Joint Cohort Normalization in a Multi-Feature Speaker Verification System

C. Sanderson and K. K. Paliwal
School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia

Abstract— **In this paper we propose a new fusion technique, termed *Joint Cohort Normalization Fusion*, where the information fusion is done prior to the likelihood ratio test in a speaker verification system. The performance of the technique is compared against two popular types of fusion: feature vector concatenation and expert opinion fusion, for fusion of Mel Frequency Cepstral Coefficients (MFCC), MFCC with Cepstral Mean Subtraction (CMS) and Maximum Auto-Correlation Values (MACV) features. In experiments on the NTIMIT database, the proposed technique is shown, in most cases, to outperform the popular methods.**

I. Introduction

Identity verification systems are now a part of our every day life. As an example, Automatic Teller Machines (ATMs) employ a simple identity verification where the user is asked to enter their Personal Identification Number (PIN), known only to the user, after inserting their ATM card. If the PIN matches the one prescribed to the card, the user is allowed access to their bank account. Similar verification systems are widely employed to restrict access to rooms and buildings.

The verification system such as the one used in the ATM only verifies the validity of the combination of a certain possession (in this case, the ATM card) and certain knowledge (the PIN). The ATM card can be lost or stolen, and the PIN can be compromised (eg. somebody looks over your shoulder while you're entering the PIN). Hence new verification methods have emerged, where the PIN has either been replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints. The use of biometrics is attractive since they cannot be lost or forgotten and vary significantly between people.

The performance of a verification system is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as:

$$\text{FA} = \frac{I_A}{I_T} \times 100\% \qquad \text{FR} = \frac{C_R}{C_T} \times 100\%$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests.

To quantify the performance into a single number, two measures can be used: Equal Error Rate (EER), where the system is configured to operate with FA = FR and Total Error (TE), defined as TE = FA + FR.

Verification systems based on speech have proven to be quite effective [1]. However their performance is still not perfect. They usually rely on only one type of feature extraction, namely Mel Frequency Cepstral Coefficients (MFCC) or MFCC with Cepstral Mean Subtraction (CMS).

In [2] information from both MFCC and MFCC-CMS features was used to reduce the error rates in a speaker identification system (from here on, MFCC-CMS features shall be referred to as CMS features).

Recently new type of features, named Maximum Auto-Correlation Values (MACV), have been proposed to augment the cepstral coefficient feature vector [3]. The MACV feature set contains both voicing and reliable pitch information. In a speaker identification scenario, this feature set was shown to reduce error rates on a variety of databases.

Two popular fusion methods, namely feature vector concatenation and expert opinion fusion, have been studied in [4] for fusion of MFCC, CMS and MACV features. In this paper we propose a new fusion technique, which we have termed as *Joint Cohort Normalization Fusion* and compare its performance against the established fusion methods.

The rest of the paper is organized as follows. In Section II, we briefly describe the MFCC, CMS and MACV features. In Section III, we describe a Gaussian Mixture Model (GMM) modality expert, which shall be used as the basis for fusion experiments. In Section IV, we describe the concatenation and opinion fusion techniques as well as the proposed method. The performance of all fusion techniques is compared in Section V.

II. Speech Feature Extraction

A. MFCC Features

The human ear processes the speech signal using a bank of non-uniformly spaced filters. Features extracted using such a filter-bank have been shown to be quite effective for speaker verification [1].

For a given speech frame (usually 20 ms in length), the spectrum is obtained using the Fast Fourier Transform (FFT). The square of the magnitude of the spectrum is taken and the result is then multiplied by a pre-emphasis filter to emphasize the high-frequency portion. 17 Mel-scale triangular filter bank energies [6] are then calculated and are expressed in logarithmic scale [7].

The upper and lower passband frequencies of each filter are the center frequencies of the adjacent filters. The frequency range of the filters was chosen to cover the telephone bandwidth. The central frequencies of the 17 filters are (in Hz): 300, 400, 500, 600, 700, 800, 900, 1000, 1149, 1320, 1516, 1741, 2000, 2297, 2639, 3031 and 3482. Since the filter bank coefficients are highly correlated, a discrete cosine transform is used to de-correlate them:

$$c_i = \frac{1}{N_F} \sum_{j=1}^{N_F} f_j \cos\{ \frac{\pi i}{N_F} (j-0.5) \} \quad i = 0, 1, ..., N_F - 1 \tag{1}$$

where N_F is the number of filters and f_j are the log filter bank energies. The MFCC feature vector is made up of $\{c_i, i = 1, 2, ..., N_F - 1\}$, ie. c_0 is omitted. c_0 represents the average value of the spectrum and hence is susceptible to varying background noise.

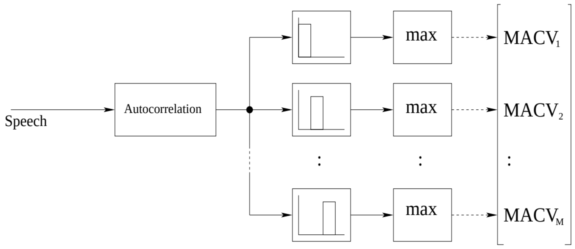

Fig. 1. MACV feature extractor (after [3])

B. CMS Features

Given a sequence of MFCC feature vectors from a speech utterance, $\{\vec{c_i}, i = 1, 2, ..., N_V\}$, we define their mean as $\vec{c_\mu}$. The mean is assumed to represent the cepstrum of the channel [9]. Thus the sequence of CMS feature vectors is obtained using:

$$\vec{d_i} = \vec{c_i} - \vec{c_\mu}, \quad i = 1, 2, ..., N_V \tag{2}$$

CMS features have been shown [5] to be significantly more immune to the effects of channel distortion. However, it has also been shown that the cepstral mean also contains the average speech cepstrum, which contains speaker information [8], [9]. Thus removal of the $\vec{c_\mu}$ from MFCC features is a double-edged sword: on one hand it makes the verification system more robust against channel mismatches, while on the other it reduces the accuracy of the system in clean conditions.

C. MACV Features

Given a speech frame $\{s(n), n = 0, 1, ..., N_S - 1\}$, the MACV features are computed as follows:

1. Compute the autocorrelation function:

$$R(k) = \frac{1}{N_S} \sum_{n=0}^{N_S-1-k} s(n)s(n+k), \ k = 0, ..., N_S - 1 \tag{3}$$

2. Normalize $\{R(k)\}$ by its value at $k = 0$, i.e., $\hat{R}(k) = \frac{R(k)}{R(0)}$
3. Divide the higher portion of $\{\hat{R}(k)\}$ into M equal parts
4. Find the maximum value of $\{\hat{R}(k)\}$ for each of the M divisions
5. The M Maximum Autocorrelation Values (MACV) form a M-dimensional feature vector

A conceptual block diagram of this process is shown in Fig. 1.

The lower portion of $\{\hat{R}(k)\}$ is not used as it contains information about the system component of speech (vocal tract). This is already used in speaker recognition systems in the form of cepstral coefficients.

III. GMM Based Modality Expert

The distribution of feature vectors for each person is modeled by a Gaussian Mixture Model (GMM). Given a set of training vectors, an N_M-mixture GMM is trained using a k-means clustering algorithm followed by 10 iterations of the Expectation Maximization (EM) algorithm [10].

Given a claim for person C's identity and a set of feature vectors $X = \{\vec{x_i}, i = 1, 2, ..., N_V\}$ supporting the claim, log likelihood of the claimant being the true claimant is calculated using:

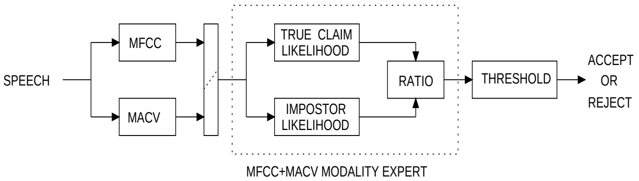

Fig. 2. Concatenation fusion based verification system

$$\log p(X|\lambda_C) = \frac{1}{N_V} \sum_{i=1}^{N_V} \log\{p(\vec{x}_i|\lambda_C)\} \tag{4}$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{i=1}^{N_M} m_i \ \mathcal{N}(\vec{x}, \vec{\mu}_m, \mathbf{\Sigma}_m) \tag{5}$$

$$\text{and} \quad \lambda = \{m_i, \vec{\mu}_i, \mathbf{\Sigma}_i, i = 1, 2, ..., N_M\} \tag{6}$$

Here λ_C is the model for person C. N_M is the number of mixtures, m_i is the weight for mixture i, and $\mathcal{N}(\vec{x}, \vec{\mu}, \mathbf{\Sigma})$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$.

Given a set of B background person models [1] (also known as cohorts) $\{\lambda_b, b = 1, 2, ..., B\}$ for person C, the log likelihood of the claimant being an impostor is found using:

$$\log p(X|\lambda_{\overline{C}}) = \log\{\frac{1}{B} \sum_{b=1}^{B} p(X|\lambda_b)\} \tag{7}$$

In practice it was observed that only one of the background speaker models usually dominates the above sum. To find out whether the claimant is a true claimant or an impostor, the following likelihood ratio is calculated:

$$r = \frac{p(X|\lambda_C)}{p(X|\lambda_{\overline{C}})} \tag{8}$$

In the log domain this becomes:

$$R = \log p(X|\lambda_C) - \log p(X|\lambda_{\overline{C}}) \tag{9}$$

In a single modality system the decision is reached as follows: given a threshold t, the claim is accepted when $R \geq t$; the claim is rejected when $R < t$. However, to use the above verification system as part of a larger system, the final thresholding is omitted. Instead an opinion, o, on the claim is generated using $o = R$. We shall refer to a verification system without the final thresholding stage as a *modality expert*.

IV. Fusion Techniques

A. Feature Vector Concatenation Fusion

In this fusion approach, two or more feature vectors are concatenated to form a single feature vector. The advantage of this approach lies in its simplicity and allows for the modeling of redundancies (for increased robustness) between different features. A block diagram of an example verification system employing concatenation fusion is shown in Fig. 2.

B. Expert Opinion Fusion

In opinion fusion, each feature type is processed independently by a modality expert. The opinions from ν modality experts then form a ν-dimensional opinion vector which is used by a *decision stage*. Since there are only two possible

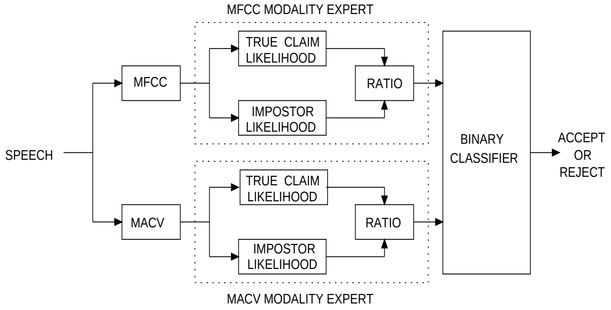

Fig. 3. Opinion fusion based verification system

outcomes (accept or reject), the decision stage can be a binary classifier [11]. The classifier is trained with example opinions of known impostors and true claimants. It then classifies a given opinion vector as belonging to either the impostor or true claimant class.

An intuitive advantage of the opinion fusion approach is that the opinions can be weighted. The weight for each modality expert can be selected according to its use for discrimination purposes and robustness.

The opinion value from each modality expert is first normalized to the $[0, 1]$ interval using an approach similar to [12]:

$$O_i = \frac{1}{1 + \exp\{-\alpha_i(o_i - t_i)\}} \tag{10}$$

where, for modality expert i, o_i is the opinion, t_i is the threshold to obtain the desired operating point for that modality and α_i indicates the interval of opinions. The normalized opinions are then fused using:

$$z = \sum_{i}^{\nu} w_i O_i \tag{11}$$

where ν is the number of modalities, w_i is the weight for modality i, with the constraint $\sum_i^{\nu} w_i = 1$. If $z < 0.5$, the claim is classified as an impostor; if $z \geq 0.5$ the claim is accepted. The normalization of opinions to the $[0, 1]$ interval is required to ensure opinions from all modalities are equally represented. This prevents any modality from dominating the fused opinion prior to weighting. An example verification system based on opinion fusion is shown in Fig. 3.

C. Joint Cohort Normalization Fusion

In the system described in Section IV-B, the information integration is done after the ratio test for each feature type. In the proposed fusion approach, which we shall term *Joint Cohort Normalization Fusion*, information integration is done prior to the ratio test.

Given a set of feature vectors of each type, $Y = \{X_i, i = 1, ..., \nu\}$ and a set of corresponding models for the claimed identity, $\lambda_D = \{\lambda_{C_i}, i = 1, ..., \nu\}$, the log likelihood of the claimant being the true claimant is calculated using [c.f. Eqn. (4)]:

$$\log P(Y|\lambda_D) = \sum_{i}^{\nu} w_i F[\log P(X_i|\lambda_{C_i})] \tag{12}$$

where ν is the number of feature vector types and w_i are the weights (with constraint $\sum_i^{\nu} w_i = 1$) for feature vectors of type i, while

$$F(x) = \frac{100}{1 + \exp\{-a(x - b)\}} \tag{13}$$

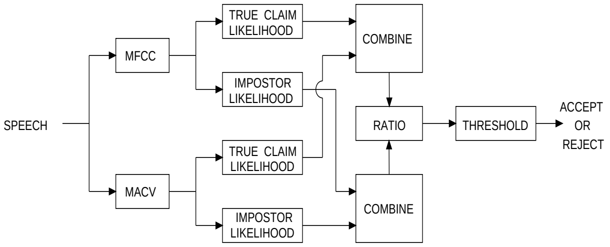

Fig. 4. Proposed Joint Cohort Normalization Fusion

normalizes $\log P(X_i|\lambda_{C_i})$ to be in the $[0, 100]$ interval. Here b is the mid point of the interval of $\log p_i(X_i|\lambda_{C_i})$, while a is the slope, selected so the above sigmoid covers the interval of $\log p_i(X_i|\lambda_{C_i})$.

The log likelihood of the claimant being an impostor is calculated using [c.f. Eqn. (7)]:

$$\log p(Y|\lambda_{\overline{D}}) = \log \frac{1}{B} \sum_{b=1}^{B} \exp \sum_{i=1}^{\nu} w_i F[\log p(X_i|\lambda_{b_i})]\} \tag{14}$$

where, for person C, λ_{b_i} is the b-th background speaker model for the i-th feature type. The normalization to the $[0, 100]$ interval is required for the same reasons as explained in Section IV-B. We use the $[0, 100]$ interval instead of $[0, 1]$ to ensure one of the background speakers dominates the sum. Finally, the opinion, o, is found using:

$$o = \log P(Y|\lambda_D) - \log p(Y|\lambda_{\overline{D}}) \tag{15}$$

The opinion is then thresholded to achieve the final accept/reject decision. A verification system utilizing the proposed fusion approach is shown in Fig. 4.

V. Fusion Experiments

The speech pre-processing and experimental setup used for experiments are similar to the work presented by Reynolds in [1]. In order to reduce modeling and detecting the environment rather than the speaker, a Speech Activity Detector (SAD) is used [7]. The detector tracks the noise floor of the signal and adapts to changing noise conditions. The portions of the signal which were marked as speech are then analyzed using a 20 ms Hamming window with a 10ms frame advance. Hence for each second of speech we extract 100 frames.

For MFCC, MACV and CMS features, the client models are 16 mixture GMMs with diagonal covariance matrices. For each speaker, 10 randomly selected background speakers were used.

For concatenated features, the number of mixtures is the sum of the number of mixtures used for each feature individually. Hence for the MFCC+MACV concatenated feature, 32 mixtures are used. This is necessary to keep the number of free parameters as similar as possible between experiments using different fusion approaches. For MACV features, we have found $M = 8$ to be optimal in preliminary experiments.

The experiments were performed on the NTIMIT database [13], which contains a phonetically balanced speech corpus transmitted over telephone lines. As in [1] only the *test* section of the database was used. For each of the 168 speakers, the 10 utterances were divided into 3 parts: train, validation and test. The first 5 utterances (sorted alphanumerically by filename) were assigned to the train part. The next 3 utterances were assigned to the validation part with the remaining 2 to the test part.

TABLE I

Performance of individual features.

feature	FA	FR	**TE**
MFCC	9.61	10.42	**20.03**
MACV	16.03	18.15	**34.18**
CMS	11.38	13.10	**24.48**

TABLE II

Performance of various fusion approaches. All results are quoted in TE. The asterix denotes the lowest TE for a particular feature combination.

fused features	concatenation	opinion	proposed
MFCC+MACV	19.67	18.84	* 16.92
MFCC+CMS	* 18.64	20.12	20.00
CMS+MACV	22.82	22.91	* 20.10
MFCC+MACV+CMS	22.25	19.57	* 17.03

The speaker models were generated from clean speech in the train part, while the validation part was used for obtaining thresholds, weights and opinions of known impostors and true claimants. For expert opinion fusion, a two step process was required for finding the thresholds and the weights. First the thresholds were found for EER performance, followed by weight selection by optimizing TE. For the proposed fusion, the weights and the threshold were optimized for EER performance. For concatenation fusion, the threshold was also optimized for EER performance.

The test part was used for final performance evaluation. For each speaker, his/her 2 test utterances were used separately as true claims, resulting in 336 true claimant tests. Impostor claims were simulated by using utterances from speakers other than the claimed speaker and his/her background speakers, resulting in 52752 impostor access tests.

To obtain baseline results, the individual performance of each feature was found. In this case, the verification system was made up of one modality expert and a thresholding stage. The results are presented in Table I.

The performance of all fusion approaches was found in four configurations: MFCC+MACV, MFCC+CMS, CMS+MACV and MFCC+MACV+CMS. The results are presented in Table II.

A. Discussion

The baseline results show that the most discriminating feature is MFCC, followed by CMS. The MACV feature obtains the worst performance, indicating that the pitch and voicing information is not sufficient by itself to distinguish speakers.

Table III shows TE reduction (in %) compared to best baseline feature in each combination. Example: compared to the MFCC feature alone, the TE for MFCC+MACV is 3.11 points lower, or reduced by 15.53%. The proposed fusion approach obtained the best performance and highest TE reduction for all bar one combination. The best performance was obtained by the MFCC+MACV combination, closely followed by the MFCC+MACV+CMS combination.

Concatenation fusion obtained the best performance for the combination of MFCC and CMS features. Compared to MFCC alone, the performance is slightly better, indicating the speaker information loss by CMS features has been diminished by MFCCs. Since the CMS features are quite similar to MFCC features, there is little complementary information. Hence the performance improvement could be due to discriminating information common to both CMS and MFCC being inadvertently emphasized during modeling.

TABLE III

TE reduction compared to best baseline feature in each combination

fused features	concatenation	opinion	proposed
MFCC+MACV	1.80%	5.94%	15.53%
MFCC+CMS	6.94%	(worse)	0.15%
CMS+MACV	6.78%	6.41%	17.89%
MFCC+MACV+CMS	(worse)	2.30%	14.98%

Ideally a fusion approach should at best provide better performance than any of the underlying features and never worse than any of them. The experimental results show that the proposed fusion approach satisfies this guideline, while the other approaches do not. The concatenation fusion approach breaches this guideline for the MFCC+MACV+CMS combination, and the opinion fusion approach for the MFCC+CMS combination.

VI. Conclusion

We have proposed a new fusion technique, termed *Joint Cohort Normalization Fusion*, for fusion of multiple speech features in a speaker verification system. In the proposed technique, the information fusion is done prior to the likelihood ratio test. The performance of the technique was compared against two popular types of fusion, feature vector concatenation and expert opinion fusion, for fusion of MFCC, CMS and MACV features. In experiments on the NTIMIT database, the proposed technique, in most cases, outperforms the popular methods.

References

[1] D. Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models”, *Speech Communication 17*, 1995.

[2] H. Altinçay, M. Demirekler, “On the use of Supra Model Information from Multiple Classifiers for Robust Speaker Identification”, *Proc. 6th European Conf. Speech Communication and Technology*, Budapest, 1999.

[3] B. Wildermoth, K. K. Paliwal, “Use of Voicing and Pitch Information for Speaker Recognition”, *Proc. 8th Australian Intern. Conf. Speech Science and Technology*, Canberra, 2000.

[4] C. Sanderson, K. K. Paliwal, “Information Fusion for Robust Speaker Verification”, *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH’01), Aalborg, 2001*

[5] D. Reynolds, “Experimental Evaluation of Features for Robust Speaker Identification”, *IEEE Trans. Speech and Audio Processing*, Vol. 2, 1994.

[6] J. Picone, “Signal Modeling Techniques in Speech Recognition”, *Proceedings of the IEEE*, Vol. 79, No. 4, 1991.

[7] D. Reynolds, “A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification”, *Technical Report 967*, Lincoln Laboratory, Massachusetts Institute of Technology, 1993.

[8] H. Gish, M. Schmidt, “Text-independent Speaker Identification”, *IEEE Signal Processing Magazine*, Oct. 1994.

[9] R. Balchandran et al, “Channel Estimation and Normalization by Coherent Spectral Averaging For Robust Speaker Verification”, *Proc. 6th European Conf. Speech Communication and Technology*, Budapest, 1999.

[10] T. K. Moon, “Expectation-maximization Algorithm”, *IEEE Signal Processing Magazine* Vol. 13, Iss. 6, 1996.

[11] S. Ben-Yacoub et al, “Fusion of Face and Speech Data for Person Identity Verification”, *Proc. IEEE Transactions on Neural Networks*, Vol. 10, No. 5, 1999.

[12] P. Jourlin et al, “Acoustic-labial speaker verification”, *Pattern Recognition Letters 18*, 1997.

[13] C. Jankowski et al, “NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database”, *Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, 1990.