2010 Asia Pacific Conference on Circuits and Systems (APCCAS 2010)
6–9 December 2010, Kuala Lumpur, Malaysia

Fusion Methods for Boosting Performance of Speaker Identification Systems

Gregory Ditzler, James Ethridge, Ravi P. Ramachandran and Robi Polikar
Department of Electrical & Computer Engineering
Rowan University
Glassboro, NJ 08028
{ditzle53,ethrid60}@students.rowan.edu, {ravi,polikar}@rowan.edu

***Abstract*—Two important components of a speaker identification system are the feature extraction and the classification tasks. First, features must be robust to noise and they must also be able to provide discriminating information that the classifier can use to determine the speaker's identity. Second, the classifier must take the features that have been extracted from a sentence and label them as corresponding to one of the enrolled speakers. However, sets of features may be even more beneficial than any single feature by itself. There may be information present in one feature that other features do not have. Therefore, we present analysis of features and fusion by employing probabilistic averaging and weighted majority voting. Weighted voting will require that the weights are determined in a non-heuristic methodology and are robust to data with a large amount of channel distortion. Results using the King database show that both fusion methods lead to enhanced performance.**

I. Introduction

Speaker recognition is the concept of using a machine that is capable of identifying an individual by the spectral properties of their voice. Speaker recognition systems typically operate in two types of modes: verification and identification [1]. Verification will validate a person's identity by comparing the captured speech to its own biometric templates that have been saved in the database, whereas the identification mode will search templates of all the users in the database for a match. This paper concentrates on a speaker identification system. In order for a speaker identification system to be effective, it must be robust to changes in channel distortion and the distortion may change with time. Therefore, robust features are sought for speaker identification that are not easily affected by changes in the channel.

Changes in the channel distortion will have a negative effect on the classification accuracy of the speaker identification system. The goal of this paper is to analyze several different techniques and features to boost the classification performance of a speaker identification system experiencing changes in channel distortion. We investigate several different robust features as well as two methods of fusion to work with the change in channel distortion. The change in channel distortion is the primary reason for using the King database as a benchmark test. This database experiences a change in channel distortion after session 6 which will allow for us to demonstrate the advantages of a particular feature(s) and a method of fusion.

The primary contributions of this work are the analysis of four different robust features and two methods for combining features to boost the predictive accuracy for a vector quantizer classifier. The rest of the paper is organized as follows: Section II will provide theory into the features used followed by Section III which will present the fusion methods. Section IV will cover the methodology of the experiments and Section V will present the results. Section VI will uncover the conclusions of this work.

II. Feature Extraction

A. Frame Selection

A pre-emphasis filter was applied to the speech signal with a transfer function $H(z) = 1 - 0.95z^{-1}$. The speech is sampled at 8 *k*Hz and the features are processed in frame sizes of 240 samples with 160 samples overlap between frames. Each frame has a Hamming window applied. Energy thresholding is performed over all frames of a sentence to determine the relatively high energy speech frames. These high energy frames are further reduced to meet the criteria of having all the roots of the linear prediction polynomial (of order 12) between 300 Hz and 3700 Hz and six of the roots having a magnitude greater than or equal to 0.88. The high energy frames with well defined formants are used to train and test with the vector quantizer (VQ).

B. Mean Removed Mel Frequency Cepstrum (MRMFCC)

The Mel cepstrum exploits auditory as well a decorrelation property in the cepstrum [2]. The magnitude of the short time Fourier transform (STFT) of the speech is logarithmically smoothed using a Mel spaced filter bank. The DCT of the output of the mel filter bank is referred to as the Mel cepstrum. The mean of all the frames, not just the high energy frames, is computed and removed from the selected high energy frames before being classified using the Vector Quantizer. This feature is referred to as the mean removed Mel frequency cepstrum *(MRMFCC)*.

C. Pole Filtered Mean Removed Cepstrum (PFMRCEP)

The PFMRCEP feature is used to remove any channel distortion that may be present in the speech signal which will corrupt the features and be detrimental to the classification performance of the identification system [3]. Before the PFMRCEP can be computed, we must compute the linear prediction cepstrum $c_{lp}(n)$. The cepstrum can be computed

using Eq. 2 where $a(k)$ are the coefficients of the linear prediction polynomial, $A(z)$ which is given by

$$A(z) = 1 - \sum_{k=1}^{p} a(k) z^{-k} \tag{1}$$

where p is the order of the prediction. The polynomial $A(z)$ is computed using the autocorrelation method [4], [5].

$$c_{lp}(n) = a(n) + \sum_{i=1}^{n-1} c_{lp}(i) a(n-i) \tag{2}$$

The transfer function for pole filtering is shown in Eq. 4 where $0 < \gamma < 1$ is a constant. If we further investigate the effect of γ, we will see that the poles of the linear prediction polynomial are being moved inward since q_k are the roots of $A(z)$.

$$C_{flp}(z) = \log\left(\frac{1}{A(z/\gamma)}\right) \tag{3}$$

$$A(z/\gamma) = \prod_{k=1}^{p} \left(1 - \gamma q_k z^{-1}\right) \tag{4}$$

Since γ is simply being multiplied by the roots, q_k, then the result of this in the cepstrum will be Eq. 5 where $n = 1, 2, \ldots, p$.

$$c_{flp}(n) = \frac{1}{n} \sum_{k=1}^{p} (\gamma q_k)^n = \gamma^n c_{lp}(n) \tag{5}$$

The PFMRCEP feature can be computed by $c_{pfmrcep}(n) = c_{lp}(n) - \mathcal{E}\left\{\gamma^n c_{lp}(n)\right\}$.

D. Pole Filtered Mean Removed Adaptive Component Weighted Cepstrum (PFMRACW)

The PFMRACW is a feature which combines Adaptive Component Weighted (ACW) cepstrum and pole filtering [6], [7] thereby leading to enhanced robustness to channel effects. The pole filtered ACW transfer function can be written as the ratio of two transfer functions $M(z)$ and $A(z/\gamma)$ as shown in Eq. 6 where $0 < \gamma < 1$. By writing the partial fraction expansion of $H_{PFACW}(z)$ as in Eq. 7 (recall that q_k are the roots of $A(z)$), the final form is given in Eq. 8. It is shown that $M(z)$ is the derivative of $A(z/\gamma)$ [7].

$$H_{PFACW}(z) = \frac{M(z)}{A(z/\gamma)} \tag{6}$$

$$= \sum_{k=1}^{p} \frac{1}{1 - \gamma q_k z^{-1}} \tag{7}$$

$$= p \frac{1 - \sum_{k=1}^{p-1} m_k z^{-k}}{1 - \sum_{k=1}^{p} \gamma^k a_k z^{-k}} \tag{8}$$

Thus, by applying the recursion equation in Eq. 2 we end up with the PFMRACW computed as $c_{pfmracw}(n) = c_{acw}(n) - \mathcal{E}\{c_{pfacw}(n)\}$ where $c_acw(n)$ is the ACW cepstrum and $c_{pfacw(n)}$ is the cepstrum corresponding to $H_{PFACW}(z)$. We compared PFMRACW with the mean removed ACW cepstrum (MRACW) which can be written as $c_{mracw}(n) = c_{acw}(n) - \mathcal{E}\{c_{acw}(n)\}$

III. Fusion Methods

We present the use of ensemble based systems for data fusion [8] to augment the performance of the PFMRCEP, PFMRACW, and MRMFCC features. The MRACW is not used for fusion as it does not perform as well as the other three features. We present two different methods of fusion: probabilistic averaging and Soong-Rosenberg fusion.

A. Probabilistic Averaging

The summed distances computed by the VQ can be converted to posterior estimates using Eq. 10 where s is a normalization factor (Eq. 9), d_i is the accumulated distance from a particular speaker's codebook and Q are the number of speakers. The posterior estimate $P(\omega_j|X)$ is the probability of class ω_j (speaker j) given a set of feature vectors X.

$$s = \sum_{i=1}^{Q} d_i \tag{9}$$

$$P\left(\omega_j|X\right) = \frac{s - d_j}{(Q-1)s} \tag{10}$$

The posterior estimates for each one of the features are averaged as a means to combine the decisions from the high energy speaker frames.

B. Weighted Voting

The weighted voting scheme used in this work was originally presented by Soong and Rosenberg [9]. However, none of the features discussed in this paper have been used in the original work. The general combination for Soong-Rosenberg fusion (SRF) is given by Eq. 11 where q is the index of the speaker ($q = 1, 2, \ldots, Q$), $d_{x,q}$ is the accumulated squared Euclidean distance for the xth feature ($x = \{f, g, h\}$) of speaker q's codebook and w_x is the weight for the xth feature.

$$\hat{d}_q = \frac{1}{N_F}\left(w_f d_{f,q} + w_{g,q} d_{g,q} + w_h d_{h,q}\right) \tag{11}$$

The weights are determined by first computing all the codebooks for each speaker and all the features. Then each speaker's utterance is tested with their own codebook and none of the others. This will results is a summed squared Euclidean distance for a speaker on his codebook which is then scaled by the number of high energy features in the speaker's sentence. The scaled distances are denoted by d_q in Fig. 1. Therefore, for each feature there will be 26 different scaled distances ($\hat{d}_q$).

The mean of the summed distances divided by the number of training frames is computed using Eq. 12. Once this has been computed the final weight for a feature is simply the inverse of D_x (see Eq. 13).

$$D_x = \frac{1}{Q} \sum_{q=1}^{Q} \frac{d_{x,q}}{N_q} \tag{12}$$

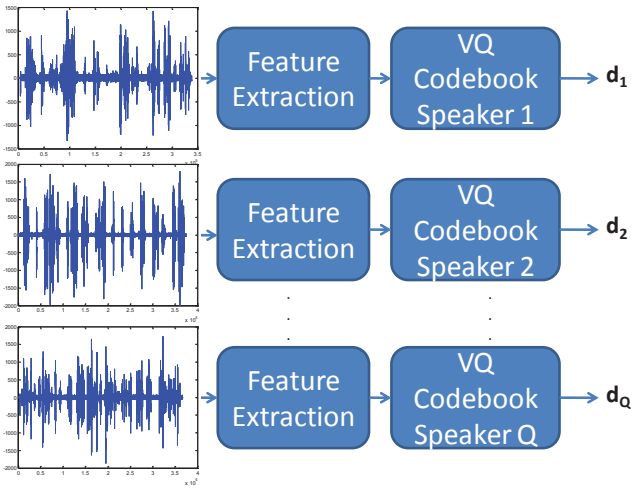

Fig. 1. Weights for Soong-Rosenberg fusion

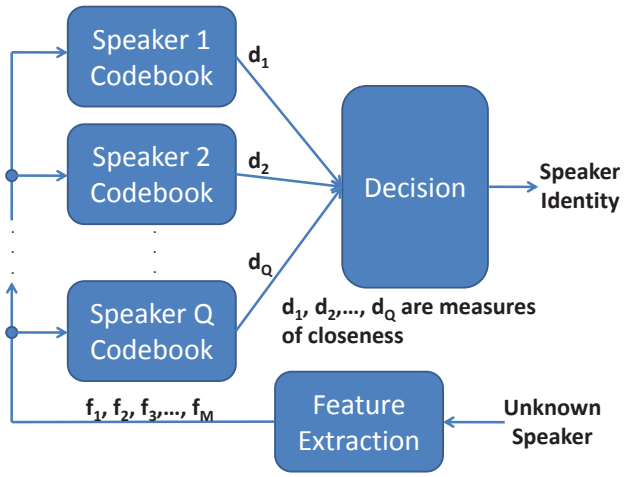

Fig. 2. Function diagram of a speaker identification system using a vector quantizer

$$w_x = \frac{1}{D_x} \tag{13}$$

N_q is the number of high energy frames for the qth speaker in the training set, $d_{x,q}$ is the summed distance from the VQ of the xth feature, and Q is the number of speakers. A weight is computed for each one of the features discussed in the previous section. This process is repeated for all features used in fusion. Note that the weights are determined from the training data only.

IV. Experiment Design

A. King Database

The King corpus was created for research in the area of speaker identification and was collected in New Jersey and San Diego [10]. There are twenty-six San Diego speakers which were used in this work. All speakers in the database are male. There are ten sessions for each speaker. The data is divided such that there is a big mismatch in the conditions between sessions 1 to 5 and sessions 6 to 10. This mismatch is due to a change in the recording equipment, which translates to a significantly changed environment. Sessions were recorded a week to a month apart. Each speaker was recorded for approximately 30 seconds speech that has been introduced to channel noise. The King database speakers are not given a controlled experiment rather they speak into the telephone when a session is being recorded and talk about different topics.

B. Procedure

The data was pre-processed and the high energy frames were extracted using the technique discussed in Section II. The order of the linear prediction polynomial was set to 12 for all trials. A VQ codebook of size 64 [11] was designed for each speaker in the database using the LBG algorithm [12]. The distance measure used is the squared Euclidean. Session 1 of the King database is used for training and sessions 2 to 9 are used for testing.

The PFMRACW and PFMRCEP use $\gamma = 0.9$ for all trials. The summed distances from the VQ are converted to probabilities for each set of feature vectors in a sentence. The probabilities are then averaged for each of the features and the result is denoted by *PA* in Section V. The summed distances from each of the features in a sentence are combined using Soong and Rosenberg fusion discussed in the previous section. These results are denoted as *SRF* in Section V. The PFMRACW, PFMRCEP and MRMFCC are used for the SRF and PA fusion.

V. Results

Table I contains the identification success rate (ISR) and is defined as the number of speech utterances for which the speaker is identified correctly divided by the total number of utterances tested (expressed as a percent). The ISR can be looked at more generally by averaging sessions 2-5, 6-10 and 2-10 (see Table II). It is useful to observe the average ISR across the different sessions because of the properties of the King database (i.e. the introduction of spectral distortion in session 6-10). The result of this spectral distortion in sessions 6-10 can clearly be observed in Table I. Note that the performance of SRF and PA are extremely similar when testing within the great divide. The only difference is on session 5 where the SRF has a boost in ISR over PA. However, the test results on the sessions that are across the great divide show that the SRF provides a better performance on average than PA. On all the sessions, the fusion methods are always performing equal to or better than the best feature tested.

Table I also shows the robustness of each of the individual features particulary with the PFMRACW and PFMRCEP. The PFMRCEP feature generally provides a large ISR when observed on sessions 2-5. However, the ISR on session 6-10 show that the PFMRACW can be used to provide a higher

TABLE I
PERFORMANCE ON INDIVIDUAL SESSIONS OF THE KING DATABASE

Identification Success Rates (%)									
Features	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8	Session 9	Session 10
SRF	84.6	73.1	80.8	84.6	57.7	46.2	53.9	53.9	50.0
PA	84.6	73.1	80.8	80.8	53.9	42.3	42.3	46.2	53.9
MRMFCC	73.1	69.2	76.9	73.1	42.3	26.9	50.0	57.7	42.3
PFMRACW	76.9	57.6	73.1	73.1	42.3	34.6	46.2	30.8	46.2
PFMRCEP	76.9	69.2	76.9	80.8	38.5	23.1	11.5	26.9	30.8
MRACW	69.2	38.4	46.2	50.0	23.1	11.5	19.2	23.1	23.1

TABLE II
AVERAGE PERFORMANCE ACROSS SESSIONS OF THE KING DATABASE

Identification Success Rates (%)			
Features	Sessions 2-10	Session 2-5	Session 6-10
SRF	65.0	80.8	52.3
PA	62.0	79.8	47.7
MRMFCC	56.8	73.1	43.8
PFMRACW	53.4	70.2	40.0
PFMRCEP	48.3	76.0	26.2
MRACW	33.8	51.0	20.0

ISR than the PFMRCEP. This indicates that the PFMRACW is more robust to changes in the channel than the PFMRCEP.

The overall ISRs are shown in Table II. The best performing individual feature is the MRMFCC followed by the PFMRACW and PFMRCEP, respectively. The SRF fusion method provides the best method of classification, although the simple probabilistic averaging works very well. The pole filtered mean removed ACW also appears to provide a more robust feature than the mean removed ACW. Its interesting to observe the high ISR of the PFMRCEP from within the great divide and how it has the largest drop in performance when testing across the great divide.

The ISRs on sessions 2-5 are quite a bit higher than when averaged over all the sessions. However, the results are similar the previous set of performances presented. The weighted vote and probabilistic average are the best performing methods of fusion. Both methods of fusion outperform all of the single features. This trend is also observed with the testing on sessions 6-10.

VI. CONCLUSION

The PFMRACW is a more robust feature than the MRACW as observed over all sessions of the the King database. This suggests that the pole filtering method gets a better channel estimation than the mean of all the feature vectors. The PFMRACW was also a better feature than PFMRCEP when tested on sessions 6-10 of the King database indicating that PFMRACW is more robust to severe changes in the channel distortion than PFMRCEP. The weighted fusion originally presented by Soong and Rosenberg [9] was compared to a simple probabilistic averaging of the PFMRACW, PFMRCEP and MRMFCC. All methods of fusion have been shown to be more robust to channel distortion than any single one of the features on average.

REFERENCES

[1] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, 2004.

[2] S. B. Davies and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] R. Ramachandran and K. Farrell, "Fast pole filtering for speaker recognition," in *IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland, 2000, pp. V49–V52.

[4] T. Quatieri, "Discrete-time speech signal processing: Principles and practice," *Prentice Hall*, 2001.

[5] P. P. Vaidyanathan, "The theory of linear prediction," *Morgan and Claypool Publishers*, 2008.

[6] A. L. Swanson, R. P. Ramachandran and S. H. Chin, "Fast adaptive component weighted cepstrum pole filtering for speaker identification," *IEEE Int. Symp. on Circuits and Systems*, pp. V–612 – V–615, 2004.

[7] M. Zilovic, R. Ramachandran, and R. Mammone, "A fast algorithm for finding the adaptive component weighted cepstrum for speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 376–382, 1997.

[8] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[9] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[10] "Brief description of the king speech database," *http://www.ldc.upenn.edu/Catalog/docs/LDC95S22/king.readme.html*.

[11] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantizer design," *Comput. Speech Lang.*, vol. 22, pp. 143–157, 1987.

[12] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization design," *IEEE Transactions on Communications*, vol. 28, pp. 84–95, 1980.