# Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking

*Alberto de Castro, Daniel Ramos and Joaquin Gonzalez-Rodriguez*

ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, Spain.
{a.castro,daniel.ramos,joaquin.gonzalez}@uam.es

## Abstract

In this paper we compare forensic speaker recognition with traditional features using two different formant tracking strategies: one performed automatically and one semi-automatic performed by human experts. The main contribution of the work is the use of an automatic method for formant tracking, which allows a much faster recognition process and the use of a much higher amount of data for modelling background population, calibration, etc. This is especially important in likelihood-ratio-based forensic speaker recognition, where the variation of features among a population of speakers must be modelled in a statistically robust way. Experiments show that, although recognition using the human-in-the-loop approach is better than using the automatic scheme, the performance of the latter is also acceptable. Moreover, we present a novel feature selection method which allows the analysis of which feature of each formant has a greater contribution to the discriminating power of the whole recognition process, which can be used by the expert in order to decide which features in the available speech material are important.

**Index Terms**: automatic formant tracking, forensic speaker recognition, traditional features, likelihood ratio.

## 1. Introduction

Forensic speaker recognition by human experts using traditional features has being increasingly important in forensic science [1], as more resources have been available to phoneticians in the form of databases and software tools. Despite such progress, the semi-automatic process for generating a result of a forensic comparison is time-demanding in general. In a typical analysis of phonetic-acoustic features, the expert has to perform several steps before a result is obtained [1]. First, the units of interest (words, diphthongs, phonemes, etc.) should be identified and accurately segmented. Second, the phonetic-acoustic features should be extracted from those units, e.g. formant frequencies and formant trajectories. Finally, with those features a comparison should be performed. This whole process may spend a considerable amount of time, as most of the tasks involved are performed manually. This problem becomes more important when the comparison process implies modelling the distribution of features among a relevant population of many speakers, as it happens in likelihood-ratio-based forensic speaker recognition, which is considered a proper way of reporting results to a court [2].

The contribution of this paper is the use of an automatic formant tracking scheme for the use of traditional features in forensic speaker recognition. This allows a much faster recognition process, and therefore, a much higher amount of data can be used as a background set for population modelling, calibration, etc. This also helps to increase the robustness and accuracy of the evidence evaluation process and the validation results from a forensic case. Thus, if an accurate segmentation of the relevant units in the speech signal is available, the rest of the proposed recognition process is automatic. In this work we have used diphthongs segmented by a human experts for comparison, but this labels could also be obtained with a speech recognition system, which would lead to a fully automatic approach of forensic speaker recognition using traditional features. The evaluation of the impact of the proposed recognition scheme with respect to an expert-based semi-automatic formant extraction method is also presented. Moreover, an analysis based on feature selection is performed with the objective of identifying the most discriminative features in the identity inference process. In order to obtain results, a likelihood ratio (LR) approach is used [3, 2]. The human expert performance is taken from the work in [4], whose database is used and experimental set-up replicated. Performance evaluation is given in terms of DET plots and measures of LR performance such as $C_{llr}$ [5].

The paper is organized as follows. In Section 2, the feature extraction approach followed by the expert in [4] will be described. In Section 3 the automatic formant tracking tool used in the paper, developed by [6], is sketched, and the proposed feature selection analysis is detailed. Experiments are presented in Section 4, where the adequacy of the methods proposed with respect to expert-based approaches is shown. Finally, conclusions are drawn in Section 5.

## 2. Expert-based traditional forensic speaker recognition

In this section we describe the expert-based approach for forensic speaker recognition, which is replicated from [4].

### 2.1. Database description

The database used in this paper includes recordings of the speech of the 27 male speakers of Australian English from a corpus described in [7] and used in [4]. Sentences are of the kind “Bide, B-I-D-E spells bide”. Such utterances contained the target diphthongs which will be used for recognition: /aɪ/, /eɪ/, /oʊ/, /aʊ/ and /ɔɪ/. Their segmentation was performed manually by the human expert by inspection of the spectrogram.

This work has been funded by the Spanish Ministry of Education under project TEC2006-13170-C02-01. We strongly thank Daniel Rudoy from Harvard University for providing the automatic formant tracking software used in this paper; Yuko Kinoshita from the University of Canberra for providing the database for the experiments in this work; and Geoffrey Stuart Morrison from the Australian National University for providing the manual segmentation and the human-supervised semi-automatic formant tracks, and for advice and inspiration.

The speech was recorded with the same microphone in the same environment, and therefore it is not in real forensic conditions, since variability and mismatch are reduced. However, it is a valuable corpus for comparison between automatic and human approaches, since there is a lack of databases segmented and analyzed by human experts.

### 2.2. Human-in-the-loop feature extraction

For each diphthong manually selected from the database, the formant tracking procedure described in [4] was applied to the first three formants (namely F1, F2 and F3). Once the formant trajectories have been determined, features are extracted by a parametric-curve fitting of the formant trajectories, either polynomial or based on the Discrete Cosine Transform (DCT). As a result, for each formant a variable number of coefficients is selected depending polynomial degree (Equation 1a) or the amount of components in the DCT (Equation 1b):

$$ax^3 + bx^2 + cx + d = 0 \rightarrow (a, b, c, d) \tag{1a}$$

$$\begin{aligned} X_c(k) = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos(\frac{k2\pi n}{N}) \rightarrow \\ \rightarrow (X_c(0), X_c(1), X_c(2), X_c(3)) \end{aligned} \tag{1b}$$

Thus, for each diphthong analyzed, the feature vector will be formed by the concatenation of the coefficients of the polynomial (e.g., $[a, b, c, d]$) or DCT fitting (e.g., $[X_c(0), X_c(1), X_c(2), X_c(3)]$) for the selected formants. Performance is improved in [4] with equalization of the duration of each diphthong and/or logarithmic frequency scaling applied prior to feature extraction.

### 2.3. Comparison, LR computation, fusion and calibration

In order to perform a comparison among coefficients, the Multivariate Likelihood Ratio (MVLR) method has been used [3]:

$$LR = \frac{p\left(\mathbf{x}, \mathbf{y} \middle| \theta_p, I\right)}{p\left(\mathbf{x}, \mathbf{y} \middle| \theta_d, I\right)} \tag{2}$$

where $\theta_p$ is the prosecution hypothesis (*The suspect is the source of the questioned recordings*), $\theta_d$ is the defense hypothesis (*Another individual in the relevant population is the source of the questioned recordings*), and $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors to be compared from questioned and control speech material. A function implementing this method in Matlab$^{TM}$ can be found in www.geoff-morrison.net, which we have used in our experiments. See [3] for details.

The comparison strategy is as follows. Every feature vector extracted from a given diphthong found in the questioned speech material (one feature vector for each diphthong occurrence) is compared to all the feature vectors for the same diphthong found in the control material coming from the suspect. Thus, for each comparison, a LR value is computed for each diphthong. Then, the logarithm of the LR values of all the diphthongs are summed (fused) for each comparison in order to improve system performance. Finally, a jackknife linear logistic regression calibration process is applied to the obtained log-LR set as described in [4]. This further calibration procedure of the final, summed log-LR is necessary, since the sum of log-LR values coming from independent sources (e.g., different diphthongs) may not be probabilistically interpretable. This last LR value after calibration will represent the weight of the evidence.

### 2.4. Performance measures

The determination of the goodness of the LR value computed is achieved by the use of the $C_{llr}$ metric [5]:

$$C_{llr} = \frac{1}{2 \cdot N_p}\sum_{i_p} \log_2\left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d}\sum_{j_d} \log_2\left(1 + LR_j\right) \tag{3}$$

where $N_p$ and $N_d$ are the number of comparisons (LR values) where $\theta_p$ and $\theta_d$ are respectively true in the experimental set, also known as target and non-target comparisons. As it can be seen, $C_{llr}$ is an average measure of performance over a given experimental set of LR values, and the higher its value the worse the given LR set.

The overall loss of performance given by $C_{llr}$ can be decomposed into a loss due to discriminating power and another loss due to calibration [5]. In order to test the discriminating power of the proposed methods alone (separation of target and non-target comparisons regardless of the range of the LR values), DET curves are used in automatic speaker recognition. Moreover, $C_{llr}^{min}$ has been also proposed as the optimization of $C_{llr}$ restricted to preserve the discriminating power of the experimental set [5]. Thus, $C_{llr}^{min}$ summarizes a DET curve with a single value, and the calibration of the experimental set is determined by $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$.

## 3. Traditional forensic speaker recognition using automatic formant tracking

In order to compare the approaches presented in this paper with respect to the one presented in [4], we have replicated the same method and experimental set-up as described in Section 2 with the use of the segmentation labels provided by the human expert, with some differences. First, the semi-automatic formant tracking procedure described in 2.2 has been replaced by a fully automatic process described below [6]. Second, the feature extraction strategies in [4] have been extended with two alternatives which improve system performance. Third, a feature selection algorithm is proposed in order to identify the most discriminant features with a phonetic-acoustic interpretation, which would aid the expert in the selection of the relevant features from the available speech.

### 3.1. Automatic formant tracking procedure

In order to automatically extract formant trajectories from the speech spectrum for each speech unit (diphthong), the formant tracking tool described in [6] was used. Figure 3.1 shows an example of using this technique. The approach is based on estimating the formant frequencies by means of a Gauss-Markov process. After a cepstral linear prediction analysis, the distribution $p(x_t|y_{1:t})$ is computed for the formant frequencies conditioned to previous waveform data observed, where $t$ is the current time frame, $x_t$ is a state vector result of parameterizing the spectral envelope at time frame $t$, and $y_{1:t}$ is a function related to the past linear prediction coefficients. Details can be found in [6]. For this work, the formant tracking software was provided by the authors in [6].

### 3.2. Feature extraction and comparison strategies

For each diphthong, the feature set may vary depending on the fitting (polynomial, DCT) and the time (equalized, not equalized) and frequency (Hz, log-Hz) transformations. In this work we have explored three different feature extraction strategies:

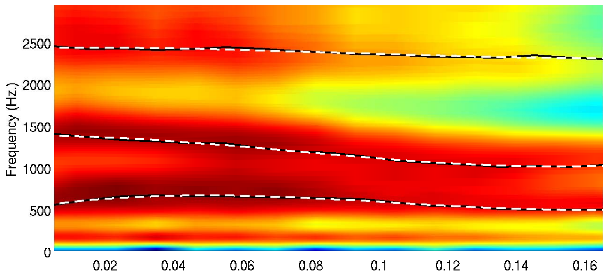

Figure 1: Example of polynomial fitting (degree 3) of /ou/ formant trajectories. Solid black lines are the estimated formant trajectories. Dashed white lines are fitted curves.

- BEST_IND_AUTO: feature set which obtained the best $C_{llr}^{min}$ for each individual diphthong, which gives a different selection for each diphthong (Table 1).

| Dipht. | Formants | Fit | *f* Scale | *t* Scale |
|---|---|---|---|---|
| /aɪ/ | F1 F2 F3 | Poly 3 | Hz | Equalized |
| /eɪ/ | F1 F2 F3 | DCT 3 | Hz | Equalized |
| /oʊ/ | F1 F2 F3 | Poly 3 | Hz | Equalized |
| /aʊ/ | F1 F2 F3 | Poly 3 | Hz | Original |
| /ɔɪ/ | F1 F2 F3 | Poly 3 | Hz | Equalized |

Table 1: BEST_IND_AUTO feature extraction scheme.

- BEST_ALL_AUTO: same feature set for all diphthongs, which obtained the best average $C_{llr}^{min}$ value across diphthongs. This strategy encourages a feature set which is more general for all types of diphthongs, being a more reasonable choice if a speech unit not analyzed before is selected for comparison due to limitations in the speech material. The feature extraction selected in this way considered polynomial fitting of degree 3 obtained form F1, F2 and F3 trajectories, natural frequency scale, and equalized duration.
- HUMAN_SEMI: feature set selected by the expert in [4] with semi-automatic human-in-the-loop formant tracking. The feature set is summarized in Table 2.
- HUMAN_AUTO: same feature set as HUMAN_SEMI (Table 2). The main objective of this strategy is the direct comparison of the automatic and the human-in-the-loop formant tracking procedures.

| Dipht. | Formants | Fit | *f* Scale | *t* Scale |
|---|---|---|---|---|
| /aɪ/ | F1 F2 F3 | Poly 3 | Hz | Equalized |
| /eɪ/ | F2 F3 | DCT 3 | Hz | Original |
| /oʊ/ | F1 F2 F3 | Poly 3 | Hz | Equalized |
| /aʊ/ | F1 F2 F3 | Poly 2 | Hz | Original |
| /ɔɪ/ | F1 F2 F3 | DCT 3 | Hz | Original |

Table 2: HUMAN_SEMI and HUMAN_AUTO feature extraction schemes.

### 3.3. Analysis based on feature selection

A feature selection scheme is proposed in order to get a deeper analysis of which specific information in the formants is discriminating. The feature selection algorithm is based on the following steps:

1. For each diphthong and feature in the original feature set, a univariate log-LR set from comparisons in the database is computed, and its $C_{llr}^{min}$ determined. This shows which feature from which formant has a better discriminating power (lowest $C_{llr}^{min}$)
2. The log-LR set from the next feature with lower $C_{llr}^{min}$ value is fused with the output log-LR and if it decreases $C_{llr}^{min}$ value the feature is selected, otherwise the feature is not selected and the sum fusion is undone.
3. The previous step is repeated for all the features in increasing $C_{llr}^{min}$ order.

## 4. Experiments

### 4.1. Results on automatic formant tracking

The experiments in this section aim at illustrating the loss of performance due to an automatic approach for formant tracking (HUMAN_AUTO) with respect to a human-in-the-loop semi-automatic formant tracking (HUMAN_SEMI). In table 3 the performance for each diphthong is shown for both strategies. It can be seen that performance in terms of discriminating power ($C_{llr}^{min}$) of the HUMAN_AUTO approach is worse than for HUMAN_SEMI. However, the HUMAN_AUTO strategy is still acceptable in terms of performance, especially considering that eliminates the need of a human expert for semi-automatic formant selection, consequently reducing the time for a comparison. Figure 2 shows the per-diphthong discriminating power of the HUMAN_AUTO strategy in terms of DET pots.

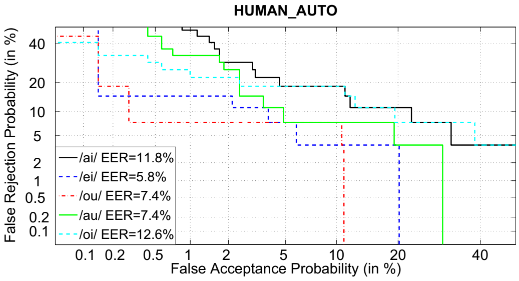

Figure 2: DET plots showing discriminating power for each diphthong with HUMAN_AUTO strategy.

| Dipht | HUMAN_SEMI | HUMAN_AUTO |
|---|---|---|
| /aɪ/ | 0.061 | 0.176 |
| /eɪ/ | 0.063 | 0.105 |
| /oʊ/ | 0.077 | 0.100 |
| /aʊ/ | 0.105 | 0.213 |
| /ɔɪ/ | 0.082 | 0.293 |

Table 3: $C_{llr}^{min}$ values showing discriminating power for each diphthong with HUMAN_AUTO strategy.

In Table 4 the results of fusing and post-calibrating the log-LR values of all the diphthongs for each comparison is shown, both as an overall performance measure $C_{llr}$ and with $C_{llr}^{min}$ as a measure of discriminating power. First, we observe that $C_{llr}$ and $C_{llr}^{min}$ values are quite close for all cases, which indicates a good calibration performance after jackknife logistic regression. This is normal, since jackknife over the test database implies highly matching conditions for calibration. Second, the HUMAN_SEMI strategy [4] achieves better performance than the rest of approaches (in fact, it gives perfect separation $C_{llr}^{min} = 0$). Moreover, although $C_{llr}$ relatively doubles for the best automatic formant tracking procedure, its value remains low in absolute terms (e.g., an increase of 0.054 from HUMAN_SEMI to BEST_IND_AUTO). Finally, the BEST_ALL_AUTO strategy performs only slightly worse than the BEST_IND_AUTO in absolute terms, which justifies the use of the same feature set for all diphthongs.

| | Before selection | | After selection | |
|---|---|---|---|---|
| Strategy | $C_{llr}^{min}$ | $C_{llr}$ | $C_{llr}^{min}$ | $C_{llr}$ |
| BEST_IND_AUTO | 0.045 | 0.110 | 0 | 0.0192 |
| BEST_ALL_AUTO | 0.074 | 0.127 | 0.0058 | 0.0273 |
| HUMAN_AUTO | 0.105 | 0.181 | 0.0074 | 0.0225 |
| HUMAN_SEMI [4] | 0 | 0.056 | - | - |

Table 4: Performance of automatic formant tracking strategies before and after feature selection.

### 4.2. Results on feature selection analysis

Table 4 shows $C_{llr}^{min}$ and $C_{llr}$ performance values for the three strategies using automatic formant tracking after the feature selection algorithm proposed in Section 2.2. It can be seen that BEST_IND_AUTO strategy outperforms the rest after feature selection, reaching perfect separation ($C_{llr}^{min} = 0$), and being also better than HUMAN_SEMI before feature selection. Moreover, $C_{llr}$ values for BEST_ALL_AUTO and HUMAN_AUTO after feature selection are also extremely low, indicating excellent performance.

It is worth noting that, due to the low number of comparisons allowed by the database used, the feature selection strategy is applied over the same data in which it is tested. Thus, it is not possible to check if the feature selection strategy improves the performance on new, unseen data. However, this analysis allows to highlight the influence of each formant in the discriminating power of the recognition process. Figure 3 shows a chart representing the final selection of features for the three proposed strategies. It can be seen that, for all cases, F2 seems to contribute with more features to the final selected set, whereas features from F3 are almost not selected, although typically F3 is assumed as significantly discriminating. This is because of the difficulty of reaching a highly accurate automatic extraction of the F3 trajectories. It is worth noting that in the semi-automatic formant tracking procedure followed by HUMAN_SEMI, the final trajectory for F3 is manually chosen among 8 different strategies [4]. This implies a much higher accuracy in F3 formant trajectories for HUMAN_SEMI than for the rest of automatic approaches. Moreover, it can be also seen that for the proposed algorithm /aɪ/ and /oʊ/ are the most feature-contributing diphthongs. These kind of studies may help the expert to decide which units and features are important from the available speech material.

## 5. Conclusions

In this paper we have presented a comparison among the performance of semi-automatic and automatic formant tracking approaches in forensic speaker recognition using traditional features. The automation of the formant tracking procedure makes the recognition process much faster. Therefore, for each comparison much more data can be used for comparison, which is especially necessary for robust modelling of a relevant population of many speakers in likelihood-ratio based forensic speaker recognition. Results show that performance with automatic formant tracking is worse, but still acceptable. Moreover, we have proposed a feature selection algorithm, which allowed us to analyze the impact of each traditional feature extracted in the discriminating power of the recognition process. Finally, it is worth noting that, although the database used is small, limited and controlled, expert analysis of a database for forensic speaker comparison is a highly demanding and time consuming process, which requires language proficiency. Thus, such corpora are extremely valuable and rare. Future work is mainly focused on the use of a speech recognizer for diphthong or phoneme segmentation, which would lead to a fully automatic approach for forensic speaker recognition using traditional features. We also plan to test the comparison of automatic and expert-based procedures in more realistic scenarios in terms of speech variability and number of speakers.

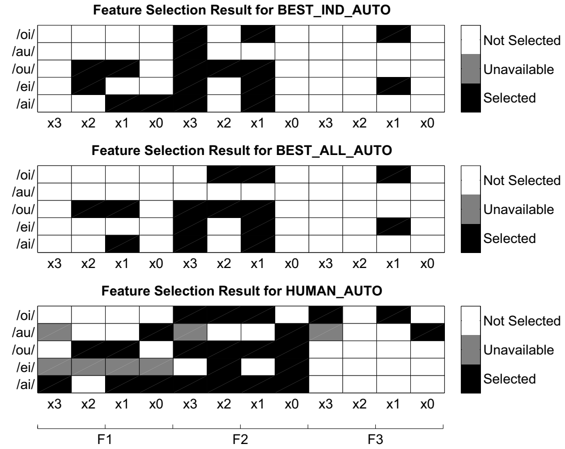

Figure 3: Features selected for automatic formant tracking strategies. Rows are diphthongs. Columns are different degrees of the polynomial or frequency index of the DCT for each formant.

## 6. References

[1] P. Rose, *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, 2002.

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[3] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

[4] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America*, vol. 125, no. 4, April 2009, In press.

[5] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[6] D. Rudoy, D. N. Spendley, and P. J. Wolfe, "Conditionally linear Gaussian models for estimating vocal tract resonances," in *Proc. of Interspeech*, Antwerp, Belgium, 2007, pp. 526–529.

[7] Y. Kinoshita and T. Osanai, "Within speaker variation in diphthongal dynamics: What can we compare?," *Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand*, 2006.