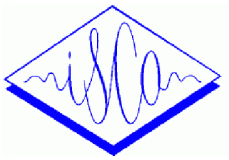
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speaker idiosyncratic rhythmic features in the speech signal

Volker Dellwo, Adrian Leemann, Marie-José Kolly

Phonetics Laboratory, University of Zurich, Switzerland
volker.dellwo@uzh.ch, {adrian.leemann, marie-jose.kolly}@pholab.uzh.ch

Abstract

Speakers' voices are to a high degree individual. In the present paper we report about an ongoing research project in which we study how temporal characteristics of human speech (e.g. segmental or prosodic timing patterns, speech rhythmic characteristics and durational patterns of voicing) contribute to speaker individuality. We report about the creation of the TEVOID-Corpus (**Te**mporal **Vo**ice **Id**iosyncrasy) that we are currently creating in our lab at Zurich University. 8 speakers producing 16 spontaneous sentences each are currently in the database which is rapidly growing. The paper gives an overview of the general ideas for the data collection and first results showing that there are significant rhythmic differences (%V, %VO, VarcoPeak) in spontaneously produced sentences between speakers of Zurich German.

Index Terms: speech rhythm, speaker idiosyncratic features, speaker identification, forensic phonetics.

10.21437/Interspeech.2012-342

1. Introduction

Voice identification has predominantly been carried out on the basis of spectral characteristics. This is particularly true in forensic applications in which the identity of voice samples is often disputed. It has been argued correctly that the amplitude spectrum, for example, is directly influenced by idiosyncratic anatomical features of a speaker's organs of speech (in particular the size of the larynx and lengths of the vocal tract cavities) which determines the range of certain spectral parameters and can thus contribute to making speakers' voices individual ([1]). The emphasis is on 'contribute' as experience has taught us that there are clear limits in identifying speakers based on spectral parameters alone. For this reason it seems plausible that the exploration of other dimensions in speech where idiosyncratic information is encoded might contribute well to speaker identification. A dimension that has been paid surprisingly little attention to in the past is 'time'. This is astonishing because research from other paradigms such as gait recognition has demonstrated convincingly that humans have highly individual ways in which they move their legs when walking and that identification of individuals based on these movements is well possible ([2]).

In the present project we argue that speech is similar to walking in that it is a highly complex brain operated series of muscle movements which may be carried out to a considerable degree in individual ways. We then go one step further and argue that such idiosyncratic motion does not need to be observed visually from the movement of the articulators but can be found in the acoustic speech signal, the immediate product of all speech articulatory movements.

For speech, strong evidence already exists on a segmental level that temporal characteristics, for example formant dynamics, can contribute strongly to speaker individuality ([3]). The present project, however, departs from segmental features and moves on to suprasegmental and prosodic characteristics of speech. Basic actions like the onset and offset of vocal fold vibration or the on and offset of vowels and consonants might as well be influenced by idiosyncratic timing of the articulators as might be the temporal distribution of syllable peaks or the temporal development of f0 rises, etc. Support for such a view can be gained from the field of speech rhythm research, where measures based on durational characteristics of vocalic and consonantal segments have been argued to be acoustic correlates of a language's rhythm class, i.e. whether a language is stress- or syllable-timed, for example ([5], [6]). Most of these measures are either based on the variability of consonantal and/or vocalic intervals (e.g. deltaC and deltaV, the standard deviation of consonantal- and vocalic intervals respectively, [5]; or rate normalized version of these measures, VarcoC: [6]; VarcoV: [7]) or are simple ratio measures between the proportion of time speech is vocalic as opposed to consonantal (%V). Other measures calculate the average difference between consecutive consonantal or vocalic intervals (Pairwise Variability Index, PVI; [5]). A detailed overview of the measures is provided in [8] and [9].

Whether such measures are strong correlates of speech rhythm and the degree to which they reveal between-language rhythmic differences is a matter of heavy debate ([10], [11], [7], [8]) as is the general question as to whether categorical rhythmic differences between languages exist at all (see discussion in [10]). This discussion, however, is irrelevant in the present context. Of relevance is the fact that there is increasing evidence revealing that rhythm measures like %V, deltaV or the PVI may vary significantly between speakers ([11], [12], [8], [13]). Based on such exemplary evidence we are currently developing a large database with which we are aiming at…

- … systematically analysing prosodic durational characteristics that vary most across speakers of a homogeneous speaker group (Zurich German) and explain the reasons for temporal variability between speakers.
- … test how robust such characteristics are towards sources of within-speaker variability (e.g. spontaneous vs. read speech or forms of voice disguise) and between speaker similarity (e.g. speakers imitating each other).

In the following we will present an overview of the ongoing database construction process and we will give an example of the type of acoustically measureable rhythmic differences between speakers that can be found in the data that is available to date.

2. The TEVOID Corpus

TEVOID stands for '**Te**mporal **Vo**ice **I**diosyncrasy' as the database this is particularly designed to study speech temporal variability across a highly homogeneous group of speakers (the

only independent variable we introduce is gender). All speakers are fluent native speakers of the same language variety (as in regional and social) and the same age group (between 20 and 30). Zurich German has been chosen for this as hardly any sociolinguistic variability is obtainable between speakers of that dialectal variety.

When typical sources of between speaker rhythmic variability like language, dialect or accent are not present, there are mainly two factors that may introduce measurable temporal differences between speakers: (a) an idiolectal use of language and (b) idiosyncratic ways to control articulatory movements in speech. The idiolectal variability may arise when speakers have phonological preferences resulting from an individual use of sounds, syllables, words or grammatical patterns, for example (speakers might differ in building sentences that consist of words with a high proportion of complex consonant clusters and thus be able to produce speech in which they spend less time at vowels but more at consonants). Such differences, however, should predominantly arise in spontaneous speech but not in read speech as speakers are bound to produce roughly the same words with roughly the same segmental content. Individual use of segment elision, for example, can easily be overcome in that only speech material is compared between speakers that produce the exact same segmental content.

2.1. Speakers and speech material

To study phonologically idiolectal and articulatory movement variability it is vital that the database consists of spontaneously produced and read speech. In total we are looking at a number of 40 speakers for the database of which 16 have been recorded and 8 have been annotated with temporal information (see below). Speakers are being recorded in a sound treated booth in our lab at Zurich University (44.1k samples/second, 16 bit). Only spontaneously produced speech of 16 speakers has been recorded so far from interview situations. 16 grammatically well-formed sentences of between 10 and 45 syllables each were extracted from the interviews of each speaker. Theses sentences were annotated as described in 2.3.

To allow a comparison between spontaneously produced and read speech we will have all speakers read the 16 spontaneously produced sentences. This will allow direct comparisons between spontaneous and read speech within and between speakers for individual utterances.

In particular, with respect to forensic phonetic applications we are interested in sources of between- and within-speaker rhythmic variability. One potential source of within-speaker variability has already been introduced by recording spontaneous and read speech. Other sources will be collected in the future by having subsets of speakers disguise their voices. To study between-speaker similarity effects we will have another subset of speakers imitate each other's voice.

2.2. Annotation for rhythmic analysis

Segment annotation is performed by human labellers to guarantee maximum correctness. The first 8 subjects were annotated by two human labellers (second and third authors). Annotation is carried out using standard SAMPA notation. Figure 1 shows a screenshot of an annotated sentence. Tier one contains the manually annotated speech segments.

To calculate temporal characteristics of vocalic and consonantal intervals, The segment tier is the basis for three other tiers containing (a) the information whether a segment is consonantal or vocalic (tier 2), (b) consecutive consonantal or vocalic segments combined to consonantal or vocalic intervals respectively (tier 3 and 4; tier 3 additionally contains the information about the number of consonantal or vocalic segments within a consonantal or vocalic interval). Tiers 2 to 4 are created automatically from tier 1. There will further be a tier with syllabic durational information.

Next to segmental durational information we find that other sources of suprasegmental characteristics might contribute to temporal differences between speakers, in particular with respect to speech rhythm. Such characteristics may be related to the temporal use of voiced as opposed to unvoiced parts of the signal ([14]) or to pulsing produced by the amplitude envelope. To study these features we use automatic annotation of voiced-

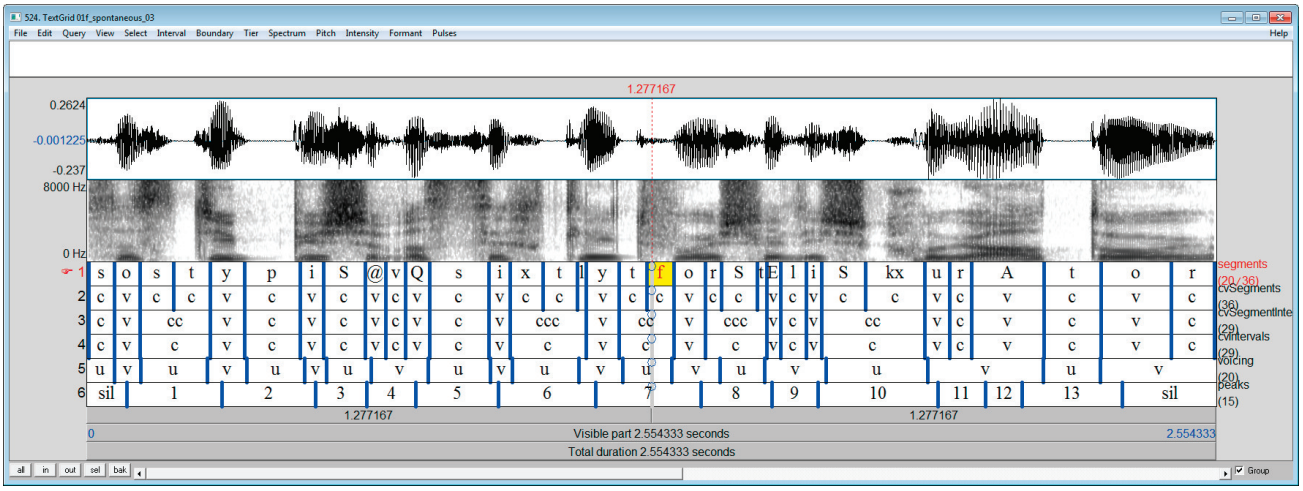


Figure 1: *The labeling tiers in the TEVOID-Corpus. Tier 1 contains manually labeled segmental information from which the consonantal and vocalic interval information from Tier 2 to 4 is derived automatically. Further derived automatically were tiers 5 (voiced unvoiced intervals) and 6 (amplitude peak points).*

unvoiced intervals in speech (tier 5) and the intervals between automatically detected peaks (tier 6, to a large degree syllabic peaks) in the amplitude envelope.

2.3. Measurement techniques

To address the question of rhythmic variability, we are using a wide variety of measurement techniques of durational variability of consonantal and vocalic intervals, voiced and unvoiced intervals or peak-to-peak intervals. Overviews of such measurement techniques are given in [9], [8] and [15].

For the present paper we give an example of temporal variability using:

- Two C:V ratio measures: (a) the percentage over which speech is vocalic (%V; [4]) and (b) the percentage over which speech is voiced (%VO; [14]).
- Two vocalic interval variability measures: (a) the rate normalized standard deviation of vocalic intervals (VarcoV; [7]) and (b) the average differences between consecutive vocalic intervals (nPVI-V; [5]).
- Two consonantal interval variability measures: (a) the rate normalized standard deviation of consonantal intervals VarcoC ([6]) and (b) the average differences between consecutive consonantal intervals (rPVI-c; [5]).
- A measure of peak-to-peak interval variability, the coefficient of variation of peak-to-peak interval durations (VarcoPeak).

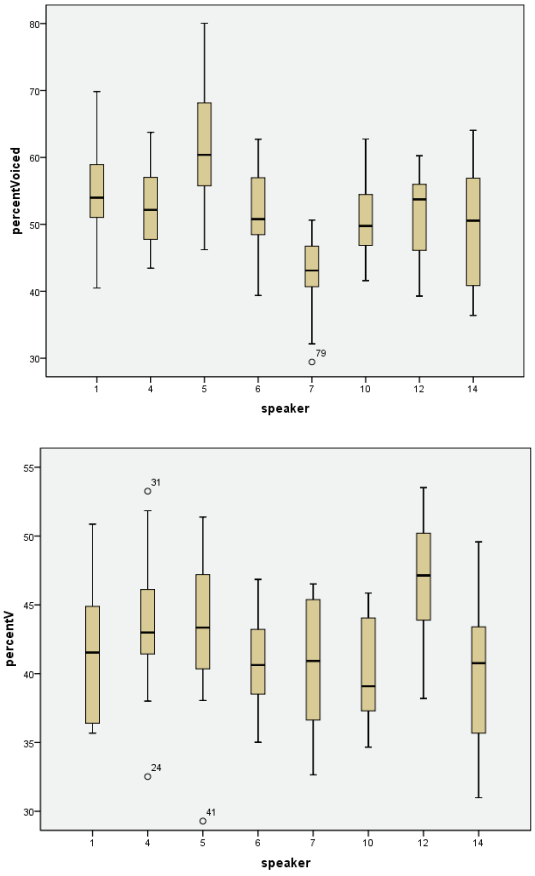


Figure 2: *Box-plots of %V (bottom) and %VO (top) for 8 speakers in the TEVOID-Corpus.*

3. Results

Figure 2 contains two box-plots showing the distributions of %V (top) %VO (bottom) for the first eight speakers in the TEVOID-Corpus. The plots reveal that there is considerable variability between speakers in both %V and %VO. An ANOVA with speakers as a factor shows that both effects are highly significant (%V: $F[7, 120]=3.86$, $p<0.001$; %VO: $F[7, 120]=9.42$, $p<0.001$). The descriptive results in the box-plots also show that the two dependent variables are not necessarily correlated. Speakers with a high %V (e.g. speaker 12) do not necessarily possess a high %VO compared to other speakers. This point is supported by the cross-plot between the two independent variables in Figure 3 where no close relationship is observable. Linear regression shows a poor but highly significant relationship ($R2=0.16$; $F[1,126]=24.8$; $p<0.001$).

The consonantal and vocalic variability measures were found to be much less variable across the eight speakers. The only significant effect we could obtain was for the average difference between consecutive vocalic intervals, nPVI-V ($F[7,120]=2,37$; $p=0.026$). The standard deviation of vocalic intervals, VarcoV, was marginally significant ($F[7,120]=2.1$; $p=0.048$). No effects could be obtained for the consonantal variability (VarcoC, $F=0.7$. $p=0.67$; rPVIC, $F= 0.98$, $p=0.46$).

The peak to peak Interval measure, VarcoPeak, revealed a highly significant effect for between speakers variability ($F[7, 120]=3.1$, $p=0.004$).

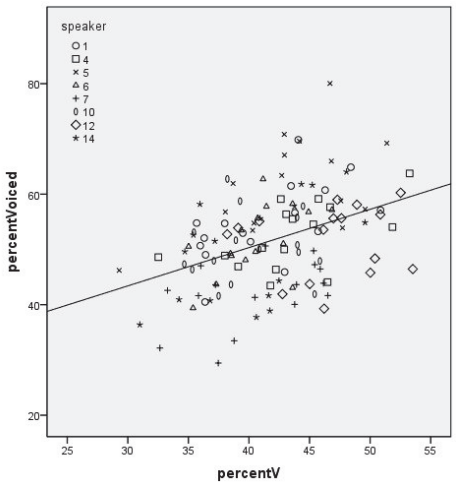


Figure 3: *Scatter-plot of %VO as a function of %V with a linear regression line superimposed for eight speakers in the TEVOID-Corpus.*

4. Discussion and Conclusions

In this paper we presented first results from 8 speakers of the TEVOID-Copus. Using spontaneously elicited interview speech we found that acoustic measures of speech rhythm based on temporal characteristics of speech intervals can reveal highly significant differences between speakers. We found that ratio measures between the overall durations of consonantal and vocalic or voiced and voiceless interval as well the variability of peak-to-peak intervals derived from the amplitude envelope of

speech revealed significant differences between speakers. Small effects were found for vocalic variability; no effects were obtainable for consonantal variability.

This initial result suggests that vocalic durational information is a more vital source of durational information between speakers than consonantal intervals. This suggestion seems plausible as consonants are short and speakers may not have as much articulatory flexibility than they have in the production of vowels. However, there is drastic within-class variability for consonantal intervals. An interval consisting of a plosive may be much more restricted in articulatory flexibility than an interval consisting of a nasal or a semi-vowel. It may also be that different types of consonants reveal very different results. For this reason it seems inevitable to distinguish between consonant types and possibly even between individual consonants when studying their durational variability characteristics between speakers.

As for the ratio measures it seems obvious that vocalic and voiced durations are used in different ways by speakers. The time speakers spend on producing voiced or vocalic intervals varies between them. Furthermore, the two variables do not correlate strongly which means that speakers who spend a relatively large proportion of their articulation time on voicing do not necessarily spend this time on vocalic proportions of speech. This should mean that the proportion of time speakers spend on voiced consonants must be relatively higher. An encouraging result was also obtained with the VarcoPeak measure, which shows that intervals between syllabic peaks show different degrees of variability between speakers.

We find it very possible that the time speakers spend on voiced, vocalic and/or voiced consonantal intervals and the time a speaker needs to arrive from one syllabic peak to the next may perhaps be related to individual ways of articulation. Such individual ways could result from the fact that speakers have individual anatomic characteristics of their organs of speech, which means that different distances need to be overcome to put the articulators into their respective positions. This may very well influence the time that speakers have between voiced and voiceless, vocalic and consonantal or voiced and unvoiced consonantal intervals, for example. It may also have an influence on the time that speakers need to arrive from one syllabic peak to the next. Such individual temporal features on the other hand could also well be related to individual ways of using the articulators that are simply based on habitual learning or a certain idiosyncratic pronunciation style that speakers have acquired. The present result needs to be especially interpreted in light of the fact that all speech was produced spontaneously and no equal sentences exist between speakers. It is therefore well possible that an individual use of syllables, words or grammatical patterns influenced the results (see discussion above) that is not necessarily related to individual organs of speech anatomy. The database is steadily growing and with read speech being available that contains comparable phonological content with comparable phonotactic sequencing we trust that we can provide answers to these questions in the near future.

5. Acknowledgements

We wish to thank Anders Erikkson for useful comments on earlier presentations of this research. This research is supported by the Swiss Science Foundation (grant number: 100015_135287).

6. References

[1] Dellwo, V., Huckvale, M., Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production & description for speaker classification. In: Müller, C. (Ed.). Speaker Classification I. Berlin: Springer Verlag, 1-20.

[2] Nixon, M. S. (2008) Automated human recognition by gait using neural network. In: First Workshops on Image Processing Theory, Tools and Applications (IPTA).

[3] McDougall, K. (2006) Dynamic Features of Speech and the Characterisation of Speakers: Towards a New Approach Using Formant Frequencies. In: International Journal of Speech, Language and the Law (13, 1), 89-126.

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

[5] Grabe, E. and Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In: C. Gussenhoven and N. Warner (eds.) Papers in Laboratory Phonology 7, Berlin, New York: Mouton de Gruyter.

[6] Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for deltaC. In: Karnowski, P., Szigeti, I. (Eds.). Language and language-processing. Frankfurt am Main: Peter Lang, 231-241.

[7] White, L., & Mattys, S.L. (2007). Calibrating rhythm: First language and second language studies. In: *Journal of Phonetics* (35) 501-522.

[8] Dellwo, v. (2010) Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. PhD-Dissertation, Universität Bonn (electronic publication: http://hss.ulb.uni-bonn.de:90/2010/2003/2003.htm).

[9] Loukina, Anastassia, Greg Kochanski, Burton Rosner, Elinor Keane, and Chilin Shih (2011). Rhythm measures and dimensions of durational variation in speech. Journal of the Acoustical Society of America 129(5), 3258-3270.

[10] Arvaniti, A. (2009) Rhythm, Timing and the Timing of Rhythm. In: Phonetica (66,1-2), 46-63.

[11] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. (2010) How stable are acoustic metrics of contrastive speech rhythm? In: Journal of the Acoustical Society of America (127,3), 1559-1569.

[12] Yoon, T.J. (2010) Capturing inter-speaker invariance using statistical measures of speech rhythm. In: Electronic proceedings of Speech Prosody, Chicago/IL, USA.

[13] Dellwo, V. and Koreman, J. (2008) How speaker idiosyncratic is measurable speech rhythm? Abstract presented at the annual IAFPA meeting 2008, Lausanne/Switzerland. Attached to this application as IAFPA2008_DellwoKoreman.pdf.

[14] Dellwo, V., Fourcin, A., Abberton, E. (2007). Rhythmical classification based on voice parameters. In: International Conference of Phonetic Sciences (ICPhS), Saarbrücken/Germany, 1129-1132

[15] Dellwo, V. (2009) Choosing the right rate normalization methods for measurements of speech rhythm. In: Proceedings of AIVS.