. The exact success rate depends on the experimental setup and languages: for studies based on low-pass filtering of the signal, the success rate for distinguishing between two languages is around 65%, with chance level at 50%. [for references see 5]. This is comparable to the success rate of our multidimensional classifier (53%, chance level: 30%).

We also found that some measures are better than others in separating specific languages. This agrees with an observation by [13] who noted complementarity between %V and VnPVI across different languages. Thus the efficiency of the measure depends on the languages in the corpus. Therefore studies based on different combinations of languages may come to different conclusions and this has to be taken into account when comparing their findings.

Our results provide evidence that rhythm is at least a two-dimensional phenomenon. While there seems to be an improvement in performance as we go from two-dimensional to high-dimensional classifiers, the increment in performance from each dimension beyond the first two clearly must be small. The possibility remains that rhythm requires more than two dimensions, but that existing RMs are strongly correlated with each other and that the set we tested are only capturing two dimensions of rhythm.

Finally, we have demonstrated the advantages of automatic segmentation, which consistently segments the data based on acoustic parameters. We have shown that acoustic properties of segments do not always match their expected phonological or even phonetic category. These differences are language-specific and provide experimental evidence that acoustic differences between phonological categories cannot be generalized across languages. This in turn raises the question of to what extent rhythm measures based on manual labelling are sensitive to potential differences in the phonological interpretation of sounds of a given language. For example, [13] note, that contrary to prediction, their intervocalic rPVI values for Japanese are similar to German and English, because they included devoiced vowel in the intervocalic regions. While it could be argued that perception of native language may be affected by the knowledge of phonological oppositions, this is certainly not true for unknown languages or pre-lexical infants. Therefore segmentation based on clearly defined acoustic parameters offers a better approximation of how rhythm is perceived in situations where segmental information is not available.

5. Acknowledgements

This project is supported by the Economic and Social Research Council (UK) via RES-062-23-1323. The authors would like to thank John Coleman for useful discussions. We acknowledge the National Science Foundation for providing support to Dr. Shih via IIS-0623805 and IIS-0534133. We also thank Speech Technology Center Ltd. (Russia) and Institute for Speech and Language Processing (Greece) for their help with automatic transcription of the data.

6. References

[1] W. Barry, B. Andreeva, M. Russo, S. Dimitrova, and T. Kostadinova, "Do rhythm measures tell us anything about language type?," in *Proceedings of the 15th ICPhS*, M. J. Solé and J. Romero, Eds. Barcelona, 2003, pp. 2693-2696.

[2] E. Keane, "Rhythmic characteristics of colloquial and formal Tamil," *Language and speech*, vol. 49, pp. 299-332, 2006.

[3] W. D. Raymond, M. Pitt, K. J. Johnson, E. Hume, M. Makashay, R. Dautricourt, and C. Hilts, "An analysis of transcription consistency in spontaneous speech from Buckeye corpus," in *ICSLP-02*. Denver, 2002.

[4] F. Ramus, "Acoustic correlates of linguistic rhythm: perspectives," in *Speech prosody* Aix-en-Provence, 2002, pp. 115-120.

[5] M. Komatsu, "Reviewing Human Language Identification," in *Speaker classification II*, C. Müller, Ed. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 206-228.

[6] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, p. 28, 1999.

[7] A. Galves, J. Garcia, D. Duarte, and C. Galves, "Sonority as a basis for rhythmic class discrimination," in *Speech Prosody 2002, Aix-en-Provence*, 2002, pp. 11--13.

[8] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *JASA*, vol. 118, pp. 1038-1054, 2005.

[9] G. Kochanski and C. Orphanidou, "What marks the beat of speech?" *JASA*, vol. 123, pp. 2780-2791, 2008.

[10] F. Bechet, "LIA_PHON : un système complet de phonétisation de textes," *Traitement Automatique des Langues*, vol. 42 numéro 1 - pp 47-67, 2001, pp. 47-67, 2001.

[11] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Oxford: Blackwell, 1996.

[12] D. Deterding, "The measurement of rhythm: a comparison of Singapore and British English," *Journal of Phonetics*, vol. 29, pp. 217-230, 2001.

[13] E. Grabe and E. L. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis," in *Laboratory Phonology, 7*, C. Gussenhoven and N. Warner, Eds.: Mouton de Gruyter, Berlin, Germany, 2002, pp. 515-46.

[14] W. Barry and M. Russo, "Measuring rhythm: is it separable from speech rate?," in *Actes des interfaces prosodiques*, A. Mettouchi and G. Ferré, Eds. Nantes: Université Nantes, 2003, pp. 15-20.

[15] E. Ferragne and F. Pellegrino, "A comparative account of the suprasegmental and rhythmic features of British English dialects," in *Modelisations pour l'Identification des Langues.* Paris, 2004.

[16] P. Wagner and V. Dellwo, "Introducing YARD (yet another rhythm determination) and re-introducing isochrony to rhythm research," in *Speech Prosody 2004*, Nara, Japan, 2004, pp. 227-230.

[17] E. L. Asu and F. Nolan, "Estonian rhythm and the pairwise variability index," in *FONETIK 2005*, Göteburg University, 2005, pp. 29-32.

[18] V. Dellwo, "Rhythm and speech rate: a variation coefficient for deltaC.," in *Language and language-processing: Proceedings of the 38th Linguistics Colloquium. Piliscsaba 2003.*, P. Karnowski and I. Szigeti, Eds. Frankfurt am Main: Peter Lang Publishing Group, 2006, pp. 231-241.