Speaker identification based on temporal information: A forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German

Volker Dellwo, Marie-José Kolly, Adrian Leemann
Phonetics Lab, University of Zurich, Switzerland
(volker.dellwo@uzh.ch)

Speakers’ voices are to a high degree individual. By now there is a considerable body of evidence revealing that acoustically measurable speech rhythmic characteristics can vary considerably between speakers (Wiged et al., 2010, Yoon, 2011) and might thus contribute to the individuality of speakers’ voices. At previous IAFPA conferences we contributed to this discussion, reporting that some parameters of acoustically measurable speech rhythm (in particular the percentage over which speech is vocalic; %V) may vary highly between speakers even when within speaker variability of speech rate is at an extreme (Dellwo & Koremann, 2008) and that such information is not easily disguised by accent or fundamental frequency levelling (Dellwo, Ramyead & Dancovicova, 2009). Based on these previous studies our lab in Zurich has now received funding from the Swiss Science Foundation to investigate speech temporal information for speaker identification in more detail. In the present abstract we give an overview about the project ideas and report first results on spontaneously elicited speech.

We start from the observation that the acoustic speech signal is made up of dynamic processes resulting from the movements of the articulators. Other scientific domains showed that humans can be identified on the basis of their movements only, e.g. by the way they walk. Our working hypothesis is that the movements of the organs of speech (e.g. jaw, lips or tongue) can be equally idiosyncratic as human gait and that idiosyncratic ways to control the organs of speech result in individual temporal characteristics in the acoustic speech signal. We will therefore study numerous durational parameters in speech from segment durations (e.g. the durations of consonants and vowels) over syllable and word to prosodic durations (e.g. durational characteristics of intonation) across a number of speakers from a homogeneous speaker group. The Zurich variety of Swiss German has been chosen for this not only for convenience reasons (speakers are easily recruitable as our lab is in Zurich) but also because this German variety shows little to no speech variability as a function of socio-economic status. For the entire project we plan the following steps.

- Find the most idiosyncratic temporal parameters (or combinations of temporal parameters) in the speech signal.
- Test how robust these idiosyncratic temporal parameters are towards influences of within-speaker variability (e.g. voice-disguise) and between-speaker similarity (e.g. voice imitation)
- Test whether the most speaker idiosyncratic temporal parameters are perceptually salient.

It is well possible that we will find that some temporal speaker idiosyncratic features are perceptually salient and others are not. We argue that the salient temporal features will help us explaining how human listeners identify speakers on the basis of their voice. Non-salient features, however, may be less prone to within-speaker variability like voice disguise as these features should be difficult to control for speakers. Such features may thus be highly suitable for acoustic voice identification of non-cooperative speakers (i.e. speakers not wishing to be identified) typically found under forensic circumstances.

At the present point we have collected spontaneously elicited speech from interview situations from 16 speakers of Zurich German (about 30 min each). From the interviews we have selected 16 sentences per speaker (256 sentences in total of about 15 syllables on average) that are grammatically well formed to the extent that speakers would be able to read these utterances back from an orthographic transcription. All 256 sentences for each speaker have
been segmentally annotated manually by the second and third authors. During the next step we will collect read speech data for which all 16 speakers will be recorded reading the 256 sentences each. This will allow us to draw direct comparisons of temporal phenomena between utterances of identical linguistic content across spontaneous and read speech, within- and between different speakers.

Results thus far have been calculated for the spontaneously elicited speech of the 16 speakers for three temporal measures extracting different types of temporal information from the signal:

- The percentage over which speech is vocalic (%V, Ramus et al., 1999), calculated based on the manually attributed segment information.
- The percentage over which speech is voiced (%VO, Dellwo et al., 2007), calculated automatically using Praat pitch tracking algorithm (www.praat.org).
- The standard deviation of peak-to-peak intervals in the amplitude envelope of speech (deltaPeak). This measure has been newly developed in our lab to test an automatic method to gain temporal information from the amplitude envelope of speech.

We found that all three measures show highly significant variability across the 16 speakers (ANOVA: %V: \( f[15, 255] = 4.71, p<0.001 \); %VO: \( f[15, 255] = 9.1, p<0.001 \), deltaPeak: \( f[15,255] = 5.24, p<0.001 \)). The distributions for the automatically derived measures for each speaker can be viewed in Figure 1 (left graph: deltaPeak; right graph: %VO). Results thus far reveal that there is considerable temporal variability at different levels (segmental, voicing patterns and amplitude envelope) even when sentences are spontaneously elicited and utterances do not share the same linguistic content. These preliminary results of the project provide further evidence for the potential of temporal information in speaker identification.

REFERENCES


