EXPLORATION OF SIMPLE METHODS OF LIKELIHOOD RATIO CALCULATION IN FORENSIC VOICE COMPARISON

Michael Jessen, Almut Braun and Sonja Menges

Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{michael.jessen | almut.braun | sonja.menges}@bka.bund.de

ABSTRACT

Two multi-speaker datasets were used, one of them from a laboratory corpus, one from forensic casework. Long-term formants were analysed, both with the maximally performing GMM-UBM approach and with simpler methods for likelihood ratio (LR) calculation. Results show that for the casework data, though not the laboratory data, GMM-UBM performance could be approximated with the simple methods. Reasons for the difference probably lie in signal quality and speech style differences which are expected to affect the complexity of the long-term formant distributions. Results support the view that simple methods of LR calculation can improve the process of estimating LRs, which is a widely practiced method in forensic science, they can be a point of departure for the development of more complex methods, or an educational device.

Keywords: likelihood ratio, forensic voice comparison, long-term formants

1. INTRODUCTION

The view that conclusions of forensic science casework should be expressed in terms of likelihood ratios (LR) has been gradually gaining acceptance internationally [1, 2]. Methods of calculating LRs are found across the forensic sciences with more progress in some than others so far. Forensic voice comparison is actually one of the very earliest forensic sciences having promoted and developed useful concepts and tools in that regard [3, 4, 5]. Methods of LR calculation are usually complex. In voice comparison, the most complex ones are from the domain of automatic speaker recognition [6]. Fairly complex also are methods used in semiautomatic speaker recognition, i.e. the application of manually performed or supervised methods of (most commonly) acoustic phonetics to voice comparison [7]. The two best known of such methods are the Multivariate Kernel Density (MVKD) approach and the Gaussian Mixture Model-Universal Background Model (GMM-UBM) approach. The former was developed for the purpose of various forensic sciences, most specifically glass evidence [8], and initially adapted to semiautomatic speaker recognition by Rose [9]. The latter was developed in automatic speaker recognition [10] and was initially adapted to semiautomatic speaker recognition by Becker et al. [11]. The MVKD approach is best suited for token-based data (e.g. twelve tokens of /a/ per recording with one set of formant measurements each) and the GMM-UBM is best suited for stream-based data (e.g. measurements of formants every 10 milliseconds scanning through the vocalic parts of a recording). Since the data analysed in this study are semiautomatic and stream-based, the GMM-UBM approach will be used to establish the top-line indicating optimum performance.

The goal of this study is to compare the performance of one of these high-end methods to the performance characteristics of simpler methods of LR calculation. There are several reasons why we think this is a worthwhile endeavour. Firstly, estimation of LRs in situations where insufficient data are available to calculate them is an option that is explicitly included in current guidelines for forensic analysis [12]. The simpler the methods the better the analyst can understand them and use them as conceptual tools in the estimation process, thereby making the estimations more concept-driven and less intuition-driven. Secondly, simple methods can form the beginning of the development of more complex ones in new areas of the voice comparison feature space or in other forensic sciences with little exposure so far to the LR framework. Thirdly, looking at simple methods can have an educational aspect. One can observe how performance can gradually change by adding more complexity to the methods.

2. METHODS

Two datasets were used as the basis for the experiments in this paper. The first one is called GFS 2.0 (German Forensic Speech). It contains anonymized speech from forensic casework involving telephone interception data. There are data from 22 male adult speakers of German, each with one questioned speaker recording and one suspect recording. There are additional single recordings from 25 adult male speakers recorded under the same conditions. That latter collection will be referred to as the UBM (Universal Background model). More detail on GFS is found in [13, 14]. The second dataset is a subset of Pool 2010 [15]. This subset contains 21 male adult speakers with two recordings each and a UBM consisting of 22 further such speakers. The same dataset was used in [16], where further description is provided. Long-term formant analysis (LTF) [17] was performed on these datasets, and the data streams containing F1 to F3 from the vowels in each recording form the input to the methods applied.

Among the LR calculation methods referred to as simple in the title of this paper, three different ones were used, M(ethod) 1 to 3. M2 was performed with two variations, called M2a and M2b.

Method 1: M1 generates pure similarity scores. In a two-dimensional matrix of 22 questioned-speaker recordings compared against 22 suspected-speaker recordings each of the 484 comparison cells contains the result of the following calculation (for the GFS dataset; for Pool 2020 it is 21 x 21 test speakers). Separately for each formant F1, F2, F3, the mean formant frequency value from the second recording is subtracted from the mean value of the first. The absolute value of the result is taken because it does not matter which of the two compared mean values is higher. Since scores represent degree of similarity, whereas subtraction results represent dissimilarity, a minus-sign is added to each of the results. [1]

Method 2: M2 generates scores that take into account both similarity and typicality, but no probabilities are involved. In addition to the step of similarity calculation in M1, each questioned-speaker value is also compared against the UBM. The result from the similarity calculation step is divided by the result from the typicality calculation step involving the UBM. Taking the absolute value and adding a minus-sign to the resulting ratio concludes the calculation. There are two versions of dealing with the UBM. In **M2a** the UBM value is the mean of the means (of F1 etc.) of all the 25 UBM-speakers (for GFS; 22 for Pool 2010). This is the value subtracted from each questioned-speaker recording in order to obtain the denominator of each of the ratios. In **M2b**, a given questioned-speaker value is compared against each UBM speaker and the mean of these 25 (or 22) comparisons is taken as the denominator of the ratio involving that questioned speaker.

Method 3: M3 generates scores that take into account both similarity and typicality, and it does so by producing probabilities and arranging them in the structure of a LR. The formula for this was taken from Morrison [18, see p. 175]. In addition to the mean value for each questioned and suspected speaker, the formula also requires standard deviations for the suspect and the UBM. In order to obtain mean and standard deviation for the UBM, all individual formant values (extracted every 10 ms) across all UBM speakers were collated into one large data column, from which mean and SD were calculated. The formula assumes one single data point from the questioned speaker (or many points with averaging after the entire procedure), but here the mean value was taken for the questioned speaker instead – a methodological option mentioned on [18, p. 177]. Although the formula contains some higher-level math, the method which it captures can be visualised and understood quite easily (see figure on p. 175). Morrison (pers. comm.) presented this formula for introductory purposes, hence in a mainly educational intent. For these reasons M3 is classified with the simple LR methods.

As the top-line method, expected to represent maximum performance, GMM-UBM was used. The same method was used in [14], where further technical detail is found (except that the option "symmetric testing" was disabled here). The software used for that purpose is VOCALISE version 1.6 [19]. The method was applied in two versions. In the one called **GMM1**, suspect model and UBM were modelled with a single Gaussian, in **GMM3** with three Gaussians. MAP (maximum a posteriori) adaptation was not applied. In contrast to the GMM method, M1 to M3 were all limited to univariate data, i.e. there is a calculation for each formant separately. There was also no modelling of more complicated than single Gaussian distributions: M1 and 2 have no distribution model at all and M3 works with single Gaussians. Both of this differs from GMM, which can work with multivariate data and several Gaussians.

Each of the methods generates scores. These are subsequently transformed into LRs with the use of logistic regression cross validation calibration or with fusion across all three formants ([20] on cross validation, [18] on calibration and fusion). Some results are shown visually as Tippett plots [7] and all reported numerically as Equal Error Rate (EER, convex hull method) and log-likelihood-ratio cost (Cllr) [21]. For calibration, fusion, plots and performance calculation the software BIO-METRICS version 1.8 [22] was used.

3. RESULTS

Figure 1 shows Tippett plots for the score results (not the calibrated LRs) of Methods 1 to 3 in the GFS dataset, applied to the parameter F3, which turns out to be the most successful among the single formants. In the Tippett plots, same-speaker scores are shown in the cumulative distributions rising from left to right and different-speaker scores are the ones rising from right to left.

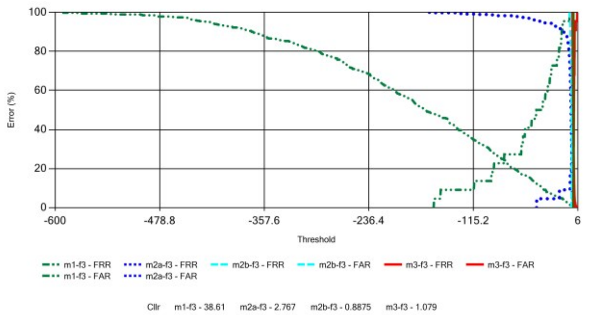

Figure 1: Tippett plots of scores for parameter F3 analysed with Methods 1, 2a, 2b and 3 on dataset GFS. (Cllr values shown are prior to calibration.)

Fig. 1 shows that M1 generates a wide range of scores, which due to the added minus-sign have the value zero as upper limit (that limit also holds for M2ab). All the other methods result in scores that are limited to a much narrower range of values. This range is expanded in Fig. 2.

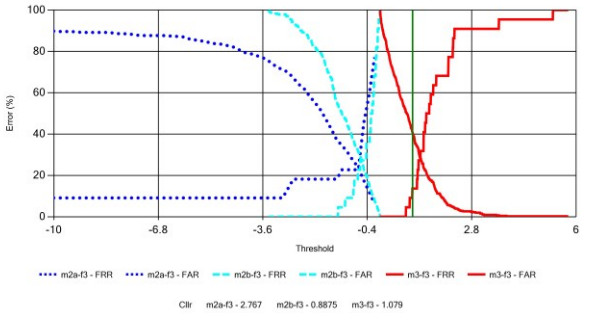

Figure 2: Tippett plots of scores for parameter F3 analysed with Methods 2a, 2b and 3 on dataset GFS. Green vertical line indicates the value 1.

Fig. 2 shows that M2a and M2b have similar patterns but there are more values extending into the lower range of the scores in M2a. As will be discussed below, this can be at least partially the result of outliers that are due to a limitation of the method. M3 generates scores that in their intersection point are quite close to the value one. This proximity to unity indicates that the purpose of the formula behind M3 to create LRs has been successful, i.e. the scores are quite well calibrated even before applying logistic regression calibration.

Speaker discrimination performance measured as EER and Cllr for GFS is shown in Table 1 (lower values indicate better performance). Addressing EER first, when the different methods are compared there are no clear systematic differences except that the results for F1 and F2 are better with GMM3 (but not GMM1) than with the simple methods. F1 and F2 capture phonological vowel distinctions, F3 only marginally so. It could therefore be expected that the distributions of F1 and F2 values across a recording are more complex than the distribution of F3, because they capture differences between different vowel categories. In previous LTF studies this effect however is stronger for F2 than F1 [16, 23 for illustration], whereas here the degree of improvement is similar, though indeed proportionally slightly stronger for F2. Three Gaussians allows for a better modelling of complex distributions than one (in GMM1 as well as M3) or none at all (other methods), which would explain the improvement.

Method	F1	F2	F3	Fusion F123	Multiv F123
M1	36.3 0.93	34.9 0.91	21.8 0.69	20.0 0.60	-
M2a	37.2 1.09	36.5 1.00	23.2 1.31	35.4 1.41	-
M2b	36.2 0.92	34.6 0.94	23.5 0.73	20.8 0.66	-
M3	38.0 0.97	39.0 0.95	26.5 0.84	26.1 0.82	-
GMM1	36.1 0.91	33.2 0.89	25.4 0.76	19.2 0.68	20.9 0.68
GMM3	32.1 0.89	29.2 0.90	25.0 0.79	20.0 0.70	21.5 0.71

Table 1: EER (upper value per cell) and Cllr (lower value) in GFS 2.0 for different methods applied to single formants, fusion of all three formants and multivariate combination of the formants, the latter only possible for GMM.

When comparing EER in different formants and their fusion/combination, F3 turns out to be stronger than the other formants across methods. Fusion of all formants compared to single formants is of benefit for most methods. In GMM, where multivariate processing is possible, multivariate combination of the formants produces very similar results as fusion of the formants.

Turning to the Cllr results now, Cllr most of the time has relative patterns similar to EER, which is expected because discrimination is part of what Cllr measures (and calibration loss, which can affect Cllr, is minimised due to cross validation calibration or fusion). There is strong discrepancy however in M2a, where values of Cllr above 1 occur (1 is chance level). These are likely due to outlier scores that originate from the circumstance that when the questioned speaker value is close to the UBM value, the difference approximates zero and this can strongly increase the value of the ratio calculated in M2a (hence reduce the final score when the minus-sign is added; see Figs. 1 and 2). A few of such problems also occurred in M2b, but less frequently or strongly so.

Method	F1	F2	F3	Fusion F123	Multiv F123
M1	44.9 1.01	30.3 0.88	26.0 0.71	23.0 0.66	-
M2a	43.5 1.00	35.8 1.00	31.3 1.91	40.8 1.81	-
M2b	43.0 1.02	28.9 0.89	28.3 0.73	24.5 0.69	-
M3	45.7 1.01	39.9 0.99	27.7 0.82	27.6 0.81	-
GMM1	34.4 0.92	30.4 0.85	28.1 0.74	22.8 0.64	20.6 0.65
GMM3	27.5 0.83	19.2 0.66	25.8 0.72	11.8 0.42	11.2 0.49

Table 2: EER (upper values) and Cllr (lower values) in Pool 2010.

Table 2 shows EER and Cllr for the Pool 2010 dataset. Pool 2010 is based on a laboratory collection and is expected to yield better performance than the real-case data in GFS. A comparison between the different methods again shows an improvement for F1 and F2 of GMM3 over the simple methods (for F1 also of GMM1). Proportionally this improvement is stronger now than it was with the GFS data. Percentage of improvement is again slightly higher for F2 than for F1. It is likely that the formant distributions are more complex in the Pool 2010 data than the GFS case data because the formants can be measured better on average and the speech style in Pool 2010 is farther to the right on a scale between casual and clear speech. This could explain why the improvement of GMM3 over the simple methods is stronger with the Pool 2010 data. When all these formants are fused or combined, EER is considerably better with Pool 2010 than GFS.

The Cllr results in Table 2 generally follow the pattern of EER, with the same kind of problems for M2a as before.

4. CONCLUSIONS

Based on two datasets involving long-term formant analysis, comparisons were made between the use of a method that is state-of-the-art for semiautomatic voice comparison and simpler methods of LR calculation. Justification for the term LR is given by the use of cross validation calibration, which turns even the results of the simplest method of pure similarity scoring (M1) into LRs. When the methods are applied to real case data (GFS 2.0) the simple methods show nearly the same speaker discrimination performance as the top-line method. When applied to a laboratory collection (Pool 2020), the top-line method clearly outperforms the simple methods, which is probably due to more complex formant value distributions caused by more defined formant structure and a clearer speech style.

The demonstrated relative success of the simple methods with case data does not imply that they should be preferred in casework over state-of-the-art methods if these are available, nor that pure similarity scores should be used in more than preliminary ways [24]. Moreover, it is hard to predict whether even within casework, conditions occur that give the more complex methods a clear advantage over simpler ones. One crucial factor seems to be whether the forensic feature analysed can be approximated by a normal distribution or whether the distribution of feature values is more complex. Features might also have a multivariate structure, which is the case with formants. As could be shown however, performance of multivariate processing (which the simple methods do not perform) is equivalent to univariate processing followed by fusion (which can be done with the simple methods).

Though it is clear that simple methods should not replace complex ones if available, results have shown that simple methods of LR calculation in casework conditions can approximate the performance of complex ones (with the provision that this study is limited to just one type of forensic-phonetic feature). This is interesting since results could have turned out to show a greater performance difference. This outcome suggests there is some substance to the reasons for this study, stated in the introduction, that simple methods of LR calculation can serve as conceptual tools for LR estimation, a starting point for the development of more complex methods or an educational device.

It should be mentioned that the method of cross validation calibration that was used to arrive at LRs from the scores, as well as fusion, are not entirely simple methods, but very common in forensic phonetics and increasingly so in other forensic sciences [1, 18]. The performance index EER however is not affected by calibration, but it can be by fusion. When thinking of LR estimation, the equivalent of numerical calibration would be to either have a good estimate of similarity and also typicality, e.g. based on experience and some research results [12 for examples], or, especially in the case of pure similarity scores (as in M1), to develop knowledge about where the turning point is between differences that support the same-origin hypothesis and those that support the different-origin hypothesis. Fusion would be about developing an idea about degree of correlation between features.

5. REFERENCES

[1] Robertson, B., Vignaux, G. A., Berger, C.E.H. 2016. *Interpreting Evidence*, 2nd ed. Wiley.

[2] Gold, E., French, P. 2019. International practices in forensic speaker comparisons: second survey. *Int. J. Speech, Lang. Law* 26, 1–20.

[3] Meuwly, D., El-Maliki, M., Drygajlo, A. 1998. Forensic speaker recognition using Gaussian Mixture Models and a Bayesian framework. *COST-250 Workshop on Speaker Recognition by Man and by Machine*, Ankara, 52–55.

[4] Rose, P. 2002. *Forensic Speaker Identification*. Taylor & Francis.

[5] Morrison, G. S. 2022. Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Sci. Int.: Synergy* 5, 100270.

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

[7] Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen J., Niemi, T. 2015. *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Verlag für Polizeiwissenschaft. http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf.

[8] Aitken, C. G. G., Lucy, D. 2004. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53, 109–122.

[9] Rose, P. (2006): The intrinsic forensic discriminatory power of diphthongs. *Proc. 11th Australian Int. Conf. Speech Science and Technology*, 64–69.

[10] Reynolds, D., Quatieri, T., Dunn, R. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19–41.

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

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

[13] Solewicz, Y.A., Jessen, M., van der Vloed, D. 2017. Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proc. Interspeech*, Stockholm, 2849–2853.

[14] Jessen, M. 2021. MAP adaptation characteristics in forensic long-term formant analysis. *Proc. Interspeech*, Brno, 411–415.

[15] Jessen, M., Köster, O., Gfroerer, S. 2005. Influence of vocal effort on average and variability of fundamental frequency. *Int. J. Speech, Lang. Law* 12, 174–213.

[16] Jessen, M. Alexander, A., Forth, O. 2014. Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proc. Audio Engineering Soc. 54th Int. Conf. Audio Forensics*, London, 28–35.

[17] Nolan F., Grigoras, C. 2005. A case for formant analysis in forensic speaker identification. *Int. J. Speech, Lang. Law* 12, 143–173.

[18] Morrison, G. S. 2013. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Aust. J. Forensic Sci.* 45, 173–197.

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

[20] Morrison, G. S. 2011. 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 Commun.* 53, 242–256.

[21] Van Leeuwen, D. A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*. Springer, 330–353.

[22] BIO-METRICS softw. https://oxfordwaveresearch.com/products/bio-metrics/

[23] Jessen, M., Becker, T. 2010. Long-term formant distribution as a forensic-phonetic feature. *J. Acoust. Soc. Am.*, 128 (4), 2378. (For illustr. see p. 6 of https://www.researchgate.net/publication/343268341_Long-Term_Formant_Distribution_as_a_forensic-_phonetic_feature.)

[24] Morrison, G. S., Enzinger, E. 2018. Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Sci. Justice* 58, 47–58.

[1] Mathematical expression of the proposed methods. Abbreviations not explained so far: s_x = score according to method x; i, j = the two recordings to be compared; f = formant results; Q = questioned speaker; S = suspected speaker; topline = mean.

Method 1

$$s_1(i,j) = -\left|\overline{f_{Q,i}} - \overline{f_{S,j}}\right|$$

Method 2a

$$s_{2a}(i,j) = -\left|\frac{\overline{f_{Q,i}} - \overline{f_{S,j}}}{\overline{f_{Q,i}} - \overline{f_{UBM}}}\right|$$

Method 2b

$$s_{2b}(i,j) = -\frac{\left|\overline{f_{Q,i}}-\overline{f_{S,j}}\right|}{\frac{1}{K}\sum_{k=1}^{K}\left|\overline{f_{Q,i}}-\overline{f_{UBM,k}}\right|}$$

Method 3

$$s_3(i,j) = \frac{\frac{1}{\sigma_S\sqrt{2\pi}}\, e^{-\frac{\left(\overline{f_{Q,i}}-\overline{f_{S,j}}\right)^2}{2\sigma_S^2}}}{\frac{1}{\sigma_{UBM}\sqrt{2\pi}}\, e^{-\frac{\left(\overline{f_{Q,i}}-\overline{f_{UBM}}\right)^2}{2\sigma_{UBM}^2}}}$$