Test speaker sample size: speaker modelling for acoustic-phonetic features in conversational speech

Heeren, W.F.L.

Test speaker sample size: speaker modelling for acoustic-phonetic features in conversational speech

Willemijn Heeren
Leiden University Centre for Linguistics, Leiden University, The Netherlands
w.f.l.heeren@hum.leidenuniv.nl

Speech samples offered for forensic speaker comparisons may be short. A relevant question then is whether the sample will hold a sufficient number of tokens of the acoustic-phonetic features of interest so that the features' strength of evidence may be estimated reliably. The current study contributes to answering this question by investigating the effects of test speaker sample size on speaker modelling and on LR system performance.

To calculate a feature's strength of evidence, an algorithm is used that models within-speaker variance using a normal distribution, whereas between-speaker variance is estimated using multivariate kernel density (Aitken and Lucy, 2004). Also, in an LR system, a development set, a reference set and a test set of speakers are used to first compute calibration parameters from the development and reference sets, and then evaluate feature and system performance on the test set. The by-speaker sample sizes of these different data sets affect system performance. Kinoshita & Ishihara (2012) investigated Japanese [e:], represented as MFCCs, and reported sample size effects for both test and reference data, with samples varying from 2 to 10 tokens per speaker. The effect seemed stronger in the test than the background set. A later study using Monte Carlo simulations (Ishihara, 2013) reported that system validity improved with sample size. Hughes (2014) investigated the number of tokens per reference speaker using between 2 and 10 (for /aɪ/) or 13 (for /uː/) tokens. Results showed that relatively stable system behavior was found from 6 tokens on. The author furthermore remarks that: "*…considerably more than 13 tokens may be required to precisely model within-speaker variation, at least for these variables*" (Hughes, 2014, p. 213). This remark is especially relevant for test speaker modelling.

In the current study, the earlier work on test speaker sample size was extended by including larger numbers of naturally produced tokens, and by including various speech sounds' acoustic-phonetic features. The goal was to assess at which sample sizes (i.e. numbers of tokens included) test speaker behavior is modelled reliably and LR system output is valid. It is expected that larger sample sizes are needed in cases of more variability; this is predicted for segments that are more strongly affected by co-articulation and for features that are less stable within a speech sound or across instances of a sound.

Method

Using spontaneous telephone conversations from the Spoken Dutch Corpus (Oostdijk, 2000), tokens of [a:], [e:], [n] and [s] were manually segmented from 63 ([a:, e:]), 57 ([n]), or 55 ([s]) male adult speakers of Standard Dutch (aged 18-50 years). Per speech sound and speaker, median numbers of 40-60 tokens of each speech sound were available, with minimum numbers at 30-32. For each of the speech sounds, and for multiple acoustic-phonetic features per sound, test speaker sample size was assessed in two ways.

First, the stabilization of each feature's mean and standard deviation by sample size was examined. Up to 10 tokens, sample sizes were increased by 2, and from 10 on in steps of 5 tokens. Tokens were always sampled in sequence, thus simulating shorter versus longer recordings. Second, same-speaker and different-speaker LRs (LRss, LRds) as well as LR system performance were computed as a function of sample size. For the vowels, the available speakers were distributed over the development, reference and test sets. For the consonants, LRs were determined using a leave-one-out method for score computation and for calibration. A MATLAB implementation (Morrison, 2009) of the Aitken & Lucy (2004) algorithm was used for the computation of calibrated LRs. Sample size was varied for the test set only, increasing the number of tokens from 2 to 20, in steps of 2. If data allowed, multiple repetitions of the same token set size were included. For within-speaker comparisons, first versus second halves of the speaker data were used. System performance was evaluated the R package *sretools* (Van Leeuwen, 2008).

Example results

Various acoustic-phonetic features and feature combinations were assessed, in the above-mentioned ways, from [a:, e:, n, s]. As an example, Figure 1 gives results for [a:]'s second formant (F2). Estimates of the mean and standard deviation seem to stabilize from 10-20 tokens on. LLRss and LLRds show increasing separation with sample size, with mean LLRds falling below zero from 10 test speaker tokens on. Validity slowly improves with sample size for [a:]'s F2.

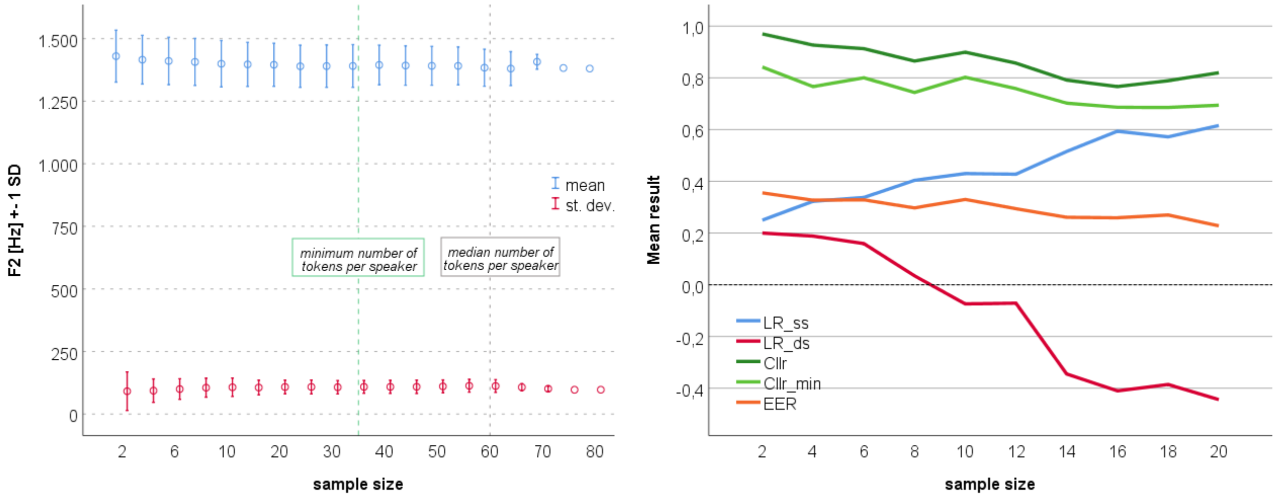

Figure 1 *Left*: Error bar plot showing stabilization of F2 mean and standard deviation by sample size (bar shows ± 1 SD). The minimum and median numbers of tokens by speaker in the dataset are indicated. *Right*: Line plot showing means of log-LRss, log-LRds, cllr, cllr$_{\text{min}}$ and EER (as a proportion), by sample size. Note that the vertical axis represents different measurement units that use a similar scale.

Acknowledgement: This work is being supported by an NWO VIDI grant (276-75-010).

References

Aitken, C. G. G. and D. Lucy. (2004) Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53:4, 109–122.

Hughes, V. S. (2014) *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. University of York: PhD dissertation.

Ishihara, S. (2013) The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations. *In Proceedings of Australasian Language Technology Association Workshop*, 25–33.

Kinoshita, Y. and S. Ishihara. (2012) The effect of sample size on the performance of likelihood ratio based forensic voice comparison. *In Proceedings of the 14th Australasian International Conference on Speech Science and Technology (Vol. 3, No. 6).*

Morrison, G. S. (2009) "train_llr_fusion_robust.m", https://geoffmorrison.net/#TrainFus (Last viewed 28-11-2019).

Oostdijk, N. H. J. (2000) Het Corpus Gesproken Nederlands [The Spoken Dutch corpus]. *Nederlandse Taalkunde*, 5, 280–284.

Van Leeuwen, D. A. (2008) SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations. http://sretools.googlepages.com (Last viewed 2-3-2020).