Big Data - The use of crowdsourced data for forensic phonetics

Ingrid Hove\textsuperscript{1} and Adrian Leemann\textsuperscript{2}

\textsuperscript{1}Phonetics Laboratory, Dep. of Comparative Linguistics, University of Zurich, Switzerland  
-ingrid.hove@uzh.ch

\textsuperscript{2}Phonetics Laboratory, Dep. of Theoretical and Applied Linguistics, University of  
Cambridge, UK  
al764@cam.ac.uk

Population statistics are crucial for forensic voice comparison since one needs to know how features are distributed in the population. Until now, most population statistics are based on data of around 50-100 speakers (cf. e.g. Künzel, Masthoff & Köster, 1995; Jessen, 2007).

In January 2015 the smartphone application Voice\textsuperscript{App} was launched. It provides its users with information on their dialect and their voice – pitch and speaking rate – as well as on general aspects of speech. Today we aim to examine whether the claim we made at the IAFPA conference in 2014, namely that the database gathered through this application could be valuable for forensics, holds true.

Known disadvantages of crowdsourced linguistic data are that the truthfulness of the self-declarations cannot be verified, that the recording condition cannot be controlled, and that there are no instructions concerning the speaking style, to name just a few. The advantage of crowdsourced data is its sheer volume – the 18,000 downloads in the first month allow us to expect a large number of recordings. Currently the database exceeds 35GB of audio data. The question which we ask is whether this mass of data can compensate for errors due to wrong, imprecise or missing data (cf. Elspaß & Möller, 2006: 147).

We intend to present

- statistics of the data gathered in the first 6 months after the launch of the application: number of recordings, distribution according to age, sex and dialect
- statistics concerning the quality of the data: proportion of useful recordings and reasons why certain recordings are not useful (e.g. background noise, non cooperative speaker etc.)
- population statistics concerning acoustic measures which can be extracted without manual segmentation such as f0 and speech rate, which have already been calculated natively in the app

By testing which kinds of analyses are possible and make sense in a large database and how much work remains to be done manually we aim at showing the advantages and disadvantages of crowdsourced data as opposed to databases containing highly controlled recordings.

References


67.