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

MAP adaptation characteristics in forensic long-term formant analysis

Michael Jessen

Bundeskriminalamt, Germany
michael.jessen@bka.bund.de

Abstract

Forensic data from long-term formant analysis were used as input to the GMM-UBM approach, which is a way of deriving Likelihood Ratios. Tests were performed running 22 same-speaker comparisons and 462 different-speaker comparisons from a corpus of anonymized casework data involving telephone-intercepted speech. In a first series of tests, the number of Gaussian modules for GMM-modeling was increased from 1 to 32. In a second series of tests the duration of formant input in the compared files was reduced from 10 seconds to 5 and then to 2.5. All tests were performed both without and with the use of MAP adaptation. Results were evaluated in terms of overall performance characteristics EER and C_{llr} and in terms of score distributions visualized as Tippett plots. The main goal of the study was to compare the use and non-use of MAP and to look at the practical forensic implications of the difference. Results show that in terms of overall performance characteristics there is little difference between the selection and de-selection of MAP. Tippett plot patterns however reveal strong differences. Application of MAP allows for more symmetric same- and different-speaker distributions and shows more robustness against duration reductions, both of which are forensically important.

Index Terms: Forensic voice comparison, long-term formants, Likelihood Ratio, MAP adaptation, semiautomatic speaker recognition, GMM-UBM approach

1. Introduction

Long-term formant analysis (LTF) is an established method of forensic voice comparison casework and is subject to forensic-phonetic research [1–5]. With LTF, vowel formants are analyzed globally, i.e. without labeling of the particular vowel categories contained in the analyzed recording. LTF can be used semiautomatically. According to the definition of the semiautomatic approach in [6], this means that formants are extracted manually (by supervising the output of a formant tracker) and then treated automatically in order to derive a (calibrated or uncalibrated) Likelihood Ratio (LR). Two major methods have been used to derive LRs with LTF data and other acoustic-phonetic features. The first one is the GMM-UBM approach (Gaussian Mixture Model - Universal Background Model), the second one the MVKD (Multivariate Kernel Density) approach. The GMM-UBM approach was originally proposed for automatic speaker recognition [7], the MVKD approach for the analysis of forensic evidence in general, e.g. the optical properties of glass fragments [8]. For the LR-analysis of LTF-data, GMM-UBM was first used by [2] and MVKD by [9]. When the choice is for GMM-UBM, there is the option of using MAP-adaptation (Maximum A Posteriori) [7]; in this paper the procedure will be referred to as MAP for brevity (i.e. without the adaptation term). In the GMM-UBM approach the raw data of the questioned speaker are compared to a GMM-model of the suspect data and also to the UBM, i.e. a model of several speakers representative of the comparison at hand. The result of the first comparison is the numerator and the result of the second the denominator of the LR. When no MAP is used, the suspect data are modeled in isolation, i.e. without connection to the UBM. When MAP is used, the UBM is taken as the point of departure for calculating the LR-numerator and the UBM is moved into the direction of the raw data of the suspect recording [10, 11 for illustration and more technical detail]. Previous research on LTF within GMM-UBM has either not used MAP [2, 12–14] or has used it [15–17]. There has however not been any direct comparison of +/-MAP within the same study. When judged from the success of MAP in automatic speaker recognition [7] it might be expected of such a direct comparison that MAP leads to improvement also in semiautomatic speaker recognition. However, acoustic-phonetics-based studies (including cepstra applied to phonetic segments) have shown no uniform advantage of GMM-UBM with MAP over MVKD, where no equivalent of MAP occurs [18–21]. This could indicate that there is no clear +/-MAP performance difference either when applied to LTF as a semiautomatic feature.

2. Method

The corpus GFS 2.0 (German Forensic Speech) was used for this study. It consists of two recordings each of 23 male adult speakers of German. The recordings are drawn from anonymized authentic forensic cases involving telephone interception in both the questioned-speaker and the suspect recordings. Further 25 recordings from different speakers of the same type of casework were used for the UBM. This corpus was also used in [22].

Long-term formant analysis was performed on all the recordings of the corpus. Net-speech extraction from the corpus was designed in a way that approximately 10 seconds of pure formant data were derived from each recording. For one of the speakers no reliable formant measurements could be made due to a combination of a very high-pitched voice and poor signal quality, which left 22 speakers for analysis.

LTF was performed using the software WAVESURFER Version 1.8.5 [23] and taking the system's default settings for formant tracking. All sections of speech except vowels with visible F1, F2 and F3 were eliminated from the signal. Formant tracking was applied to the remaining speech (as mentioned, about 10s long). The formant tracks superimposed on spectrograms were examined visually and any tracking errors were corrected manually by locally re-drawing formant contours where necessary.

LR-analysis was performed using the software VOCALISE version 1.6 [24], which is a tool for automatic and semiautomatic speaker recognition in the GMM-UBM

http://dx.doi.org/10.21437/Interspeech.2021-1697

approach. A previous version of this software is addressed in [14]. Formant frequencies F1, F2 and F3 were used as input data. For the GMMs, diagonal covariance matrices were used and ten iterations with the EM (Expectation Maximization) algorithm were applied. For MAP, adaptation of the UBM towards the suspect data involved means, weights and covariance matrices. Symmetric testing was applied, i.e. the roles of questioned speaker and suspect were analyzed in both directions and the average of the two scores taken.

Tests with 22 same-speaker comparisons and 462 different-speaker comparisons were performed. In one series of tests the number of Gaussian modules was varied from 1 to 32. These numbers apply to both the GMM of the suspect and the UBM. In a second series of tests, duration of the compared files was reduced from 10 to 5 and then 2.5 seconds. In all the tests, separate analyses were made with MAP selected and de-selected. Calculations of EER (Equal Error Rate) using the Convex Hull method and C_{llr} (log-likelihood-ratio cost) as well as creating Tippett plots were performed with the software BIO-METRICS version 1.8 [25]. For information on EER and C_{llr} see [26]. Various forms of Tippett plots are shown in [6]; the particular ones used here are also referred to as Equal Error Graphs in the biometrics literature [27, pp. 33f.].

3. Results

3.1. Varying number of Gaussians

Figure 1 shows EER for tests with the number of Gaussians increasing from 1 to 32, separately for processing without and with MAP. It can be seen that neither the number of Gaussians nor MAP have any relevant effect on speaker discrimination expressed with EER. In all conditions EER clusters around about 16 percent.

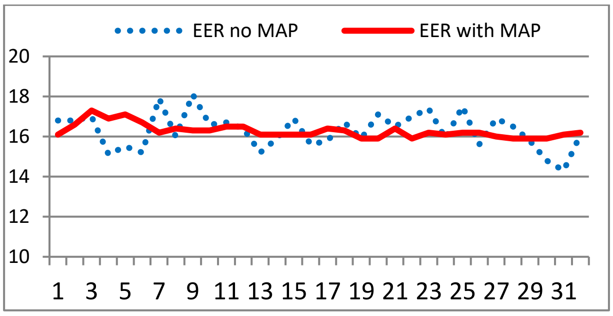

Figure 1: *Equal Error Rate in long-term formant-based tests with increasing number of Gaussians, with/without MAP*

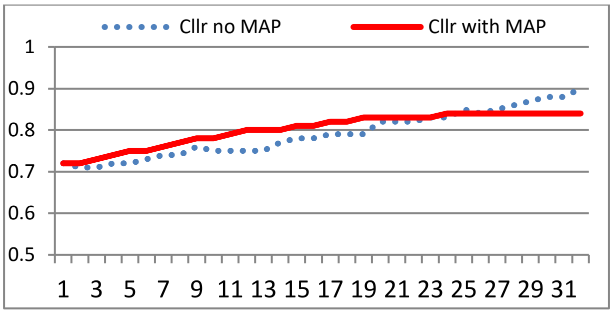

Figure 2: C_{llr} *in long-term formant tests with increasing number of Gaussians, with/without MAP*

Figure 2 shows C_{llr} for the tests. Increasing the number of Gaussians has a negative effect: C_{llr} becomes gradually worse (increasing values) when adding more Gaussian modules. Though there are differences in the details, there is little overall difference in C_{llr} between using and not using MAP.

Figure 3 shows Tippett plots of the tests without MAP (x-axis expressing $\log_e$ LR; y-axis expressing error rate). The curves rising from right to left show the different-speaker comparisons and the ones rising from left to right show the same-speaker comparisons.

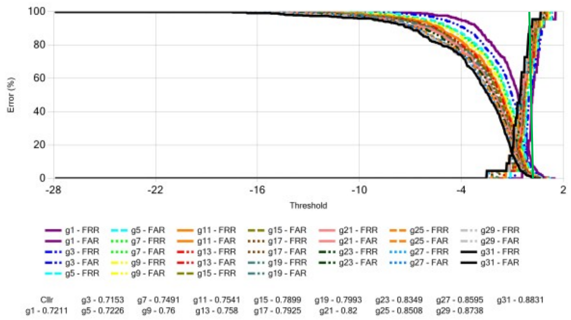

Figure 3: *Tippett plots of long-term formant-based tests without MAP and varying numbers of Gaussians*

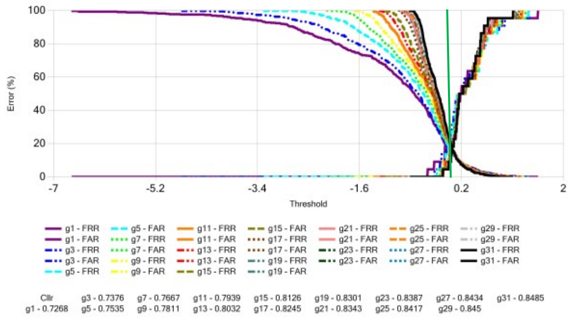

Figure 4: *Tippett plots of long-term formant-based tests with MAP and varying numbers of Gaussians*

Figure 3 (no MAP applied) shows that the same-speaker curves exhibit a steep rising pattern, whereas the different-speaker curves show less steepness. Moreover, as the number of Gaussians is increased (every second increase displayed), both the same-speaker and the different-speaker distributions undergo a left-shift towards lower score values. The probability value on the y-axis at the point where the two distributions intersect for any given number of Gaussians corresponds to the EER. Since both the same-speaker and the different speaker-distributions are left-shifted by about the same amount in the lower part of the plot, the error probability value of the intersection point, i.e. EER, stays the same, which corresponds to the results in Figure 1. The left-shift has the consequence that the scores gradually move away from the log Likelihood Ratio (LLR) value of zero, which is indicated by a green vertical line. This has the consequence that C_{llr} is rising, as shown in Figure 2, because the same-speaker distributions are increasingly penalized in the C_{llr} formula (penalty for same-speaker comparisons that occur to the left of LLR=zero).

Figure 4 (MAP applied) shows that the same-speaker comparisons are not affected much by changes in the number of Gaussians. The different-speaker results, on the other hand are affected strikingly, exhibiting a folding-fan-like pattern. At the lower end of the probability values on the y-axis results are barely affected by number of Gaussians. Since same-speaker pairs are not much affected either, the intersection point stays essentially constant, which results in near-constant EER throughout (Figure 1). But the higher the probability values on the y-axis the more the different-speaker curves diverge with changes in the number of Gaussians.

The LLR value of zero, shown by the green vertical line, is very close to the intersection points of the distributions. This shows that application of MAP leads to good calibration, independently of the number of Gaussians. Stable calibration when judged from the small distance of the intersection to LLR=0 does not have the effect, though, that C_{llr} stays constant across the number of Gaussians, instead it deteriorates. Deterioration of C_{llr} with MAP is probably due to the following effect. According to [19, p. 246] the C_{llr} formula assigns a weak penalty for LLR values between about LLR -1 and +1 even when the evidence is not misleading (misleading evidence would be same-speaker scores < LLR zero and different-speaker scores > LLR zero). As the number of Gaussians is increased the percentage of scores that lie between LLR -1 and 0 is lowest with low number of Gaussians and gets higher with higher numbers, resulting in increasing penalty values that lead to an increase in C_{llr}.

It was mentioned above that without using MAP the slopes for the same-speaker distributions are steeper than the ones for the different-speaker distributions, hence there is asymmetry in the Tippett plots. When using MAP, symmetry in the Tippett plots depends on the number of Gaussians. With the extremes of one Gaussian and 32 Gaussians there is asymmetry – the different-speaker curve being either shallower (G1) or steeper (G32) than the same-speaker curves. But symmetry can be found somewhere in-between those extremes. That symmetry in Tippett plots is a desirable situation in forensic analysis will be argued in the Discussion. For the data reduction tests reported in the following section a value of Gaussians was selected that upon visual inspection of the Tippett plots leads to maximum symmetry. This selected value is 17 Gaussians.

3.2. Reducing formant file durations

Table 1 shows the results of the tests with reduced duration of the long-term formant input files. The baseline is called q10s10, which means that the questioned-speaker recording file is 10 seconds long and the suspect recording file is also 10 seconds long. (The UBM files are kept at 10 seconds throughout all tests.)

Table 1: *Impact of duration for processing long-term formants with and without using MAP*

	No MAP			**With MAP**		
Duration pattern	EER	C_{llr}	C_{llr} after CV	EER	C_{llr}	C_{llr} after CV
q10s10	15.8	0.79	0.56	16.4	0.82	0.61
q5s10	17.2	0.92	0.61	16.5	0.83	0.63
q2.5s10	25.7	1.58	0.83	15.7	0.86	0.69
q5s5	21.9	1.11	0.64	21.5	0.85	0.69
q2.5s2.5	24.2	2.40	0.83	23.0	0.90	0.78

The entries q5s10 and q2.5s10 mean that the questioned-speaker-recording files were reduced to 5 seconds and then to 2.5 seconds, while keeping the suspect files at 10 seconds. That questioned-speaker recordings are shorter than suspect recordings is a common situation in forensic speech analysis. In q5s5 and q2.5s2.5 both questioned and suspect files were reduced by an equal amount to first 5 and then 2.5 seconds. The table shows the results in terms of EER, C_{llr}, as well as C_{llr} after logistic regression calibration [28] using cross validation (CV) [19] was applied (with the BIO-METRICS software mentioned above).

Focusing on EER for now, when reducing the duration of just the questioned speaker recording, there is almost no effect when MAP is used. When no MAP is used, the first reduction step from 10 to 5s has not much of a negative effect on EER either, but the one from 5 to 2.5s has. The remaining two duration reduction patterns have a clear negative impact on EER which is about equally strong without as with MAP.

Tippett plots of the first three duration patterns of Table 1 are shown in Figure 5 (no MAP) and Figure 6 (MAP).

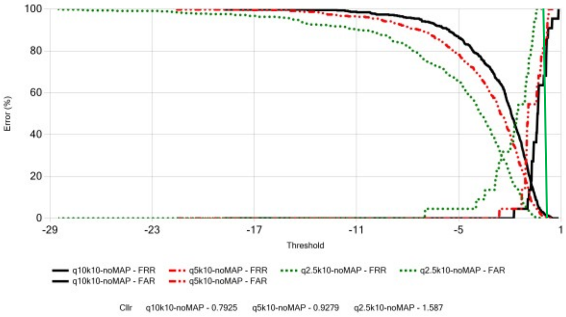

Figure 5: *Tippett plots of long-term formant-based tests with reductions of duration and without MAP*

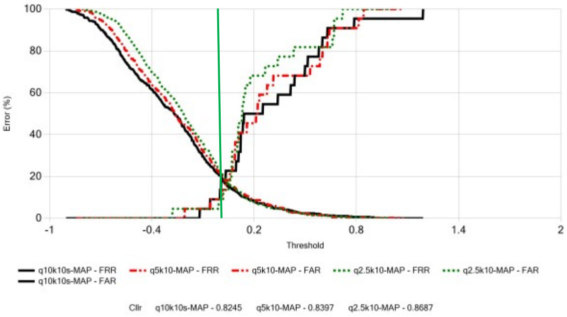

Figure 6: *Tippett plots of long-term formant-based tests with reductions of duration and with MAP*

Figure 5 (no MAP applied) shows that as the duration of the questioned-speaker files is reduced, the scores of both the same-speaker and the different-speaker distributions shift towards lower values. Since LLR=0 (green line) occurs at the right end of the plots, any left-shift has the effect of reducing calibration, which (along with some discrimination reduction) is reflected in the increasing C_{llr} values in Table 1.

Figure 6 (MAP applied) shows very little change depending on the duration of the questioned-speaker files. It also shows that LLR=0 is close to the intersection of the same-and different speaker distributions across conditions. Consequently, C_{llr} is comparatively low and uniform across the three duration conditions shown in Figures 5 and 6. This is in contrast to the tests without MAP, where C_{llr} became higher with every reduction step.

When logistic regression calibration with cross validation (CV) is applied, most of the C_{llr} differences due to +/- MAP in the five duration conditions of Table 1 disappear and there is a further boost in performance (lower values). The former occurs because CV calibrates in each duration conditions separately and therefore the duration-depending score shifts are effectively compensated for. The improvement of C_{llr} occurs because the logistic regression calibration method has some further optimizing characteristics of its own.

4. Discussion and Conclusions

The goal of this paper was to compare the effects of MAP when long-term formant data are processed within the GMM-UBM approach. The difference in behavior between the use and non-use of MAP was investigated by using varying number of Gaussian modules and by varying the duration of the input material.

The part of the study where Gaussian modules were added has shown that – contrary to the expectation from automatic speaker recognition but in line with other semiautomatic research – there is very little difference between the use and non-use of MAP when the results are evaluated in terms of the global performance characteristics EER and C_{llr}. Clear differences however are revealed when the scores are presented in the form of Tippett plots. When no MAP is used, the same-speaker scores exhibit a steep rising pattern, which means that the maximum score value that can be obtained by the formant analysis method is reached relatively quickly. The different-speaker scores exhibit a shallower pattern. Here, very low score values can be obtained and this effect becomes stronger the more Gaussians are added. This behavior results in an asymmetry between the same-speaker and different-speaker comparisons. This kind of asymmetry seen in the Tippett plots in Figure 3 and 5 is not unique to this study or the analysis of long-term formants. It has been shown in many studies (most of them using MVKD) when phonetic features such as formants and fundamental frequency are analyzed within the numeric Likelihood Ratio framework and no MAP is involved [e.g. 5, 13, 29–31]. In contrast, no such asymmetry occurs when case-authentic recordings similar to the ones used here are analyzed with automatic speaker recognition using MFCC (Mel Frequency Cepstral Coefficients) and no MAP [14, p. 34]; see also [32, p. 66 and chap. 6] for Tippett plot near-symmetry with cepstral coefficients and no MAP. A possible explanation for the asymmetry is as follows. Acoustic phonetic features such as formants have a more limited amount of information compared to MFCC. For example, formant analysis mostly focuses on vowels, whereas the use of MFCC covers all sounds. Formants also have fewer dimensions (here: 3) compared to MFCC (about 13 plus 13 deltas) and more correlation (F1,2,3 correlated; MFCCs essentially uncorrelated). As a result, the amount of similarity between recordings that can be achieved is limited (cf. [29] for a similar argument). The amount of *differences*, on the other hand, is much less limited. When the formant (or other acoustic-phonetic) patterns do not fit at all between two recordings, very low scores will result. The effect becomes the stronger the more specific (high number of Gaussians) the models of the suspect become, i.e. there are increasingly more ways in which the questioned-speaker data do not fit into the suspect model.

When MAP is applied the range of possible negative scores is reduced (compare the x-axes in Figure 3 and 4 left of the vertical lines). With MAP, the suspect data are not GMM-modeled in isolation, but are anchored on the UBM. This probably has the effect that the amount of differences that can be obtained if questioned and suspect data do not fit well is more constrained compared to the non-use of MAP. This effect implies the potential of obtaining a much better symmetry in the score distribution of same-speaker and different-speaker comparisons. (Symmetric patterns with GMM-UBM and MAP are also shown in the segmental-phonetic studies of [18–19 and partially 20].) With a number of Gaussians of about 17, full symmetry is reached. Such a symmetry between the range of same-speaker scores and different-speaker scores is useful in a forensic context. Conclusion scales in the forensic sciences that are based on Log Likelihood Ratios (verbal or numerical) have such a type of symmetry when reporting strength of evidence in support of the prosecution hypothesis and the defense hypothesis [33]. Another advantage of MAP is that it leads to better calibration than no MAP when looking at the closeness of the intersection of the distributions to the point of LLR=0.

The duration study was performed in order to test the expectation that the use of MAP can deal quite well with short data input: When no MAP is used, short duration of input data might be insufficient to provide enough information for the suspect models, i.e. these models might be unstable. When suspect models are anchored on the UBM it is ensured that they are stable, even though they might show little difference to the UBM, in which case LLR of the comparison is close to zero. This expectation was confirmed partially. In terms of EER, there was in fact a clear benefit of MAP in the condition q2.5s10 (Table 1), but less so in other conditions. It seems that the danger or getting erratic results with short data input when no MAP is used is limited for long-term formants. This again might be the result of limited information density in formants, i.e. even few data can produce stable isolated suspect models that do not generally lead to less successful discrimination than MAP-adapted models. Figures 5 and 6 show that the structure of the Tippett plots is much less sensitive to duration changes when MAP is used than when MAP is not used. This stability in the Tippett plots is forensically important. In forensic voice comparison the comparison result of a case is interpreted with reference to the relevant Tippett plot [6, chap. 6]. Since in casework the duration of formant input cannot be controlled as much as in the current study, robustness against duration differences in the Tippett plot is a welcomed practical advantage.

Taken together, the tests have shown that whereas the use of MAP has no or little advantage over its non-use as far as overall performance characteristics are concerned, there is an advantage of MAP in the patterns of the score distributions that turns out to be of benefit for forensic casework. There is still room for discussion about the mutual advantages and limits of +/- MAP though. In terms of forensic reasoning, there is an advantage of not using MAP: The suspect model represents the prosecution hypothesis, the UBM the defense hypothesis. MAP could be seen as a procedure that dilutes this clear distinction. Here engineering principles and forensic science seem to be in some conflict. On the other hand, it has also been shown that the symmetry pattern obtainable with MAP and the duration robustness in the plots are clear advantages from a forensic point of view.

5. References

[1] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *The International Journal of Speech, Language and the Law*, vol. 12, pp. 143–173, 2005.

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

[3] A. Moos, "Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech," *The Phonetician*, vol. 101/102, pp. 7–24, 2010.

[4] H. Cao and J. Kong, "Speech length threshold in forensic speaker comparison by using long-term cumulative formant (LTCF) analysis," in *Second International Conference on Instrumentation & Measurement, Computer, Communication and Control 2012*, Harbin, pp. 418–421.

[5] E. Gold, *Calculating Likelihood Ratios for Forensic Speaker Comparisons Using Phonetic and Linguistic Parameters*. PhD dissertation, University of York, UK, 2014.

[6] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Frankfurt: Verlag für Polizeiwissenschaft, 2015. [also accessible at http:// enfsi.eu/wp-content/uploads/2016/09/guidelines _fasr_and_fsasr_0.pdf].

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[8] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics*, vol. 53, pp. 109–122, 2004

[9] 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 *Proceedings of Meetings on Acoustics* (Acoustical Society of America), vol. 19, 060041, 2013.

[10] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans," *IEEE Signal Processing Magazine*, vol. 32, pp. 74–99, 2015.

[11] G. S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, and A. Lozano-Díez, "Statistical models in forensic voice comparison," in D. L. Banks, K. Kafadar, D. H. Kaye, and M. Tackett (Eds.), *Handbook of Forensic Statistics*. Boca Raton: CRC, pp. 451–497, 2020.

[12] T. Becker, M. Jessen, and C. Grigoras, "Speaker verification based on formants using Gaussian mixture models," in *Proceedings of NAG/DAGA 2009*, Rotterdam, pp. 1640–1643.

[13] T. Becker, *Automatischer forensischer Stimmenvergleich*. Norderstedt: Books on Demand, 2012.

[14] M. Jessen, A. Alexander, and O. Forth, "Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software," in *Proceedings of the Audio Engineering Society 54th International Conference on Audio Forensics*, London, pp. 28–35, 2014.

[15] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing," in *Proceedings INTERSPEECH 2017,* Stockholm, pp. 3893–3896.

[16] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, and E. San Segundo, "The individual and the system: assessing the stability of the output of a semiautomatic forensic voice comparison system," in *Proceedings INTERSPEECH 2018,* Hyderabad, pp. 227–231.

[17] V. Hughes, P. Harrison, P. Foulkes, P. French, and A. J. Gully, "Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison," in *Proceedings of the International Congress of Phonetic Sciences 2019,* Melbourne.

[18] P. Rose and E. Winter, "Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analysis," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, Melbourne, pp. 42–45, 2010.

[19] G. S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM)," *Speech Communication*, vol. 53, pp. 242–256, 2011.

[20] P. Rose, "Forensic voice comparison with secular shibboleths – A hybrid fused GMM-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2011*, Prague, pp. 5900–5903.

[21] P. Rose, "More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends," *The International Journal of Speech, Language and the Law*, vol. 20, pp. 77–116, 2013.

[22] Y. A. Solewicz, M. Jessen, and D. van der Vloed, "Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition," in *Proceedings INTERSPEECH 2017*, Stockholm, pp. 2849–2853.

[23] WAVESURFER software. https://www.speech.kth.se/wavesurfer/index2.html

[24] VOCALISE software. https://oxfordwaveresearch.com/products/vocalise/

[25] BIO-METRICS software. https://oxfordwaveresearch.com/products/bio-metrics/

[26] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin: Springer, pp. 330–353, 2007.

[27] T. Dunstone and N. Yager, *Biometric Systems – Design, Evaluation, and Data Mining*. New York: Springer, 2009.

[28] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, pp. 173–197, 2013

[29] P. Rose, Y. Kinoshita, and T. Alderman, "Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/," in *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, Canberra, pp. 329–334, 2006.

[30] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," *Journal of the Acoustical Society of America*, vol. 125, pp. 2387–2397, 2009.

[31] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition," *The International Journal of Speech, Language and the Law*, vol. 16, pp. 91–111, 2009.

[32] V. Hughes, *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*. PhD dissertation, University of York, UK, 2014.

[33] European Network of Forensic Science Institutes (ENFSI), *ENFSI Guideline for Evaluative Reporting in Forensic Science*, 2015 http://enfsi.eu/wp-content/uploads/2016/09/m1_ guideline.pdf.