Can phonetic theories predict speaker discrimination performance?

Ricky K. W. Chan[1], Bruce Xiao Wang[2]
[1]Speech, Language and Cognition Laboratory, School of English, University of Hong Kong
[2]Department of English and Communication, The Hong Kong Polytechnic University

Abstract

Phonetic theories provide frameworks that help us understand how speech sounds are produced, transmitted, and perceived. Whether phonetic theories can predict speaker discrimination performance in the context of forensic voice comparison is an important issue that requires investigation. This paper evaluates the speaker-discriminatory power of long-term acoustic-phonetic features, mel-frequency cepstral coefficients, and their combinations under the Bayesian likelihood ratio framework. Results suggest that long-term source and filter features are not necessarily complementary for distinguishing speakers even though they are largely independent in speech production and perceived voice quality. Moreover, our analysis of long-term features provides little support for the complementarity of acoustic-phonetic and ASR approaches to forensic voice comparison. Suggestions for future research are provided.

1 Introduction

Speech conveys a certain degree of speaker-related information such as gender and regional background, and it often allows us to distinguish familiar speakers from unfamiliar ones (Nolan, 1999). While it is scientifically interesting to determine the extent to which a person's voice is unique, there are specific situations where it is crucial to identify or discriminate speakers based solely on their speech. One context where explicit comparison of speech features is crucial is forensic voice comparison (FVC). FVC typically involves comparing two speech samples in forensic situations, such as hoax emergency calls, ransom demands, or conversations with accomplices, to help determine whether the voices belong to the same speaker (French & Stevens, 2013). With the widespread availability of speech recordings, law enforcement and courts increasingly rely on specialists to analyse speech samples and provide expert opinions during court proceedings or as part of investigations.

The past two decades have witnessed an increasing number of research papers exploring the speaker-discriminatory power of individual speech features such as vowel formants (e.g. McDougall 2004; Nolan & Grigoras, 2005; Rose, 2007), laryngeal voice quality (e.g. Chan, 2023; Hughes et al., 2019), lexical tones (e.g. Chan, 2016; 2020; accepted; Chan & Wang, 2024a; Pingjai, 2019; Rose & Wang, 2016), and F0 (Hudson et al., 2007; Jessen et al., 2005; Kinoshita et al., 2009). Recent research in FVC has focused on combining different speech features to optimize speaker-discriminatory performance. Ideally, these features should not uncorrelated and provide independent speaker-related information. Phonetic theories, which provide models for understanding the production, transmission, and perception of speech sounds, may help us make more informed predictions in this regard. For instance, the source-filter theory (Fant, 1960) assumes the independence between source and filter in speech production. Source features such as F0 and laryngeal voice quality may thus provide speaker-related information that is independent from filter features such as vowel formants. Hughes et al. (2023) found that, based on contemporaneous speech data of the hesitation marker *um* in Southern British English, combining source and filter features has the potential to yield the strongest speaker discrimination performance. On the other hand, according to the psychoacoustic model of voice quality proposed by Kreiman et al. (2014), harmonic source spectral shape, inharmonic source excitation, time-varying source characteristics, and vocal tract transfer function are distinct components that are both necessary and sufficient for modelling perceived voice quality. This suggests that these components have different characteristics for distinguishing voices.

However, predictions from phonetic theories do not always translate into speaker discrimination performance in real-world data, especially for forensically relevant speech. Empirical tests are needed to determine how various speech features can be combined for optimal speaker discrimination. This study seeks to test the speaker-discriminatory power of long-term F0 (LTF0), long-term formant distributions (LTFDs), long-term laryngeal voice quality (LTLVQ), long-term mel-frequency cepstral coefficients (MFCCs), and combinations thereof.

2 Method

2.1 Speech data

We analysed 75 male Australian English speakers from Sydney and other areas within New South Wales (see Chan, 2023 and Chan & Wang, 2024b for details). The speech data were sourced from a forensic-realistic database developed by Morrison et al. (2015). For each speaker in this study, two recordings involving two speech styles commonly found in forensic casework—a casual telephone conversation with a friend (CNV), a mock police interview (INT)—were analysed. The speakers were recorded on two separate sessions with a a two-week interval between each session. The recordings were thus coded as CNV1 and INT2. We analysed non-contemporaneous recordings with speech style mismatch because these conditions are common in forensic casework (Morrison et al., 2015).

2.2 Speech feature extraction

Only the vocalic portions of the speech recordings were manually segmented using a *TextGrid* in Praat (Boersma & Weenink, 2023), resulting in approximately 33 seconds of net vocalic material per speaker per recording session for analysis. Acoustic-phonetic features (i.e. LTF0, LTFDs and LTLVQ) were extracted using VoiceSauce (Shue et al., 2011), and MFCCs were derived using the *librosa* Python library (McFee et al., 2015) with a 20ms window length and 10ms window shift. Details of system configuration for extracting these speech features are as follows:

1) Long-term fundamental frequency (LTF0): fundamental frequency was extracted using the Straight algorithm (Kawahara et al., 2016).

2) Long-term laryngeal voice quality (LTLVQ): five spectral tilt measures (H1-H2, H2-H4, H1-A1, H1-A2, H1-A3, with harmonic/spectral amplitudes corrected for formant frequencies and bandwidths) and five additive noise measures (cepstral peak prominence (CPP) and harmonic-to-noise ratio (HNR) at four frequency ranges: 0-500 Hz, 0-1500 Hz, 0-2500 Hz, and 0-3500 Hz) were extracted.

3) Long-term formant distributions (LTFDs): the first three formants (F1-F3) were extracted using the algorithm in the Snack Sound Toolkit (Sjölander, 2004), with a 6000 Hz ceiling for four formants and a pre-emphasis of 0.96 and 12 LPC order.

These three features are often analysed in the acoustic-phonetic approach to FVC.

4) Mel-frequency cepstral coefficients (MFCCs): the first 13 MFCCs, alongside their corresponding delta and delta-delta coefficients (39 coefficients in total), were derived with a frequency range from 0 to 11025 Hz. MFCCs are often analysed in classical automatic speaker recognition (ASR) approach to FVC.

With reference to the psychoacoustic model of voice quality proposed by Kreiman et al. (2014), LTLVQ correspond to the harmonic source spectral shape (spectral tilt parameters) and the inharmonic source excitation (additive noise parameters) components. LTF0 and LTFDs are relevant to time-varying source characteristics and vocal tract transfer function respectively. Besides, according to the source-filter theory, LTF0 and LTLVQ can be categorized as 'source' features whereas LTFDs as 'filter' features. Thus, LTLVQ, LTF0, and LTFDs are expected to provide different and considerable complementary information for speaker discrimination. MFCCs are often assumed to mostly capture vocal tract filter information, and it is often claimed that source information is removed by smoothing out rapid local changes in the spectrum that are caused by harmonics or noise in the signal (Hughes et al., 2023; Jurafsky & Martin, 2008). Nonetheless, the degree of source-filter decoupling in MFCCs hinges on the number of coefficients involved in the analysis: fewer cepstral coefficients lead to a smoother spectral representation, which results in less source information being captured (Hughes et al., 2023). With the use of 13 coefficients in the present study, our MFCC data are expected to carry both source and filter information. Therefore, we hypothesize that the addition of LTF0, LTLVQ, and/or LTFDs will not considerably improve MFCCs-based system performance.

2.3 Statistical analysis

Speaker discrimination performance of the speech features was assessed using Bayesian likelihood ratio (LR)-based testing, which is standard in FVC. The LR quantifies the probability of the evidence under two opposing hypotheses: 1) the two speech samples are from the same speaker, and 2) the samples are from different speakers. In a forensic context, the LR reflects the extent to which the evidence favours the prosecution's hypothesis over the defence hypothesis, or vice versa (Aitken & Taroni, 2004). To evaluate speaker discrimination performance, pairs of samples with known ground truth regarding whether they originate from the same speaker (SS) or different speakers (DS) are needed.

The 75 speakers were randomly assigned to the training, test, and reference sets (25 speakers each set), and speaker models were built using Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al., 2000). Multiple SS and DS comparisons were performed for the training and test sets, and each comparison produced an LR-like score which quantified the similarity between the two sets of data, and the typicality of the data based on a model created by the reference set. The scores were calibrated and converted to log LRs using logistic regression (Brümmer et al., 2007). This process involved shifting and scaling the test scores using calibration coefficients learnt from the training scores to enhance their comprehensibility, comparability, and interpretability (Morrison et al., 2013). Scores from different features were combined using logistic regression fusion (Pigeon, Druyts, & Verlinde, 2000), a technique that accounts for the underlying correlations in the scores when combining results. This procedure was repeated 30 times with different speakers in the training, test and reference sets to account for variability due to different configurations of the three sets (Wang et al., 2019). System validity was evaluated based on the distributions of log-LR cost (C_{llr}) across the 30 replications. A C_{llr} value close to zero imply better performance with fewer and less severe speaker-discriminatory errors. A C_{llr} value of 1 or above implies that the system yields no speaker-discriminatory information.

3 Results and discussion

Figure 1 shows the mean C_{llr} values of individual long-term features as a function of the number of Gaussians (1, 2, 4, 8, 16, 32, 64) across 30 repetitions. This serves as a pre-test for identifying the optimal number of Gaussians for modelling each long-term feature. The number of Gaussians that yielded the lowest C_{llr} values (i.e.

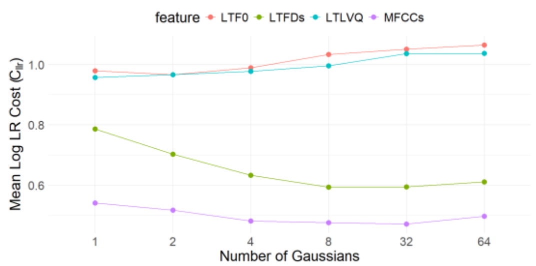

Figure 1:Mean C_{llr} values of systems based on individual long-term features for CNV1 vs. INT2.

better system validity) clearly depends on the features being modelled. Specifically, more Gaussians are required for optimal modelling of LTFDs and MFCCs which mainly capture vocal tract filter information, whereas fewer Gaussians (1 to 8) seem to be sufficient for LTF0 and LTLVQ which are source features. However, speech style and the time gap between recordings may also play a role in causing the fluctuations observed. The optimal number of Gaussians is 32, 32, 2 and 1 for MFCCs, LTFDs, LTF0, and LTLVQ respectively and the subsequent analysis was based on these optimal number of Gaussians.

Figure 2 illustrates the distributions of C_{llr} values for both individual features and their various combinations across 30 repetitions, and Table 1 gives the descriptive statistics of the corresponding C_{llr} values. In general, all individual features and their combinations yielded small standard deviation in C_{llr} values ranging from 0.03 to 0.13, indicating relatively high system reliability across 30 repetitions.

	C_{llr}			
Input Feature(s)	**Min**	**Max**	**Mean**	**SD**
MFCCs	0.25	0.70	0.37	0.09
LTFDs	0.37	0.78	0.53	0.10
LTLVQ	0.87	0.99	0.92	0.03
LTF0	0.73	1.00	0.89	0.06
LTLVQ+LTF0	0.71	0.98	0.85	0.06
LTFDs+LTF0	0.38	0.92	0.54	0.10
LTFDs+LTLVQ	0.43	0.82	0.55	0.07
LTFDs+LTF0+LTLVQ	0.37	0.90	0.54	0.10
MFCCs+LTF0	0.23	0.77	0.38	0.11
MFCCs+LTLVQ	0.17	0.72	0.37	0.10
MFCCs+LTLVQ+LTF0	0.13	0.79	0.38	0.12
MFCCs+LTFDs	0.19	0.70	0.30	0.10
MFCCs+LTFDs+LTF0	0.19	0.70	0.32	0.12
MFCCs+LTFDs+LTLVQ	0.20	0.88	0.32	0.13
All four features	**0.13**	**0.95**	**0.33**	**0.16**

Table 1: Descriptive statistics of Cllr values for MFCCs, LTF0, LTFDs, LTLVQ, and combinations thereof across 30 repetitions.

Among the individual acoustic-phonetic features, LTFDs performed the best (mean C_{llr} = 0.53), but LTF0 or LTLVQ alone provided limited speaker-discriminatory information (mean C_{llr} values = 0.89 and 0.92 respectively). These results align with previous findings that LTFDs, which are supposed to capture vocal tract filter information, are a reasonably good speaker discriminant (e.g. French et al., 2015; Gold et al., 2013; Jessen et al., 2014; Moos, 2010), but not so much for source features, especially when speech style mismatch and non-contemporaneous recordings are involved (e.g. Chan, 2023; Jessen et al., 2023; Rose & Zhang, 2018). On the other hand, MFCCs performed best in speaker discrimination (mean C_{llr} = 0.37).

The combinations of acoustic-phonetic features (i.e. LTFDs, LTF0, and LTLVQ) did not lead to a considerable drop in mean C_{llr} (i.e. less than 0.1) when compared with using a corresponding feature. LTLVQ + LTF0 yielded a mean C_{llr} of 0.85, which is only slightly better than LTF0 alone (0.89). Adding source features—LTF0 and/or LTLVQ—to an LTFDs system even led to slightly worse performance, with the mean C_{llr} 0.53 increasing to 0.54-0.55. This suggests that LTLVQ, LTF0 and LTFDs, despite assumed to be largely independent in speech production and perceived voice quality, do not necessarily provide complementary information for discriminating speakers. Furthermore, the addition of source features to MFCCs-based systems did not improve performance either. A possible reason is that LTLVQ and LTF0 performed rather poorly and did not have much speaker-discriminatory information to add to LTFDs or MFCCs-based systems. Finally, adding LTFDs to an MFCCs-based system resulted in a small decrease in mean by C_{llr} 0.7. This is consistent with our prediction and the findings by Becker (2012) and Hughes et al. (2017). Overall, our analysis of long-term speech features provides no support for the claim that the acoustic-phonetic and automatic speaker recognition (ASR) approaches may be complementary for speaker discrimination (cf. French & Stevens, 2013; Hughes et al., 2019; 2023).

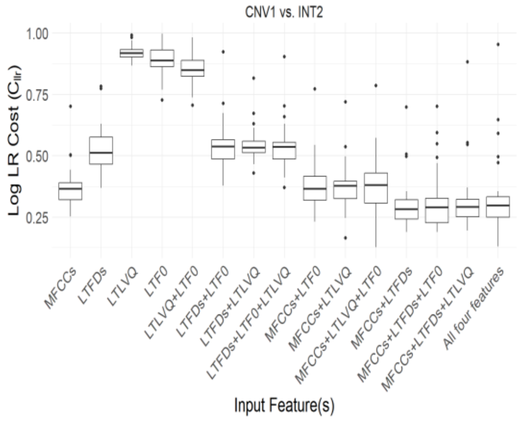

Figure 2: Boxplots of C_{llr} values of individual long-term features and combinations thereof.

4 Conclusion

This paper investigates whether phonetic theories can predict the complementarity of speech features in speaker discrimination for FVC. Results suggest that although LTLVQ, LTF0, and LTFDs are generally considered largely independent in speech production and perceived voice quality, they do not necessarily offer complementary information for distinguishing between speakers. Additionally, our analysis of long-term speech features provides no strong support for the complementarity of acoustic-phonetic and ASR approaches to FVC. Future studies could test the predictions by other phonetic theories on speaker discrimination, and explore other ways in which phonetic analysis and ASR systems might be complementary for optimal speaker discriminatory performance.

5 References

Aitken, C., Roberts, P., & Jackson, G. (2010). *Fundamentals of Probability and Statistical*

Evidence in Criminal Proceedings. https://rss.org.uk/news-publication/publications/law-guides/

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*(3), 161–164. https://doi.org/10.1016/j.scijus.2009.07.004

Becker, T. (2012). *Automatischer forensischer Stimmenvergleich*. BoD–Books on Demand.

Boersma, P., & Weenink, D. (2023). *Praat:Praat: Doing phonetics by computer [Computer program].* (Version 6.3.08) [Computer software]. http://www.praat.org/

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

Chan, R. (2016). Speaker variability in the realization of lexical tones. *International Journal of Speech, Language and the Law, 23*(2), 195-214.

Chan, R. (2020). Speaker discrimination: citation tones vs. coarticulated tones. *Speech Communication*, *117*, 38-50.

Chan, R. (2023). Evidential value of voice quality acoustics in forensic voice comparison. *Forensic Science International*, *348*, 111725.

Chan, R. (accepted). Tone languages. In F. Nolan, K. McDougall & T. Hudson (Eds), *Oxford Handbook of Forensic Phonetics*. Oxford University Press.

Chan, R. & Wang, B. (2024a). Modeling Lexical Tones for Speaker Discrimination. Language and Speech, 0(0). https://doi.org/10.1177/00238309241261702

Chan, R., Wang, B. (2024b). Do long-term acoustic-phonetic features and mel-frequency coefficients provede complementary speaker-specific information for forensic voice comparison. *Forensic Science International,* 112119. https://doi.org/10.1016/j.forsciint.2024.112199

Enzinger, E., Zhang, C., & Morrison, G. S. (2012). Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop.*, 78–85.

Fant, G. (1971). *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. Walter de Gruyter.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

French, P. & Stevens, S (2013) Forensic speech science. In Jones, M. & Knight, R. (eds.) *The Bloomsbury Companion to Phonetics*. London: Continuum.

French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E., & Stevens, L. (2015). The vocal tract as a biometric: Output measures, interrelationships, and efficacy. *Proceedings of the 18th International Congress of Phonetic Sciences*. International Congress of Phonetic Sciences, University of Glasgow.

Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, *18*(2), Article 2. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech Language and the Law*, *26*(1), Article 1. https://doi.org/10.1558/ijsll.38028

Gold, E., French, P., & Harrison, P. (2013). *Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework*. 060041–060041. https://doi.org/10.1121/1.4800285

Hudson, T., de Jong, G., McDougall, K. & Nolan, F. (2007). f0 statistics for 100 young male speakers of standard Southern British English. In Trouvain, J. & Barry, W. J. (eds.) In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrucken, Germany, pp. 1809–1812.

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

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker-specificity in speech production: The contribution of source and filter. *Journal of Phonetics*, *97*, 101224. https://doi.org/10.1016/j.wocn.2023.101224

Hughes, V., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Effects of formant analysis settings and channel mismatch on semiautomatic forensic voice comparison. *International Congress of Phonetic Sciences*, 3080–3084.

Hughes, V., Harrison, P., Foulkes, P., French, J. P., Kavanagh, C. & San Segundo, E. (2017). Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech*, Stockholm, Sweden, 3892–3896.

Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass*, *2*(4), 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Jessen, M. (2021). MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis. *Interspeech 2021*, 411–415. https://doi.org/10.21437/Interspeech.2021-1697

Jessen, M., Alexander, A., & Forth, O. (2014). Forensic Voice Comparisons in German with Phonetic and Automatic Features using VOCALISE Software. *Audio Engineering Society*. Audio Engineering Society Conference: 54th International Conference: Audio Forensics.

Jessen M., Konrat, C., & Horn, J. (2023). Voice comparison analysis of forensic recordings using the VoiceSauce program. In *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*. Prague, Czech Republic.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*, 174–213.

Jurafsky, D. & Martin, J. (2008). *Speech and Language Processing.* Tracy Dunkelberger.

Kafadar, K., Stern, H., Ceullar, M., Curran, J. M., Lancaster, M., Neumann, C., Saunders, C., Weir, B., & Zabell, S. (2019). *American Statistical Association Position on Statistical Statements for Forensic Evidence*. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf

Kawahara, H., Agiomyrgiannakis, Y., & Zen, H. (2016). Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. *arXiv preprint arXiv:1605.07809*.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law*, *16*, 91–111.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *loquens*, *1*(1), e009.

Ladefoged, P., & Bladon, A. (1982). Attempts by human speakers to reproduce Fant's nomograms. *Speech Communication*, *1*(3), 185–198. https://doi.org/10.1016/0167-6393(82)90016-4

McFee, B., Raffel, C., Liang, C., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*, *8*, 18–25.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician*, *101*, 7–24.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173–197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), Article 2. https://doi.org/10.1080/00450618.2011.630412

Morrison, G. S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B. K., De Souza, S., Cummins, N., & Chow, D. (2015). *Forensic database of voice recordings of 500+ Australian English speakers*. https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/index/index/docId/67850

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English/aI. *International Journal of Speech, Language and the Law*, *11*(1), 103-130.

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92-100.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, *12*(2), 143-173.

Nolan F. (1999). Speaker identification and forensic phonetics. In Hardcastle W. J., Laver J. (Eds.), *Handbook of phonetic sciences* (pp. 519–533). Blackwell.

Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, *10*(1-3), 237-248.

Pingjai, S. (2019). *A Likelihood-Ratio Based Forensic Voice Comparison in Standard Thai*. PhD Thesis. Australian National University.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

Shue, Y.-L., Keating, P., & Vicenik, C. (2011). VOICESAUCE: A program for voice analysis. *International Congress of Phonetic Sciences*, *126*, 1846–1849. https://doi.org/10.1121/1.3248865

Sjölander, K. (2004). The snack sound toolkit [computer program].

Stevens, K. N., & House, A. S. (1995). Development of a Quantitative Description of Vowel Articulation. *The Journal of the Acoustical Society of America*, *27*(3), 484–493. https://doi.org/10.1121/1.1907943

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

Willis, S. M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, A., Nordgaard, A., Berger, C. E. H., Sjerps, M. J., Lucena-Molina, J. J., Zadora, G., Aitken, C. G. G., Lunt, L., Champod, C., Biedermann, A., Hicks, T. N., & Taroni, F. (2015). *ENFSI guideline for evaluative reporting in forensic science*.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication*, *55*(6), Article 6. https://doi.org/10.1016/j.specom.2013.01.011

Rose, P. (2007). Forensic speaker discrimination with Australian English vowel acoustics. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp. 1817–1820.

Rose, P., & Wang, 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. *Odyssey 2016*, 326-333.

Rose, P., & Zhang, C. (2018). Conversational style mismatch: its effect on the evidential strength of long-term F0 in forensic voice comparison. *Proceedings of ASSTA*, 157-160.