Optimizing the strength of evidence: combining segmental speech features

Heeren, W.F.L.; Smorenburg, B.J.L.; Gold, E.

Optimizing the strength of evidence: Combining segmental speech features

Willemijn Heeren[1], *Laura Smorenburg*[1], *and Erica Gold*[2]
[1]*Leiden University Centre for Linguistics, Leiden University, The Netherlands*
{w.f.l.heeren, b.j.l.smorenburg}@hum.leidenuniv.nl
[2]*California State University San Marcos, California, USA.*
egold@csusm.edu

In research, speaker specificity is often investigated at the level of individual speech sounds. In casework, however, conclusions are drawn by evaluating multiple features (e.g., Gold & French, 2011). Gold & Hughes (2015) compared several ways of combining different acoustic-phonetic features into one overall likelihood ratio (LR). They argued for the evaluation of correlations between speech features prior to combining evidence from various phonetic features. The current study considers segmental correlations prior to combining various Dutch speech sounds into a joint strength of evidence. It is expected that the combination of different speech sounds will support stronger conclusions by an LR system with higher validity.

For the current study, spontaneous conversational telephone speech from adult male speakers was used. In the first phase of the study correlations between speech features from the same sounds and across all six speech sounds were computed. In the second phase, an overall LR was computed, taking the correlations into account.

Method

Landline telephone data (300-3400 Hz) were taken from Heeren (2020, [a:, e:]), Smorenburg & Heeren (2020, [s, x]), and Smorenburg & Heeren (2021, [m, n]). Per speech segment, two well-performing acoustic-phonetic features from each segment were selected: F2 and F3 for the vowels, N2 and N3 for the nasals, and CoG and spectral standard deviation for the fricatives. The same set of 60 speakers contributed 20 tokens per speech sound[1].

Correlations between features within a speech sound and across speech sounds were computed using distance correlations implemented in the *R* package *Energy* (https://cran.r-project.org/web/packages/energy/) to assess non-linear relationships and Pearson's *r* to assess a linear relationship.

An overall LR was computed by developing separate MVKD LR systems (Aitken & Lucy, 2004) for non-correlating features using the MATLAB implementation by Morrison (2007), and by then multiplying the resulting LRs per system. The 60 speakers were randomly divided into equally-sized groups for development, reference, and test data, which was repeated 10 times per system (using fixed grouping per repetition). After score-to-LR conversion, ELUB limiting with 1 CMLR (Vergeer et al., 2016) was applied to each of the intermediate results, and also to the overall LRs after multiplication. Each system's validity was evaluated by computing Cllr, Cllr_{min} and EER (Brümmer & Du Preez, 2006, in Van Leeuwen, 2008).

Results

Within speech sounds and within speech sound classes (vowels, nasals, fricatives), significant correlations between speech features (e.g. F2, F3) were found ($r >= .33$, $p <= .016$). The only significant correlations between speech sounds from different classes were found between N2 of nasal [m] and the vowel [a:]'s formants F2 ($r = .42$, $p = .001$) and F3 ($r = .45$, $p = .002$). Given this result, two LR systems were built: one for vowels+nasals and one for fricatives.

The LR results are summarized in Table 1, presenting medians and interquartile ranges across 10 repetitions per system.

[1] One speaker had only 17 [s] tokens.

	[a:] + [e:] + [m] + [n]	[s] + [x]	*combined*
$LLR_{same\text{-}speaker}$	1.15 [1.0, 1.3]	0.55 [0.44, 0.55]	1.45 [1.40, 1.45]
$LLR_{different\text{-}speaker}$	−1.25 [−1.25, −0.95]	−0.50 [−0.61, −0.35]	−1.10 [−1.25, −1.10]
Cllr	0.53 [0.49,0.56]	0.84 [0.82, 0.86]	0.51 [0.47, 0.53]
$Cllr_{min}$	0.28 [0.24, 0.36]	0.71 [0.66, 0.79]	0.22 [0.18, 0.27]
EER (%)	9.11 [7.44, 9.52]	25.08 [23.31, 26.55]	6.28 [5.06, 7.70]

Table 1. Results of the LR analysis, per system and for the combined result.

Discussion

Results show that an acoustic-phonetic system for Dutch may perform well using features from just six, carefully-selected segments. Without limiting, comparable results were obtained for combined features in English (formants, F0, articulation rate, Gold, 2014). Even though Dutch [s, x] may not appear to be a strong system on its own, when combined with [a:, e:, n, m] it is very helpful in increasing strength of evidence and validity. Accounting for the correlations between and within features allows us to avoid miscarriages of justice that would traditionally over-estimate strength of evidence. The results in this study show that accounting for correlations within and between just six phonetic parameters provides appropriate same-speaker and different-speaker strengths of evidence. We also have a respectable EER for the system and the overall Cllr is not too high.

References

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. of the Royal Stat. Soc. Series C: Applied Statistics*, 53(1), 109–122.

Brümmer, N., a& Du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3), 230–275.

Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD Dissertation, University of York.

Gold, E. & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18(2), 293–307.

Gold, E., & Hughes, V. (2015) Frontend approaches to the issue of correlations in forensic speaker comparison. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Heeren, W.F.L. (2020). The Effect of Word Class on Speaker-dependent Information in the Standard Dutch Vowel /aː/. *Journal of the Acoustical Society of America*, 148(4), 2028–2039.

Smorenburg, B.J.L., & Heeren, W.F.L. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America*, 147(2), 949–960.

Smorenburg, B.J.L. & Heeren, W.F.L. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *Journal of the Acoustical Society of America*, 150(2), 979–989.

Morrison, G.S. (2007). *Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation*. [software]

Van Leeuwen, D. A. (2008). SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations. Downloaded from http://sretools.googlepages.com/. [software]

Vergeer, P., Van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? *Science & Justice*, 56(6), 482–491.