Vocale - A Semi-Automatic Annotation Tool for Prosodic Research

Jesus Garcia, Ulrike Gut & Antonio Galves

Institute of Mathematics and Statistics, University of São Paulo, Brazil
Department of Linguistics and Literary Studies, University of Bielefeld, Germany
{jesusg; galves}@ime.usp.br, gut@spectrum.uni-bielefeld.de

Abstract

Large annotated speech corpora are a critical component of research in prosody. The classification of languages according to their speech rhythm, for example, requires a great number of annotated sentences by different speakers in different languages. We have developed *Vocale*, a tool for the semi-automatic annotation of vocalic and consonantal parts of speech because in recent models these units have been identified as reliable acoustic correlates of speech rhythm. *Vocale* is based on relative entropy and uses various additional classifiers such as energy and length for the annotation of vowels and consonants. It runs using Praat speech analysis facilities and gives a Praat label file as an output. *Vocale* is open source software and is available to the scientific community under http://www.ime.usp.br/∼tycho/tipal/prosody/vocale/.

1. Introduction

Research on speech prosody relies on large annotated corpora. This is especially true for typological comparisons of prosodic systems. It has been claimed since [5] and [1] that the languages of the world differ in their speech rhythm. Originally, three classes of speech rhythm were proposed: syllable-timed rhythm, where the time interval between sllables was supposed to be equal; stress-timed languages, where isochrony was proposed for stress beats; and mora-timing, where moras are produced with equal timing. No acoustic correlates for these claims were ever found [7].

Current approaches to the measurement of speech rhythm assume no discrete classes of languages any more but propose a single dimension ranging from stress-timing to syllable-timing on which languages can be grouped. As a phonetic correlate of speech rhythm, the proportion and standard deviation of vocalic and consonantal intervals in the speech signal are calculated [6]. In a comparison of ten languages it was found that they grouped along this dimension in clusters very similar to the original classification of stress-timed and syllable-timed classes.

Due to the time-consuming work of annotating speech by hand, no comprehensive comparisons of languages which comprise more than a few sentences and more than a very limited number of speakers have yet been carried out. Furthermore, human annotation is error-prone and implicitly incoherent. For both reasons, the development of an automatic prosodic annotation system is becoming increasingly important.

Currently available automatic annotation tools either require a transcription of the spoken text as input [8], [9] or do not annotate any phonological units smaller than the word. The underlying algorithms are either Hidden Markov Models (HMM) or Euclidian distance [2].

With *Vocale* we are developing a tool for a semi- automatic speaker-independent annotation of vocalic and consonantal intervals in large amounts of speech data that does not reqire a previous transcription of the speech material. It is based on relative entropy, in combination with other measurements of the acoustic properties of speech. In the following we describe the layout and the algorithm of the programme as well as the first results we have obtained for English and Polish.

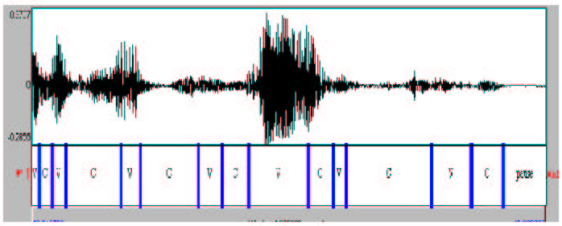

Figure 1: A Praat label file with the annotated vocalic and consonantal intervals automaticallt produced by Vocale.

2. Vocale - basic layout and algorithm

2.1. Basic layout features

Vocale takes a wav speech file as input. In order to allow the processing of speech data with low sampling rates such as television recordings, we are using a sample rate of 11025 Hz. *Vocale* then uses the Praat [10] software to create a Gaussian spectrogram with the parameters 0 to 5000 Hz, frequency step 50 Hz and time step 2ms. The output file is a Praat label file, with the option of converting this into an ESPS/waves+ label file [4].

This file serves as a basis for the calculation of the vocalic proportion ($V\%$) and standard deviation of consonantal intervals (ΔC).

2.2. The algorhithm

Our computations are based on the analysis of the spectrogram of the signal. We denote by $c_t(f)$ the Fourier coefficient for the frequency f around t estimated with a 25ms window, the value of the spectrogram in time t is $(c_t(f))^2$.

We define the renormalized spectrogram by

$$p_t(i) = \frac{(c_t(i))^2}{\sum_f (c_t(f))^2} \qquad (1)$$

As a first step, we divide speech from pauses by means of the total energy of each column of the spectrogram and the length of the event. We have set a threshold of 0.002 for the detection of a speech event in contrast to pause. Paired with a time constraint of > 27ms we avoid annotating closure parts of stops as pauses.

In the second step, in the range of 0 to 1000 Hz, the relative

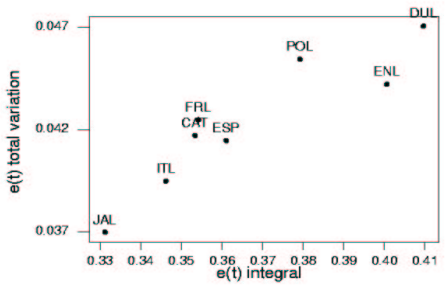

Figure 2: Classification of rhythmic classes equivalent to Ramus's findings obtained with the measurements of the integral and the total variation of $e(t)$.

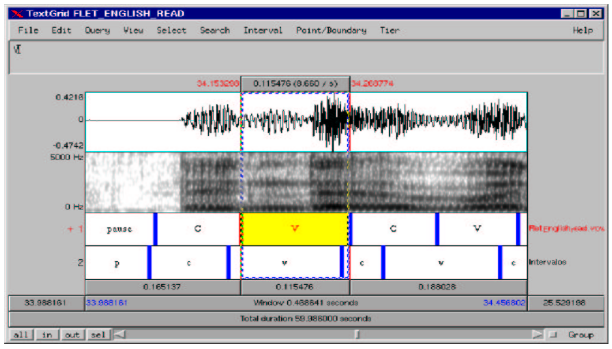

Figure 3: Comparison of the human annotations (above) with the automatic annotations (below).

entropy is calculated for four consecutive columns. The formula for the relative entropy for the column p_t with respect to the column p_s is

$$h\left(p_t|p_s\right) = \sum_i p_t\left(i\right) \log\left(\frac{p_t\left(i\right)}{p_s\left(i\right)}\right) \tag{2}$$

It indicates how different the two columns are from each other as probability measures. We use the mean value of three relative entropies:

$$h(t) = \frac{1}{3}\sum_{i=1}^{3} h\left(p_t|p_{t+i}\right) \tag{3}$$

Thus, vowels and nasals as well as voiced stops will have a low entropy value, whereas voiceless stops, fricatives and flaps have a high relative entropy. Time constraints for vowels and consonants are applied. In addition, a smoothing factor is introduced

$$e(t) = \frac{1}{4}\sum_{i=-1}^{1} h(t+i) \tag{4}$$

It has been shown that using $e(t)$ as a rough measure of sonority it is possible to replicate the result obtained by [6] (see figure 2).

In order to annotate nasals, we include the measurement of the Euclidian distance and set a threshold for the detection of nasals. Voiceless fricatives are annotated with the additional classifier of low energy in the band between 0 and 1000 Hz.

In sum, *Vocale* uses the following classifiers:

- Pauses: low total energy and a minimal length of 30 ms.
- Vowels: low relative entropy between 0 - 1000 Hz, high energy between 0 and 5000 Hz and at least 16 ms length.
- Consonants: in general high relative entropy between 0 - 1000 Hz and at least 8ms length.

  Special cases:

  - nasals: low relative entropy and very low energy above 1000 Hz
  - voiceless fricatives: low energy between 0 and 1000 Hz
  - voiceless stops: high relative entropy between 0 - 1000 Hz and a minimal length of 8 ms

3. First Results

We are currently testing *Vocale* against hand-labelled data in various languages. First, it was applied to British English speech, which constitutes part of the LeaP corpus (http://www.spectrum.uni-bielefeld.de/LeaP/) and which was hand-labelled and cross-checked by trained phoneticians. We selected four speech recordings with a female and a male speaker, one read speech and one re-telling of a story for each speaker. In total, the human labellers annotated 398 vocalic intervals 429 consonantal intervals and 45 pauses in the two recordings with speaker 1 and 570 vocalic intervals, 606 consonantal intervals and 72 pauses for speaker 2. These annotations were compared with the annotations created automatically by *Vocale*.

Second, we tested *Vocale* against ten sentences of Polish, read by three different female speakers. The annotations were done by Franck Ramus [6] and comprise 158 vocalic intervals, 168 consonantal intervals and 12 pauses. Table 1 presents the results for British English. Recognition rate for vocalic intervals was 67.8%, for consonantal intervals 88.9% and for pauses 71.1%. Criterion for a match between human and automatic annotation was the agreement of boundaries with less than 10ms difference in the time stamp.

Table 1: Recognition rate (in %) of vocalic and consonantal intervals and pauses by Vocale for the English data compared to human annotations.

speaker	vocalic intervals	consonantal intervals	pauses
speaker1	67.8	88.9	71.1
speaker2	49.9	80	97.4

Table 2 presents the results for the Polish data. Recognition rate for vocalic intervals was 50 %, for consonantal intervals 56% and for pauses 92%.

Table 2: Recognition rate (in %) of vocalic and consonantal intervals and pauses by Vocale for the Polish data compared to human annotations.

vocalic intervals	consonantal intervals	pauses
50	56	92

4. Conclusions and Outlook

In this paper we showed that *Vocale* is a good alternative for the time-consuming and error-prone task of annotating speech data by hand. Although at the current early stage of development the program still produces erroneous output for some speakers, it has the advantage over hand labelled data that errors are systematic and implicitly coherent. Correction by hand is therefore a relatively simple task.

Future developments of *Vocale* will include the integration of an automatic learning model to find more classifiers and their optimal values. This will optimize our current approach of determining the values by hand. Furthermore, as shown in tables 1 and 2 and in figure 2, the mean value of the entropy changes with the rhythmic class of the language. Therefore different languages probably require different parameters, and the learning model will achieve finding these new parameters and assigning their optimal values with speed and precision. We are also planning to integrate a probabilistic model in the algorithm because it has been shown that the length of consonantal and vocalic intervals are gamma-distributed [3].

4.1. Acknowledgements

This work was partially supported by FAPESP (Projeto Tematico Rhythmic patterns, parameter setting and language change, grant 98/33820 and grant 00/07959-1), CNPq (Project Probabilistics Tools for Pattern Identification Applied to Linguistics, grant 465928/2000-5, and *Agreement Brasil-France*, grant 69.0014/01-5), agreement USP-COFECUB and CAPES/PICDT and is part of the activities of the Nucleo de Excelencia Critical phenomena in probability and stochastic processes (grant 66.2177/19966).

5. References

[1] Abercrombie, D., 1967. *Elements of general phonetics.* Chicago: Aldine.

[2] Cosi, P., 1997. SLAM v1.0 for Windows: A Simple PC-Based Tool for Segmentation and Labeling. *Proceedings of ICSPAT* , San Diego.

[3] Duarte, D., 2001. Statistical evidence and the rhythmic class hypothesis. http://www.physik.uni-bielefeld.de/complexity/duarte.pdf

[4] Milde, J.-T., & Gut, U., 2001. The TASX-environment: an XML-based corpus for time-aligned language data. *Proceedings of the IRCS Workshop on Linguistic Databases* , Philadelphia.

[5] Pike, K., 1945. *The intonation of American English.* Ann Arbor: University of Michigan Press. *2nd ESCA Workshop on Speech Synthesis.* , New York: Mohonk, 155-158.

[6] Ramus, F., Nespor, M., & Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* , 73(3), 265-292.

[7] Roach, P., 1982. On the distinction between stress-timed and syllable-timed languages. In: D. Crystal (ed.) *Linguistic Controversies. Essays in Linguistic Theory and Practice.* London: Arnold, 73-79.

[8] http://www.sys.uea.ac.uk/∼ sjc/annotation.html

[9] http://poseidon.itc.it:7117/∼ssi/SRS_project/papers/EUROSPEECH93_4.html

[10] http://www.praat.org