



High level forensic voice comparison based on fused long-term fundamental frequency and word n -gram features

Michael Carne¹, Shunichi Ishihara¹ & Yuko Kinoshita¹

¹Speech and Language Laboratory, the Australian National University

michael.carne@anu.edu.au, shunichi.ishihara@anu.edu.au, yuko.kinoshita@anu.edu.au

Abstract

Feature robustness is particularly important in forensic applications of speaker recognition, where there are often significant differences in the recording conditions between forensic samples. For this reason, high level features have previously been recommended for use in forensic systems, since they tend to be more robust to the acoustic variability introduced by recording conditions [1]. A drawback of high level features though is their poor performance relative to low-level cepstral features. We suggest, however, it may be possible to improve the performance of high feature systems by combining acoustic and idiolectal information, and this may deliver a better trade-off with respect to robustness, interpretability and discrimination performance. In this paper we evaluate a likelihood ratio-based (LR) forensic voice comparison (FVC) system fusing two high level feature subsystems: word n -grams and long-term fundamental frequency ($LT F_0$). Preliminary experiments demonstrate some promising performance gains. We also examine how the duration of speech data impacts on this proposed system.

Index Terms: high level features, forensic phonetics, forensic voice comparison, n -grams, likelihood ratios.

1. Introduction

High level features are based on linguistic or long-term information [2]. These feature types are attractive to forensic applications because of their relative robustness to acoustic variability introduced by transmission effects, background noise and other artefacts of recording [3], which are almost inevitable in forensic recordings. High level features are also thought to be more easily interpreted by juries and lay-persons than low-level cepstral-based features [4]; perhaps since linguistic features are accessible to perception, more conducive to visual representation [1], and can be related to aspects of speech production in a fairly straight forward manner [5]. Interpretability is important in many jurisdictions, the admissibility of evidence based on automatic features has been questioned in some on the basis of its lack of interpretability of cepstral features [6]. The trade-off is performance. High level feature systems are less discriminative relative to low-level cepstral-based ones [7], and appear best employed to provide complementary information to cepstral-based systems [8]. In automatic speaker recognition (ASR) systems quantification of high level information is based mainly on discrete representations of speech derived from an orthographic or phone transcription of recordings e.g. word [9] or phone n -grams [3], or stylised F_0 contours [10][11]. High level information can be also quantified continuously via region conditioned cepstra [2] or acoustic-phonetic features (e.g. formant frequencies, fundamental frequency). Use of acoustic-phonetic features is not common in ASR, but is one of the main approaches to likelihood ratio-based forensic voice comparison (LR-based FVC) analyses [12]. Likelihood ratios are increas-

ingly widespread in the forensic comparison sciences. This includes applications in DNA [13], fingerprint [14][15], gasoline particle [16] and illicit drug [17] comparison; there is a large body of literature demonstrating LR based approaches can be effectively applied to voice evidence too (e.g. [4] [18][19]). The reasons the LR is increasingly regarded as fundamental to forensic science inference are described in several works e.g. [20] [21] [22] [23] [24].

$LT F_0$ has been previously evaluated using LR-based approaches [4] [25]. Despite possessing many desirable characteristics forensically (it's availability in speech, ease of extraction and robustness to the effects of telephony [4]), F_0 is also subject to large degree of within-speaker variation, which limits its discriminatory power [4]. The use of idiolectal based features, such as lexical frequency, has received almost no attention in LR-based FVC. This paper examines the performance of a high level FVC system combining $LT F_0$ and word n -grams. We first examine the performance of the systems separately, and then compare with that of a fused system. Finally, we test the effect of the amount of net speech on performance.

2. Experiments

2.1. Word-level n -gram system

2.1.1. Data

This study used spontaneous conversational speech from 90 male speakers recorded on two separate occasions. Speakers were selected from a speech database of Australian English speakers built for FVC research [26]. A detailed description of collection procedures and tasks can be found here [27]. The recordings are approximately 10 minutes in duration with 3-5 minutes of speech available for each conversation participant. For these experiments only the first 3 minutes of each participant's recording was orthographically transcribed. Data was partitioned into test, training and reference datasets with 30 speakers per partition, two recordings per speaker. The training data was used to calculate weights for the logistic-regression calibration/fusion procedure [28]; the reference data to estimate the typicality term in the feature-based LRs; and the test data to simulate same- and different-speaker forensic voice comparisons. Given the relatively small number of speakers, 3-fold cross-validation was applied in these experiments.

2.1.2. Feature extraction

Punctuation was removed from the raw transcriptions and all characters converted to lower case, but no stemming was applied so as to maintain morpho-syntactic idiosyncrasies [29]. The transcription for each speaker and session was tokenised and used to construct bag-of-grams models. The top N most frequently occurring n -grams included in the models was incremented by 25, up to 1000. This resulted in 40 different feature

vectors (one each for 1- and 2-gram). The FVC performance based on each of the resulting feature vectors was separately examined to determine the optimum number of feature to be included in the system. Since it is inevitable that some 1-gram and 2-gram counts will be zero, resulting in zero probabilities for unseen words, smoothing is essential. For simplicity, Laplace (add-one) smoothing was applied for these preliminary experiments. The resulting smoothed unigram and bigram speaker vectors were used as inputs for LR estimation.

2.1.3. LR estimation & evaluation

LRs were estimated using multinomial LR models, which are a natural choice for feature-based LR estimation using multivariate count data. We evaluated two approaches for estimating the multinomial model parameters: one- and two-level. In the one-level model, the parameters were obtained by means of the maximum likelihood approach with the background data. In the two-level model, the prior can be assumed for the model parameters, and the conjugate prior for the multinomial is a Dirichlet model. The hyper parameter in our model was obtained from the reference data. A detailed exposition of multinomial LR models can be found in [30]. All likelihood ratios were calibrated using a logistic regression procedure [31]. We evaluated performance using the log-likelihood ratio cost function (C_{l_r}) [28], which is a gradient measure of accuracy used for evaluating LR-based FVC systems [32]. The greater the magnitude of counterfactual LR, the higher the penalty C_{l_r} applies. A value of 0 indicates perfect accuracy, which deteriorates as C_{l_r} approaches 1. C_{l_r} values < 1 indicate that the system is delivering useful evidential information. C_{l_r} of 1 means the evidence is not informative and $C_{l_r} > 1$ means the system is better off without it. C_{l_r} can be decomposed of two components $C_{l_r}^{min} + C_{l_r}^{cal}$ which give measures of discrimination and calibration loss respectively.

2.1.4. Results

Table 1. shows C_{l_r} and its decomposed values ($C_{l_r}^{min}$, $C_{l_r}^{cal}$) for the best performing features for each model variant and n -gram size.

Table 1: Performance: Log-likelihood cost using 180 seconds of data

n -gram	model	C_{l_r}	$C_{l_r}^{min}$	$C_{l_r}^{cal}$
1-gram	one-level multi LR	0.57	0.47	0.10
	two-level multi LR	0.49	0.42	0.07
2-gram	one-level multi LR	0.75	0.64	0.11
	two-level multi LR	0.57	0.49	0.09

Results show that the two-level multinomial LR outperforms the one-level multinomial LR model, with the greater difference seen for the 2-gram feature vector ($C_{l_r} = 0.75$ vs. $C_{l_r} = 0.57$). This is not unexpected, since the two-level models also takes into account prior information about variation of the parameters in the reference population.

Another performance characteristic we observe is the 1-gram consistently outperforms to the 2-gram vectorisation. This is somewhat surprising, since we might expect information about a speaker's sequential word usage to be more individualising than individual words (i.e. 1-gram). However, it is a result that is consistent with authorship attribution studies [33].

There are also differences between 1- and 2-grams in terms of the strength of evidence obtained, as seen in the Tippett plots in Figure 3. Panels A and B present plots for the best performing (i.e. the two-level multinomial) 1- and 2-gram models. $\text{Log}_{10}LRs$ are shown on the x -axis and the cumulative proportion of same- (SS) (solid orange lines) and different-speaker (DS) (solid blue lines) trials on the y -axis. The 1-gram model yields greater magnitude LRs for the DS relative to the 2-gram model, the opposite situation LRs occurs for the SS comparisons.

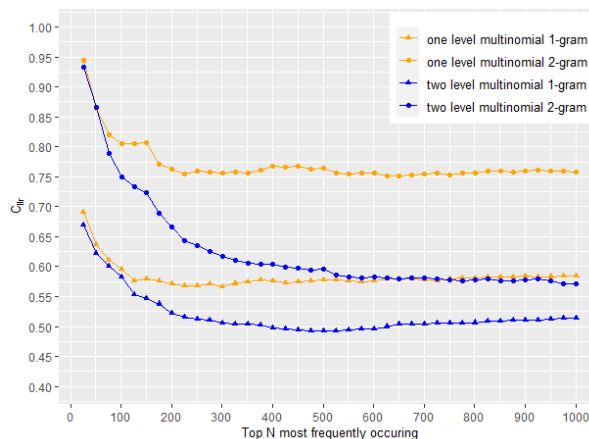


Figure 1: Performance (C_{l_r}) as function of n -gram frequency

Next, we examine the effect of n -gram frequency on discrimination performance. We tested performance as a function of the top N most frequently occurring n -grams from the corpus. These were incremented in steps of 25 up to 1000. Figure 1 shows C_{l_r} (y -axis) plotted as a function of the top 1000 most frequently occurring n -grams (x -axis). Multinomial model variants are plotted (one-level = orange, two-level = blue), as well as 1- and 2-grams (triangles and circles respectively). Similar effects are observed across the word n -gram models. There is a rapid improvement in C_{l_r} values (i.e. in the discrimination accuracy) as N increases, before performance plateaus and begins to degrade; notably for the 2-gram parametrisation this effect is delayed. This result is broadly consistent with [9].

2.2. $LT F_0$ system

2.2.1. Data

The same 90 speakers and recordings were used for the acoustic-phonetic system. We also maintained the same data partitions for test, training and reference data, so the results obtained are comparable to the previous system.

2.2.2. Feature extraction

$LT F_0$ can be parametrised simply via the mean and standard deviation [18], however incorporating additional information from a speaker's $LT F_0$ distribution has been shown to improve FVC performance [19]. In these experiments we test both parametrisations. In the feature extraction stage, silences were automatically removed via a voice activity detection algorithm based on short-term energy. Speaker's F_0 is extracted using an autocorrelation method described in [34] with pitch floor and ceiling set between 75 Hz and 400 Hz. Raw F_0 values were then fitted to a kernel density distribution. After [19] feature vectors are constructed containing six parameters characterising

speaker’s $LT F_0$: mean, standard deviation (sd), kurtosis, skew, modal F_0 ($modal f_0$), maximum probability density ($maxpdf$). Examples of kernel density estimates for two speakers from the database (recordings 1 and 2) are shown in Figure 2. There are some obvious within- and between-speaker differences in shape of the speaker’s F_0 distributions.

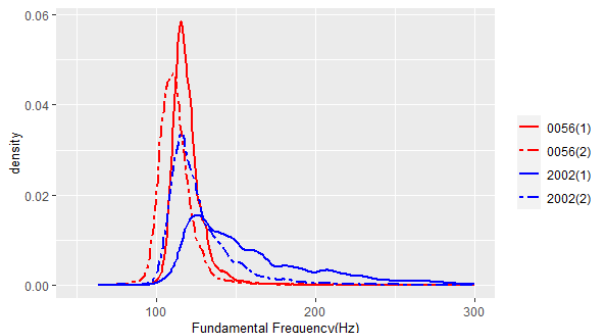


Figure 2: $LT F_0$ distribution from speakers 0056 and 2002 from 180 seconds of net speech. Solid lines = recording session 1, dot-dash = session 2.

First we look at the difference between the two speakers (Speakers 0056 and 2002). The two distributions from Speaker 0056 have much higher peaks than those of Speaker 2002. This indicates that Speaker 2002 generally speaks more monotonically (in a narrow pitch range) than Speaker 2002. Comparisons between the two distributions from each speaker, we also see that there is considerable within-speaker variability. The two distributions from Speaker 0056 both have relatively narrow and peaky shapes, but their modal F_0 is quite different from one occasion to another. With Speaker 2002, within-speaker variability is even greater, as two distributions have markedly different shapes. The first recording from this speaker shows a much wider pitch range, suggesting that Speaker 2002 spoke with much wider pitch variation – perhaps a more animated fashion – on this occasion. How well speakers can be discriminated on the basis of differences in their $LT F_0$ distributions will depend to a large extent on whether the within-speaker variation across the two occasions is sufficiently contained relative to the between-speaker variation.

2.2.3. LR estimation & evaluation

The distributional parameters described in the previous section were used as inputs for a multivariate kernel density likelihood ratio estimation (MVKD) procedure [35] widely used in acoustic-phonetic approaches to FVC. MVKD models within-speaker variance via a Gaussian distribution and while between-speaker variance is modelled via a kernel density function. A detailed mathematical exposition is in [35]. Again, logistic regression calibration was applied to the raw LRs, the log-likelihood ratio cost function used to evaluate performance and results are 3-fold cross-validated.

2.2.4. Results

Results in (Table 2.) show $LT F_0$ yields considerably weaker accuracy (C_{Ur}) relative to the best performing word-level n -grams system in the previous section (0.73 vs. 0.49).

The C_{Ur} differences between using sd and mean vs. all six parameters appear negligible. Discrimination loss is less for

Table 2: Performance: log-likelihood cost 180 seconds of data. all = mean, sd , $maxpdf$, $modal f_0$, $kurt$, $skew$.

features	model	C_{Ur}	C_{Ur}^{min}	C_{Ur}^{cal}
mean + sdev	MVKD	0.74	0.58	0.16
all	MVKD	0.73	0.52	0.20

the latter ($C_{Ur}^{min} = 0.52$ vs. 0.58), but this comes with poorer calibration (C_{Ur}^{cal}). The Tippett (Panel C, Figure 3) shows the $LT F_0$ system yields poor strength of evidence for the same-speaker hypotheses, as the low magnitude of the LRs shows (e.g the largest magnitude SSLR is only $Log_{10} LR = 1.5$); its LR is slightly stronger for the DSLRs but a large portion of these (about half) yield contrary-to-fact LRs. The relatively poor discrimination performance is more or less consistent with previous FVC studies using $LT F_0$ [18] [19] [36]; in particular where only a few minutes of speech data are used.

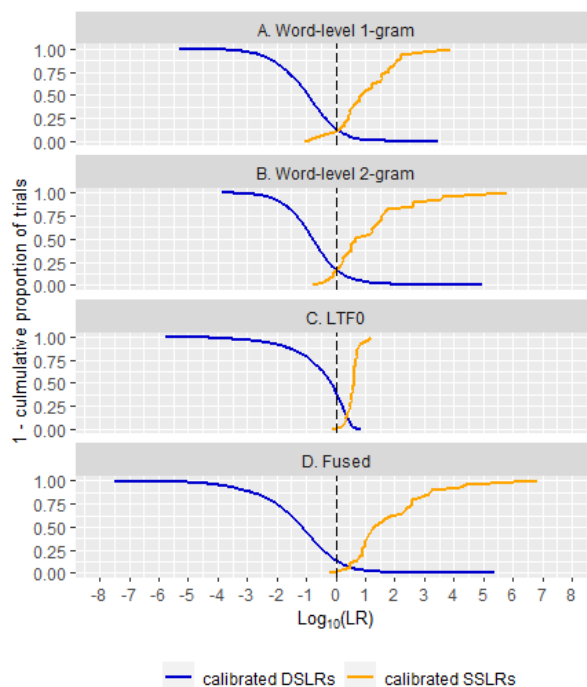


Figure 3: Tippett plots. Blue lines are different-speaker trials (DSLRs) and orange lines are same-speaker trials (SSLRs)

2.3. Fused system

Now we fuse the systems evaluated in the previous sections. Logistic-regression fusion [28] is a procedure that combines parallel sets of raw likelihood ratios (scores) from different forensic comparison systems [31]. Logistic-regression fusion simultaneously calibrates sets of scores, as well as discounts the correlation between scores in the process of fusing. While it is not expected that $LT F_0$ and word n -grams will be correlated, we do expect 1- and 2-gram to be. Logistic-regression fusion provides a convenient way of calibrating the $LT F_0$ scores and dealing with the n -gram correlation simultaneously. The logistic-regression model is trained using a set of known speakers, which should be a dataset independent of the test and back-

ground model data. As mentioned earlier we reserved a set of 30 speakers in the training database for this purpose. Coefficient values from the logistic regression fitted to the training data provide the linear weights. These are then used to fuse the best performing $LTFO$ and best performing word-level n -gram systems. Table 3 presents the results from the fused system. It shows that a modest performance improvement can be achieved by fusing the $LTFO$ system to word-level n -gram systems (C_{U_r} : $LTFO$ 0.73, 1-gram 0.49, fused 0.44).

Table 3: Performance metrics for the fused system

features	C_{U_r}	$C_{U_r}^{min}$	$C_{U_r}^{cal}$
word-level n -grams + $LTFO$	0.44	0.31	0.13

Although we only witness a modest improvement in the fused system relative to the best word-level n -gram system (1-gram = 0.49 vs. fused = 0.44) comparison of the Tippett plots in Figure 3 show a considerable improvement in magnitude of LRs in both same-speaker (SSLRs) and different-speaker (DSLRS) trials in the fused system (Figure 3, Panel D). There is the possibility however that the strength of evidence is overstated. Given the relatively small amount of speakers, sampling variability may be affecting the precision of the LR estimates [37]. It may be possible to resolve this in future work by shrinking the estimated LRs [37].

2.4. Net speech effects

Three minutes of speech is relatively generous in a forensic context. Forensic speech samples are often far shorter in real world situations. How well the two subsystems and fused systems perform as a function of the amount of net speech available is therefore of interest. To investigate this, we repeated the experiments using the best performing models, but varied the amount of net speech used to train the models (between 30 and 180 seconds, in 30 second increments). The results are shown in the barplot in Figure 4. C_{U_r} values are plotted on the y-axis as a function of net speech for the 1-gram and 2-gram (multinomial LR) systems, $LTFO$ and fused system. Not surprisingly, in all four systems performance positively correlates with the duration of the available speech. With the 30 second increment, fusing does not improve the performance of 1-gram and 2-gram systems. This is simply because of the very poor performance of $LTFO$; its C_{U_r} is > 1 , indicating that discrimination accuracy is worse than random. With all the other durations, however, the best result is achieved by the fused system. This suggests two things: combining high level features is a promising way forward; and net speech of 30 seconds is too short to use $LTFO$ in speaker characterisation, at least for Australian English.

With the exception of 30 second increment the fused system always offers an improvement in performance over the individual systems. We also observed that the performance of $LTFO$ in this increment is furthermore highly dependant on the speakers which make up the reference, training and test data. The C_{U_r} across iterations of the cross validation for $LTFO$ was highly variable ranging from 0.83 to 1.44. Moreover, while 1- and 2-gram subsystems steadily improve as a function of additional data, there is $LTFO$ performance characteristics are less stable. For example, $LTFO$ has a lower C_{U_r} value at 90 seconds vs. 120; at 180 performance is more or less the same.

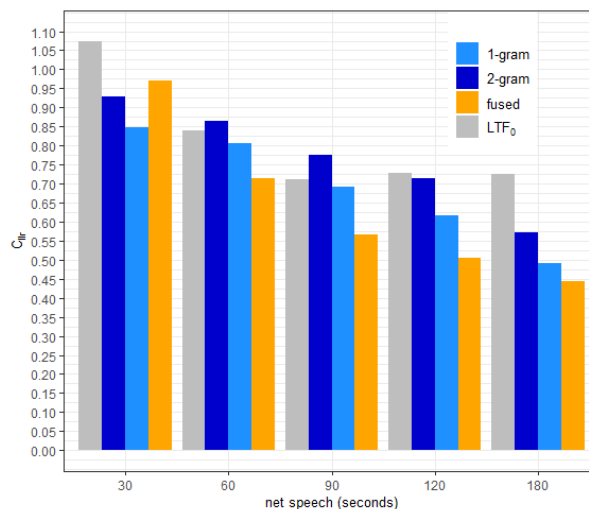


Figure 4: Barplot of C_{U_r} values as function of net speech data used in the acoustic-phonetic, n -gram and fused FVC systems

3. Conclusions

This paper has demonstrated that fusing features not conventionally combined – acoustic-phonetic features and word n -gram – could bring some improvements in the accuracy of LR-based FVC systems. Although $LTFO$ performance becomes less stable with short speech, word n -grams appear relatively robust against reduction of speech duration. Thus, while fusion almost always improves overall accuracy, we should not fuse $LTFO$ with the word n -grams where sufficient duration of speech material is not available. More generally, based on these results the inclusion of $LTFO$ in fused high level feature FVC is perhaps questionable, given its lack of stability. However, further investigation of impact of duration and other factors known to affect F_0 , such as speaking style, language, background noise and emotion to determine the most effective use of F_0 forensically. It may be better to pursue other acoustic-phonetic features, such as formant frequencies. This is an area for future investigation. Another is to investigate the robustness of high level feature systems to mismatches between speaking style. This is particularly pertinent in forensic contexts where this kind of mismatch frequently occurs between a recording of a suspect made in an interview vs. an offender’s telephone conversation.

4. Acknowledgements

We would like to thank the reviewers for their valuable comments. The first author is supported by an Australian Government Research Training Scholarship and ANU Supplementary Scholarship.

5. References

- [1] E. Shriberg and A. Stolcke, “The case for automatic higher-level features in forensic speaker recognition,” in *Proc. INTERSPEECH 2008*, Brisbane, Australia, September 2008, pp. 1509–1512.
- [2] E. Shriberg, “Higher-level features in speaker recognition,” in *Speaker Classification I: Fundamentals, Features, and Methods*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Springer, 2007, vol. 4343, pp. 241–259.
- [3] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek,

- “Phonetic speaker recognition with support vector machines,” in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. MIT Press, 2004, pp. 1377–1384.
- [4] P. Rose, *Forensic speaker identification*. London: CRC Press, 2002.
- [5] —, “Technical forensic speaker recognition: Evaluation, types and testing of evidence,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 159–191, 2006.
- [6] T. Coy, V. Hughes, P. Harrison, and A. J. Gully, “A Comparison of the Accuracy of Disen and Keshet’s (2016) DeepFormants and Traditional LPC Methods for Semi-Automatic Speaker Recognition,” in *Proc. INTERSPEECH 2021*, Brno, Czech Republic, Aug./Sep. 2021, pp. 406–410.
- [7] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
- [8] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt, “The contribution of cepstral and stylistic features to SRI’s 2005 NIST speaker recognition evaluation system,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.
- [9] G. Doddington, “Speaker recognition based on idiolectal differences between speakers,” in *Proc. EUROSPEECH*, P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, Ed., Aalborg, Denmark, September 2001, p. 2521–2524.
- [10] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, “Modeling prosodic dynamics for speaker recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03)*, vol. 4. IEEE, 2003, pp. IV–788.
- [11] M. K. Sönmez, E. Shriberg, L. P. Heck, and M. Weintraub, “Modeling dynamic prosodic variation for speaker verification,” in *IC-SLP*, vol. 7, 1998, pp. 3189–3192.
- [12] E. Gold and P. French, “International practices in forensic speaker comparison,” *International Journal of Speech Language and the Law*, vol. 18, no. 2, 2011.
- [13] I. W. Evett and B. S. Weir, *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sinauer Associates Sunderland, MA, 1998, vol. 244.
- [14] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthoiz, and A. Bromage-Griffiths, “Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae,” *Journal of Forensic Sciences*, vol. 52, no. 1, pp. 54–64, 2007.
- [15] D. Ramos, R. Haraksim, and D. Meuwly, “Likelihood ratio data to report the validation of a forensic fingerprint evaluation method,” *Data in brief*, vol. 10, pp. 75–92, 2017.
- [16] P. Vergeer, A. Bolck, L. J. Peschier, C. E. Berger, and J. N. Hendrikse, “Likelihood ratio methods for forensic comparison of evaporated gasoline residues,” *Science & Justice*, vol. 54, no. 6, pp. 401–411, 2014.
- [17] A. Bolck, H. Ni, and M. Lopatka, “Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison,” *Law, Probability and Risk*, vol. 14, no. 3, pp. 243–266, 2015.
- [18] Y. Kinoshita, “LR estimation using long term F0 as a parameter: good, bad or useless? Initial investigation using Japanese data,” in *Proc. SST 2004*, 2004, pp. 498–503.
- [19] Y. Kinoshita, S. Ishihara, and P. Rose, “Beyond the long-term mean: exploring the potential of f0 distribution parameters in forensic speaker recognition,” in *Odyssey 2008*. International Speech Communication Association, 2008, pp. 1–8.
- [20] G. S. Morrison, “Forensic voice comparison and the paradigm shift,” *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.
- [21] I. W. Evett, “Towards a uniform framework for reporting opinions in forensic science casework,” *Science & Justice*, vol. 3, no. 38, pp. 198–202, 1998.
- [22] I. W. Evett, C. Berger, J. Buckleton, C. Champod, and G. Jackson, “Finding the way forward for forensic science in the us—a commentary on the pcast report,” *Forensic Science International*, vol. 278, pp. 16–23, 2017.
- [23] C. Champod and D. Meuwly, “The inference of identity in forensic speaker recognition,” *Speech Communication*, vol. 31, no. 2-3, pp. 193–203, 2000.
- [24] J. Buckleton, “A framework for interpreting evidence,” *Forensic DNA evidence interpretation*, pp. 27–63, 2005.
- [25] Y. Kinoshita, S. Ishihara, and P. Rose, “Exploring the discriminatory potential of f0 distribution parameters in traditional forensic speaker recognition,” *International Journal of Speech, Language & the Law*, vol. 16, no. 1, 2009.
- [26] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, “Forensic database of voice recordings of 500+ Australian English speakers, 2015,” URL: <http://databases.forensic-voice-comparison.net/>.
- [27] G. S. Morrison, P. Rose, and C. Zhang, “Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice,” *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155–167, 2012.
- [28] N. Brümmner and J. Du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.
- [29] A. Omar and W. I. Hamouda, “The effectiveness of stemming in the stylometric authorship attribution in arabic,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 116–121, 2020.
- [30] A. Bolck and A. Stamouli, “Likelihood Ratios for categorical evidence; Comparison of LR models applied to gunshot residue data,” *Law, Probability and Risk*, vol. 16, no. 2-3, pp. 71–90, 2017.
- [31] G. S. Morrison, “Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio,” *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, 2013.
- [32] —, “Measuring the validity and reliability of forensic likelihood-ratio systems,” *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011.
- [33] E. Stamatatos, “A survey of modern authorship attribution methods,” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.
- [34] P. Boersma *et al.*, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proc. of the Institute of Phonetic Sciences*, vol. 17, no. 1193. Citeseer, 1993, pp. 97–110.
- [35] C. G. Aitken and D. Lucy, “Evaluation of trace evidence in the form of multivariate data,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 53, no. 1, pp. 109–122, 2004.
- [36] P. Rose and C. Zhang, “Conversational style mismatch: its effect on the evidential strength of longterm f0 in forensic voice comparison,” *Proc. ASSTA 2018*, pp. 157–160, 2018.
- [37] G. S. Morrison and N. Poh, “Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/bayes factors,” *Science & Justice*, vol. 58, no. 3, pp. 200–218, 2018.