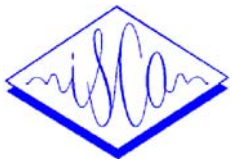
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate

Volker Dellwo[1,2], Ingmar Steiner[2], Bianca Aschenberner[2], Jana Dankovičová[3], Petra Wagner[2]

[1]Department of Phonetics and Linguistics, University College London, UK
[2]Institut für Kommunikationsforschung und Phonetik, Universtität Bonn, Germany
[3]Department of Human Communication Sciences, University College London, UK
volker@phon.ucl.ac.uk

Abstract

Work is currently being carried out on a speech database constructed in order to study speech rhythm in connection with speech rate. The database, BonnTempo-Corpus, and the Praat based analysis tools, BonnTempo-Tools, are a powerful instrument for examining various aspects of recently proposed rhythm measures (e.g. %V, ΔC, nPVI, rPVI, etc.) in relation to speech rate among a wide range of languages and speakers. First observations pose new problems on traditionally not well classifiable languages like Czech.

10.21437/Interspeech.2004-294

1. Introduction

In recent years new promising measures have been proposed ([1], [2]) to capture speech rhythm, amongst all the classic rhythm distinction between 'stress-timing' and 'syllable-timing'. Unlike the traditional measures, based on the duration and standard deviations of syllables and inter-stress units and now widely thought to be inadequate as acoustic correlates of speech rhythm [3], these measures are based on the duration of vocalic and intervocalic (or consonantal) intervals (henceforth: V- and C-intervals). [1] propose to measure the standard deviation of C-intervals (ΔC) and the percentage of V-intervals (%V) while [2] propose to measure normalized variation of consecutive V-intervals (nPVI) and non-normalized variation of consecutive C-intervals (rPVI). The rationale behind this type of measures has been widely discussed by the respective authors: they take into account a higher degree of vowel reduction in languages traditionally classified as stress-timed (leading to a lower %V and a higher nPVI) and a lower degree of syllable complexity in syllable timed languages (leading to a lower ΔC and a higher rPVI).

A general characteristic of numerous studies on speech rhythm, including [1] and [2], is a limited amount of data, i.e. a small number of speakers per language (sometimes only 1) and a small number of speech samples. For example, in a classical and often cited study [3], 2 minutes of spontaneous speech (a description of a picture) by one speaker represent the results for the whole language.

In [1] and [2] slightly more but still rather little speech material was used: In [1] 8 languages are represented by 4 speakers each, while 5 sentences of 15 to 19 syllables (n=17) have been recorded for each speaker. This sums up to 2720 syllables for the whole database (8*4*5*17) or 340 syllables per language. [2] examined 16 languages, each represented by 1 speaker reading the original or a respective translation of 'The North Wind and the Sun', which contains 141 syllables in the English version. Assuming the average number of syllables in each language version is around 150 syllables/text, the total number of syllables for the whole data analyzed sums up to 2256 syllables (16*150).

The use of a limited amount of data especially in rhythm studies may lead to artifacts in the results. For this reason studies based on [1] and [2]'s measures, such as [4], [5] and [6] started using larger data collections. [6], who use an earlier form of the corpus presented in the current article, base their study on three languages or 6420 syllables (about 2140 syllables/language). [5] uses the database described in the current article (without Czech).

So far [4] have apparently used the largest data set in rhythm research with more than 5000 interpause stretches for 3 languages. Since the authors do not specify the number of syllables in their database a straightforward comparison is difficult but considering the number of interpause stretches the number of syllables in [4]'s database must sum up to several thousands.

Apart from a limited size of experimental data the discussion about the reliability of recent rhythm measures is concerned with their dependency on speech rate ([2], [4], [5], [6], [7]) since with in- or decreasing rate relative durations of sound segments (the basis for the rhythm measurements) change drastically. Attempts to normalize the PVI measure (nPVI) according to speech rate have already been criticized [4]. Attempts to standardize syllable rate by comparing sentences of roughly equal duration and number of syllables [1] may lead for instance to comparing speech rates which are normal for speakers of one language but fast for speakers of another language [6]. However the rhythmic pattern may be affected by a change in speech rate [5].

In order to examine recent and traditional rhythm measures on a larger data set under different speech rates in L1 and L2 conditions, we have started compiling a database, the BonnTempo-Corpus (BTC). To be able to obtain these measures, the database has been labelled for syllables and C-/V-intervals. The analysis of the BTC with respect to a wide variety of rhythm and speech rate measures is facilitated by a collection of software tools, the BonnTempo-Tools (BTT). The first versions of the BTC (BTC 1.0) and the BTT (BTT 1.0) are now available from the first author. The present article gives an overview of content and construction of the BTC (section 2) and explains the facilities of the BTT (section 3) as well as examples of first observations (section 4).

2. BonnTempo-Corpus (BTC)

The BTC 1.0 is one of the largest databases currently available for the study of the recently established rhythm measures of the type of [1] and [2], and is expected to grow further. It currently consists of 24 070 syllables and 43 227 C- and V-intervals (c = 22 705; v = 21 522) for 5 languages and 4 L2

conditions, while the absolute number of speakers (and thus syllables) per language still varies considerably (cf. 2.2).

2.1. Speech Material

The speech material in BonnTempo currently consists only of read speech but we plan to include spontaneous speech in the near future. The text is a short passage from a novel by Bernhard Schlink ('Selbs Betrug') with 76 syllables in the German version. This text has been translated into the other languages under investigation by philologically educated native speakers of the target languages, Czech (93 syllables), English (77 syllables), French (93 syllables), Italian (106 syllables).
The languages were selected to represent both traditional rhythmic classes. 'Stress-timing' is represented by English and German, 'syllable-timing' by French and Italian. Rhythmic classification of Slavic languages has been widely disputed, and Czech has been included in the database as a good example of this. It has traditionally been classified as syllable-timed (e.g. [8]), but later shown to have a tendency towards stress-isochrony ([9]), or towards either rhythmic patters, depending on the type of measure ([10]). Polish is due to be included in BTC as well (recordings have been made but labeling is yet to be completed).

2.2. Speakers/Languages

The BTC 1.0 contains examples from speakers of the following languages (in brackets: number of labeled speakers, number of total syllables, respectively):

- Czech (4, 1855)
- English (7, 2684)
- French (6, 2732)
- German (15, 5699)
- Italian (3, 1619)
- French from Cameroon (1, 343)

The database further contains L2 speakers:

- English speaking German (3, 1140)
- French speaking English (2, 776)
- French speaking German (1, 380)
- Germans speaking French (8, 3503)
- Germans speaking English (8, 3087)

Recordings of the following languages already exist and are in the process of being labelled:

- Polish
- Brazilian Portuguese
- Portuguese Portuguese

The authors intend to extend the database to include more languages in the near future, especially languages of the third recognised rhythm type - mora-timed languages (e.g. Japanese).

2.3. Recording Procedure

Recordings have been carried out mainly in the sound proof booth of the Institut für Kommunikationsforschung und Phonetik at Bonn University with a large membrane condenser microphone directly on PC in wav file format. Most of the recordings for French have been carried out in Bordeaux in private homes on mini-disc, all recordings for Czech have been carried out in private homes in the Prague region on DAT. Since the basic interest of the authors in the database is in segment durations potential differences in different recording techniques and places only play a minor role. The authors recommend to treat possible measurements of RMS or intensity in the mini-disc and DAT recordings with care since recording levels were automatically controlled by the respective devices.
During the recording process speakers (Ss) first familiarized themselves with the text by reading it aloud while the recording levels were set. Ss were allowed to practice the text as many times as they wanted (on average four times) before the actual recording started. For the first recording they were asked to read the text in a way they considered 'normal reading' (no). After that they were asked to read the same text at different intended speech rates (isr): First Ss were asked to read it slowly (s1) and then even more slowly (s2). Following the recordings at slow rates, they were required to read the text fast (f1) and then they consecutively had to increase their reading speed until they found themselves unable to speak any faster, or until reading quality became so poor that the labeling would be impossible and recording was stopped (f2). Ss varied between three and eight attempts of the fast versions.

2.4. Labeling

Labeling of syllables as well as C-/V-intervals has been carried out by human labelers (all authors) to the normal version (no), the two slow versions (s1 & s2), the first fast version (f1) and the fastest version (f2); f2 being the version with the highest articulation rate (syllables/second) and a the lowest amount of syllable elisions (typically no more than three syllables). Syllables have been labeled as phonological syllables unless no acoustic trace of the syllable could be found (elision). A C-interval is defined as a consonant or a stretch of consonants between vowels or vowel and pause. A V-interval is a vowel or a stretch of vowels between consonants or consonant and pause.
The authors want to point out that hand labeling of the data is an important issue in the construction of databases of the type presented. Although we support strongly work on automatic labeling tools, a currently available tool to label C and V-intervals automatically have produced very poor results (especially for the fast versions) in BTC and was therefore not considered (cf. [11] for a detailed account). Alternative tools are currently under observation and may be considered to assist future labeling work.

2.5. Filing System

All five isr-versions (s2, s1, no, f1, f2) of each speaker have been saved in wav format in one file each. The file names contain information about the native language of the speaker in capital letter (e.g. 'E' for English), the language the speaker used for reading the text in small letters (e.g. 'f' for French), an abbreviation for the name of the speaker (e.g. 'Ji' for Jim), and the isr version (e.g. 's2'). Language, speaker's name, and isr information are separated by an underscore (e.g.: Ef_Ji_f1.wav = English native speaker, Jim, intending to read French fast).
The labelling work for each speaker has been saved in Praat label files of the type 'TextGrid'. For each wav file there exists one TextGrid file with the same file name but respective extension (e.g. Ef_Ji_f1.TextGrid).

3. BonnTempo-Tools (BTT)

The BonnTempo-Tools (BTT) is a collection of Praat (cf. [12]) based software (Praat scripts) to facilitate access and analysis of the BTC. An installation of the Praat speech analysis software (obtainable for free under www.praat.org) is therefore necessary in order to use BTT. The tools are independent of the corpus and are not necessarily needed to perform analysis. In BTT 1.0 the following major tools are available:

3.1. Tool: *Get Content*

This tool displays the actual content of the BTC. Since the corpus will grow in the future and users are able to add their own data to BTC (thus creating an individual version) it is useful to have a tool that informs the user about the actual content of the database (e.g. total number of languages, the number of speakers for the whole database as well as for each language, number of syllables, number of C- and V-intervals, etc). *Get Content* prints all values of the 'current state of the database in the Praat info window.

3.2. Tool: *Open Subject*

This is a tool with which the labeled files (TextGrid) and sound files (wav) of a particular speaker in the BTC can be added as an object in Praat list of objects. It facilitates a quick and easy access to individual files or groups of files of each speaker in the BTC. An interface lets the user choose the speaker, the type of file (all, wav, TextGrid) and the language (all, particular L1s, particular L2s) that is to be opened.

3.3. Tool: *Analysis*

With this tool the results for all available analysis parameters are extracted into text files of the Praat object type TableOfReal.
Amongst the analysis parameters are basically all mean values (mean), standard deviation of mean values (stdev) and the variation coefficient of the stdevs (varco) of respective segment durations (syllables, consonantal and vocalic intervals, and also pauses). The variation coefficient is defined as the percentual stdev of the mean (cf. [5] for an in depth account). In addition to theses values the percentage of V-intervals in a speech signal and the laboratory measurable speech rate (lsr) or articulation rate (syllables/second without pauses), as well as the nPVI and rPVI are available. The following list gives an overview of all analysis parameters currently available:

- Laboratory measurable speech rate (lsr)
- Mean syllable duration (meanSyllable)
- Mean duration of consonantal intervals (meanC)
- Mean duration of vocalic intervals (meanV)
- Mean pause duration (meanPause)
- Stdev of syllable duration (deltaSyllable)
- Stdev of the duration of C-intervals (deltaC)
- Stdev of the duration of V-intervals (deltaV)
- Stdev of pause durations (deltaPause)
- Varco of syllable durations (varcoSyllable)
- Varco of durations of C-intervals (varcoC)
- Varco of durations of V-intervals (varcoV)
- Varco of pause durations (varcoPause)
- Percentage of vocalic intervals (percentV)
- Normalized pairwise variability index (nPVI)
- Raw pairwise variability index (rPVI)

3.4. Tool: *Display Results*

With this tool analysis of the results can be presented in graphical form. It allows the user to display either any two of the above specified analysis parameters along a two dimensional graph or any of the analysis parameters along an axis containing the five isr-versions (s2, s1, no, f1, f2). Figure 1 (a and b) gives an example of nPVI (y-axis) displayed along rPVI (x-axis) for two languages: French (dotted line) and German (plain line). Mean values for each isr version of a speaker are consecutively connected from the slowest version (s2) to the fastest version (f2), while version s2 is accompanied by the language classification (cf. 2.5), here: Ff (French reading French) and Dd (German reading German). X and y standard deviations of the respective mean values are displayable as arrows or circles surrounding the means (cf. figure 1b). Numeric results for all x and y values are additionally printed into the Praat info window. A further option to display single speakers instead of language mean values is currently under construction.

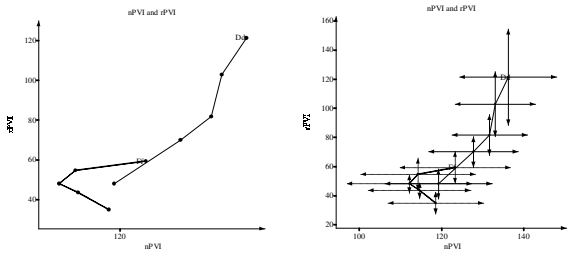


Figure 1: nPVI against rPVI for the five isr versions of German (plain line) and French (dotted line). Left: without and right: with standard deviation arrows.

4. Examples of observations based on BTC and BTT

This section gives a flavor of the type of observations that can be achieved by using BTT and BTC:
a) While most languages show a rather low variation as a function of speech rate according to %V [5, 6], Italian seems to be the first language showing high variation in this respect (cf. figure 2, centre). An explanation for this is yet to be found. Closer investigations about which particular segments are causing this are in progress.
b) Relationships between the ΔC/%V dimensions and the nPVI/rPVI are visible (cf. figure 2, top and centre) and may be much larger than described by [2], who compared the two measurements on the basis of their database and came to the conclusion that nPVI/rPVI are more robust acoustic correlates of rhythm classes than ΔC/%V.
Czech seems to be a very interesting case, showing a considerable mismatch in relation to different rhythm measures. This language clusters rather with syllable-timed languages along the nPVI/rPVI measure (figure 2, top), while it is in between the stress-timed and syllable-timed clusters along the ΔC/%V measure (figure 2, centre). However, with respect to the varcoC/%V measure, it joins clearly stress-timed languages (figure 2, bottom). This observation is coherent with [10] but is yet to be explained.
These are theoretically interesting results that will need to be explored further. A number of other results have already been

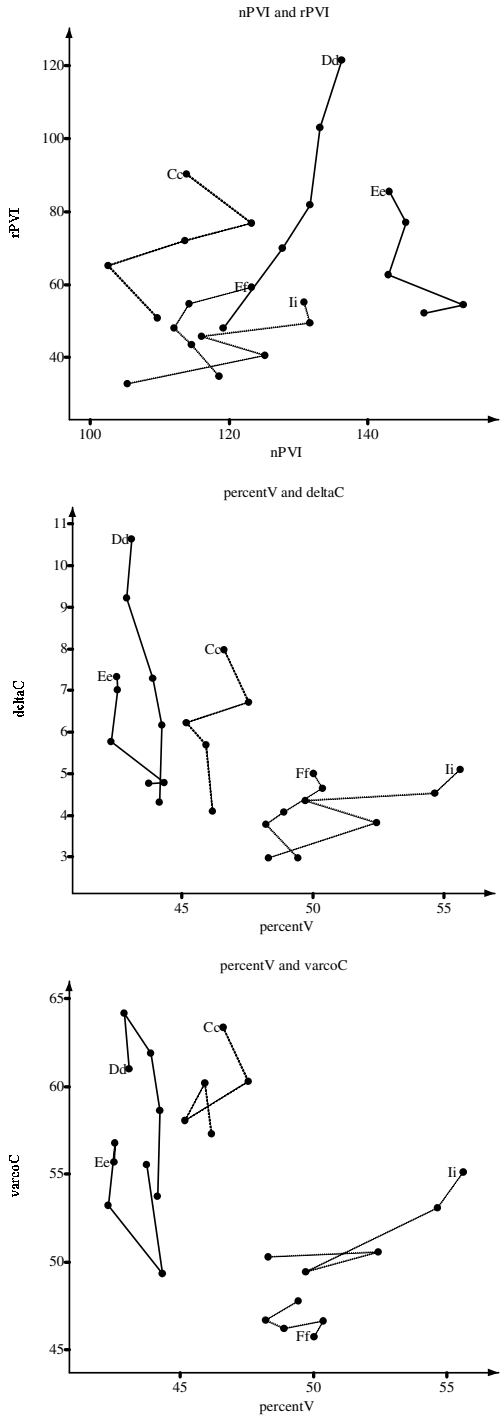


Figure 2: Stress-timed languages (English, Ee, and German, Dd, in plain lines), syllable-timed languages (French, Ff, and Italian, Ii, in dotted lines) and Czech (Cc, dashed line) along the dimensions nPVI/rPVI (top), ΔC/%V (middle), and varcoC/%V (bottom).

published, eg. initial results on the relationships between rhythm and speech rate as well as results for within- and between-language variation of lsr as a function of isr for an earlier version of the BTC in [6]. Results for varcoC of the current version of the BTC (without Czech) can be found in [5].

5. Conclusions

BTC and BTT are a powerful tool for the analysis of speech rhythm. Both corpus and tools are expected to be developed further in the future. For BTC the current number of speakers is expected to increase and to be equalized across languages (approximately 15 speakers per language). Also, a greater variety of languages will be available in the future. Suggestions for collaborations from people wishing to contribute languages to the BTC or whishing further analysis parameters to be added to the BTT will be more than welcome.

6. Acknowledgements

The authors wish to thank Judith Adrien, Stacy Dellwo, Franco Ruina and Maciej Karpinski for their translations, and to all speakers who contributed to the database with their voice.

7. References

[1] F. Ramus, Nespor, M., Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition*, 73: 265-292, 1999.

[2] Grabe, E. and Low, E. L., "Durational variability in speech and the rhythm class hypothesis", *Papers in laboratory phonology*, 7: 515-546, to appear.

[3] Roach, P., "On the distinction between 'stress-timed' and 'syllable-timed' languages", *Linguistic Controversies*, 73-79, 1982.

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

[5] Dellwo, V., "Rhythm and Speech rate: A variation coefficient for ΔC", *UCL Working Papers in Phonetics and Linguistics*, forthcoming.

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

[7] Ramus, F., "Acoustic correlates of linguistic rhythm: Perspectives", *Proceedings of speech prosody*, 115-120, 2002.

[8] Palková, Z., "*Fonetika a fonologie čestiny"*. Praha: Karolinum, 1994.

[9] Dankovičová, J., "*The linguistic basis of articulation rate variation in Czech", Forum Phoneticum* (71), Hector, Frankfurt am Main, 2001.

[10] Duběda, T., "K izoslabičnosti a izochronnosti v češtině", *Proceedings of the conference of the Czech section of ISPhS*, pp. 19-28. Faculty of Arts, Charles University, Prague, 2004.

[11] Steiner, I., "*Zur Rhythmusanalyse mittels akustischer Parameter"*, Magisterarbeit, Bonn, Institut für Kommunikationsforschung und Phonetik, 2004.

[12] Boersma, P. and Weenink, D., "Praat, a system for doing phonetics by computer", *Glot International*, 5: 341-345, 2001.

[13] Wagner, P. and Dellwo, V. "Introducing YARD and re-introducing isochrony to rhythm research", *Proceedings of Speech Prosody*, 227-230 , 2004.