Adapted Gaussian Mixture Model in Likelihood Ratio based Forensic Voice Comparison using Long Term Fundamental Frequency

Carolin Elisabeth Buncle Diesner[1], Shunichi Ishihara[2]

[1] School of Literature, Languages and Linguistics, the Australian National University, Australia
[2] Department of Linguistics, the Australian National University, Australia
u5820042@anu.edu.au, shunichi.ishihara@anu.edu.au

Abstract

In this paper, the Gaussian Mixture Model – Universal Background Model (GMM-UBM) is applied to one-dimensional speech data, namely the distribution of long term fundamental frequency (LTF0) in likelihood ratio based forensic voice comparison. A series of experiments were conducted using varying numbers of Gaussians, differing adaptation rates to a UBM, and different lengths of speech samples. The results of the GMM-UBM procedure are compared to two previously proposed procedures for LTF0. All three procedures exhibited unique characteristics in their performances. Thus, there was no consistency in performance in that no one procedure constantly outperformed the others.
Index Terms: forensic voice comparison, likelihood ratio, GMM-UBM, long-term F0 distribution

1. Introduction and previous studies

The most common task carried out by forensic phoneticians is forensic voice comparison (FVC) [1]. Different parameters have been used for this task based on, for example, formant frequency, fundamental frequency (F0) contour and the distribution of long-term F0 (LTF0). Formant frequencies have been widely used in FVC, however, when the recording is of poor quality, formant analysis can be difficult. Using formant frequencies can also be rather time consuming. In comparison, [2] suggests F0 as a parameter, as it is generally easy to display and to quantify throughout an utterance. It can be relatively easily extracted even from poor quality recordings, and it is not badly affected by telephone transmission [2]. F0 is also freely available as there are more voiced than unvoiced segments in speech [2]. As a result, the F0-based parameters have been attractive in traditional FVC for some time, although it is important to note that within-speaker variation in F0 can be fairly susceptible to, for example, physiological factors such as age or intoxication, and psychological factors, such as time of day or emotional state [2], [3] and [4].

There are different ways of parameterising the individual unique use of F0 in forensic situations, one of which is modelling F0 contours. This has been proposed in [5] using the *Fujisaki* model. A more commonly used method is looking at distributional patterns of LTF0, the focus of this research. [6] claims that thus far, it was predominantly mean and standard deviation (sd) that have been used to compare the distributional patterns of individual speakers' LTF0. Since this approach (based on mean and sd) seems rather primitive and has provided unsatisfying results, this paper proposed to not only include mean and sd of LTF0 in its likelihood ratio (LR) based FVC approach, but also skewness, kurtosis, modal F0 and probability density of modal F0 (hereafter, the six parameter procedure). All of these parameters relate to the shape of a distribution. The experiments, which used the six parameter procedure with non-contemporary speech samples of 201 Japanese male speakers, succeeded in achieving an overall equal error rate (EER) of 10.7%.

In the above experiments, [6] used running speech between 10 and 25 minutes and did the LR calculations using the multivariate kernel density (MVKD) method [7] in a non-cross-validation manner. Even though this study came up with promising results, the authors criticised the six parameter procedure, as it does not capture the characteristics of bimodal distribution which was mentioned as being very common due to creakiness in voice. Two of the authors went on to do further research and attempted to capture the shape of the LTF0 distribution based on percentiles [8] (hereafter, the percentile procedure). This procedure successfully improved the performance of the system.

Approaches based on the Gaussian Mixture Model (GMM) are commonly used in FVC [9], and, in particular, it was reported that the adapted version of the GMM procedure, namely the Gaussian Mixture Model – Universal Background Model (GMM-UBM) procedure performs well in both automatic [10] and traditional FVC [11].

Despite the fact that the GMM-UBM is almost considered as one of the standard procedures in FVC, to the best of our knowledge, it has never been applied to one-dimensional speech data, such as LTF0. Thus, motivated by [6] and [8], the current study seeks to find out how well the GMM-UBM procedure works with LTF0 within the LR framework. Further experiments using the same database as used in the GMM-UBM procedure, but applied to the procedures proposed in [6] and [8] (the six parameter and percentile procedures), will allow direct comparison of the three tactics.

2. Likelihood ratio

This study is an LR-based FVC study. In the context of forensic science, as given in (1), an LR is the probability (P) of the evidence (E) occurring if an assertion is true (e.g. the prosecution hypothesis (H_p) is true), divided by the probability that the same evidence would occur if the assertion is not true (e.g. the defence hypothesis (H_d) is true) [12].

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad (1)$$

In FVC, the LR value would indicate the probability of viewing the difference between two speech samples (e.g. the offender and suspect speech samples) if they had come from

the same speaker, relative to the probability of viewing the same evidence if the two speech samples had come from different speakers. The LRs that are higher than one (LR > 1) support the H_p, and those that are lower than one (LR < 1) support the H_d. The further away the LR is from unity (LR = 1), the stronger it supports either hypothesis.

3. Experiments

3.1. Database

For the experiments, the monologues of 201 speakers were selected from the Corpus of Spontaneous Japanese (CSJ) [13], which consists of different types of speech samples from 1464 speakers. The selected recordings, the same which have been used in [6], belong to level four or five of the so-called 'spontaneous' scale of either Academic Presentation Speech or Simulated Public Speech. CSJ uses a five-scale evaluation rating of different aspects of the speech used in the recordings. In this case it means that the chosen recordings sound as if they had not been read out, but spoken freely. Using speech samples that involve natural speech rather than speech that is read out, is important as this better simulates forensic casework. Another selection criteria used to determine speech samples was the availability of non-contemporaneous recordings of the same speaker in order to perform within-speaker comparison.

The 201 speakers were further separated into three mutually exclusive databases of test, background and development databases, each of which is made up of 67 speakers. Out of the 67 speakers of the test database, 67 same speaker (SS) comparisons and 4422 independent DS comparisons are possible.

F0 was sampled from each recording at every 0.01 second with the ESPS routine of the Snack Sound Toolkit (http://www.speech.kth.se/snack/). The measured F0 samples were all pooled together for each recording, and they were used to create eight different lengths of samples: 5, 10, 20, 40, 60, 80, 100 and 120 seconds. These different lengths of sample are to investigate the relationship between the performance of a system and the length of samples.

3.2. Likelihood ratio estimation: GMM-UBM

A GMM, which is a parametric probability density function represented as a weighted sum of M component Gaussian densities, is claimed to be the most effective likelihood function in text-independent speaker recognition [10]. GMM parameters (mixture weight, mixture mean and mixture variance/covariance) are estimated from a training database (e.g. suspect samples) using the iterative Expectation-Maximisation (EM) algorithm with the maximum likelihood (ML) estimation. The main idea of the GMM-UBM is that the GMM, which was built by the above process for a suspect, is adapted to a UBM which was built based on the background database. This way of estimating GMM parameters is called Maximum A Posterior (MAP) estimation. Please refer to [10] for a mathematical exposition of the GMM-UBM. In this study, a series of experiments were carried out by altering the number of Gaussian components (2, 3, and 4) and the relevance factor (adaptation weight) (0, 16, 32, 64, and 128). The relevance factor = 0 means no adaptation.

3.3. Likelihood ratio estimation: MVKD

In order to compare the performance of the GMM-UBM procedure with the performances of the procedures proposed in [6] and [8], further experiments were done by replicating [6] and [8] with the same sets of databases. In short, the distribution of the LTF0 was modelled using the mean, sd, skewness, kurtosis, and modal F0 and modal density in [6] (the six parameter procedure), while using the F0 and density values of some percentile points in [8] (the percentile procedure). The MVKD method was used to calculate LRs for both of these procedures. The percentile F0 and its density values were obtained at 5%, 30%, 50%, 70% and 95% for the current study.

3.4. Calibration

A logistic-regression calibration [14] was applied to the derived scores from the three different procedures. The FoCal toolkit (https://sites.google.com/site/nikobrummer/focal) was used for the logistic-regression calibration in this study [14]. The logistic-regression weight was obtained from the development database.

3.5. Performance Assessment

Log-likelihood-ratio cost (C_{llr}) was used to assess the outcomes of the experiments and to gain an overall view of the performance of the systems.

Table 1: *C_{llr} values from the GMM-UBM experiments. Numerals in bold face = lowest C_{llr} value per time unit. Underlined numeral in bold face = lowest C_{llr} value overall*

Gaussians	Adaptation	5	10	20	40	60	80	100	120
2	0	0.881	0.810	0.755	0.710	0.712	0.719	0.705	0.703
2	16	0.868	0.807	0.755	0.711	0.712	0.720	0.705	0.703
2	32	0.860	0.805	0.756	0.712	0.713	0.720	0.706	0.704
2	64	0.853	0.804	0.757	0.713	0.713	0.720	0.706	0.685
2	128	0.852	0.804	0.761	0.715	0.715	0.721	0.707	0.705
3	0	0.873	0.806	0.738	0.714	**0.688**	**0.695**	0.704	0.679
3	16	0.848	0.801	0.745	0.704	0.716	0.719	0.705	0.702
3	32	0.874	0.802	0.756	0.717	0.713	0.710	0.706	0.705
3	64	0.873	0.811	0.761	0.706	0.706	0.724	0.708	0.692
3	128	**0.847**	0.810	0.765	0.723	0.720	0.724	0.710	0.706
4	0	0.854	0.795	**0.733**	0.703	0.709	0.718	**0.670**	**0.667**
4	16	0.873	0.798	0.753	**0.698**	0.708	0.699	0.680	0.678
4	32	0.856	**0.783**	0.749	**0.698**	0.696	0.715	0.684	0.681
4	64	0.895	0.795	0.779	0.707	0.699	0.724	0.689	0.685
4	128	0.889	0.808	0.779	0.714	0.707	0.710	0.718	0.702

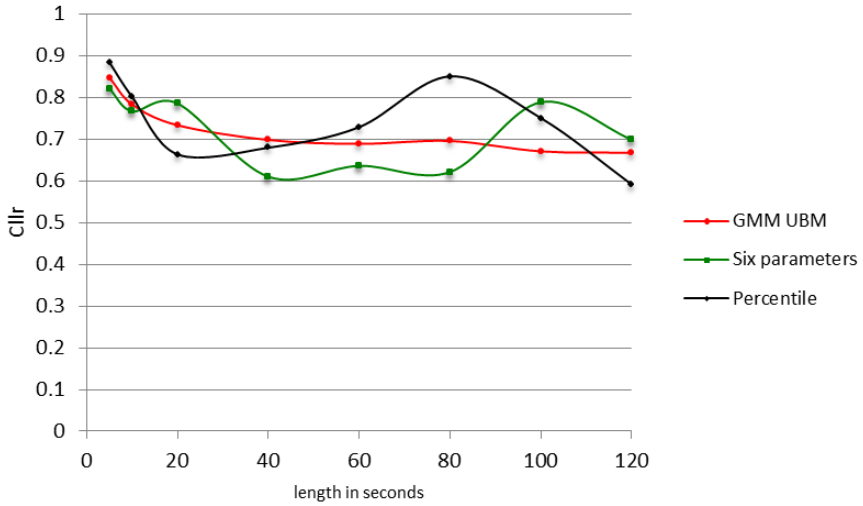

Figure 1: C_{llr} *comparison of all three procedures.*

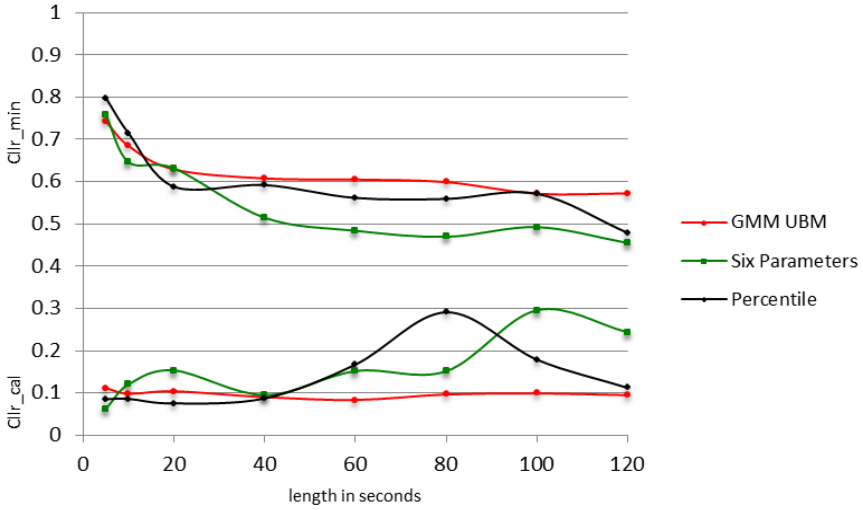

Figure 2: C_{llr_min} *and* C_{llr_cal} *comparison of all three procedures.*

The final C_{llr} value is the sum of two different values: C_{llr_min} and C_{llr_cal}. The former indicates the system's discriminability when it is ideally calibrated and the latter shows the loss caused by its calibration component.

4. Results and Discussion

4.1. GMM-UBM

Table 1 presents the C_{llr} values from the 120 experiments of the GMM-UBM procedure conducted (= 3 Gaussian numbers * 5 relevant factors * 8 sample lengths). Each time unit (0, 5, 10, 20, 40, 60, 80, 100 and 120 seconds) is divided into two, three and four Gaussians with each indicating different adaptation rates starting at zero and rising to 128. The performance improves from 5 to 40 seconds, after which the performance remains relatively stable with some very minor ups and downs in C_{llr}; this can be seen in Table 1 as well as in Figure 1 (red curve), in which the best results (C_{llr} values) per time unit have been plotted. The longer the sample, the more accurately individualising information can be naturally extracted for comparison. Thus, the observation that the performance improves as the sample length becomes longer (in particular from 5 to 40 seconds) is not surprising.

Table 1 further indicates the overall improvement of performance with the increase of Gaussians used per unit of time; the best performance was achieved with the Gaussian number of four, except the sample lengths of 5, 60 and 80 seconds. This was to be expected as an increase in Gaussians increases the chance of better capturing the actual shape of the F0 distribution. Furthermore, it can be observed from Table 1 that within shorter speech samples, a higher amount of adaptation generally improves performance. This comes as no surprise, as the UBM will fill in missing information due to the shorter speech samples. With longer speech samples, a higher amount of adaptation decreases the performance as it then generalises individual information with its adaptation to the UBM. As a whole, the lowest C_{llr} value of 0.667 was achieved with the combination of four Gaussians, zero adaptation and 120 seconds of speech.

4.2. GMM-UBM vs. MVKD based procedures

In Figure 1, the C_{llr} values of the two MVKD-based procedures: the six parameter procedure and the percentile procedure, are plotted against the different lengths of samples. Overall it can be said that these two MVKD-based procedures (green and black curves) also come up with useful results.

However, they are not as stable as the GMM-UBM (red curve), and the results overall do not necessarily improve as the length of speech samples increases, as one would expect. It can be observed from Figure 1 that the C_{llr} values of the MVKD-based procedures fluctuate a lot from 40 to 120 seconds. Partly because of this instability of the MVKD-based procedures, there is no consistency in performance in that one procedure constantly outperforms the others.

It is interesting to observe that the percentile procedure performed better than the six parameter procedure, only when the sample length is 100 seconds or longer, except for the sample length of 20 seconds. This result contradicts that of [8], which reported that the percentile procedure outperformed the six parameter procedure. This may be due to the fact that [8] used very long sample length (10-25 minutes). Judging from the results of the current study and that of [8], the percentile procedure may work better than the six parameter procedure when the sample is large (100 seconds or longer).

In order to further investigate the instability of the two MVKD-based procedures, the C_{llr} values given in Figure 1 are decomposed into C_{llr_min} and C_{llr_cal}, and they are plotted in Figure 2 against the different lengths of speech samples. If only C_{llr_min} is considered, the three procedures exhibit more or less the same trend; there is a large improvement in performance from 5 to 20 seconds, after which the performance continues to improve overall (with some minor ups and downs) as a function of the sample length, but to a lesser degree. This is something which we expected.

However, unlike C_{llr_min}, it is evident from Figure 2 that the values of C_{llr_cal} fluctuate largely between 0.1 and 0.3 for the two MVKD-based procedures (green and black curves), in particular, with longer samples (60 seconds or longer). It is clear that the instability of C_{llr_cal} is responsible for the irregular pattern observed in the C_{llr} values of the MVKD-based procedures. Yet, the reason behind it is not clear at this stage, and warrants further research.

5. Conclusion and further directions

This paper investigated how well the GMM-UBM is applied to one-dimensional speech data, namely the distribution of LTF0 in LR based FVC. The outcomes of the GMM-UBM procedure were compared to the outcomes of two other, previously proposed procedures: one that models the distribution of LTF0 using the mean, sd, skewness and kurtosis, modal F0 and the density of modal F0 (the six parameter procedure) [6], and the other that models the distribution the F0 and density values of some percentile points (the percentile procedure) [8]. The six parameter procedure and the percentile procedure used the MVKD formula to calculate LRs. The performance was assessed in terms of C_{llr}, including C_{llr_min} and C_{llr_cal}.

As for the performance of the GMM-UBM procedure, it is observed that i) the performance generally improves with the increase of Gaussians used per unit of time, and that ii) with shorter speech samples, a higher amount of adaptation generally improves performance with the combination of four Gaussians.

As for the comparisons between the three procedures, the results (C_{llr}) showed that all three procedures yielded relatively similar results overall. A reason for the similar results could be that they all look at the same thing, that being the distributional pattern of LTF0.

The GMM-UBM seems to be more stable than its competitors, but it does not impress with overall discriminability (C_{llr_min}). The unstable performance of the MVKD-based procedures (the six parameter procedure and the percentile procedure) is due to the inconsistent performance of calibration (C_{llr_cal}) with different lengths of samples. Overall, there is no consistency in performance across the three different procedures; in some cases, one system performed better than the others, in other situations a different system came up with better results depending on the circumstances.

For forensic casework, it is recommended that all three methods be tried to observe how they perform under the particular circumstances of the case.

It is also important to point out that even though the database that was used consisted of highly spontaneous recordings (yet, they are monologues), it would be of high value to do further experiments with more forensically realistic data.

6. Acknowledgements

We would like to thank our anonymous reviewers for their comments. This paper is based on the first author's project report submitted for LING6535 at the ANU.

7. References

[1] P. Foulkes and P. French, "Forensic phonetic speaker comparison," *Oxford Handbook of Language and Law,* pp. 557-572, 2012.

[2] P. Rose, in *Forensic speaker identification*, ed London: Taylor & Francis, 2002, pp. 244-253.

[3] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press, 1983.

[4] A. Braun, *Fundamental frequency - how speaker-specific is it?* Trier: Wissenschaftlicher Verlag, 1995.

[5] A. Leemann, H. Mixdorff, M. O'Reilly, M.-J. Kolly, and V. Dellwo, "Speaker - individuality in Fujisaki model f0 featuers: implications for forensic voice comparison," *The International Journal of Speech, Language and the Law,* vol. 21, pp. 343-370, 2014.

[6] 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 and the Law,* vol. 16, pp. 91-111, 2009.

[7] 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.

[8] Y. Kinoshita and S. Ishihara, "F0 can tell us more: speaker verification using the long term distribution," presented at the Speech Science and Technology Conference, Melbourne, 2010.

[9] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)," *Proceedings of 2001 Odyssey-The Speaker Recognition Workshop,* 2001.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processin,* vol. 10, pp. 19-41, 2000.

[11] 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, Feb 2011.

[12] B. Robertson and G. A. Vignaux, *Interpreting Evidence*. Chichester: Wiley, 1995.

[13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," presented at the The Second International Conference of Language Resources and Evaluation, Athens.

[14] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language,* vol. 20, pp. 230-275, Apr-Jul 2006.