System performance and speaker individuality in LR-based forensic voice comparison

Bruce X. Wang, Vincent Hughes and Paul Foulkes
{xw961, vincent.hughes, paul.foulkes}@york.ac.uk
Department of language and linguistic science, University of York

1. Introduction

The speaker-discriminatory power of segmental phonetic features has been explored in numerous studies within the likelihood-ratio (LR) framework (Hughes et al., 2016; Morrison, 2009; Rose & Wang, 2016; Zhang et al., 2011). Many studies use vowel formants to test overall system validity. In general, combining more segmental phonetic features (e.g. multiple vowels) or parameters (e.g. multiple formants) improves system validity. However, few studies have explored system stability using different phonetic parameters, and how individual speakers are affected when different systems are used (but see Morrison et al. 2011; Wang et al., 2019). The current study addresses these issues by considering the performance of systems and the effects on individual speakers (i.e. validity and stability).

2. Methods

Filled pauses (FPs) (*um*) from 90 SSBE speakers (DyViS corpus, Nolan et al., 2009) were used. Quadratic polynomials were fitted to the F0, F1, F2 and F3 trajectories of the vowel portion of *um.* The coefficients and both vocalic and nasal durations were used for LR comparisons. Five systems were tested: F1, F2, F3, vocalic/nasal durations, and the combination of all four. In each system, 25 speakers were randomly sampled into the test, training and reference sets. The multivariate kernel density formula (MVKD, Aitken & Lucy, 2004) was used to compare the same-speaker (SS) and different-speaker (DS) pairs of test and training data to produce a series of test and training scores. The training scores were then used to build a logistic regression model (Brümmer et al., 2007; Morrison, 2011) that was applied to test scores to produce calibrated Log_{10} LRs (LLR). The experiment was replicated 100 times, sampling training and reference speakers. This provides insight into system stability, because training and reference speakers were treated as the system, and test speakers were used as suspect and offender samples. Using the same test speakers across 100 replications enables us to explore the effect of different systems (i.e. different configurations of training and reference speakers) on individual speakers. System performance was evaluated using the log LR cost function (C_{llr}, Brümmer & du Preez, 2006), while results for individual speakers were assessed using mean LLRs with a modified zoo plot (Doddington et al., 1998) and root-mean-square error (RMSE). Animal groups, i.e. *chameleons, phantoms, doves,* and *worms*, adapted from Dunstone and Yager (2009) were used in the zoo plot. Instead of upper and lower quartiles of scores that are traditionally used in ASR, the zoo plot thresholds were adjusted based on the LLR verbal expression (Champod and Evett, 2000; Table 1). This is because calibrated LLRs were used for zoo plots, and LLRs are comparable between speakers and across systems, which are different from comparison scores used in ASR systems.

Animal group	SS LLR	DS LLR
Phantoms	≤ 0	≤ -1
Worms	≤ 0	≥ -1
Doves	≥ 1	≤ -1
Chameleons	≥ 1	≥ 0

Table 1. LLR threshold for animal groups

3. Results

Figure 1 shows the C_{llr} range across 100 replications in five systems. Combining all parameters produces the best validity, with the lowest C_{llr} across replications of 0.14. This is followed by F2 (min. C_{llr} = 0.37), F3 (min. C_{llr} = 0.60), F0 (min. C_{llr} = 0.65), F1 (min. C_{llr} = 0.69) and duration (min. C_{llr} = 0.72). The F1 system yields the largest overall range (OR = 0.47), followed by the combined system (C_{llr} OR = 0.17), F3 (C_{llr} OR = 0.12), Duration (C_{llr} OR = 0.13) and F2 (C_{llr} OR = 0.04). Figure 2 reveals a lack of *worm* and *chameleon* groups across all systems. Speakers in general yield better performance with more parameters, as the majority of speakers (18 out of 25) shift to the *dove* group (right top) in the combined system. Speakers' performance varies across single parameter systems; speakers 72, 54 and 114 fall in the *phantom* group in the Duration and F1 systems, while they shift to the *dove* group in the combined

system. However, speaker 120 is classified in the *phantom* group in the F2 system but does not shift to the *dove* group in the combined system. Some other speakers, e.g. 48 and 53, can be well-separated using F2 alone, but yield misleading LLRs in the F1 and F3 systems. Moreover, the patterns of speakers 48 and 53 show that using combined parameters does not necessarily contribute to the magnitude of the strength of evidence. For SS comparisons, all speakers tend to fluctuate most in the combined system and least in single parameter systems. Meanwhile, the majority of speakers show little fluctuation (RMSE values vary between ca. 0.1-0.2) across single parameter systems (e.g. speakers 8, 13, 20). For DS comparisons, speakers fluctuate much more in their DS LLRs across different systems. This is due to the fact that between-speaker variance is likely to be larger than within-speaker variance (Rose & Morrison, 2009). Only speakers 36 and 90 seem comparatively stable across different systems, where the DS RMSE values vary between 0 and 2.5. In DS comparisons, the 19 speakers tend to fluctuate most in F2 systems (e.g. speakers 13, 20, 21), while the other six - 8, 46, 48, 51, 53 and 77 - yield the most variable performance in the combined system.

4. Conclusion

The current study used replications to explore system performance and individual speakers' behaviour, which provides novel insights for forensic speech science. The experiments show that combining multiple parameters contributes to overall system performance, and the performance of individual speakers is system-specific, i.e. training/reference speakers and parameters used.

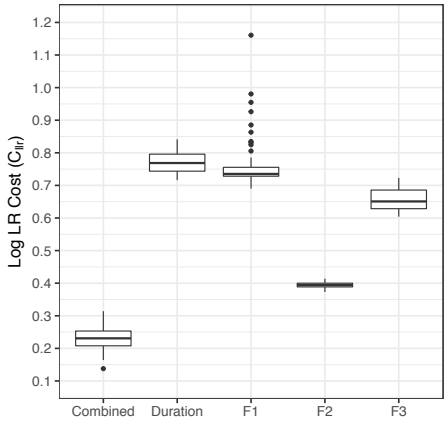

Figure 1. C_{llr}s in six systems

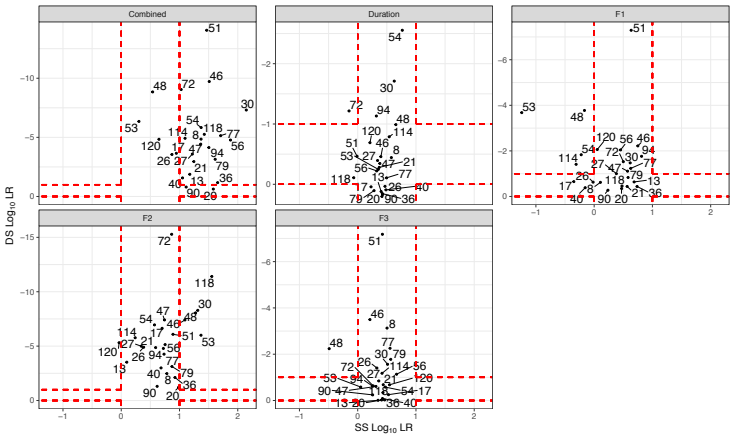

Figure 2. Zoo plot of 25 test speakers in six systems.

References

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. *Proc. Int'l Conf. Spoken Language Processing*, 4.

Dunstone, T., & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. Springer.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, *23*(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

Morrison, G. S. (2009). Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International Journal of Speech Language & the Law*, *15*(2), 249–266. https://doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, *53*(2), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G. S., Zhang, C., & Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, *208*(1–3), 59–65. https://doi.org/10.1016/j.forsciint.2010.11.001

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language & the Law*, *16*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31

Rose, P., & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech Language & the Law*, *16*(1), 139–163. https://doi.org/10.1558/ijsll.v16i1.139

Rose, P., & Wang, B. X. (2016). *Cantonese forensic voice comparison with higher-level features: Likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone*. 326–333. https://doi.org/10.21437/Odyssey.2016-47

Wang, X. B., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech Language and the Law*, *26*(1), 97–120. https://doi.org/10.1558/ijsll.38046

Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). Forensic voice comparison using Chinese /iau/. *Hong Kong*, ICPhS (pp. 2280-2283).