TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING GMM-UBM AND FRAME LEVEL LIKELIHOOD NORMALIZATION

Rong Zheng, Shuwu Zhang, Bo Xu

High Technology and Innovation Center, National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing
{rzheng, swzhang, xubo}@hitic.ia.ac.cn

ABSTRACT

In this paper, we describe a Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker identification system. In this GMM-UBM system, we derive the hypothesized speaker model by adapting the parameters of UBM using the speaker's training speech and a form of Bayesian adaptation. The UBM technique is incorporated into the GMM speaker identification system to reduce the time requirement for recognition significantly. The paper also presents a new frame level likelihood score normalization for adjusting different scores of speaker models to get more robust scores in final decision. Experiments on the 2000 NIST Speaker Recognition Evaluation corpus show that GMM-UBM and frame level likelihood score normalization yield better performance. Compared to the baseline system, around 31.2% relative error reduction is obtained from the combination of both techniques.

1. INTRODUCTION

Over the past ten years, Gaussian mixture models (GMM) for the modeling of speaker spectral characteristics has become the dominant approach for speaker identification systems which use untranscribed training data [1].

Reynolds et al. [2] presented GMM-based speaker verification system which uses a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from UBM.

In this paper, a speaker model based on Gaussian Mixture Model-Universal Background Model (GMM-UBM) is introduced into text-independent speaker identification. Our work focuses on applications which require high identification rates using short utterance from unconstrained (text-independent) conversational speech and robustness to degradations produced by transmission over a telephone channel.

For speaker recognition over the telephone, one of the largest challenges is dealing with channel variability. There are different acoustic environment and transmission channel. In addition, different training and testing conditions result in a low match score. Typically a speaker enrolls his/her voice using one microphone (or handset) and then to be verified using a different microphone. Different microphones impose different characteristics on the acoustic signal [3]. Current research has gone some way towards reducing channel effects.

Compensation techniques for channel effects can be classified into three broad categories: feature-based compensation, model-based compensation and score-based compensation. In feature domain, methods for feature compensation use some form of linear or nonlinear channel compensation at speech analysis and feature extraction stages, such as cepstral mean subtraction (CMS) [1], RASTA filtering [4] or artificial neural networks [5]. Model-based techniques attempt to reduce the effect of channel variations by learning channel characteristics or enhancing the speaker probability distribution models. Methods for model compensation contain Speaker Model Synthesis [6], Synthetic Variance Distribution [7] etc. There are also some robust scoring methods and normalization compensation techniques that applied in match score domains [8][9]. Score-based normalization techniques apply some form of compensation transformation to the final likelihood scores.

In this paper, we present a new frame level likelihood normalization in match score domain for robust speaker identification using GMM speaker models, which can be viewed as a special kind of score compensation.

The paper is organized as follows. In next part, an overview of the GMM-UBM speaker identification system is described, which includes the UBM technique and Bayesian adaptation of speaker models. In section 3 we describe the new frame level likelihood score normalization. The experimental results are discussed in Section 4. Finally, we will draw a conclusion.

2. GMM-UBM SPEAKER IDENTIFICATION SYSTEM

0-7803-8678-7/04/$20.00 ©2004 IEEE

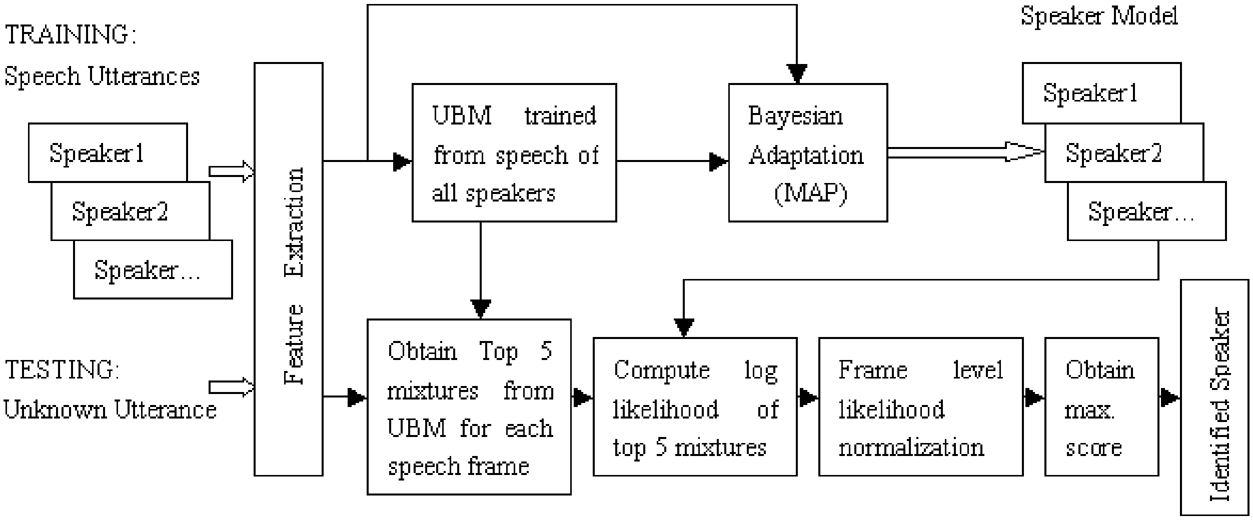
TRAINING:
Speech Utterances
Speaker1
Speaker2
Speaker...
TESTING:
Unknown Utterance
Feature Extraction
UBM trained from speech of all speakers
Bayesian Adaptation (MAP)
Speaker Model
Speaker1
Speaker2
Speaker...
Obtain Top 5 mixtures from UBM for each speech frame
Compute log likelihood of top 5 mixtures
Frame level likelihood normalization
Obtain max. score
Identified Speaker

Figure 1: Block Diagram of the GMM-UBM SID System

Fig. 1 shows a block diagram of the GMM-UBM Speaker Identification (SID) system. It can be mainly divided into three parts: UBM training, Bayesian adaptation of speaker models and speaker identification.

2.1. The Universal Background Model Technique

A UBM in the SID system is a GMM representing the characteristic of all different speakers. Instead of employing the Maximum-Likelihood training, each speaker model can be created by performing Bayesian adaptation from the UBM using speaker-specific training speech. The training operation is illustrated at the upper part of Fig. 1.

From previous experiments conducted for speaker recognition, Reynolds et al. [2] have found that only a few of the mixtures of a GMM contributes significantly to the likelihood value for a speech feature vector. In addition, the mixture components of each adapted speaker model retain a certain correspondence with the UBM, therefore log-likelihood score of the speaker model can be computed by scoring only the more significant mixtures. In our SID system, the top 5 mixtures are used. Because of the correspondence of mixtures between the UBM and the speaker models, these significant mixtures can be obtained by calculating the mixtures from the UBM that have the highest score. The computation requirement for recognition is reduced significantly by employing this mixture scoring strategy. The procedure is shown at the lower part of Fig. 1.

2.2. Bayesian Adaptation of Speaker Model

For each hypothesized speaker in the GMM-UBM system, we derive the hypothesized speaker model by adapting the parameters of the UBM using the speaker's training speech and maximum a posteriori (MAP) adaptation. A speaker-specific GMM with diagonal covariance matrices is trained. Unlike the standard approach of Maximum-Likelihood training of a speaker model independently of the UBM, the basic idea of adaptation approach is to derive the speaker model by updating the well-trained parameters in the UBM. This provides a better coupling between the speaker's model and UBM, which not only produces better performance than decoupled models, but also allows for a fast-scoring technique.

The specifics of the adaptation are as follows [2]. Given T feature vectors $X = \{x_1, ..., x_T\}$, the mixture weights satisfy the constraint $\sum_{i=1}^{M} \omega_i = 1$. Collectively, the parameters of the density model are denoted as, $\lambda = \{\omega_i, \mu_i, \Sigma_i\}$, where $i = 1, ..., M$:

$$\Pr(i \mid x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)} \tag{1}$$

$$n_i = \sum_{t=1}^{T} \Pr(i \mid x_t) \tag{2}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t) x_t \tag{3}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t) x_t^2 \tag{4}$$

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i / T + (1 - \alpha_i^{\omega}) \omega_i] \gamma \tag{5}$$

$$\hat{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m) \mu_i \tag{6}$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{7}$$

In this GMM-UBM system, we use a single adaptation coefficient for all parameters ($\alpha_i^{\omega} = \alpha_i^{m} = \alpha_i^{v} = n_i / (n_i + r)$) with a relevance factor of r = 50.

3. FRAME LEVEL LIKELIHOOD NONLINEAR SCORE NORMALIZATION

In this paper, we present a new scoring approach, which is a nonlinear score compensation transformation based on GMM and can be viewed as a special kind of frame level likelihood normalization [9]. Test speech utterance is processed by all hypothesized speaker models in parallel in frame by frame manner. In order to enlarge the score difference among different speakers at the same frame and reduce the score difference for the same speaker at different frames, the likelihoods are processed using the following nonlinear normalization to transform them into new scores, that is, we employ a confidence-based weighting scheme that updates the scores of speaker models at frame level. Then transformed likelihoods are accumulated over all test frames to form a final score for each speaker model. The unknown speaker is identified as the speaker, whose model gives the best score.

A gaussian mixture density is a weighted sum of M component densities and is given by the form:

$$p(x \mid \lambda) = \sum_{i=1}^{M} \omega_i \, p_i(x) \qquad (8)$$

where, x is a D-dimensional feature vector, $p_i(x), i = 1, ...M$, is the component density and $\omega_{i,} i = 1, ..., M$, is the mixture weight. Each component density is a D-variate gaussian funtion of the form:

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T (\Sigma_i)^{-1} (x-\mu_i)\} \quad (9)$$

with mean vector μ_i and covariance matrix Σ_i . The mixture weights satisfy the constraint that:

$$\sum\nolimits_{i=1}^{M} \omega_i = 1 \qquad (10)$$

The parameters of density model are denoted as

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\}, \; i = 1, ..., M \qquad (11)$$

Given T feature vectors $X = \{x_1, ..., x_T\}$, where $t = 1, ..., T$, and N speaker models $\lambda = \{\lambda_1, ..., \lambda_N\}$, where $n = 1, ..., N$. The scoring procedure is as follows.

For each test vector x_t :

Step1. Calculate the log-likelihood $\log p(x_t \mid \lambda_n)$ for each speaker model λ_n . We detect the maximal and minimal log-likelihood value as,

$$\max\log = \arg\max_n \log p(x_t \mid \lambda_n), \;\; \min\log = \arg\min_n \log p(x_t \mid \lambda_n) \quad (12)$$

Step2. Normalize $\log p(x_t \mid \lambda_n)$ between 0 and 1,

$$\bar{p}(x_t \mid \lambda_n) = \frac{\log p(x_t \mid \lambda_n) - \min\log}{\max\log - \min\log} \qquad (13)$$

Step3. Adjust the score of test vector x_t with integer q and A which is close to 1:

$$S(x_t, \lambda_n) = \frac{\{\bar{p}(x_t \mid \lambda_n)\}^q}{\{\bar{p}(x_t \mid \lambda_n)\}^q + A} \qquad (14)$$

Step4. Compute the score at time t over previous K frames. K is trade-off between robustness and computational load.

$$S'(x_t, \lambda_n) = \frac{1}{K} \sum_{k=1}^{K} S(x_{t-k}, \lambda_n) \qquad (15)$$

Then we get the transformed score of test vector x_t is,

$$\bar{S}(x_t, \lambda_n) = S(x_t, \lambda_n) + \alpha * sign[S(x_t, \lambda_n) - S'(x_t, \lambda_n)] \qquad (16)$$

Where, $sign[x-y] = \begin{cases} 1 \;\; if\,(x-y) < -\alpha/2 \\ -1 \;\; if\,(x-y) > \alpha/2 \\ 0 \;\; otherwise \end{cases}$

For each speaker model, the final score is:

$$S(X, \lambda_n) = \sum_{t=1}^{T} \bar{S}(x_t, \lambda_n) \qquad (17)$$

In our experiments, q and α are speaker-independent parameters and optimized on a registration population. How to tune the parameter α for maximal generalization is non-trivial issue. The search is done by stepping through α from 0.0 to 1.0 with each step incremented by 0.01. In our simulations, we use α =0.05.

4. EXPERIMENTS

The experiments described in this section are conducted as part of the participation in the 2000 NIST speaker recognition evaluation. The data includes conversational telephone-quality speech taken from the Switchboard 2 corpus. All of the test segments are recorded from calls made from a telephone number that is different from the one used to enroll. Therefore, all test utterances may be considered to be collected using a different handset than the one used for training the speaker model. Each speaker is trained using a single two-minute session, while test utterances range between few seconds and a minute (with a primary focus on utterances with 15s~45s). A detailed description of the evaluation corpus may be found in [10].

Feature extraction process is performed as follows:

Divided into 24ms frames, shifted by 12ms, high-emphasis filtering with filter $1/(1 - 0.97z^{-1})$, hamming windowing, ignoring low energy frames which do not contain much speaker information (about 10%~15%), 17 Mel-Frequency Cepstral Coefficients (MFCCs), including energy, and 17 delta coefficients are calculated from useful telephone bandwidth (approximately 300~3400Hz), CMS is applied to mitigate linear channel effects. We finally obtain 34-dimensional feature vectors.

For the UBM, training data are selected from the 2000 SRE to have two hours of male speech and two hours of female speech, both equally distributed over the Carbon-button and Electron handset types [3]. A single GMM with 512 mixtures is trained using the iterative expectation-maximization (EM) algorithm [11] by pooling all the training data.

We have used 50 speakers (25males and 25 females) from NIST2000 SRE, which are different from the UBM training data. Training segments are 2 minutes. Since test utterances mainly last between 15 and 45 seconds, performance is computed on test segments that have a duration of 12 seconds (1000 frames extracted from the beginning of the test utterances). All speaker models are 512 mixtures trained by Bayesian adaptation of the UBM. There are totally 315 test segments.

The baseline system is based on a GMM with 128 mixtures trained by performing Maximum-Likelihood training. Each GMM has a diagonal covariance matrix.

The results of the experiments on highly mismatched telephone speech are reported in Table 1:

System used	Test results (% error)
Baseline system	43.3
Baseline + Frame level likelihood normalization	35.8
GMM-UBM	32.5
GMM-UBM + Frame level likelihood normalization	29.8

Table 1: Speaker identification performance

The results show that frame level likelihood normalization, GMM-UBM, and GMM-UBM + frame level likelihood normalization achieve 35.8%, 32.5%, 29.8% of error rate respectively, which achieve a considerable improvement 17.3%, 24.9%, 31.2% over the baseline system. Moreover, by employing mixture scoring strategy in GMM-UBM, the computation requirement for recognition is reduced significantly.

5. CONCLUSIONS

This paper has introduced GMM-UBM into speaker identification system. In addition, we present a new frame level likelihood normalization in score domain to improve the recognition accuracy. It has been shown that a significant gain in performance can be obtained for highly mismatched telephone speech. Also, by employing mixture scoring strategy in GMM-UBM, the computation requirement for recognition is reduced significantly. Compared to the baseline system, Frame level likelihood normalization, GMM-UBM respectively bring 17.3% and 24.9% relative error reductions, whereas the combination of both techniques yields 31.2% relative reduction.

6. REFERENCES

[1] D.A.Reynolds, and R.C.Rose, “Robust Text-idependent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. SAP, pp.72-83, Jan.1995

[2] D.A.Reynolds, T.F.Quatieri, and R.B.Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, vol.10, pp.19-41, Jan.2000

[3] T.F.Quatieri, D.A.Reynolds, and G.C.O’Leary, “Estimation of Handset Nonlinearity With Application to Speaker Recognition”, IEEE Trans. SAP, pp.567-584, Sep.2000

[4] H.Hermansky, N.Morgan, A.Bayya, and P.Kohn, “Compensation for the Effects of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)”, EUROSPEECH’91, pp.1367-1370, 1991

[5] L.P.Heck, Y.Konig, M.K.Sonmez, and M.Weintraub, “Robustness to Telephone Handset Distortion in Speaker Recognition by Discriminative Feature Design”, Speech Communication, vol.31, pp.181-192, 2000

[6] R.Teunen, B.Shahshahani, and L.Heck, “A Model-based Transformational Approach to Robust Speaker Recognition”, ICSLP’00, pp.495-498, 2000

[7] H.A.Murthy, F.Beaufays, L.P.Heck, and M.Weintraub, “Robust Text-Independent Speaker Identification over Telephone Channels”, IEEE Trans. SAP, pp.554-568, Sep.1999

[8] H.Gish, and M.Schmidt, “Text-independent Speaker Identification”, IEEE Signal Processing Magazine, pp.18-32, Oct.1994

[9] K.Markov, and S.Nakagawa, “Frame Level Likelihood Normalization for Text-Independent Speaker Identification Using Gaussian Mixture Models”, ICSLP’96, pp.1764-1767, 1996

[10] http://www.nist.gov/speech/tests/spk/2000/index.htm

[11] A.P.Dempster, N.M.Laird, and D.B.Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society Series, Vol.39, pp.1-38, Nov.1977