INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cross-linguistic Speaker Individuality of Long-term Formant Distributions: Phonetic and Forensic Perspectives

Justin J. H. Lo

Department of Language and Linguistic Science, University of York, UK
jl2355@york.ac.uk

Abstract

This study considers issues of language- and speaker-specificity in long-term formant distributions (LTFDs) from phonetic and forensic perspectives and examines their potential value in cases of cross-language forensic voice comparison. Acoustic analysis of 60 male English–French bilinguals revealed systematic differences in LTFDs between the two languages, with higher LTF2–4 in French than in English. Cross-linguistic differences in the shapes of LTFDs were also found. These differences are argued to reflect not only vowel inventories of each language but also language-specific phonetic settings. At the same time, a high degree of within-speaker consistency was found across languages. Likelihood ratio based testing was carried out to examine the effect of language mismatch on the utility of LTFDs as speaker discriminants. Results showed that while the performance of LTFDs was worse in cross-language comparisons than in same-language comparisons, they were still capable of providing speaker-specific information. These findings demonstrate that, in spite of deteriorated performance, LTFDs are still potentially useful speaker discriminants in cases of language mismatch. These findings thus call for further empirical investigation into the use of linguistic-phonetic features in cross-language comparisons.

Index Terms: long-term formant distributions, bilingual, forensic voice comparison, speaker specificity

1. Introduction

In cases of forensic voice comparison (FVC), it is not uncommon for analysts to encounter speech materials in different languages, especially in multilingual communities. Language mismatch between the known sample and the questioned sample presents particular challenges to forensic work, as speech in different languages entails, among other things, distinct sound inventories and phonetic targets. In its Code of Practice, the International Association for Forensic Phonetics and Acoustics recommends that members "exercise particular caution with cross-language comparisons" [1]. While automatic speaker recognition (ASR) systems claim to be only minimally affected by the issue of language mismatch [2], the efficacy of linguistic-phonetic features in such cases remains under-researched. This study focuses on the variable of long-term formant distributions (LTFDs), which measure the overall collection of formant estimates taken from a speech sample, and explores their discriminatory value in cross-language comparison.

1.1. LTFDs as speaker discriminants

The viability of LTFDs as an acoustic-phonetic speaker discriminant was first proposed in [3]. By considering the aggregate distribution of formant frequencies over whole speech samples instead of individual sounds, LTFDs are argued to not only reflect the physiology of the individual vocal tract, but also capture the overall articulatory habits of the speaker. Indeed, LTF1 means have been found to correlate with raised or lowered larynx, and higher LTF2 means have been shown to correlate with a fronted tongue body, thus giving evidence in support of the relationship between LTFDs and idiosyncratic vocal settings [4].

Studies conducted within the framework of likelihood ratios (LRs) provide empirical corroboration for the discriminatory potential of LTFDs, with low error rates reported in both English and German [4]–[6]. Higher formants (LTF3–4) generally outperform lower formants (LTF1–2) [6], and even stronger performance can be obtained through combining multiple LTFDs and their corresponding formant bandwidths [5].

1.2. Cross-linguistic LTFDs

While the suprasegmental, holistic nature of LTFDs lends itself to high speaker-discriminatory power in the absence of language mismatch, their discriminatory potential in cross-language comparisons may be limited if LTFDs are language-dependent. To date, the effect of language on LTFDs has been rarely examined in phonetic studies and results have so far been mixed. On the one hand, cross-linguistic comparisons of LTFDs from spontaneous telephone speech in German, Albanian and Russian have found LTF2 and LTF3 distributions to be comparable across languages [7]. On the other hand, small-scale studies examining intraspeaker variability of LTFDs in bilinguals have found them to display language-specific tendencies. Korean–English bilinguals, for example, largely maintain the shapes of LTFDs across languages, but produce LTF2 with lower peaks in Korean than in English [8]. Dutch–Turkish bilinguals do not produce LTF2 and LTF3 with significantly different means in each language, but the shape of their LTF2 distributions exhibits cross-linguistic differences [9]. These findings thus suggest that while LTFDs remain largely speaker-specific in terms of their central tendencies and shapes, they also appear to convey language-dependent information.

Related findings on phonetic settings also lend support to the idea that LTFDs may be subject to language-specific effects. Articulatory studies on inter-speech postures (ISPs) demonstrate language specificity in the phonetic settings adopted by monolingual speakers of different languages [10], as well as bilinguals in each of their languages [11]. Similar tendencies have been found in long-term average spectra of bilinguals [12], pointing to the differential use of voice quality according to the language spoken. The precise nature and origin of language specificity in phonetic settings, however,

http://dx.doi.org/10.21437/Interspeech.2021-1699

remain to be explored. Similar to the case of ISPs [10], language specificity of LTFDs may emerge as a consequence of the sound inventory in each language or as otherwise learned targets.

1.3. Current study

The work reviewed above points to the need to investigate further how factors of language- and speaker-specificity interact in LTFDs, in order to ascertain how useful LTFDs may be as speaker discriminants in cross-linguistic contexts. Therefore, the current study first examines the effect of language on the production of LTFDs, using a corpus of Canadian English–French bilinguals. The use of bilinguals controls for effects of speaker idiosyncrasy and allows language to be isolated as the variable of interest.

Building on the results from the acoustic analysis, this study also examines the impact of language mismatch on the performance of LTFDs in FVC within the framework of likelihood ratios (LRs), now widely accepted across fields of forensic science for evidence evaluation. If LTFDs are highly speaker-specific and language-independent, their speaker-discriminatory potential in cross-language comparisons is not expected to be adversely affected in relation to same-language comparisons. If LTFDs are largely constrained by language-specific effects, then they will unlikely offer much discriminatory value in cross-language comparisons.

2. Method

2.1. Materials

60 Canadian male bilinguals of English and French from the Voice ID Database [13] were selected for this study. Each speaker contributed one high-quality recording in English and another in French. In each recording, the speaker read 20 phonetically balanced sentences. A small number of recordings also contained a short, phonetically balanced reading passage. While the quality of the recordings may result in more optimistic performance than forensically realistic materials, the primary focus of the present study is on the effect of language itself. As long-term measures depend not only on the inventory of sounds and their phonetic implementation in, but also on the frequency of occurrence of each sound within each language and each sample [14], using uniform set of materials ensured that variation in LTFDs could be attributed to factors of speaker and language rather than speech content.

Out of all 60 speakers, 23 and 31 speakers reported English and French as their first language (L1) respectively, and the remaining six reported to have acquired both languages simultaneously. No speaker reported speaking any other languages. While no information on level of L2 proficiency was available, preliminary auditory analysis suggested mixed levels of proficiency. This set of speakers was thus considered to be representative of the relevant bilingual community, which is far from homogeneous.

2.2. Data extraction

All recordings were segmented by forced alignment using the Montreal Forced Aligner [15] and then manually checked and corrected where necessary. F1–F4 estimates were extracted every 10ms from all instances of vowels and glides (Table 1) in Praat [16]. Formant settings were determined by preliminary testing and were fixed at a maximum of 6 formants up to 5500 Hz for all speakers.

Table 1: *Target phonemes for formant extraction.*

Language	Vowels	Glides
English	i ɪ e ɛ æ ə ɜ ɑ ʌ ɔ o ʊ u aɪ aʊ ɔɪ	j w
French	Oral: i y e ø ɛ œ ə a o u	j w ɥ
	Nasal: ɛ̃ œ̃ ɑ̃ õ	

2.3. Acoustic analysis

To assess the effect of language on LTFDs, each LTFD was fitted with a separate linear mixed-effects model, using the *lme4* package [17] in R [18]. Language spoken (English vs French) and speaker group (L1 English, L1 French, and simultaneous bilingual) were included as effect-coded fixed effects along with their interaction. Random by-speaker intercepts and language-by-speaker slopes were included to model individual speaker variation. Significance testing of predictors was done by means of likelihood ratio tests (LRTs) implemented using ANOVAs, where the full model was compared against a reduced model with the predictor excluded to see if its inclusion led to an improved fit.

Further correlation analysis was conducted to examine the cross-linguistic within-speaker consistency of LTFDs. To allow examination of both central tendencies and shapes of LTFDs, the means and SDs of each LTFD were calculated for each speaker in English and in French. Pearson correlation coefficients were then calculated between English and French for each measure.

2.4. Likelihood ratio-based testing

To assess the effect of language mismatch on the performance of LTFDs as speaker discriminants, system testing was conducted in two conditions. In the *same-language* condition, the language in the known and questioned samples were matched; in the *cross-language* condition, the questioned sample was in one language and the known sample in the other. In both conditions, the language of the reference (background) data was matched with that of the questioned sample, as the relevant population is logically defined in accordance with relevant characteristics of the questioned sample and not of the known sample.

Five sets of systems were tested within each condition. Each of LTF1-4 was tested individually as input parameters, as well as the combination of all four LTFDs. To test each system, the 60 speakers were first randomly divided into *test*, *training* and *background* sets, each comprising 20 speakers. Comparisons were conducted between all pairs of speakers in the *test* set, with each speaker acting in turn as the known and questioned speaker for all speakers. As only one recording per language was available from each speaker, the first half was taken to be the known sample and the second half to be the questioned sample. This was maintained in both conditions to keep the amount of input data constant.

LTFDs were modelled and compared using the GMM–UBM approach [19], implemented with the *mclust* package [20] in R. The reference population was modelled with a single GMM (the UBM) composed of 12 Gaussians, calculated using data pooled from all 20 speakers in the *background* set. Data from the known sample in each comparison was then used to adapt the UBM to derive the

GMM for the known speaker. The process was repeated with the *training* set, the scores from which were used to calibrate the *test* scores to obtain $\log_{10}$LRs (LLRs) through a logistic regression procedure. The performance of each system was then evaluated with the log LR cost function (C_{llr}) [21], a standard metric of system validity. A C_{llr} of below 1 indicates that a system is well-calibrated and provides useful speaker-specific information. The closer C_{llr} is to 0, the stronger the system is judged to be performing. The procedure of sampling and testing was repeated 100 times to compare all speaker-pairs and minimise effects of speaker random sampling [22].

The effect of language mismatch on Cllr was analysed using linear mixed-effects models. One model was fitted to C_{llr} obtained from systems with all four LTFDs combined as input, with condition, language of the questioned sample (QS) and their interaction included as fixed effects. Random by-repetition intercepts were also included. A separate model was fitted to systems with individual LTFDs as input. The same fixed and random effects were included, with the addition of input parameter and its interaction with the other predictors. As in 2.3, predictors were effect-coded and their significance was determined by LRTs.

3. Results

3.1. Language specificity of LTFDs

Figure 1 shows the overall distribution of LTFDs in each language. Compared to LTFDs produced in English, the distributions in French showed a similar peak frequency in LTF1 but higher peak frequencies in LTF2-4. Some differences in the shapes of the distributions could also be observed between the two languages. LTF1 was more sharply peaked in English, while LTF2 showed a heavier tail in lower frequencies in English. In the case of LTF3, both a flatter peak and heavier lower tail could be found.

Mixed-effects models fitted to the formant data confirmed cross-linguistic differences in LTFDs. The effect of language was not significant for LTF1 (p = .65), but were significant for LTF2-4 (ps < .0005), indicating that speakers produced higher LTF2-4 means in French than in English. The estimated difference between the two languages was the largest for LTF2 (133 Hz), but smaller for LTF3 (92 Hz) and LTF4 (34 Hz). Neither speaker group nor its interaction with language was significant for any of the LTFDs (ps > .10).

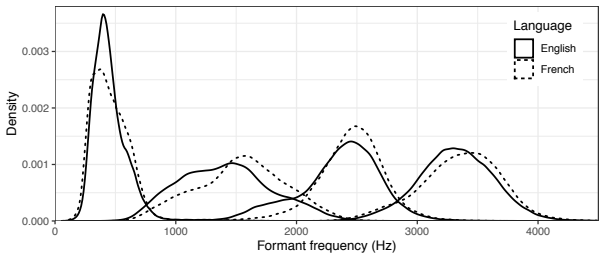

Figure 1: *LTF1-4 from all 60 speakers by language.*

3.2. Speaker specificity

Figure 2 illustrates LTFD means for each speaker in English and French. LTFD SDs for each speaker are illustrated in Figure 3. The cross-linguistic differences in the means and shapes of LTFDs, as described in 3.1, are clearly evidenced here. The vast majority of speakers reported higher LTF2–4 means in French in Figure 2, whereas in Figure 3 all speakers produced lower SDs for LTF3 in French than in English, indicating a lower variability that is consistent with the sharper peak in Figure 1. Considerable variability between speakers could be found, but individuals remained consistent across languages. As summarised in Table 2, strong correlations of speaker means between English and French were found for LTF1, LTF3 and LTF4, while LTF2 means showed a weaker correlation. Similarly, SDs for each LTFD were strongly correlated, most strongly in the case of LTF1 and LTF4. Taken together, these results show that, in spite of the effect of language, the speaker specificity of LTFDs is largely maintained across languages.

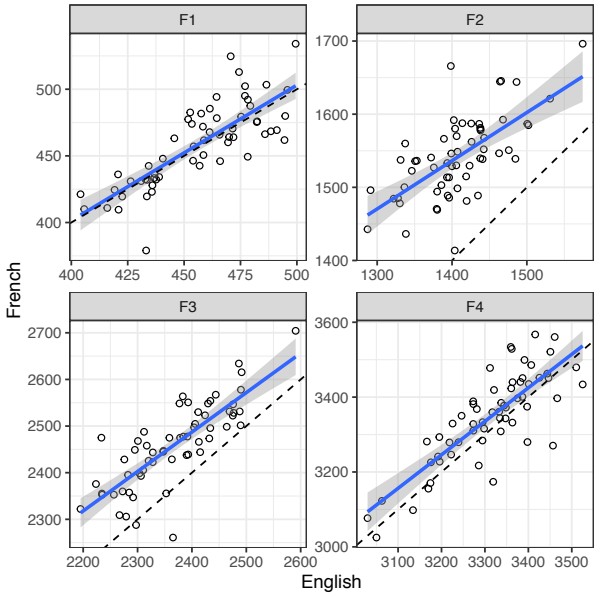

Figure 2: *LTF1-4 means by speaker with best-fitting line (solid) and line of equality (dashed).*

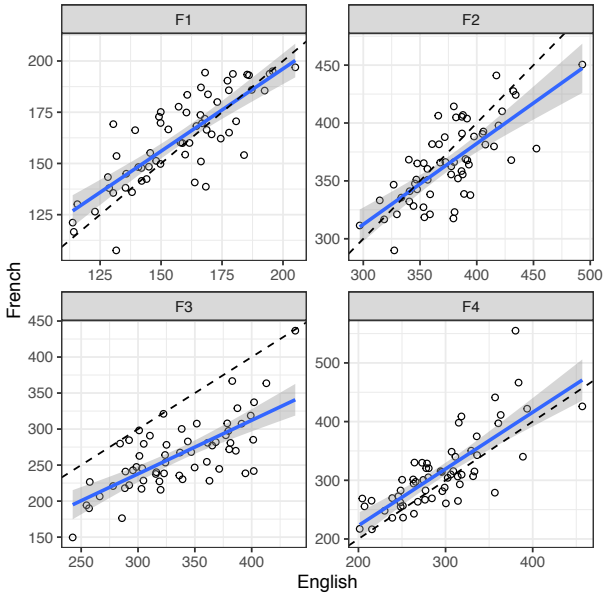

Figure 3: *LTF1-4 SDs by speaker with best-fitting line (solid) and line of equality (dashed).*

Table 2: *Pearson's r for LTFD speaker means and SDs (p < 0.0001 in all cases).*

	LTF1	LTF2	LTF3	LTF4
Mean	0.80	0.66	0.81	0.81
SD	0.80	0.73	0.71	0.78

3.3. System performance

Figure 4 illustrates the performance of LTFD-based systems in different conditions. In the same-language condition, Systems with individual LTFDs as input all performed on broadly similar levels, with mean C_{llr} between 0.61 and 0.74. In the cross-language condition, mean C_{llr} increased to 0.72–0.94, as confirmed by a significant main effect of condition (p < .0001). Figure 4 further shows that the effect of condition on different systems was not uniform, as confirmed by significant interactions between condition and both QS language and parameter (ps < .0001). Notably, even in the case of language mismatch, mean C_{llr} (and indeed C_{llr} in nearly all repetitions) remained below 1, meaning that each system still contains some useful speaker-discriminatory information.

When all four LTFDs were combined as input, much lower C_{llr} was obtained, attesting to a clearly improved performance over individual LTFDs. As in the case of individual LTFDs, mean C_{llr} for systems combining all four LTFDs was significantly higher (p < .0001) in the cross-language condition (0.46) than in the same-language condition (0.29). A significant main effect of QS language was found (p < .0001), such that mean C_{llr} for systems with QS in French was slightly lower than those in English. There is a differential effect of condition on C_{llr} across QS languages, as shown by a significant interaction between the two factors (p < .0001).

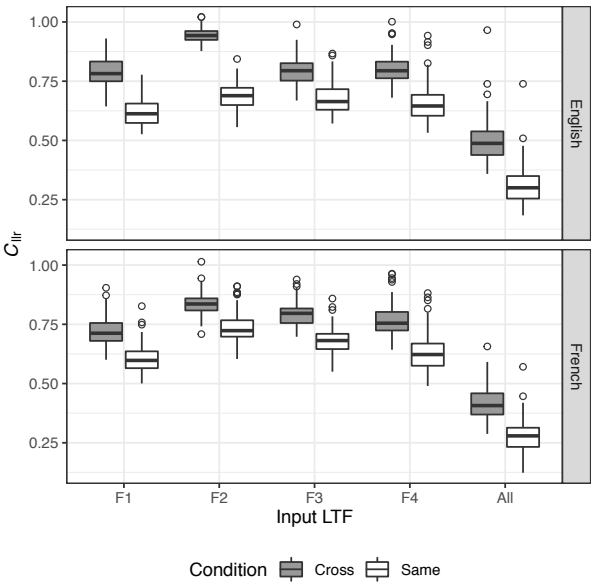

Figure 4: *C_{llr} from all systems by QS language.*

4. Discussion

Results from the present study reveal systematic effects of language on LTFDs, in line with earlier studies on phonetic settings [11]. When compared to English, LTFDs in French exhibited a general shift towards higher frequencies regardless of the linguistic background of the speaker. This shift was found most prominently in LTF2, which may be attributed to a more crowded front vowel space in French, where pairs of unrounded and rounded front vowels can be found. The upward shift was not limited to LTF2, but also extended to the higher formants, albeit in smaller magnitudes. Higher formants are generally considered to be less constrained to encode linguistic information than lower formants, but instead more indicative of a speaker's individual voice quality [23]–[25]. Therefore, it may be the case that the factor of vowel inventory alone cannot fully account for the differences in LTFDs between English and French, and that bilinguals make subtle changes across languages in their setting of the supralaryngeal vocal tract. Future research designed to tease apart these sources of language specificity in LTFDs would be necessary to test such a possibility. Nevertheless, strong correlations in LTFD means and SDs indicate that, relative to the whole population, individuals remained consistent across languages in both central tendencies and shapes of LTFDs.

Both language- and speaker-specific aspects in the phonetics of LTFDs were reflected in their speaker-discriminatory power, as evidenced by the empirical results from LR-based testing. While system performance in the same-language condition was generally stable across languages, the cross-language condition saw a consistent increase in C_{llr} across all systems tested, thus illustrating the adverse impact of language mismatch on the discriminatory potential of LTFDs. It is acknowledged, however, that limitations in the materials are likely to have also contributed to the discrepancy in performance, as within-speaker variation is more constrained in the same-language condition by the use of single recordings than in the cross-language condition where samples were derived from separate recordings [26]. Given the varying degree to which each LTFD differed across languages, these findings provide an indication of the sensitivity of LTFDs to minor shifts in the distribution of individuals within the population. While poorer performance is expected, given the phonetic evidence of language dependence, systems in the cross-language condition were still reasonably well calibrated, as mean C_{llr} remained below 1 in all cases. Therefore, even though the discriminatory potential of LTFDs is diminished in cases of language mismatch, they can still offer useful speaker-specific information, especially when multiple LTFDs are analysed in combination.

5. Conclusions

This study has considered the interplay of language specificity and speaker specificity in LTFDs and its implications for forensic casework. Within a bilingual population, LTFDs were found to be language-dependent, which served to mediate their speaker specificity. The practical impact of these findings was demonstrated in LR-based testing, which found weaker performance of LTFDs as speaker discriminants in the presence of language mismatch between the speech samples being compared. LTFDs were shown to contain useful speaker-specific information even in cases of cross-language comparisons, thus raising the possibility of using linguistic-phonetic variables in such cases. However, as testing was conducted in optimal conditions, the effect of language mismatch in FVC warrants further investigation. Future studies using forensically realistic materials would shed light on how performance is degraded in such scenarios.

6. References

[1] *Code of Practice*, International Association for Forensic Phonetics and Acoustics, 2020. [Online]. Available: https://www.iafpa.net/the-association/code-of-practice/

[2] H. J. Künzel, "Automatic speaker recognition with cross-language speech material," *Int. J. Speech Lang. Law*, vol. 20, no. 1, pp. 21–44, 2013.

[3] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *Int. J. Speech Lang. Law*, vol. 12, no. 2, pp. 143–173, 2005.

[4] P. French, P. Foulkes, P. Harrison, V. Hughes, E. San Segundo, and L. Stevens, "The vocal tract as a biometric: Output measures, interrelationships, and efficacy," in *Proc. 18th Int. Congr. Phonetic Sci.*, 2015, Paper 0817.

[5] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models," in *Proc. Interspeech 2008*, pp. 1505–1508.

[6] E. Gold, P. French, and P. Harrison, "Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework," in *Proc. Meeting Acoust.*, vol. 19, 2013, 060041.

[7] M. Jessen and T. Becker, "Long-term formant distribution as a forensic-phonetic feature," presented at the 2nd Pan-Amer./Iberian Meeting Acoust., Cancún, Mexico, Nov. 15–19, 2010.

[8] S. Cho and M. J. Munro, "F0, long-term formants and LTAS in Korean-English bilinguals," in *Proc. 31st General Meeting Phonetic Soc. Japan*, 2017, pp. 188–193.

[9] W. Heeren, D. van der Vloed, and J. Vermeulen, "Exploring long-term formants in bilingual speakers," presented at the 23rd Int. Asso. Forensic Phonetics Acoust. Annu. Conf., Zurich, Switzerland, Aug. 31–Sep. 3, 2014.

[10] B. Gick, I. Wilson, K. Koch, and C. Cook, "Language-specific articulatory settings: Evidence from inter-utterance rest position," *Phonetica*, vol. 61, pp. 220–233, 2004.

[11] I. Wilson and B. Gick, "Bilinguals use language-specific articulatory settings," *J. Speech Lang. Hear. Res.*, vol. 57, no. 2, pp. 361–373, 2014.

[12] M. L. Ng, Y. Chen, and E. Y. K. Chan, "Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—A long-term average spectral analysis," *J. Voice*, vol. 26, no. 4, pp. e171–e176, 2012.

[13] Royal Canadian Mounted Police, "Voice ID Database." Unpublished audio corpus. 2010–2016.

[14] I. Mennen, J. M. Scobbie, E. de Leeuw, S. Schaeffler, and F. Schaeffler, "Measuring language-specific phonetic settings," *Second Lang. Res.*, vol. 26, no. 1, pp. 13–41, 2010.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech 2017*, pp. 498–502.

[16] Praat: Doing phonetics by computer (Version 6.0.19). (2016). P. Boersma & D. Weenink. Available: http://www.praat.org

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[18] R Core Team, "R: A language and environment for statistical computing," version 3.5.1. R Foundation for Statistical Computing, 2018, Vienna, Austria. Available:

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[20] L. Scrucca, M. Fop, T. Brendan Murphy, and A. E. Raftery, "mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models," *R J.*, vol. 8, no. 1, pp. 289–317, 2016.

[21] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 230–275, 2006.

[22] B. X. Wang, V. Hughes, and P. Foulkes, "The effect of speaker sampling in likelihood ratio based forensic voice comparison," *Int. J. Speech Lang. Law*, vol. 26, no. 1, pp. 97–120, 2019.

[23] K. McDougall, "Speaker-specific formant dynamics: An experiment on Australian English /aɪ/," *Int. J. Speech Lang. Law*, vol. 11, no. 1, pp. 103–130, 2004.

[24] P. Rose, *Forensic speaker identification*. New York, NY, USA: Taylor & Francis, 2002.

[25] P. Ladefoged and K. Johnson, *A course in phonetics*, 7th ed. Stamford, CT, USA: Cengage Learning, 2015.

[26] E. Enzinger and G. S. Morrison, "The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems," in *Proc. 14th Australas. Int. Conf. Speech Sci. Technol.*, 2012, pp. 137-140.