Rhythmical classification of languages based on voice parameters

Volker Dellwo, Adrian Fourcin and Evelyn Abberton

Department of Phonetics and Linguistics, University College London
v.dellwo@ucl.ac.uk

ABSTRACT

It has been demonstrated that speech rhythm classes (e.g. stress-timed, syllable-timed) can be distinguished acoustically and perceptually on the basis of the variability of consonantal and vocalic interval durations. It has moreover been shown that even infants are able to use these cues to distinguish between languages from different rhythm classes. Here we demonstrate that the same classification is possible in the acoustic domain based simply on the durational variability of voiced and voiceless intervals in speech. The advantages of such a procedure will be discussed and we will argue that 'voice' possibly offers a more plausible cue for infants to distinguish between languages of different rhythmic class.

Keywords: voice, speech rhythm, rhythm measures, infant speech perception, laryngography

1. INTRODUCTION

It has been demonstrated exhaustively from the 1970s to the 1990s that the traditional classification of languages into rhythmic classes like stress-timed and syllable-timed is not manifested in isochronous inter-stress-intervals or isochronous syllable-durations, respectively, on an acoustic level. However, languages can be classified acoustically and perceptually into traditional rhythm classes on the basis of the variability of their consonantal and vocalic intervals (see [6] who defines a consonantal interval as the consonant/s between two vowels and a v-interval the vowel/s between two consonants). [6] found further that on an acoustic level this variability is reflected in the overall percentage over which speech is vocalic (%V) and the standard deviation of consonantal interval durations (ΔC). Plotting these two parameters along two dimensions shows that stress-timed languages cluster differently from syllable-timed languages (see Figure 1). The rationale underlying these measures is the assumption that speech rhythm is a product of the phonotactic complexity of a language. Languages traditionally classified as stress-timed show phenomena like vocalic reductions and complex consonant clusters. It is assumed that the presence of vocalic reductions in the speech signal leads to an overall lesser percentage over which speech is vocalic and that the presence of complex consonant intervals leads to a greater variability of consonantal interval durations.

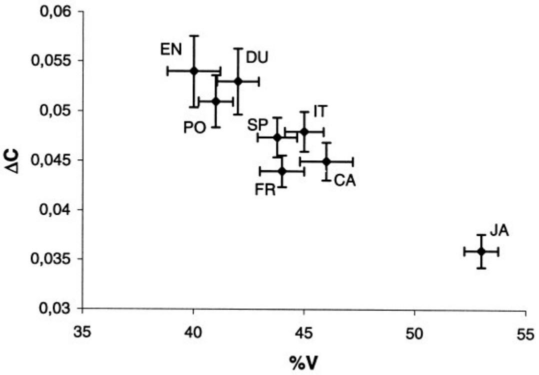

Figure 1: Results from [6] for cv-variability represented by ΔC and %V respectively: Stress-timed languages (EN: English, DU: Dutch, PO: Polish) can be distinguished from syllable-timed languages (SP: Spanish, IT: Italian, FR: French, CA: Catalan)

[6] claimed that listeners use parameters like %V and ΔC to distinguish languages belonging to different rhythmic classes. They further argue that infants are able to distinguish different languages on the basis of this type of acoustic information before they actually have any knowledge of a language's phonological structure.

In this paper we take a different approach. We argue that if infants are able auditorily to use acoustic information about the speech signal simply to distinguish different rhythmic classes from each other, then this information should even be less complex than 'vocalic' and 'consonantal'. Why should an infant for example for French be able to distinguish auditorily between a nasal (consonantal interval) and nasal vowel (vocalic interval)? We therefore assume that there are probably rhythmical aspects in the use of voice alone on the basis of which listeners are able to distinguish languages.

For this reason we adopted the measures %V and ΔC developed by [6] and calculated them for 'voiced' and 'unvoiced' stretches of speech, replacing 'vocalic' and 'consonantal' respectively (thus %V is applied to voiced intervals, ΔC is applied to unvoiced intervals). Some of the main differences between these segmentation techniques are that most voiced consonants are part of voiced stretches of speech and no longer consonantal. However, voiced plosives with a voiceless stop gap will contain a short unvoiced interval at this point. In the present paper we report on the results of these acoustic measurements. A perceptual evaluation of our results is currently in progress.

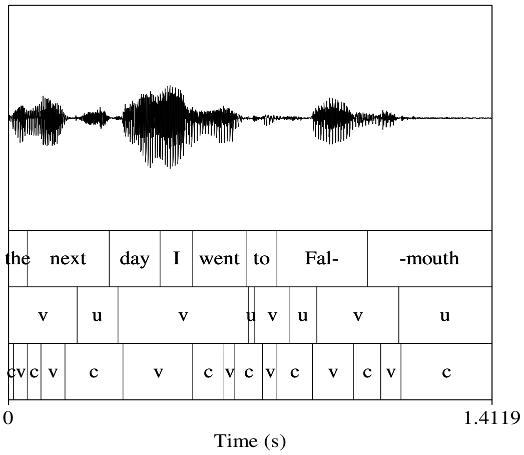

Figure 2: Waveform of the inter-pause interval ‘The next day I went to Falmouth’ segmented on three tiers 1) syllable durations (top tier), 2) voiced (v) and unvoiced (u) intervals (middle tier), 3) consonantal (c) and vocalic (v) intervals (bottom tier).

2. Experiment

The speech material used for the present experiment has been taken from the BonnTempo-Corpus since languages of different rhythmic classes in this database have been shown to be well separated by %V and ΔC (see [4]). Voiced and voiceless stretches of speech were labeled automatically with Praat [1] (see below). Consonantal and vocalic intervals have been manually annotated by [4] in the BonnTempo-Corpus since automatic annotation failed to produce adequate accuracy (see [4]).

2.1. Method

2.1.1. Languages, Speakers & Speech Material

Languages traditionally classified as stress-timed, English (E) and German (G), and syllable-timed, French (F) and Italian (I), have been chosen from the BonnTempo-Corpus. The number of speakers available for these languages were E: 7, G: 13, F: 5, and I: 3.

The analysis is based on inter-pause intervals of speech, which are intervals between two pauses in speech discourse as performed by the speaker. BonnTempo offers a number of different intended speech rates for which speakers either intended to speak slow, normal, or fast. Only the normal intended speech rates are used for the current experiment since a different use of voicing mechanisms could be the case for slow and fast speech. The number of inter-pause intervals obtainable for each language are E: 48, G: 104, F: 67, and I: 24.

2.1.2. Procedure

All inter-pause intervals have been labeled automatically in voiced and voiceless intervals with a Praat script written for the purpose of this experiment. The script first produced a pitch tier of the speech waveform that was smoothed with a bandwidth of 10 Hz using the Praat smoothing function. This pitch tier was then resynthesised into a waveform (hum). For a 1 msec frame the root-mean-square (rms) was calculated in steps throughout the whole signal. When a strong change in rms was detected between two consecutive frames, a boundary was placed. After that intervals were checked for energy and intervals with energy below a certain threshold were labeled unvoiced, above the threshold voiced. The automatic segmentation procedure was checked manually and a near 100% precision was found.

2.2. Results

2.2.1. ‘Voiced’ versus ‘vocalic’ intervals

Figure 2 shows the waveform of an English inter-pause interval ‘The next day I went to Falmouth’ with three types of segmentations, a) syllables, b) voiced and unvoiced intervals, c) consonantal and vocalic intervals. The figure illustrates nicely typical differences that can be observed between voiced/unvoiced stretches and consonantal and vocalic intervals. While there are only 4 voiced stretches in this signal there are 7 vocalic intervals. The first voiced interval includes two consonantal and two vocalic intervals. The total number of voiced-intervals as opposed to vocalic-intervals is about 1 to 4 for all languages in the data. From this it can be concluded that there is a large quantitative and distributional difference between voiced and vocalic intervals.

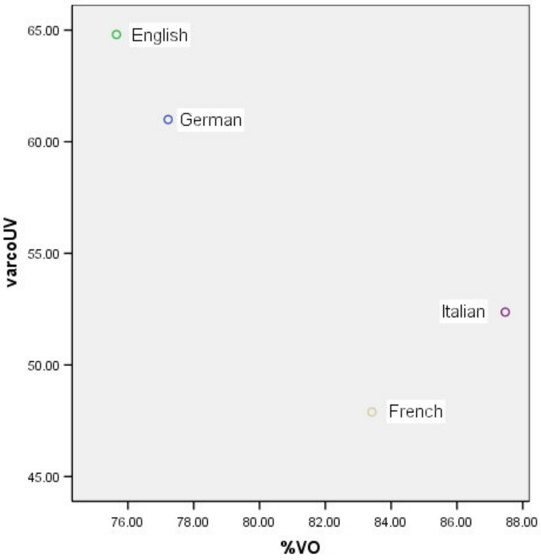

Figure 3: Cross plotted mean values for rate normalized variability of unvoiced intervals (varcoUV) and percentage over which speech is voiced (VO).

2.2.2. *Variability of voiced and unvoiced intervals*

In this section we process the variability measures %V and ΔC to voiced and unvoiced sections of the signal respectively and call it %VO (percentage over which speech is voiced) and ΔUV (standard deviation of unvoiced intervals). [4] showed that there is considerable variation of ΔC as an effect of speech rate and developed a rate normalized measure by calculating the variation coefficient of the standard deviation of consonantal intervals (ΔC*100/meanC; varcoC). Since a similar variability is expectable for voiced and unvoiced intervals the rate normalized variety of ΔC is used for monitoring the standard deviation of unvoiced intervals (varcoUV = ΔUV*100/meanUV).

Figure 3 shows the rate normalized measure varcoUV as a function of %VO. The graph shows that a pattern can be observed that is like the one [6] obtained for consonantal and vocalic variability (ΔC and %V respectively). Languages traditionally classified as stress-timed languages (here: English and German) have a higher variability of unvoiced intervals than syllable-timed languages (here: French and Italian). Also, an overall shorter percentage over which speech is voiced can be observed for English and German compared to French and Italian. The total percentages for this parameter lie far higher than for the comparable parameter %V which is an effect of a large number of voiced consonants being part of %VO but not of %V.

An ANOVA (univariate procedure) with %VO and varcoUV as the dependent variables shows that there is highly significant variability between the four distributions (F[3, 239]=17.41, p<0.001).

A Tukey’s post-hoc test reveals details about within and between rhythmic class variability. For %VO within class comparison is represented by the pairs G-E and F-I which have p values of .717 and .191 respectively, i.e. the variation between groups E and G is due to chance. This is not the case for between rhythmic class variability represented by G-F, G-I, E-F, and E-I which all have p values smaller than .001, i.e. there is highly significant variability between rhythmic class.

For varcoUV we receive the same quality of results (ANOVA: F[3, 239]=5.17, p<0.005). P values from a Tukey’s post-hoc test reveal again non-significant within class variability (G-E: .839, F-I: .890), however, between class variability is not as clear: While G-F and E-F is significant at .009 and .004 levels respectively, G-I and E-I are non-significant at .467 and .232 respectively.

In conclusion it can be said that, apart from the case of unvoiced interval variability in Italian, rhythmic classes are well separated in the data.

3. Discussion

The results of this research have shown that stress- and syllable-timed languages can be distinguished on the basis of voiced and unvoiced intervals. In the following we will discuss the advantages of such a segmentation procedure.

The main advantage of the present method is that rhythmic classification of languages can be carried out with much less effort. Manual labeling of consonantal and vocalic intervals is labor intensive and because of the considerable level of phonological knowledge involved in this process (e.g. is a retroflex approximant /r/ vocalic or consonantal?) automatic procedures have so far revealed unsatisfactory results. Such procedures would require specific training for individual languages when applied cross linguistically. Also, because of the level of phonological knowledge involved in the distinction of vocalic and consonantal intervals between-labeler disagreement can be significant. This disagreement is even stronger across different languages or when accentual pronunciation variability occurs.

Detecting voiced and voiceless parts of the signal is a much easier and more reliable method and it is applicable on a cross language basis with fewer assumptions. To obtain additional precision obtaining the 'voice'-data, technology monitoring vocal fold activity directly can be used (e.g. laryngograph, see discussion in the next section).

Since fewer assumptions are required to distinguish stress- and syllable-timed languages on the basis of voiced and voiceless cues this may also have implications on how infants distinguish between rhythm-class. After all infants receive most of their familiarization with speech acoustics in the mother's womb [6] where they are exposed to a highly low pass filtered signal (<300Hz) and no visual cues are not available. In such an environment voice cues are much more salient than any other acoustic feature of speech. For this reason we raise the assumption that infants may prefer voice variability cues over consonantal and vocalic interval variability cues to distinguish between speech rhythm class.

4. Prospects

The present results pose a number of questions. It may be that the results are only valid for the languages under investigation which is why we plan to extent our studies to a wider variety in the future.

So far the acoustic analysis has been based on an estimation of voiced stretches in speech with standard pitch tracking algorithms (here the pitch tracking algorithm in Praat). These algorithms produce a considerable number of artifacts and our results have possibly been influenced by them. More reliable results for identifying voiced stretches in the speech signal can be obtained with a laryngograph (see [5]) which plots a function of vocal fold contact area derived from a small current passing through the larynx via two attached electrodes. Periodic variability in vocal fold contact area then reliably distinguishes stretches during which periodic vocal fold activity is present or not. We are currently collecting data from a number of speakers of languages traditionally classified as typically stress- or syllable-timed and carrying out recordings of the acoustic and laryngographic waveforms. We expect to report on this data soon.

Another major stage in the progress of this work will be to conduct auditory experiments without visual cues. If our theory holds we need to find a way to demonstrate that infants actually prefer voice cues over consonantal and vocalic interval cues to distinguish between languages. We are currently thinking of experiments for which we use stimuli from languages of different rhythmic classes that have similar consonantal and vocalic interval variability but are distinguished by voiced interval variability. If infants were still able to distinguish rhythmic class, voice would demonstrate to be a reliable cue even if consonantal and vocalic variability fails.

5. Conclusion

The current research has demonstrated that, on an acoustic basis, languages are distinguishable in rhythmic classes on the basis of their use of voiced intervals. We argue that this is an easier and more relevant method for cross language analysis than the commonly used segmentation technique of consonantal and vocalic intervals and may be preferred by infants in order to distinguish languages of different rhythmic class.

6. REFERENCES

[1] Boersma, P. 2001 *Praat, a system for doing phonetics by computer.*In: *Glot International* 5:9/10, 341-345.

[2] Dellwo, V. 2006 *Rhythm and speech rate: A variation coefficient for* ΔC. Pawel Karnowski & Imre Syigeti (eds.) *Language and Language-processing.* Frankfurt am Main: Peter Lang, 231-241.

[3] Dellwo, V. and Wagner, P. 2003 *Relationships between speech rhythm and rate. Proceedings of the 15th ICPhS,* 471-474.

[4] Dellwo, V. Aschenberner, B., Dancovicova, J. and Wagner, P. 2004 *The BonnTempo-Copus and Tools: A database for the combined study of speech rhythm and rate. Proceedings of the 8th ICSLP.*

[5] Fourcin, A. 2000 *Voice Quality and Eletrolaryngography.* Kent, R. D. and Ball, M. J. (eds.) *Voice Quality Measurements.* San Diego: Singular.

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