Information Fusion for Robust Speaker Verification

Conrad Sanderson and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, Brisbane, QLD 4111, Australia

Abstract

In this paper we have studied two information fusion approaches, namely feature vector concatenation and decision fusion, for the task of reducing error rates in a speaker verification system used in mismatched conditions. Three types of features are fused: Mel Frequency Cepstral Coefficients (MFCC), MFCC with Cepstral Mean Subtraction (CMS) and Maximum Auto-Correlation Values (MACV). We have used the mismatch sensitivity of Linear Prediction Cepstral Coefficients (LPCC) as a speech quality measure for selecting the weight of the contribution of the MFCC modality in the adaptive decision fusion approach. We show that in most cases concatenation fusion is superior to decision fusion. The results lead us to propose a hybrid fusion approach in which two combinations of concatenation fusion are further fused using adaptive decision fusion. The hybrid system is shown to have the lowest error rates on both clean and noisy speech.

1. Introduction

It is well known that the performance of a speaker verification system easily degrades in the presence of a mismatch between training and testing conditions. Usually this is in the form of a channel distortion and/or ambient noise.

One popular method to alleviate the effects of channel mismatch is Cepstral Mean Subtraction (CMS). Unfortunately it has been shown [1, 2] that CMS also removes speaker information. In [3] information from both Mel Frequency Cepstral Coefficients (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 a new type of front-end, Maximum Auto-Correlation Values (MACV), has been proposed in [4] to augment the cepstral coefficient feature vector. The MACV feature contains both voicing and reliable pitch information. In a speaker identification scenario, the feature was shown to reduce error rates on a variety of databases.

The performance of a verification system is often presented as a Receiver Operating Characteristic (ROC) or Detection Error Trade-off graph [5]. The graph is obtained by varying the decision threshold and obtaining an operating point in terms of False Acceptance [FA(%)] rate and False Rejection [FR(%)] rate. While the graph useful for finding the discrimination ability of the system, it doesn't convey information on how the system will perform in real life applications. In mismatched conditions it can be easily observed that the distribution of impostor and true claimant scores change, hence the threshold found for a particular operating point on the training data corresponds to a different operating point on noisy test data. This is an additional source of performance degradation - we shall refer to this phenomenon as *operating point shift*.

In this paper we shall investigate two fusion techniques of combining MFCC, CMS and MACV features to alleviate the effects of the above mentioned sources of degradation. Linear Prediction Cepstral Coefficients (LPCC) will be used to detect the quality of the speech, which in turn is used to select the weight of the contribution of MFCC features.

The paper is structured as follows: in Section 2 MACV features are briefly explained. Section 3 shows two information fusion techniques. Section 4 is devoted to experiments evaluating the two techniques. The results are discussed in Section 5, which leads us to propose a hybrid fusion system, presented in Section 6, where both fusion techniques are used.

2. MACV Features

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

1. Compute the autocorrelation function from the speech signal using:

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

2. Normalise $\{R(k)\}$ by its value at $k = 0$, i.e., $\hat{R}(k) = \frac{R(k)}{R(0)}$

3. Discard the lower portion of $\{\hat{R}(k)\}$ as it contains the information about the system component of speech and is used in speaker recognition systems in the form of cepstral coefficients. Using only the higher portion (from 2 ms to 16 ms) of $\{\hat{R}(k)\}$:

   i. Divide the higher portion of $\{\hat{R}(k)\}$ into M equal parts.

   ii. Find the maximum value of $\{\hat{R}(k)\}$ for each of the M divisions.

   iii. The M Maximum Autocorrelation Values (MACV) form a M-dimensional feature vector.

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

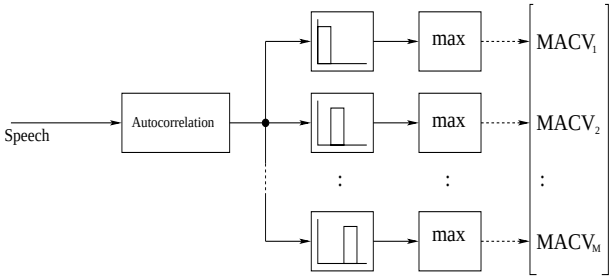

Figure 1: *MACV feature extractor (after [4]).*

3. Information Fusion

We shall investigate two types of information fusion: feature vector fusion and decision fusion. In the former approach, two or more different feature vector types are simply concatenated together to form a new feature vector. We shall refer to this approach as *concatenation fusion*.

In the latter approach, each feature type is processed independently by a *modality expert* (a verification system without the final thresholding decision stage) which produces an opinion on the claim. A relatively high opinion indicates the person is a true claimant, while a relatively low opinion suggests the person is an impostor. The opinions from n modality experts then form a n-dimensional opinion vector which is used by a *decision stage*. Since there are only 2 possible outcomes (accept or reject), the decision stage can be a binary classifier [6].

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 decision 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 mismatch susceptibility.

Many different binary classifiers can be used for the decision fusion approach - a prime example is the Support Vector Machine (SVM) [7]. However, SVM is not easily amenable to weight inclusion and has shown little performance advantage over a simple linear classifier [8]. For these reasons we have used the weighted linear classifier described in Section 3.1.

A simple method to obtain a measure of the condition of the speech signal is described in Section 3.2. We shall use this measure to modify the weights of each modality expert according to the amount of mismatch detected.

3.1. Weighted Linear Classifier

In this approach, the opinion value from each modality expert is first normalised to the $[0, 1]$ interval using [9]:

$$y_i = \frac{1}{1 + \exp\{-(x_i - t_i)\}} \tag{2}$$

where x_i is the opinion from modality expert i and t_i is the threshold to obtain the desired operating point for that modality expert. The normalised opinions are then fused using:

$$z = \sum_{i}^{n} w_i y_i \tag{3}$$

where w_i is the weight for modality i, with the constraint $\sum_i^n w_i = 1$. If $z < 0.5$, the claim is classified as an impostor; if $z \geq 0.5$ the claim is accepted.

3.2. Speech Mismatch Measure and Weight Adjustment

Previous work [10] has shown that LPCC features are sensitive to even a small amount of mismatch. While they're not useful for robust speaker verification, they can be employed to detect the amount of mismatch between the training and testing conditions.

Let us model speech parameterized using LPCC from all the speakers enrolled in the verification system by a 256 mixture Gaussian Mixture Model (GMM) and refer to it as the *clean* speech model. Given a set of LPCC feature vectors $\{\vec{v_i},\ i = 1,\ ...,\ N\}$ from the claimant's speech utterance, we work out the mismatch using:

$$q = \frac{1}{N} \sum_{i=1}^{N} \log[p(\vec{v_i}\,|\lambda_{clean})] \tag{4}$$

We convert q into the weight assigned to the MFCC modality using:

$$w = w_{min} + \frac{w_{max} - w_{min}}{1 + \exp\{-a(q - b)\}} \tag{5}$$

where w_{min} and w_{max} are minimum and maximum values of w respectively, while a and b are prior knowledge on how q changes according to the amount of mismatch. Weights for other modalities are then adjusted to take into account the $\sum_i w_i = 1$ constraint.

4. Experiments

The verification system and speech pre-processing used for experiments are similar to the work presented by Reynolds in [2]. The changes are as follows:

i. For MFCC, MACV and CMS features, the client models are 16 mixture GMMs with diagonal covariance matrices. We have found little improvement in using more mixtures.

ii. For concatenated features, the number of mixtures is the sum of 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.

iii. For each speaker, 10 random background speakers were used for the likelihood ratio test [2, 5].

Speech was analyzed every 10ms with a frame width of 20ms. For MFCC features, only the filters which fall in the telephone passband (0.3 - 3.4 kHz) were used. Cepstral coefficient $c[0]$ was omitted, resulting in a 16-dimensional feature vector. For MACV features, we have found $M = 8$ to be optimal in preliminary experiments. For LPCC features, 10th order analysis was used and 10 cepstral coefficients were derived. In all cases the speaker models were trained with k-means initialization followed by 10 iterations of the Expectation Maximization (EM) algorithm [11].

The experiments were performed on the NTIMIT database [12] which has various channel mismatches. Ambient noise was simulated with additive white Gaussian noise. The Signal to Noise Ratio (SNR) was varied from 30 dB to 0 dB in steps of 2 dB. As in [2] 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 alpha-numerically 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.

The speaker models were generated from clean speech in the train part, while the validation part was used for obtaining thresholds, weights and example opinion vectors for impostors and true claimants. Thresholds were found for Equal Error Rate (EER) performance on clean speech.

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 observe the effects of operating point shift it would be ideal to report the performance in terms of both FA and FR. However, due to space limitations we have quantified the performance into a single figure using TE = FA + FR, where TE stands

for Total Error. It is found that when the difference between FA and FR is small, TE/2 is a good approximation of EER.

The following experiments were performed:

1. In this experiment we found the individual performance of each feature. The results are presented in Fig. 2.
2. Here we have found the performance for concatenation fusion in four configurations: MFCC+MACV, MFCC+CMS, CMS+MACV and MFCC+MACV+CMS. The results are presented in Fig. 3.
3. In this experiment we have investigated the differences between linear fusion and concatenation fusion. Here the weights were equivalent to the proportion of each feature in concatenation fusion, eg.: the dimensionality of MFCC and MACV features is 16 and 8 respectively, hence the contribution of MACV in the MFCC+MACV feature is approx. 33.3%. The results are presented in Fig. 4.
4. Here the weights for linear fusion were found by optimizing performance (lowest TE) on clean speech in the validation part. The results are presented in Fig. 5.
5. In this experiment we have adapted the weight for the MFCC modality using Eqn. (5). Parameters a and b were set to 1.1 and 1.0 respectively by observing how q in Eqn. (4) decreases according to the decreasing SNR of speech in the validation set.

   For each modality combination, w_{max} was set to the weight of the MFCC modality in Experiment 4, while w_{min} was set to 0. For the 3 modality combination, weights of the MACV and CMS modalities were adjusted as follows:

   i. let $x = 1 - w_{1,adj}$
   ii. let $r = w_{2,orig}/(w_{2,orig} + w_{3,orig})$
   iii. $w_{2,adj} = rx$
   iv. $w_{3,adj} = x - w_{2,adj}$

   where w_1, w_2 and w_3 refer to the weights for MFCC, MACV and CMS modalities respectively, the adj subscript refers to the adjusted weight while the $orig$ subscript refers to original weight, as found in Experiment 4. The results are presented in Fig. 6.

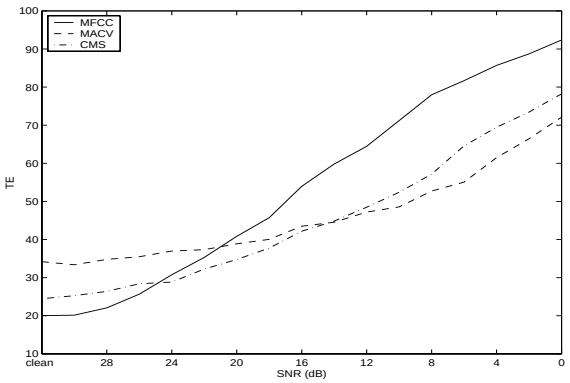

Figure 2: *Performance of each feature using a priori threshold*

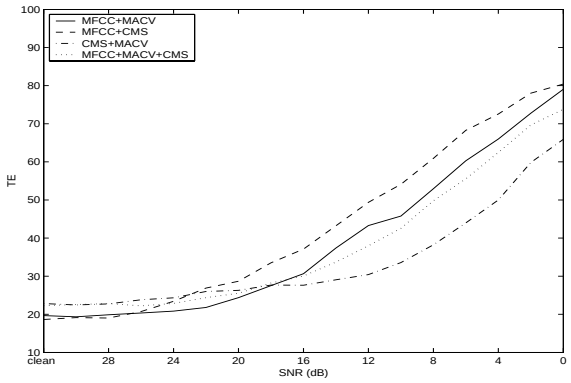

Figure 3: *Concatenation fusion*

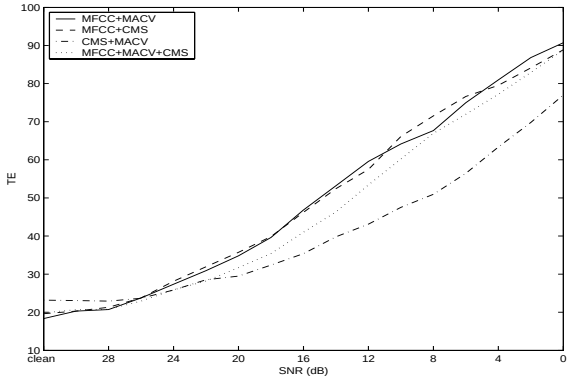

Figure 4: *Linear fusion, weights equivalent to proportion of each feature in concatenation fusion*

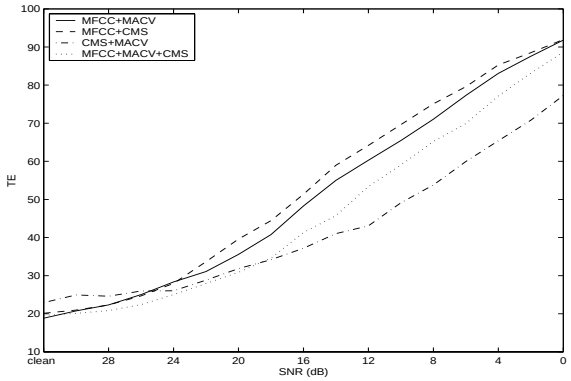

Figure 5: *Linear fusion, best weights for clean data*

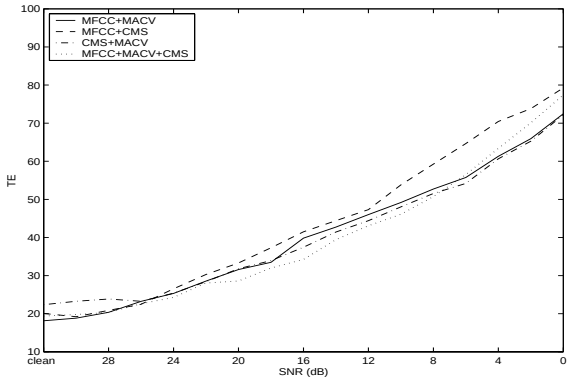

Figure 6: *Linear fusion, adaptive weights*

5. Discussion

It was observed that for all the experiments, the FR increased as noise increased, while FA decreased slightly. Hence for highly noisy cases the TE is dominated by FR.

From the results shown in Fig. 2 we can see that the MFCC feature obtains the best performance for clean data. However the performance rapidly degrades as the SNR is lowered.

The CMS feature is significantly more robust with respect to noise as compared with the MFCC feature. This robustness comes with a performance sacrifice for clean speech. The most robust feature is MACV which is found yielding the worst performance on clean speech.

In concatenation fusion (Fig. 3) the combination of MFCC and CMS features produced the best performance on clean speech, and not surprisingly, the worst in noisy speech. Compared to MFCC alone, the performance on clean speech is slightly better while in noisy speech the degradation is significantly smaller. The MFCC+MACV combination produced second best results on clean data, with performance in noisy data better than MFCC+CMS. The CMS+MACV combination is the most noise immune, with performance on clean data worse than MFCC+CMS, but better than just using the CMS feature alone. Interestingly, the performance in noisy data is significantly better than either CMS or MACV alone.

Combination of all three features, when compared to MFCC alone, produced significantly better performance in noisy speech, but surprisingly worse performance in clean speech. When compared to CMS+MACV, the performance is comparable on clean data, but significantly worse in noisy data. Hence the addition of the MFCC feature to the CMS+MACV feature vector did not result in better performance on clean speech and instead caused a degradation on noisy speech.

Using linear fusion, with the weights setup to mimic the contribution of each feature in concatenation fusion, the performance (Fig. 4) is similar on clean data, but degrades much quicker as the SNR is lowered. Compared to MFCC, the use of all three features resulted in slightly better performance on both clean and noisy data. Again, CMS+MACV is the most noise immune at the expense of worst performance on clean speech.

Modifying the weights to obtain the best performance on clean speech (Fig. 5) produced little difference to the previous setup. From these results we draw the conclusion that for non-adaptive fusion, feature vector concatenation is the preferred approach.

Use of adaptive weights (Fig. 6) significantly reduced the errors in noisy speech for all combinations, with MFCC+MACV arguably having the best overall performance. Adaptive linear fusion, in most cases, has better results in highly noisy cases (SNR $<$ 10 dB) when compared to the concatenation approach. The main exception is the CMS+MACV combination where the concatenation approach is visibly superior.

6. Hybrid Fusion

In Fig. 3 we can see that for MFCC+MACV and CMS+MACV combinations the use of MACV always makes the system more robust against noise as well as slightly reduce the error on clean data. MFCC+MACV provides the best overall performance in low noise conditions (SNR $\geq$ 18 dB) while CMS+MACV provides the best performance in moderate to high noise conditions (SNR $<$ 18 dB). Hence we propose a *hybrid fusion* approach where both concatenation and adaptive decision fusion is used. Here the MFCC+MACV and CMS+MACV concatenated features are processed by their own experts. Decision fusion is similar to Experiment 5, with the weight for the MFCC+MACV modality being adaptive. A block diagram of the proposed hybrid fusion system is shown in Fig. 7, and its performance is shown in Fig. 8. As expected, the performance is the best of both MFCC+MACV and CMS+MACV features. Compared to using just the MFCC feature, performance is slightly better in clean speech and significantly better in noisy speech.

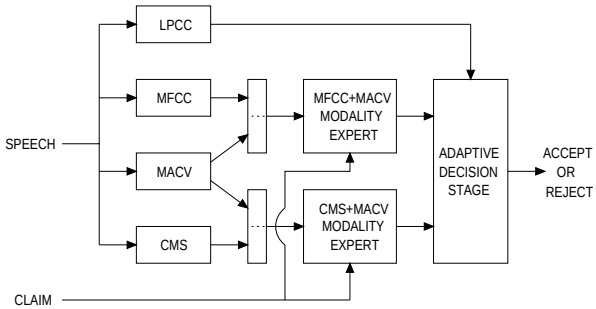

Figure 7: *Block diagram of the proposed hybrid fusion system*

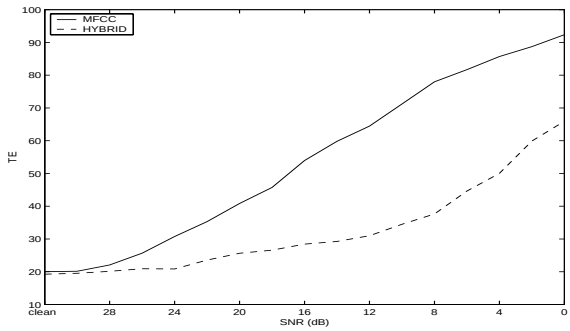

Figure 8: *Hybrid fusion with comparison to MFCC feature*

7. References

[1] H. Gish, M. Schmidt, "Text-independent speaker identification", *IEEE Signal Processing Magazine*, Oct. 1994.

[2] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication 17*, 1995.

[3] H. Altinçay, M. Demirekler, "On the use of Supra Model Information from Multiple Classifiers for Robust Speaker Identification", *Proc. EUROSPEECH'99*, Budapest, 1999.

[4] 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.

[5] G. R. Doddington et al, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective", *Speech Communication 31*, 2000.

[6] C. Sanderson, K. K. Paliwal, "Noise Compensation in a Multi-Modal Verification System", *Proc. ICASSP 2001*, Salt Lake City, 2001.

[7] V. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[8] C. Sanderson, K.K. Paliwal, "Adaptive Multi-Modal Person Verification System", First IEEE Pacific-Rim Conf. on Multimedia (IEEE-PCM2000), Sydney, 2000.

[9] P. Jourlin et al, "Acoustic-labial speaker verification", *Pattern Recognition Letters 18*, 1997.

[10] C. Sanderson, K. K. Paliwal, "Multi-Modal Person Verification System Based on Face Profiles and Speech", *Proc. Fifth Intern. Symposium on Signal Proc. Applications*, Brisbane, 1999.

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

[12] C. Jankowski et al, "NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database", *Proc. ICASSP 90*, Albuquerque, 1990.