Likelihood Ratio-based Forensic Semi-automatic Speaker Identification with Alveolar Fricative Spectra in a Real-world Case.

Phil Rose

Independent Researcher, Australian National University Emeritus Faculty

https://philjohnrose.net

Abstract

A real-world forensic voice identification is described in a case involving the blowing-up of a car, three suspects, and a miniscule amount of speech evidence. Necessary stages in the estimation of a likelihood ratio are described, based on the bandlimited cepstral spectral acoustics of two alveolar fricatives /s/ and /z/ in the questioned utterance. The speech evidence is shown to be very much more likely assuming one of the suspects said the utterance.

Index Terms: Forensic voice comparison, alveolar fricative, Bayes' theorem, bandlimited segmental cepstrum, validation.

1. Introduction

One evening in 2017 three young men, K M and C, drove to the Australian town of Wagga Wagga where, with apparently no more sinister motive than amusement, they blew up a stationary unoccupied car belonging to an acquaintance. With a view to posting on social media (!), the incident was videoed from within their car, with verbal commentary, on a mobile phone. At the time the recording starts, it was agreed that two of the three were inside their stationary car looking on, and the offender was outside laying the timed explosive device. The offender then rejoined the other two and the video captures the explosion through the rear window as they drive away. The audio contains recordings of several short utterances, both before and after the time the offender returned to the car.

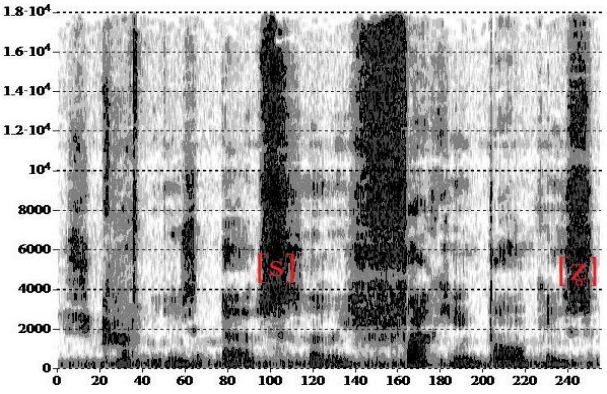

Figure 1: *Spectrogram to 18k of questioned utterance* this cunt is absolutely fucking bonkers *showing durational and spectral extent of noise from alveolar fricatives* [s] and [z̪]. *X-axis = duration (csec.), y-axis = frequency (Hz).*

When questioned by police, each participant denied laying the explosives. C maintained he had sat in the rear seat of the car making and commenting on the video. M said he had sat silent in the front while K made the video. K said he sat in the front seat, watching C make the video. The Wagga police specifically wanted to know, of course, whose was the voice on the video before the offender returned the car.

In addition to the fact that the aim is to test exoneration rather than inculpation, this case is unusual and worth documenting for several other reasons. Because the questioned speaker was one of three, an initial estimate of the prior probabilities is easy and the mandating authority can be instructed as to how Bayes' theorem can be used to estimate a posterior probability. The case also shows how – sometimes! – forensic voice comparison is possible with a miniscule amount of questioned data backed up by phonetic knowledge. The parametrisation, with a bandlimited segmental cepstrum, involved a nice demonstration of the power of combining acoustic phonetics and signal processing. Finally, it shows the crucial element of validation – demonstrating that your system actually does what you claim it to do. All the case details may be found in the anonymised full report at [1].

2. Questioned Data

Logically, the only material relevant to the voice comparison were two short utterances recorded in the absence of the offender, before he returned to the car (there were no utterances after the offender returned to the car which were incriminating by content). The second utterance was said *sotto voce* and was of no use. The first utterance – *This cunt is absolutely fucking bonkers* – contained no usable vocalic material. Likelihood ratio-based testing in [2] showed all its stressed vowels – /ae/ and / ʉː/ in *absolutely*, /a/ in *cunt* and *fucking*; and /o/ in *bonkers* – to have rather poor expected evidential strength, with equal error rates ranging from 25% (/æ/) to 36.6% (/o/). However, previous research into the speaker-identification potential of fricatives with various parametrisations [3-8] suggested a forensic voice comparison might be feasible using the spectra of /s/ and /z/ in *absolutely* and *bonkers*. (Subsequent work [9-12] has confirmed this).

Figure 1 shows a 60dB dynamic range wide-band spectrogram to 18.5 kHz of *This cunt is absolutely fucking bonkers*, which can be seen to last for about 2.5 seconds. The /s/ is at ca. csec. 100; the /z/, which shows typical word-final devoicing, at ca. csec. 240. Both last for about 10 centiseconds. The high intensity noise centered at csec. 150 is from an emphatically produced /f/ in *fucking*. The fricative energy extends to about 18 kHz, and well-defined vocalic formant structure can be seen below 4 kHz.

Figure 2 shows the questioned alveolar spectral acoustics (FFT and 14th order LPC) to 8 kHz. The tokens are unremarkable, showing spectral properties predicted from the acoustic theory of speech production [13 pp.379-389, 398-403]. The peak between 3 and 4 kHz is the quarter-wavelength resonance of the front cavity, the peak around 6 kHz is the half-wavelength resonance of the constriction. Even the back cavity resonance is visible just below 2 kHz, attesting to the quality of the recording.

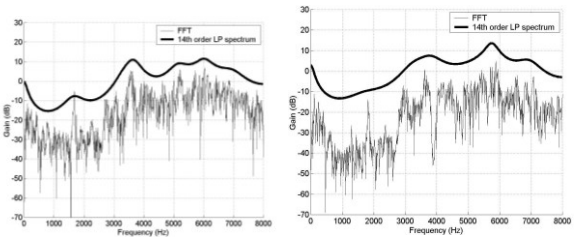

Figure 2: *Spectral acoustics of questioned /s/ (left), and /z/. X-axis = frequency (Hz.), y-axis = gain (dB).*

3. Suspect Data

Recordings were available from C K and M during their police interviews. M's recording was audibly slightly echoic in places, otherwise, for once!, the recordings were of rather good quality. Tokens of /s/ and /z/ in the suspects' speech during their police interviews were identified. To preserve comparability with the questioned tokens, suspects' tokens in the vicinity of rounded vowels were omitted to avoid coarticulation effects from lip-rounding, which will in particular lower the resonance associated with the front cavity. Because of the well-known utterance-final conditioning of duration and voicing of voiced fricatives in English, only utterance-final tokens of /z/ were selected. One of the forensically useful aspects of /s/ and /z/ is that they occur frequently in speech: the recordings of C K and M contained 48, 43 and 65 useable tokens of /s/, and 21, 34 and 16 useable tokens of /z/ respectively.

4. Processing

Suspects' /s/ and /z/ tokens were extracted and saved with *Praat*. Figure 3 is an example of one of C's prepausal /s/ tokens showing the high frequency noise portion saved. As can be seen, the broadband energy between about 3 and 7 kHz lasts for about 20 centiseconds.

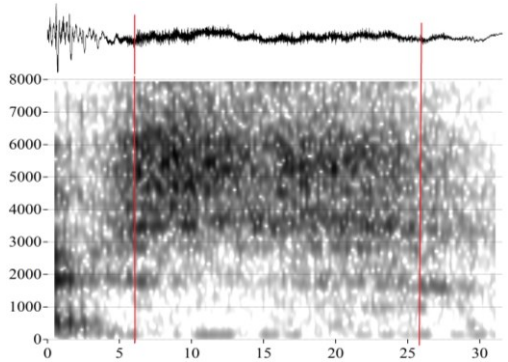

Figure 3: *Spectrogram and waveform of a prepausal /s/ token showing portion of high frequency energy associated with the fricative noise. X-axis = duration (csec.), y-axis = frequency (Hz)*

The tokens were then further processed in Matlab, where a set of eight linear prediction cepstral coefficients was extracted. A hamming widow was applied over the whole of the fricative: in this way one obtains a good estimate for the central portion of the noise. This so-called *segmental cepstrum* procedure smooths the spectrum to an extent which facilitates forensic voice comparison [6 8 10 11]. Figure 4 shows the resulting 8th order cepstral spectra of the individual /s/ and /z/ tokens of each of the three speakers (in blue), and their mean cepstral spectrum (in red). The cepstral spectra of the questioned /s/ and /z/ are plotted in black. Figure 4 shows some clear spectral differences between the speakers. With a prominent peak at about 4 kHz, K's spectra are the most different. This may be related to the fact that he sometimes whistled his /s/ and /z/. This peak will mean that K will have a relatively large second cepstral coefficient. C and M's spectra are more similar, but still differ in amplitude range, with C having a greater range than K. This will be reflected in the amplitude of the first cepstral coefficient: C's will be bigger. The questioned spectra are visually closer to C than to K and M.

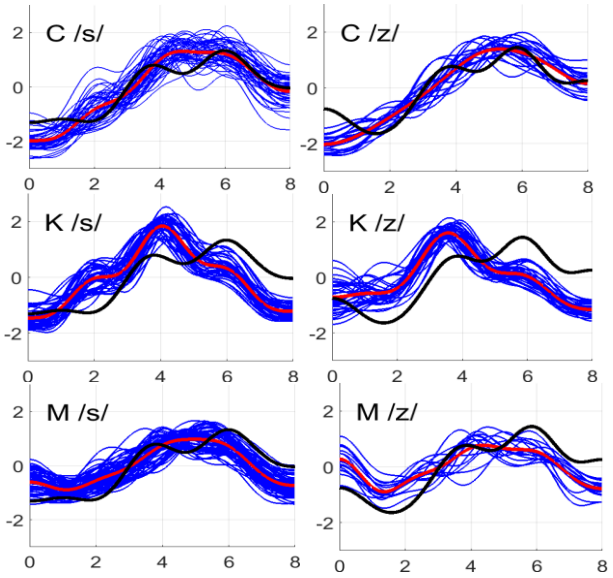

Figure 4: *Cepstral spectra for known and questioned tokens. Blue = individual known tokens, red = known mean, black = questioned. X-axis = frequency (kHz.), y-axis = gain (dB).*

5. Evaluation of Evidence

We know the questioned utterance (U*q*) comes from one of three speakers, thus three separate hypotheses – H_C H_K H_M – must be evaluated in this case: that U*q* was said by suspect C, K and M respectively. This is, of course, a classic identification task, but from a forensic point of view its solution is not simply a question of finding the speaker who has the closest acoustics to U*q*. As is by now well-known, the legally and logically correct approach must be to estimate the strength of evidence, *aka* likelihood ratio, for the three hypotheses [14 15]. The likelihood ratio (LR) is the ratio of the probabilities of the evidence under competing hypotheses: it quantifies how much more likely you are to get the evidence under one hypothesis than under an alternative. If you have the strength of evidence and know the probability of the hypothesis before that evidence is adduced, it is possible to estimate, with Bayes' theorem, what everyone wants to know: the probability of the hypothesis, given the evidence.

With all eight cepstral coefficients the results were pretty unequivocal. The LR for /s/ assuming the unknown speaker was C was 29.6; for K and M it was 6e-09 and 6e-04 respectively, indicating that the questioned /s/ spectrum is far more likely assuming it had come from C than the other two. However, for several reasons these results probably overestimate the strength of evidence that can reasonably inferred. These reasons are addressed below.

6. Validation

The first step in estimating the strength of evidence is validation: determining how reliable one's system is in delivering an accurate LR. The question is especially obvious in this case. After all, just how reliable can a system actually be that uses comparisons involving *only a single questioned [s] and [z] token*?! Validation is, or should be, an essential part of any forensic case work [16 17].

Validation is done by the simple, time-honoured method of seeing what happens with *known* data in circumstances as close as possible to those of the actual case. The car recordings could not be used as there was of course no indication of who said what. Therefore, the suspects' police interviews were used and for each of the three speakers, one of their [s] tokens was extracted in turn to serve as the single questioned datum, and its LR estimated as the ratio of two values: (1) the probability of getting the token's cepstral spectrum assuming it had come from the given speaker (i.e. a known same-speaker comparison); and (2) the probability of getting the token's spectrum assuming it had come from either of the other two speakers (i.e. a known different-speaker comparison). Some additional preprocessing was required to improve the accuracy of the strength of evidence estimate. This is now described.

6.1. Bandlimiting

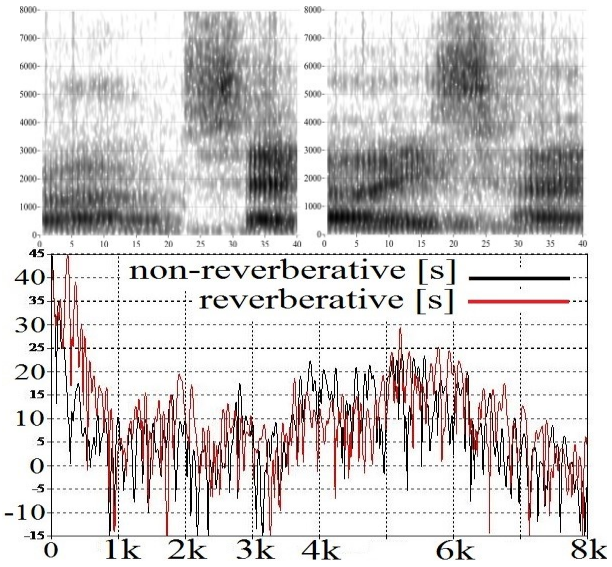

Figure 5: Top = *Spectrograms to 8 kHz of two of M's [s] tokens illustrating possible reverberation in right panel. Bottom = FFT of reverberative and non-reverberative [s] tokens. X-axis = duration (csec.)*

As already mentioned, although the police recording of C and K was of good quality, M's police recording was audibly more echoic. This has a potentially deleterious effect on the spectrum which should be controlled for. Figure 5 illustrates this. Its top left panel shows a spectrogram of one of M's [s] tokens (in *I said*) that had no clear reverb. The top right panel shows a spectrogram of one of his [s] tokens (in *me saying*), showing reverberation from at least the F1 of the preceding [i] in *me* continuing into the [s]. The bottom panel compares the FFT spectra of the two [s] tokens, where it can be seen that the token with putative [i] F1 reverberation appears to have higher energy over about the first kilohertz. This needs to be controlled for, otherwise it might make M's /s/ tokens appear more different from the questioned recording than he actually is, thus favouring identification with C or K. Therefore, firstly, /s/ tokens from M were chosen that sounded minimally reverberative. Secondly, comparisons were done using cepstral coefficients extracted over the range between 1 and 8 kHz, thus removing the spectral portion below 1 kHz where reverberation might have the greatest effect. These so-called *bandlimited cepstral coefficients* (blccs) allow for parametric specification of any cepstral sub-band within the Nyquist interval [18 - 21]. They have great potential in forensic voice comparison because they allow one to focus on the frequency ranges suspected to contain the most speaker-specific information.

6.2. Dimensionality

Estimating the likelihood ratio with a segmental cepstrum involves estimating the probability density of multivariate data. The accurate estimation of a probability density depends crucially on the number of observations, and the more dimensions, the more observations are needed to be sure that all parts of the high dimensional distribution are adequately sampled. In order to adequately sample a 7-dimensional object one would require a hair-raisingly high number of about 43,700 /s/ tokens [22 p.94]. The highest dimensionality supported with the available sample size – between about 40 and 65 /s/ tokens – is three [22 p.94]. This means that the safest choice is one based on comparison with three blccs, and only the first three were used.

6.3. Calculation and calibration of LR-scores

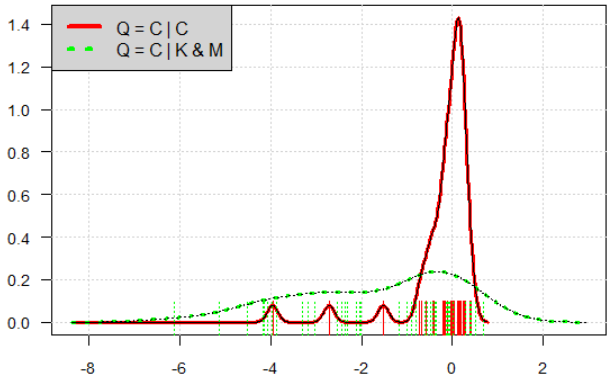

Figure 6: *Kernel density probability density distributions and rugplots for known same-speaker (red) and different-speaker (green) $\log_{10}$ LR-scores for /s/ spectrum validation assuming C is suspect. X-axis = $\log_{10}$LR, y-axis = probability density.*

As cepstral coefficients are distributed normally and the data is multivariate, LR-scores were estimated using *R*'s *mvdnorm*. (*LR-score* is a useful term for an uncalibrated measure of distance that takes both similarity and typicality into account). Thus for each /s/ token of a given speaker, the multivariate probability of its three blccs was estimated assuming it had come from that speaker, and assuming it had come from either of the other two. The LR-score was the $\log_{10}$ ratio of these multivariate conditional probabilities. Given the number of known tokens, this involved upwards of 40 same-speaker and 1560 different-speaker comparisons for /s/, and upwards of 16 same-speaker and 120 different-speaker comparisons for /z/. Figure 6 shows the results of this process treating C as the suspect. The red line (its peak centered

around 0) indicates the probability density distribution for same-speaker LR-scores, i.e. when single questioned /s/ tokens from C were compared with C's distribution. The green line indicates the probability density distribution for the different-speaker LR-scores, i.e. when single questioned /s/ tokens from C were compared with the distributions from K & M. It can be seen that the same-speaker LR-scores are generally bigger than the different-speaker LR-scores. In particular, the probability of getting a $\log_{10}$LR-score between about +/- 0.5 is very much greater if the token had come from C than from the other two. It is also obvious, however, that there is considerable overlap, with counterfactual LRs. There are even two different-speaker comparisons which have LR-scores actually higher than any of the same-speaker comparisons! This variability, perhaps, is to be expected when you are only looking at a single token, rather than the mean of several, but the reality of the case is such that only a single token is available, and so the validation has to reflect that.

The same-speaker and different-speaker LR-scores were then calibrated to convert them to likelihood ratios [23 24]. Two different calibrations were applied. One used the PAV algorithm in the *Focal* tool-kit [25]. In the other, the log-odds of the same-speaker and different-speaker scores were estimated over the range of LR scores. LRs could then be estimated from the kernel density of the log odds.

Figure 7 shows the resulting Tippett plot for the two types of calibration in comparisons where C is the suspect. The plot for the uncalibrated data is also shown. The likelihood-ratio cost metric Cllr for quantifying the validity of a LR-based detection system [26] reflects how much the system is capable of changing the user's prior belief at the most advantageous prior of 1:1. The Cllr for the uncalibrated system (thin brown line) is greater than unity, indicating that it does not reduce the user's uncertainty. The Cllrs for the calibrated data, on the other hand, are 0.77 (PAV) and 0.58 (log-odds kernel density). This indicates that, with comparisons involving C as the suspect, a calibrated system with just a single questioned /s/ token is clearly capable of giving useful information. It can be also seen that the system has about a 20% equal error rate, comprising an error rate of about 5% for same-speaker comparisons and 35% for different-speaker comparisons.

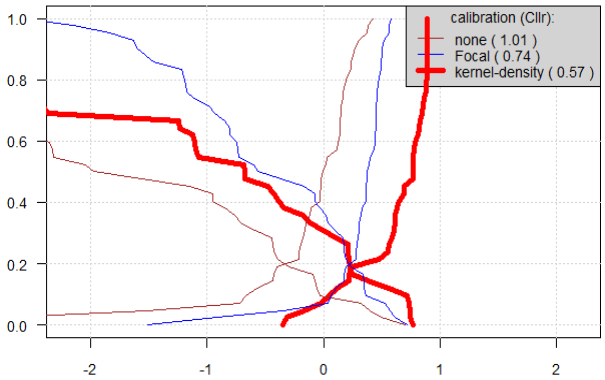

Figure 7: *Tippett plots and associated Cllrs for uncalibrated and calibrated systems assuming C is suspect. X-axis = Log_{10}LR greater than … (different-speaker comparisons) ~ $\log_{10}$LR smaller than … (same-speaker comparisons). Y-axis = cumulative proportion of comparisons.*

Given the validation results, it is then possible to estimate the LR for the actual questioned data against the calibrated distributions. Table 1 shows the final calibrated LR values for both questioned /s/ and /z/. It can be seen that the questioned /s/ and /z/ spectra are far more likely assuming the questioned utterance had come from C than the other two. The enormous value for the /z/ LR under H_C is worth noting. It is because there was rather good separation between his same-speaker and different-speaker validation LR-scores which resulted in Cllrs of 0.34 and 0.32 for both calibration methods. Evidently there was more speaker-specificity hidden in the three blccs of the [z̪] spectrum than the [s].

Table 1. Final calibrated *LR estimates for Questioned /s/ and /z/. H = hypothesis*

H	**/s/**	/z/
H_C	14	332
H_M	6e-05	0.03
H_K	-inf	5e-11

We are left with the problem of how to combine the very different strengths of evidence from /s/ and /z/. Their LRs cannot be combined à la naïve Bayes, as they are sure to be highly correlated. Neither can they be fused as they do not constitute multivariate data. The solution – a crude one – was to re-validate and recalculate with pooled /s/ and /z/ data. This gave a calibrated LR value of ca. 25 for H_C.

7. Posterior Probabilities and Outcome

Normally, the expert is not privy to the prior probability and so cannot logically combine it with their LR to estimate the posterior. They can, however, show the mandating authority how it is done: an example from a Spanish forensic voice comparison case is at [27]. This case is no exception. Even though there are only three individuals who could have said the questioned utterance, the prosecution may have grounds for assuming other than equal priors of 33.3% each. If, for example, they considered K highly unlikely to have said it, their priors might have been more like 45% each for C and M, and 10% for K. With equal priors, and a LR of 25, however, the police could be shown that the posterior odds would be (25/1 * 1/2 =) 12.5 / 1, giving a probability of {12.5 / (12.5 + 1) = } ca. 93% that C was the questioned speaker, thus narrowing it down to either K or M as the perpetrator. For reasons known only to them, the police charged M, not K. He waited until just before start of trial to confess that he had indeed set the bomb.

8. Summary

This paper has described the processing required in applying the LR framework to a real case in an example of what is now called *forensic semi-automatic speaker recognition* [28]. There was exceedingly little evidence: about 20 centiseconds. But otherwise atypically favorable conditions conspired to allow an outcome of use to the mandating authority. Three things are worth emphasising: the forensic potential of voiceless fricative spectra, especially when quantified cepstrally; the complex processing required to achieve a demonstrably valid LR estimate; and the sheer luck of having suspects who differed sufficiently in their speech acoustics for the same sound. It is a pity that not all cases are like this, but that does not excuse us from using the likelihood ratio framework, which, by dint of its explicitness, may very well actually make it possible to do such case-work in the first place.

9. Acknowledgements

Many thanks to my three anonymous reviewers for taking their time to help improve the clarity of this paper. They made some extremely useful suggestions, nearly all of which I was able to incorporate. I also owe thanks to Frantz Clermont both for providing the Matlab script for the bandlimited cepstrum, and for coming up with the concept in the first place.

10. References

[1] Rose, P., Anonymised forensic report, http://philjohnrose.net/pubs/FVC_pubs/index.html, 2017.

[2] Rose, P., "Forensic Speaker Discrimination with Australian English Vowel Acoustics", *Proc. Intl. Congr. of Phonetic Sciences,* Saarbruecken, 2011.

[3] Wolf, J.J., "Efficient Acoustic Parameters for Speaker Recognition", JASA 51: 2044–2056, 1972.

[4] Nolan, F., *Problems and Methods of Speaker Identification*. Unpublished Dip. Linguistics Dissertation, Cambridge University, 1975.

[5] Hillcoat, T.O. *An Evaluation of Selected Sibilant and Nasal Parameters for use in Forensic Speaker Identification*. Unpublished Masters of Letters Dissertation, University of New England, 1994.

[6] Rose, P., "Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate likelihood-ratio-based approach using alveolo-palatal fricative cepstral spectra", *Proc. Int'l Conference on Acoustics Speech & Signal Processing* Prague, 5900–5903, 2011.

[7] Kavanagh, C.M., *New Consonantal Acoustic Parameters for Forensic Speaker Comparison*, Ph.D. thesis, University of York, 2012.

[8] Rose, P., "More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends**",** Int'l Journal of Speech Language and the Law 20(1): 77-116, 2013.

[9] Lo, J. J. H., "One population, two languages: How does language choice affect /s/peaker di/s/crimination?" IAFPA conference poster, 2018.

[10] Zhang C. 张翠玲 and Ding P.丁盼, 擦音LPC倒谱特征在法庭说话人识别中的应用研究 [A study on the application of fricative cepstral features in forensic speaker recognition.] 中国刑警学院学报 5, 117–121, 2019.

[11] Zhang C. 张翠玲 and Ding, P. 丁盼基于LPC倒谱特征融合的法庭说话人识别方法[Forensic speaker recognition based on fused LPC cepstral coefficients],中国刑警学院 2(5), 117–121, 2020.

[12] Smorenburg, L., and Heeren, W., "The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues", JASA , 147: 949–960, 2020.

[13] Stevens, K., Acoustic Phonetics, MIT Press, 1998.

[14] Morrison, G.S., Enzinger, E. and Zhang, C. "Forensic Speech Science", in I. Freckelton and H. Selby [Eds] Expert Evidence 99, Thomson Reuters, 1051-6102, 2018.

[15] Rose, P., **"**Likelihood ratio-based forensic voice comparison with higher level features: research and reality**",** in E. Lleida & L. J. Rodriguez-Fuentes [Eds], Recent Advances in Speaker and Language Recognition and Characterisation, Computer Speech and Language Special Issue, 476-502, 2017.

[16] Holdren, J.P., Lander, E.S. et.al. "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," Science and technology advisory body to the President of the United States, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf, 2007.

[17] Lander, E.S., "Response to the ANZFSS council statement on the President's Council of Advisors on Science and Technology Report", Australian J. Forensic Science, 2017.

[18] Clermont, F. and Mokhtari, P., "Frequency-Band Specification in Cepstral Distance Computation", Proc. Australian Int'l Conf. on Speech Science and Technology, 354-359, 1994.

[19] Khodai-Joopari, M., Clermont, F. and Barlow, M., "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels", *Proc. Australian International Conf. on Speech Science and Technology*, 505-509, 2004.

[20] Clermont, F., Kinoshita, Y. and Osanai, T., "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation", *Proc. Australasian Int'l Conf. on Speech Science and Technology*, 317-320, 2016.

[21] Clermont, F. "Linear transformation from full-band to sub-band cepstrum", *Proc. 18th Australasian International Conf. on Speech Science & Technology*, Canberra, 2022.

[22] Silverman, B.W., Density Estimation for Statistic and Data Analysis, Chapman and Hall, 1986.

[23] Morrison, G.S., "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", Australian J. Forensic Sci., 25(2), 173-197, 2013.

[24] Ramos, D., "Reliable Support: Measuring Calibration of LRs", Keynote at EAFS, The Hague, http://arantxa.ii.uam.es/~dramos/files/2012_08_22_EAFS_Ramos_keynoteReliableSupport_v4.pptx.pdf , 2012.

[25] Brümmer, N., "Focal Toolkit", http://www.dsp.sun.ac.za/nbrummer/focal

[26] Brümmer, N. and du Preez, J., "Application independent evaluation of speaker detection", Computer Speech and Language IEEE Odyssey 2004 Issue 20(2-3), 230-275, 2006.

[27] Lucena-Molina, J., Gascon-Abellan, M. and Pardo-Iranzo, V., "Technical support for a judge when assessing a priori odds", Law, Probability, Risk 14(2), 147-168, 2015.

[28] Drygaylo, A., Jessen, M., Gfroerer, S., Wagner, I.,Vermeulen, J. and Niemi, T., "Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition". Verlag fuer Polizeiwissenschaft, 2015.