A first attempt at compensating for effects due to recording-condition mismatch in formant-trajectory-based forensic voice comparison

Ewald Enzinger[1,2]

[1]School of Elec. Eng. & Telecom., University of New South Wales, Sydney, Australia
[2]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

e.enzinger@student.unsw.edu.au

Abstract

This paper reports the results of a first attempt at compensating for variability in formant-trajectory representations due to differences in suspect and offender recording conditions. Formant-trajectory measurements were made on tokens of /iau/ in a database of high-quality and mobile-to-landline-transmitted recordings of 60 female speakers of Chinese. Discrete cosine transforms (DCT) were fitted to the formant trajectories. DCTs were transformed using feature mapping and using canonical linear discriminant functions. Transformed coefficients were used to calculate likelihood ratios via the multivariate kernel density formula. Application of the compensation procedures did not lead to substantial improvement in validity or reliability.
Index Terms: Forensic voice comparison, formant trajectories, mismatch compensation, likelihood ratios

1. Introduction

A common scenario in forensic-voice-comparison (FVC) casework is that a recording of the voice of an offender obtained from a telephone call is being compared with a direct-microphone recording of the voice of a suspect obtained during a police interview. Transmission and recording systems can have profound effects on the speech signal on these recordings, affecting the measurement of formant frequencies and their trajectories. A number of studies have drawn attention to the effects of different telephone-transmission systems on formant measurement, with average differences in formant frequency measurements of up to 23% in landline and 29% in mobile-telephone-transmitted speech (see [1] for a review). When performing forensic voice comparison based on formant-trajectory measurements on offender and suspect samples with mismatched recording conditions, an increase in variability in the measurements is expected. Assuming that the effect on within-speaker variability is greater than that on between-speaker variability, the overall ratio of between- and within-speaker variability is expected to decrease, thereby negatively affecting forensic-voice-comparison performance [2, p. 10]. A study investigating the impact of mismatch in recording conditions with respect to telephone transmission on the performance of formant-trajectory-based forensic-voice-comparison systems after fusion with a mel-frequency-cepstral-coefficient (MFCC) Gaussian-mixture-model universal-background-model (GMM-UBM) based system found substantial deterioration in system validity as compared to matched high-quality recording conditions, in particular for test conditions involving transmission over mobile-telephone networks [1].

Given these results we would like to have a procedure to compensate for the mismatch between the recording conditions of the offender and the suspect recordings. To the best of our knowledge no prior work has investigated compensation for effects on formant or formant-trajectory measurement due to recording-condition mismatch in forensic voice comparison. Here we consider statistical approaches to counter increased variability in formant-trajectory representations due to differences in suspect and offender recording conditions. Human-supervised formant-trajectory measurements were made on tokens of /iau/ in high-quality and mobile-to-landline-transmitted recordings. Discrete cosine transforms (DCT) were fitted to the formant trajectories. Three methods for compensation were investigated:

1. mapping DCT coefficients in the offender condition (*mobile-to-landline*) towards the distribution of DCT coefficients in the suspect condition (*high-quality*),
2. transforming DCT coefficients using canonical linear discriminant functions [3], discarding dimensions that are assumed to capture variability due to mismatched conditions, and
3. combining both methods, first applying feature mapping followed by canonical linear discriminant function transformation.

The validity and reliability of a forensic-voice-comparison system incorporating mismatch compensation are assessed and compared with that of a baseline forensic-voice-comparison system using the original DCT coefficients. In addition, all systems are fused with a MFCC GMM-UBM based system, and the relative change in performance is assessed.

2. Methodology

2.1. Database

The data were extracted from a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese [4]. See [5] for details of the data collection protocol. The recordings used were from an information-exchange task conducted over the telephone: Each of a pair of speakers received a "badly transmitted fax" including some illegible information, and had to ask the other speaker to provide them with the missing information. The original recordings were approximately 10 minutes long. The first and second recording sessions were separated by 2-3 weeks. High-quality recordings were made at 44,100 samples per second 16 bit quantization using flat-frequency-response lapel microphones (Sennheiser MKE 2 P-C) and an external soundcard (Roland® UA-25 EX), with one speaker on each of the two recording channels. Stressed tokens of /iau/ on tone 1 were manually located and marked.

In addition to the original high-quality recordings, degraded sets of recordings were created by passing the high-quality set of recordings through a mobile-to-landline transmission channel. The details of the procedure are described in [1]. The degraded recording was aligned with the high-quality recording by sliding the degraded signal past the high-quality signal in the time domain and calculating the correlation between the two signals at each sample displacement. At the displacement with the highest correlation, the degraded signal was truncated to the same start and end points as the high-quality signal. Alignment allowed the use of the same /iau/ markers as were created using the high-quality recordings. The high-quality condition was treated as the condition of the suspect (known identity) recording, and the mobile-to-landline condition was treated as the condition for the offender (questioned identity) recording.

2.2. Formant-trajectory-based system

Human-supervised measurements of the trajectories of the first three formants (F1, F2, and F3) of /iau/ tokens were made using FORMANTMEASURER [6]. See [1, 7] for details on the procedure for human-supervised formant-trajectory measurement. All tokens in all conditions were measured by the same supervisor. Discrete cosine transforms (DCTs) were fitted to the measured formant trajectories of all the /iau/ tokens. On the basis of tests made on the development set, the zeroth through fourth DCT coefficient values from F2 and F3 were used as variables in the present study. The trajectory of the first formant was excluded as it was expected to be greatly affected by the telephone bandpass, therefore negatively affecting performance. This was also confirmed in a preliminary test on the development set using trajectories of all three formants. Likelihood ratios were calculated using the multivariate kernel density (MVKD) formula [8]. These were then calibrated using logistic-regression (see Section 2.4). This system is henceforth referred to as the *baseline* system.

2.3. MFCC GMM-UBM system

The forensic-voice-comparison system extracted 16 mel-frequency-cepstral-coefficients (MFCCs) every 10 ms over the entire speech-active portion of each recording using a 20 ms wide Hamming window. Delta coefficient values were calculated and included in the statistical modelling [9]. Feature warping [10] using Gaussian cumulative distribution function matching with a 3 second sliding window was applied to the MFCCs and deltas before subsequent modelling. A Gaussian mixture model (GMM) with diagonal covariance matrices was trained using the background data as a background model [11]. Suspect speaker GMMs are adapted from the background model using maximum a-posteriori (MAP) adaptation. Following tests on the development set using a range of values, the number of Gaussians was set to 256 and the number of MAP iterations was set to 1. Training and adaptation of GMMs was performed using an implementation provided by the Hidden Markov Toolkit [12]. A score was calculated as

$$\text{score} = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{p(\mathbf{x}_i|\lambda_{\text{suspect}})}{p(\mathbf{x}_i|\lambda_{\text{UBM}})}\right), \qquad (1)$$

where $\mathbf{x}_i$ is a MFCC feature vector, N is the number of feature vectors in the offender data, and λ_{suspect} and λ_{UBM} represent the models of the suspect and the background, respectively. The scores were then converted into likelihood ratios using logistic-regression calibration (see Section 2.4)

2.4. Use of background, development, and test sets

In the tests of forensic-voice-comparison systems described below, tokens from the first 20 speakers were used as background data, tokens from the next 20 speakers were used as development data, and tokens from the last 20 speakers were used as test data. In both the development and test sets, every speaker's Session 1 recording (offender recording) was compared with their own Session 2 recording (suspect recording) for a same-speaker comparison and with every other speaker's Session 1 as well as Session 2 recording (suspect recordings) as different-speaker comparisons. The offender recordings were mobile-to-landline transmitted recordings, and the suspect recordings and the background were high-quality recordings. Both Session 1 and Session 2 recordings were included in the background. The development set was used to calculate scores which were then used to calculate weights for logistic-regression calibration [13, 14, 15] which was applied to convert the scores from the test set to likelihood ratios. Logistic regression was also used to fuse the scores from multiple individual systems and convert them to likelihood ratios [16]. In tests on the development set, scores were calibrated in a cross-validation procedure.

3. Mismatch compensation

3.1. Method 1: Feature mapping

In the first method DCT coefficients obtained from formant-trajectory measurements of /iau/ tokens of the offender sample (*mobile-to-landline*) are mapped towards the distribution of DCT coefficients in the suspect condition (*high-quality*). Figure 1 illustrates the approach. The solid blue curve shows the distribution of features in the mobile-to-landline condition. These features were then shifted by an offset $\bar{\boldsymbol{\delta}}$ (Eq. 2a) (blue arrow) so that the mean of their distribution (dashed blue curve) approximates that of their theoretical distribution in the high-quality condition (red curve).

The offset values $\bar{\boldsymbol{\delta}}$ are estimated using training data. The training data consist of DCT coefficients obtained from formant-trajectory measurements of /iau/ tokens in mobile-to-landline and high-quality recordings of speakers in the background set. For each set of DCT coefficients of a /iau/ token in high-quality condition there exists another set of DCT coefficients of the same time-aligned /iau/ token in mobile-to-landline condition (see Section 2.1). To obtain $\bar{\boldsymbol{\delta}}$ we first calculated the average of the signed differences between each pair

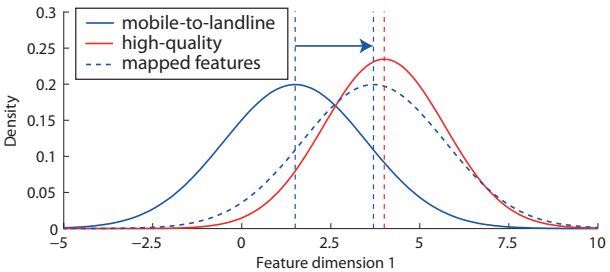

Figure 1: Example of feature mapping. The solid blue curve shows the distributions of features of a speaker in mobile-to-landline condition. The red curve shows their theoretical distribution in high-quality condition. The dashed blue line shows the shifted distribution of mapped mobile-to-landline-condition features.

of N_s high-quality and mobile-to-landline sets of DCT coefficients $\mathbf{x}_{j,\text{suspect}}$ and $\mathbf{x}_{j,\text{offender}}$ of a speaker s as a recording-condition-dependent offset $\boldsymbol{\delta}_s$ (Eq. 2c). Then, the offset $\bar{\boldsymbol{\delta}}$ is calculated as average of the per-speaker offsets $\boldsymbol{\delta}_s$ of each of S speakers in the background set (Eq. 2b).

$$\mathbf{x}_{i,\text{mapped}} = \mathbf{x}_{i,\text{offender}} + \bar{\boldsymbol{\delta}} \tag{2a}$$

$$\bar{\boldsymbol{\delta}} = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\delta}_s \tag{2b}$$

$$\boldsymbol{\delta}_s = \frac{1}{N_s}\sum_{j=1}^{N_s}(\mathbf{x}_{j,\text{suspect}} - \mathbf{x}_{j,\text{offender}}) \tag{2c}$$

Additional scaling by the ratio of the pooled variance estimates of sets of DCT coefficients from high-quality and mobile-to-landline formant-trajectory measurements resulted in deterioration in performance in tests on the development set as compared to only shifting by $\bar{\boldsymbol{\delta}}$.

3.2. Method 2: Canonical linear discriminant functions

Canonical linear discriminant functions (CLDF) are linear combinations of variables that are derived so that the groups in the training data are maximally separated on the new dimensions described by the functions [3]. A series of orthogonal functions are derived with the first accounting for more of the between-group variation than the second, the second accounting for more of the between-group variation than the third, etc. Here, the variables are DCT coefficients obtained from formant-trajectory measurements and the groups to be separated are the speakers. In the estimation of the CLDF coefficients both within- and between-group variation are taken into account.

To estimate the CLDF coefficients we first obtained the within- ($\mathbf{S}_w$) and between-speaker ($\mathbf{S}_b$) scatter matrices from DCT coefficients, pooled from suspect and offender conditions,

$$\mathbf{S}_w = \sum_{s=1}^{S}\sum_{i=1}^{N_s}(\mathbf{x}_{s,i} - \boldsymbol{\mu}_s)(\mathbf{x}_{s,i} - \boldsymbol{\mu}_s)^T \tag{3}$$

$$\mathbf{S}_b = \sum_{s=1}^{S}(\boldsymbol{\mu}_s - \boldsymbol{\mu})(\boldsymbol{\mu}_s - \boldsymbol{\mu})^T, \tag{4}$$

where $\mathbf{x}_{s,i}$ are the DCT coefficients obtained from formant-trajectory measurements of a /iau/ token i of speaker s,

$$\boldsymbol{\mu}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_{s,i} \tag{5}$$

are the mean DCT coefficients for each speaker s, and $\boldsymbol{\mu}$ are the overall mean DCT coefficients. The between-to-within-class variability is maximized by solving the generalized eigenvalue problem:

$$\mathbf{S}_w\mathbf{v} = \boldsymbol{\lambda}\mathbf{S}_b\mathbf{v} \tag{6}$$

A transformation matrix $\mathbf{P}$ composed of the eigenvectors $\mathbf{v}$ with the highest k eigenvalues λ (sorted in descending order according to their corresponding eigenvalue) is then used to transform DCT coefficients, discarding dimensions that predominantly capture variability due to mismatched distances while retaining those capturing speaker-specific information:

$$\mathbf{y}_i = \mathbf{P}\mathbf{x}_i \tag{7}$$

Using a geometric analogy, we estimate a transformation from the space of DCT coefficients to a lower-dimensional space of canonical linear discriminant functions. Figure 2 gives a graphical example of the method. The blue and red data points are projected onto a line so that the two groups minimally overlap. The orientation of the line is determined using the methods outlined above. In reality we had d dimensions projected to $d-1$ canonical linear discriminant functions, but this example only shows two dimensions and one canonical linear discriminant function.

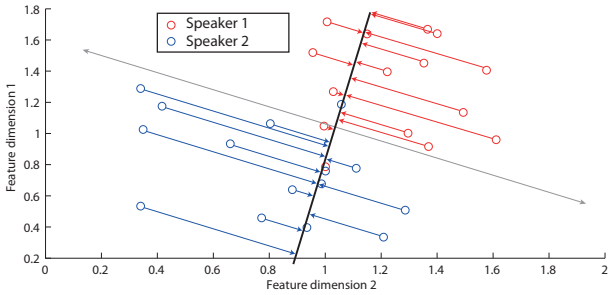

Figure 2: Example of canonical linear discriminant function transformation of data points of two speakers in two dimensions.

The number of eigenvectors $k = 9$ to retain in the CLDF transformation was experimentally set based on the lowest pooled C_{llr} value obtained in tests on the development set.

Similar techniques are commonly applied for channel and session compensation to reduce mismatch in i-vectors in state-of-the-art automatic speaker recognition systems [17].

3.3. Method 3: Combination of feature mapping and CLDF

In the third method we first apply feature mapping by shifting DCT coefficients in the offender condition towards the distribution of DCT coefficients in the suspect condition. As a second step DCT coefficients are transformed using canonical linear discriminant functions. The rationale for combining both approaches is that the CLDF transformation may be able to discard dimensions representing residual within-speaker variability.

4. Results

The validity and reliability of the systems was evaluated using the log likelihood-ratio cost (C_{llr}) as a metric of validity (accuracy), and an estimate of the 95% credible interval (95% CI) as a metric of reliability (precision) [18, 19] (C_{llr} was calculated using the mean procedure [18, §3.3] and the 95% CI using the parametric procedure [19, §2.3]).

Figure 3 shows the results of the baseline system and systems incorporating mismatch compensation. All three methods show improvements in validity. Methods 2 and 3 further increase reliability as compared to the baseline system.

Figure 4 shows the results of the baseline system and systems incorporating mismatch compensation after fusion with the MFCC GMM-UBM system. Method 3 shows minor improvements in both validity and reliability. Methods 1 and 2 show an increase in validity while reliability deteriorates.

5. Discussion & Conclusion

The present paper investigates three methods to compensate for variability due to recording-condition mismatch in formant-trajectory-based forensic voice comparison. The first method

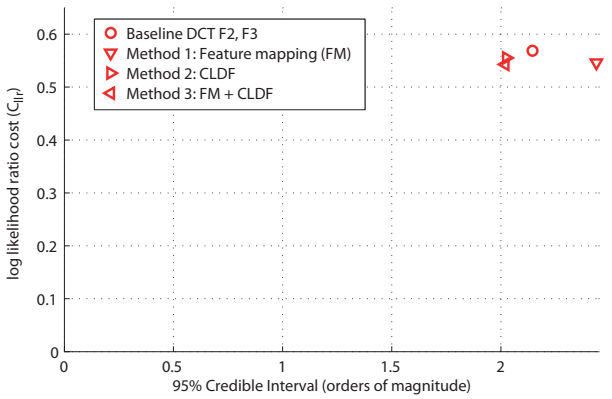

Figure 3: Measures for validity (C_{llr}) and reliability ($\log_{10}$ 95% CI) for systems based on the original DCT coefficients of F2 and F3 (red circle) as well as after applying feature mapping (FM, $\triangledown$), canonical linear discriminant function transform (CLDF, $\triangleright$), and a combination of both FM and CLDF approaches ($\triangleleft$).

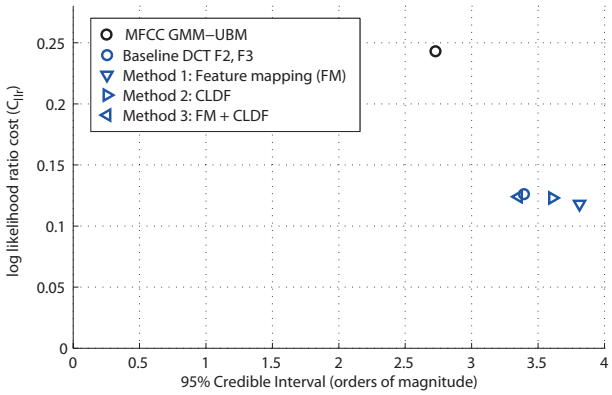

Figure 4: Measures for validity (C_{llr}) and reliability ($\log_{10}$ 95% CI) for the MFCC GMM–UBM system (black circle) and fusion of the MFCC GMM–UBM system with systems based on the original DCT coefficients of F2 and F3 (blue circle) as well as after applying feature mapping (FM, $\triangledown$), canonical linear discriminant function transform (CLDF, $\triangleright$), and a combination of both FM and CLDF approaches ($\triangleleft$).

maps DCT coefficients in the offender condition towards the distribution of DCT coefficients in the suspect condition. The second method transforms DCT coefficients using canonical linear discriminant functions, discarding dimensions that are assumed to capture variability due to mismatched conditions. The third method combines both approaches, first applying feature mapping followed by canonical linear discriminant function transformation.

While improvements in both validity and reliability were observed compared to the baseline system, none of the methods achieved to substantially reduce the effects due to recording-condition mismatch on system performance after fusion with the MFCC GMM-UBM system. One potential reason for failing to find improvement could be that differences in formant-trajectory measurements caused by the mobile-telephone transmission channel cause non-linear changes in fitted DCT coefficients which could not be reduced by the proposed methods.

6. Acknowledgements

Thanks to Geoffrey Stewart Morrison (Department of Linguistics, University of Alberta, Edmonton, Canada) and Cuiling Zhang (Department of Forensic Science & Technology, China Criminal Police University).

7. References

[1] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices," *Speech Commun.*, vol. 55, pp. 796–813, 2013.

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

[3] W. R. Klecka, *Discriminant analysis*. Beverly Hills, CA: Sage Publications, 1980.

[4] C. Zhang and G. S. Morrison. (2011) Forensic database of audio recordings of 68 female speakers of standard chinese. [Online]. Available: http://databases.forensic-voice-comparison.net/

[5] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Aus J Forensic Sci.*, vol. 44, no. 2, pp. 155–167, 2012.

[6] G. S. Morrison and T. M. Nearey. (2011) FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories. [Online]. Available: http://geoff-morrison.net/#FrmMes

[7] C. Zhang, G. S. Morrison, F. Ochoa, and E. Enzinger, "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," *J. Acoust. Soc. Amer.*, vol. 133, pp. EL54–EL60, 2013.

[8] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics*, vol. 53, no. 1, pp. 109–122, 2004.

[9] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 34, pp. 52–59, 1986.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey*, 2001.

[11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[12] S. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Department of Engineering, Cambridge University, U.K., Tech. Rep., 1993.

[13] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, pp. 230–275, 2006.

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

[15] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Aus J Forensic Sci.*, vol. 45, pp. 173–197, 2013.

[16] S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Process.*, vol. 10, pp. 237–248, 2000.

[17] M. McLaren and D. van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources," *IEEE Trans. Audio, Speech Lang. Proc.*, vol. 20, no. 3, pp. 755–766, 2012.

[18] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Sci. Justice*, vol. 51, pp. 91–98, 2011.

[19] G. S. Morrison, T. Thiruvaran, and J. Epps, "Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system," in *Proc. Odyssey 2010: The Language and Speaker Recognition Workshop*, 2010, pp. 63–70.