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, US

anastassia.loukina@phon.ox.ac.uk, greg.kochanski@phon.ox.ac.uk, cls@illinois.edu, elinor.keane@phon.ox.ac.uk, ian.watson@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 to compute rhythm measures over a large corpus. It may also approximate more closely the segmentation available to pre-lexical infants, who apparently can discriminate between languages. We find that within-language variation is large and comparable to the between-languages differences we observed. We evaluate the success of different measures in separating languages and show that the efficiency of measures depends on the 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, various quantitative statistical indices have been proposed to capture the rhythmic properties of languages. We follow Barry et al. [1] and collectively call these indices rhythm measures (RMs).

To date, observed differences in RMs have generally been interpreted as differences between languages or groups, but 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 have 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 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 permits application of rhythm measures to a larger corpus of data than previously used in these studies.

2. Data and methodology

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

Speakers were 20-28 years old; all had been 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 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 could repeat any text if they were not satisfied with their reading. Overall 15% of the recordings were repeated, though the fraction was highly variable from speaker to speaker. The recordings took place in two or three sessions on separate days.

2.1. Automatic and manual segmentation

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]. Many of the published rhythm measures are calculated on the basis of vocalic and intervocalic intervals. There is also an increasing interest in segmentation based on acoustics and not on phonological units. For example, Ramus [6] noted 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].

For 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. This is broadly consistent with most published rhythm measures which are defined on the basis of vocalic and intervocalic intervals.

Our algorithm[1] computes time series of specific loudness and aperiodicity [8, 9]. These values are smoothed and then compared against thresholds (see 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 (the smoothing process tends to suppress very short segments); [b] the normalised[2] loudness of the silence-to-nonsilence transition; [c] the normalised loudness of the nonsilence-to-silence transition (i. e. the transitions have hysteriesis); [d and e], irregularity thresholds for the **2→1** and **1→2** transitions respectively.

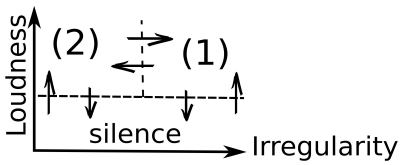

Figure 1: *Transitions between the three states.*

The parameters are set by an optimization procedure and apply to the entire corpus. They are adjusted to minimize the mean-squared difference between the number of regions generated by the segmentation and the number predicted from the phoneme-level transcriptions of utterances[3]. Based on expected transcription of the text, the number of occurrences of state (**2**) is matched to the number of sequences of vowels

[1] Source code is available at http://sourceforge. net in the "speechresearch" project under z2009aesopRM.
[2] Normalisation involves subtracting an estimated noise floor, and then scaling so that the average loudness is unity.
[3] We used [10] to transcribe French texts.

and sonorants; state (**1**) to the remaining phonemes; and silences are weakly constrained to appear about 10% as often as the other regions. The resulting parameters for our corpus are: 0. 022 (seconds), 0. 622, 0. 0001, 0. 382, and 0. 479, respectively.

A comparison of the results against 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 consistently matched silences in all languages, there were 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'.

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.

2.2. Rhythm measures

Based on the segmentation described above, we computed for each text the rhythm measures listed in Table 1. Although we follow the literature in using V and C in our labels, these really refer to states (**2**) and (**1**) respectively.

Previous studies of RMs differed in their treatment of pauses and pre-pausal syllables. To estimate the effect of such differences, we computed three alternatives for each measure. First, we calculated the scores for each interpause stretch (IPS) then averaged over all the IPSs within a text (The average was weighted by the duration of each IPS). Second, we applied the same algorithm but omitted the final consonantal and vocalic intervals of each IPS. Third, scores were computed across the whole text including intervals spanning a pause.

2.3. Classifier

To compare intra-group variation in RMs to inter-group variation, we apply classifier techniques as used in [8]. The classifier[4] is an algorithm that will optimally predict which language was most likely to have produced the observed RMs. We measure how often it can correctly predict the language, based on one or more RMs. 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. Each language then forms a convex polygonal region in the space of 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.

[4] Source code is available at http://sourceforge. net in the "speechresearch" project under "g_classifiers-0. 28. 0".

Table 1. *Rhythm measures used in this study*

RM	**Description**
%V	Percentage of vocalic intervals [6]
ΔV	Std. deviation of vocalic intervals [6]
ΔC	Std. deviation of consonantal intervals [6]
VI	Variability index of syllable durations [12]
CrPVI	Raw pairwise variability index (PVI) ofconsonantal intervals [13]
VnPVI	Normalised PVI of vocalic intervals [13]
CnPVI	Normalised consonantal PVI [13]
Vdur/Cdur	Ratio of vowel duration to consonant duration [14]
PVI-CV	PVI of consonant+vowel groups [1]
med_CrPVI	median CrPVI [15]
med_VnPVI	median VnPVI [15]
YARD	Variability of syllable durations [16]
nCVPVI	Normalised PVI of consonant+vowel groups [17]
VarcoΔC	ΔV/mean vocalic duration [18]
VarcoΔV	ΔC/mean consonantal duration [18]

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 one measurement of a single RM). The classifier was making a 5-way choice for each paragraph.

The 3 alternative ways of handling pauses had little effect on separating languages overall. Classifiers based on RMs computed without pre-pausal intervals performed slightly better than those based on RMs computed using the other two algorithms. However these small differences did not affect the overall ranking of measures.

The success rates for each measure appear in Table 2. The values are 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 consonantal regions. Differences between the classifiers that are larger than 3% are significant at P<0. 01.

Table 2. *Results for classifiers based on one rhythm measure.*

RM	PCorrect	RM	**%Correct**
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%		

Classifiers based on single measures could not distinguish between languages traditionally assigned to different rhythm classes (e. g. English and French) any better than between

languages from the same rhythm class (e. g. English and Russian). However these classifiers did relatively well at distinguishing Mandarin from other languages.

We also built and tested classifiers on pairs of languages (45 one-dimensional classifiers for each of 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. However, 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 built classifiers that made a five-way language prediction based on pairs of RMs. We tested all 120 pairs of RMs with three pause-handling alternatives for each pair, a total of 360 two-dimensional classifiers. While pairs of RMs were more effective than singletons, no pair correctly classified more than half of the data. The most efficient combinations 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.

Table 3. *% of correctly classified data, by language*[5].

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

Although the most efficient pairs of measures achieved a similar success rate, they differed in how well they identified specific languages. Table 3 shows the percentage of correct identification achieved by these pairs for each language[6].

We also ran 6 three-dimensional classifiers which combined the most successful pairs of measures and singletons. The success rate of the most efficient of these classifiers did not exceed the success rate of the 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 three 15-dimensional classifiers (one for each of the pause-handling alternatives) 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 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.

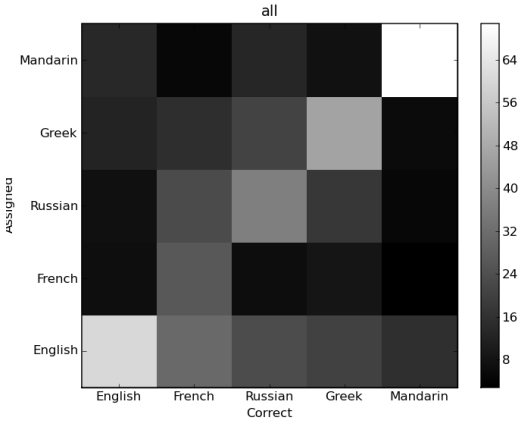

Figure 2: *Confusion matrix for 45-dimensional classifier.*

3.4. Speech rate

Speech rate may be an important factor that affects rhythm measures [cf. 4]. However, a classifier based on the average duration of underlying syllables was not significantly better than chance, which proves that differences in average speech rate do not account for perceived differences between languages.

To gain a better understanding of the effect of speech rate, we then built 45 two-dimensional classifiers based on speech rate combined with one or another RM. The most efficient combinations are presented in Table 4. As in Table 2, the reported values are computed for inter-pause stretches without final syllables. The numbers in parentheses indicate the success rate for the measure without including the speech rate information. Significant differences are marked with an asterisk[7].

Table 4. *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 normalised by speech rate. Therefore, even though speech rate 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 measures based on automatic segmentation reveal 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 lacking segmental information, listeners often cannot correctly identify the language. The exact success rate depends on the experimental setup and languages: in 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

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

[6] Performance for English was high because English was the largest fraction of the training set.

[7] 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.

references see 5]. The success rate of our multidimensional classifier (53%, vs. a chance level of 30%) is at least comparable.

We also found that some measures are better than others in separating specific languages. This agrees with an observation by [13] who noted a 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 appears to be a a two- or three-dimensional phenomenon[8]. 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 strength of this argument is limited by the set of published RMs. 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 is only capturing two dimensions of rhythm.

Finally, we have demonstrated the advantages of automatic segmentation, which consistently segments speech 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 may vary across languages. In turn this suggests that 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 are similar for Japanese, German and English, because they included devoiced vowels 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 and Burt Rosner 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. (St.-Petersburg, Russia) and Institute for Speech and Language Processing (Athens, 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. mith, J. O. and Abel, J. S. , "Bark and ERB Bilinear Transforms", IEEE Trans. Speech and Audio Proc. , 7(6):697-708, 1999.

[19] Soquet, A. , Saerens, M. and Jospa, P. , "Acoustic-articulatory inversion", in T. Kohonen [Ed], Artificial Neural Networks, 371-376, Elsevier, 1991.

[20] Stone, H. S. , "On the uniqueness of the convolution theorem for the Fourier transform", NEC Labs. Amer. Princeton, NJ. Online: http://citeseer. ist. psu. edu/176038. html, accessed on 19 Mar 2008.

[8] Speech rate might count as an additional dimension.