Robust Local Scoring Function for Text-Independent Speaker Verification

Ming Liu, Thomas S. Huang
IFP, Beckman Institute
University of Illinois at Urbana-Champaign
[mingliu1,huang]@ifp.uiuc.edu

Zhengyou Zhang
Multimedia Collaboration Group
Microsoft Research
zhang@microsoft.com

Abstract

Traditionally, the Universal Background Model (UBM) is viewed as the background model of the entire acoustic feature space. We propose a novel interpretation of the UBM model, and consider it as a mapping function that transforms the variable length observations (speech utterances) into a fixed dimensional feature vector (sufficient statistics). After this mapping, a similarity measurement is computed on the fixed dimensional features. With this novel interpretation, we proposed a new similarity measurement which produces more than 10% *relative improvement over the conventional UBM-MAP framework in both equal error rate and detection cost function.*

1 Introduction

Speaker verification is the process of verifying the claimed identity of a speaker based on the speech signal from the speaker (voiceprint). After 9-11, speaker verification, as well as other biometric techniques, gains considerable attention. There are two types of speaker verification systems: Text-Independent Speaker Verification (TI-SV) and Text-Dependent Speaker Verification (TD-SV). Text-independent Speaker Verification is a process of verifying the identity without constraint on the speech content. Compared to TD-SV, it is more convenient because the user can speak freely to the system. However, it requires longer training and testing utterances to achieve good performance. This paper focus on Text-independent Speaker Verification.

The UBM-MAP framework [8] is a state-of-the-art method for Text-independent Speaker Verification. Traditionally, the Universal Background Model (UBM) is viewed as the background model of the entire acoustic feature space. However, in this paper, we propose a novel interpretation of the UBM model and consider it as a mapping function that transforms the variable length observations (speech utterances) into a fixed dimensional feature vector. After this mapping, a novel similarity measurement is proposed to compute the similarity between the training and testing utterances based on these fixed dimensional features. With this similarity measurement, the system outperforms the conventional UBM-MAP framework by more than 10% relatively in both Equal Error Rate (EER) and Detection Cost Function (DCF).

2 UBM-MAP Framework

For TI-SV, the UBM-MAP framework is the state-of-the-art in terms of performance and speed. The UBM is a Gaussian Mixture Model (GMM) which can be represented as follows:

$$P(x|\lambda) = \sum_{i=1}^{M} w_i P_i(x|\lambda) = \sum_{i=1}^{M} w_i N(x : m_i, \Sigma_i) \quad (1)$$

where x is the feature vector of each frame and λ is the parameter of Gaussian Mixture Model. Parameter λ includes the prior probability of each component w_i, the mean vector of each component m_i and the covariance matrix of each component Σ_i. $P_i(\cdot|\lambda)$ denotes the likelihood function of the ith component which is a multivariant Gaussian in our scenario. For simplicity, the covariance matrix is assumed to be diagonal for lower computation load. EM algorithm is used to obtain the maximum likelihood estimation of these parameters.

In the UBM-MAP framework, the target speaker model is generated by the Maximu A Posterior (MAP) adaptation [5, 8]. The mean-only MAP adaptation was the best method compared with other types of MAP adaption such as the fully MAP adaptation [8]. The procedure of the mean-only MAP is listed as follows.

$$\gamma(i|x_t) = \frac{w_i P_i(x_t|\lambda)}{\sum_{j=1}^{M} w_j P_j(x_t|\lambda)} \quad (2)$$

$$\gamma(i) = \sum_{t=1}^{T} \gamma(i|x_t) \quad (3)$$

$$\bar{m}_i = \frac{1}{\gamma(i)} \sum_{t=1}^{T} \gamma(i|x_t) x_t \quad (4)$$

$$\hat{m}_i = m_i + \frac{\gamma(i)}{\gamma(i)+\alpha}(\bar{m}_i - m_i) \quad (5)$$

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

where x_t is the feature vector of frame t, $\gamma(i|x_t)$ is the probability of x_t drawn from the ith component, $\gamma(i)$ is the soft count of frames belonging to the ith component, $\bar{m}_i$ is the sample mean of observations at the ith component, and α is a smoothing factor to incorporate the prior parameter m_i from the UBM model. Because of the smoothing factor α, when few frames are observed at component i, the prior m_i is more favorable, while when enough frames are observed the sample mean $\bar{m}_i$ is more favorable. Usually α is set empirically ($\alpha = 16$ in this paper).

After the target speaker model is generated, a log-likelihood ratio between the target speaker model and the UBM model is then used to evaluate testing utterances. The log-likelihood ratio is computed as follows

$$LLR(x_1^T) = \frac{1}{T}\sum_{t=1}^{T}\log\frac{P(x_t|\lambda_1)}{P(x_t|\lambda_0)}$$

where x_1^T is the feature vectors of the observed utterance, λ_0 is the parameter of UBM and λ_1 is the parameter of the target model. Essentially, the verification task is to construct a generalized likelihood ratio test between hypothesis H_1 (observation drawn from the target) and hypothesis H_0 (observation not drawn the target).

$$GLR(x_1^T) = \frac{P(x_1^T|H_1)}{P(x_1^T|H_0)}$$

The UBM model is usually considered as a background model which provides a description of acoustic feature space. Therefore, the likelihood of testing utterances on this UBM model $P(x_1^T|\lambda_0)$ can serve as an estimation of $P(x_1^T|H_0)$.

3 Sufficient Statistics of Speech Utterances

Besides treating UBM as a background model, the Equation (3) and Equation (4) already suggest another point of view. In MAP adaptation, the training utterances can be represented by $(\gamma(i), \bar{m}_i)_{i=1}^M$ without loosing any information which suggests that the UBM is a mapping function. This mapping function is to transfer the variable length speech utterances into a fixed dimensional super-vector $(\gamma(i), \bar{m}_i)_{i=1}^M$, where $\gamma(i)$ and $\bar{m}_i$ are the same as Equation (3) and Equation (4). And this super-vector is the sufficient statistics of the speech utterance. To simply the following derivation, we define $\delta_i = \bar{m}_i - m_i$ which is the adjustment between sample mean $\bar{m}_i$ and background mean m_i. Then we have following mapping:

$$\{x_1^T\} \quad \xrightarrow{\text{UBM}} \quad (\gamma(i), \delta_i)_{i=1}^M$$

Although the test utterance is not mapped into its sufficient statistic explicitly in UBM-MAP framework, the log-likelihood score of a test utterance can be bounded by a similarity measure between the sufficient statistics of the training and testing utterances. The derivation is as follows.

$$\begin{aligned}
\log\frac{P(x_t|\lambda_1)}{P(x_t|\lambda_0)} &= \log\frac{\sum_{i=1}^M w_i P_i(x_t|\lambda_1)}{\sum_{k=1}^M w_k P_k(x_t|\lambda_0)} \\
&= \log\sum_{i=1}^M \frac{w_i P_i(x_t|\lambda_1)}{\sum_{k=1}^M w_k P_k(x_t|\lambda_0)} \\
&= \log\sum_{i=1}^M \frac{w_i P_i(x_t|\lambda_1)}{\sum_{k=1}^M w_k P_k(x_t|\lambda_0)}\frac{w_i P_i(x_t|\lambda_0)}{w_i P_i(x_t|\lambda_0)} \\
&\geq \sum_{i=1}^M \gamma(i|x_t)\log\frac{P_i(x_t|\lambda_1)}{P_i(x_t|\lambda_0)} \\
&= \sum_{i=1}^M \gamma(i|x_t)(\hat{m}_i - m_i)^T\Sigma_i^{-1}(x_t - \frac{\hat{m}_i + m_i}{2})
\end{aligned}$$

where $\hat{m}_i$ is the mean of the target model and m_i is the mean of UBM. Let $b_i = \frac{\hat{m}_i + m_i}{2}$. Then the log-likelihood of whole utterances is bounded by a function depends only on two super-vectors.

$$\begin{aligned}
LLR(x_1^T) &= \frac{1}{T}\sum_{t=1}^T \log\frac{P(x_t|\lambda_1)}{P(x_t|\lambda_0)} \\
&\geq \frac{1}{T}\sum_{t=1}^T\sum_{i=1}^M \gamma(i|x_t)(\hat{m}_i - m_i)^T\Sigma_i^{-1}(x_t - b_i) \\
&= \frac{1}{T}\sum_{i=1}^M (\hat{m}_i - m_i)^T\Sigma_i^{-1}\sum_{t=1}^T \gamma(i|x_t)(x_t - b_i) \\
&= \frac{1}{T}\sum_{i=1}^M \gamma(i)(\hat{m}_i - m_i)^T\Sigma_i^{-1}(\bar{m}_i - b_i)
\end{aligned}$$

where $\gamma(i)$ is the soft-count for the ith component from the testing utterance, $\hat{m}_i$ is the adapted mean in Equation (5), m_i is the background mean, and $\bar{m}_i$ is the sample mean of the testing utterance. Recall that $\delta_i = \bar{m}_i - m_i$ and let $(\hat{\gamma}(i), \hat{\delta}_i)_{i=1}^M$ denote the sufficient statistics of the training utterance and $(\gamma(i), \delta_i)_{i=1}^M$ denote the sufficient statistics of the testing utterance. Then the final bound of the log-likelihood ratio is given by Equation (6).

$$LLR(x_1^T) \leq \frac{\sum_{i=1}^M \gamma(i)\frac{\hat{\gamma}(i)}{\hat{\gamma}(i)+\alpha}\hat{\delta}_i\Sigma_i^{-1}(\delta_i - \frac{\hat{\gamma}(i)}{\hat{\gamma}(i)+\alpha}\frac{\hat{\delta}_i}{2})}{\sum_{i=1}^M \gamma(i)} \tag{6}$$

This bound, which is a similarity measure between two sufficient statistics, is now used to replace the LLR score of the testing utterance. Table 1 shows the experimental results of the LLR baseline and the lower-bound system. The system based on this lower-bound is actually slightly better than the LLR baseline.

4 Robust Local Scoring Function

Although the lower-bound based system can achieve slightly better performance than UBM-MAP baseline, the complicated form of this lower-bound suggests that it may not be the optimal formulation. However, it does provide the main cues for designing new similarity measure. In the nominator, there are mainly two parts: $\gamma(i)\frac{\hat{\gamma}(i)}{\hat{\gamma}(i)+\alpha}\hat{\delta}_i\Sigma_i^{-1}\delta_i$ and $\gamma(i)\frac{\hat{\gamma}(i)}{\hat{\gamma}(i)+\alpha}\hat{\delta}_i\Sigma_i^{-1}\frac{\hat{\gamma}(i)}{\hat{\gamma}(i)+\alpha}\frac{\hat{\delta}_i}{2}$. The first part is a scaled cross-correlation between the adjustments at the ith component of the training and testing utterances. The second part is a threshold which only depends on $\hat{\delta}_i$ of the training utterance. Inspired by this lower-bound formulation, we propose a symmetric similarity measurement which only contains the cross-correlation part as follows

$$LLR_0 = \frac{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)\hat{\delta}_i\Sigma_i^{-1}\delta_i}{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)} \tag{7}$$

In LLR_0, the training and testing utterances are treated equally. The factor $\gamma(i)\hat{\gamma}(i)$ serves as a weight factor for each component. By this weight factor, those components which has enough observations from both training and testing utterances are emphasis. The denominator $\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)$ is a normalized term. LLR_0 is simpler than the lower-bound of LLR. However, the performance of this measure lags behind the baseline system. This is due to the dependence of LLR_0 on both training and testing data. To achieve better verification performance, LLR_0's data dependency needs to be reduced. In UBM-MAP framework, the ratio operation between the target model and the UBM model reduces the data dependency on the testing data to some extent. T-Norm [3] reduces the data dependency on the training data to some extent. In this paper, we would like to treat the data dependency on training data and testing data in a principle way.

To reduce the data dependency of LLR_0, the term $\hat{\delta}_i\Sigma_i^{-1}\delta_i$ needs to be normalized. In this paper, the normalization is done by carefully choosing thresholds. Recall the generalized likelihood ratio test $\frac{P(x_1^T|H_1)}{P(x_1^T|H_0)}$. These thresholds can be viewed as the likelihood of H_0 at each component. Thresholding at each component squeezes the distribution of target scores, and makes the score of target data at each component more consistent. This leads to better discrimination between target and imposter. Thinking the threshold as the reference point of hypothesis H_0 suggests that we can estimate them from a cohort speakers set [3, 6, 4, 9]. Firstly, for each target speaker, similarity LLR_0 is used to select the cohort speaker set from a cohort speaker pool (174 speakers in this paper). After the cohort speaker set is selected, the threshold of each component is the average score of the cohort speaker data for this component. The threshold of the ith component for the training utterance is therefore given by

$$\hat{t}_i^0 = \frac{1}{N_{cohort}}\sum_{m=1}^{N_{cohort}}\hat{\delta}_i\Sigma_i^{-1}\delta_i^m \tag{8}$$

where N_{cohort} is the number of cohort speakers and is set to 10 in this paper, $\hat{\delta}_i$ is the adjustment of the ith component of the training utterance, and δ_i^m is the adjustment of the ith component of the cohort speaker m.

In form of Equation (7), the role of training and testing utterances is symmetric, so the cohort speaker set selection can also be conducted for the testing utterance. This procedure is considered as online cohort speaker set selection [2, 10]. Notice that the cohort speaker set for the training utterance and that for the testing utterance may not be the same. The threshold of the ith component for the test utterance, t_i^0, is computed in the same way as in Equation (8).

After getting the threshold for each component, the sufficient statistics for each utterance is turned into

$$\{x_1^T\} \xrightarrow[\text{cohort speaker set}]{\text{UBM}} (\gamma(i), \delta_i, t_i^0)_{i=1}^M$$

The normalized LLR_0 is turned into

$$LLR_1 = \frac{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)[\hat{\delta}_i\Sigma_i^{-1}\delta_i - (\hat{t}_i^0 + t_i^0)/2]}{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)} \tag{9}$$

Following the same idea, we can use LLR_1 to select a new cohort speaker set from the same cohort speaker pool and compute the threshold for LLR_1 at each component.

$$\hat{t}_i^1 = \frac{1}{N_{cohort}}\sum_{m=1}^{N_{cohort}}[\hat{\delta}_i\Sigma_i^{-1}\delta_i^m - (\hat{t}_i^0 + t_i^0)/2] \tag{10}$$

Since t_i^0 is usually computed from a different set of cohort speakers from the entire cohort speaker pool, threshold t_i^1 incorporates information from a larger set of cohort speakers. We call this large cohort speaker set *the implicit cohort set*. The normalized LLR_1 is now turned into

$$LLR_2 = \frac{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)[\hat{\delta}_i\Sigma_i^{-1}\delta_i - (\hat{t}_i^0+t_i^0)/2 - (\hat{t}_i^1+t_i^1)/2]}{\sum_{i=1}^{M}\gamma(i)\hat{\gamma}(i)} \tag{11}$$

This type of iteration can be repeated as many as needed. However, the whole cohort speaker pool has a limited number of speakers (174 in this paper). After two iterations, the implicit cohort set is almost equal to the whole cohort pool, so the performance gains very little after LLR_2.

EER / DCF$\times 10^{-3}$	w/o TNorm	w TNorm
baseline	10.47% / 49.24	9.60% / 40.50
lower-bound.	9.63% / 46.94	9.36% / 39.41

Table 1. *Comparison between lower-bound based system and UBM-MAP baseline on NIST2002 corpus. The first row is the performance of baseline system without and with TNorm respectively, the second row is the performance of lower-bound based system without and with TNorm*

5 Experiments and Results

In order to show that the lower-bound in Equation (6) will not degrade the performance, we conducted the speaker verification experiments on NIST 2002 speaker recognition corpus [1] to verify this claim. The training data of the UBM model is the NIST 2001 develop set which contains 60 speakers. The 174 speakers in the NIST 2001 evaluation task is used as the cohort speaker pool. For T-Norm [3], these 174 speakers are used as the T-Norm Speaker pool. The UBM model contains 2048 components. Acoustic feature used in this paper is a warped version[7] of a 36-dimensional vector which combines static and first order derivatives of MFCC. The NIST 2002 corpus contains 330 speakers and 39105 trials. The training utterance of each speaker is a telephone conversation that lasts from 60 sec. to 120 sec. The testing utterance lasts from 3 sec. to 120 sec. The baseline system is a UBM-MAP system with the log-likelihood ratio scoring and the compared system is the UBM-MAP system with the lower-bound scoring. Table 1 shows the experimental results. Without T-Norm, the lower-bound based system outperforms the baseline system by about 8% in terms of DCF($49.24*10^{-3} \rightarrow 46.94*10^{-3}$) and by about 5% in terms of EER($10.47\% \rightarrow 9.63\%$); after T-Norm the approximation is still slightly better than baseline system.

Figure 1 shows the experimental results of the novel local scoring function on the NIST02 database. It clearly shows that this new type of measures, especially LLR_2, outperforms the lower-bound based system. The relative improvement is 12.28% in terms of EER ($9.36\% \rightarrow 8.21\%$) and is 11.42% in terms of DCF ($39.4*10^{-3} \rightarrow 34.9*10^{-3}$).

6 Conclusion and Future Work

In this paper, we consider the Universal Background Model as a mapping function which transforms the variable length observations (speech utterances) into a fixed dimensional feature vector (sufficient statistics). A novel similarity measurement between two feature vectors is proposed for text-independent speaker verification. The experiments with the NIST02 database indicate that this new local scoring function outperforms the conventional UBM-MAP framework by more than 10% both in Equal Error Rate

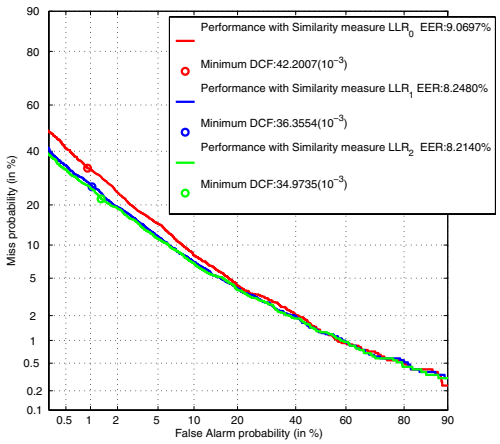

Figure 1. *Comparison of new Similarity measure for speaker verification on NIST2002 corpus. The blue curve is the performance with LLR_0 measure. The red curve is the performance with LLR_1. The green curve is the performance with LLR_2*

(EER) and Detection Cost Function (DCF). In the future, the new measurement will be extended in a time dependent fashion in order to incorporate the temporal information of the speech signals.

References

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

[2] A. Ariyaeeinia and P. Sivakumaran. Analysis and comparison of score normalization methods for text dependent speaker verification. In *Eurospeech*, pp.1379–1382, 1997.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. 10:42–44, 2000.

[4] R. Finan, A. Sapeluk, and R. Damper. Impostor cohort selection for score normalization in speaker verification. 18:881-888, 1997.

[5] J. Gauvain and C. Lee. Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions SAP*, 2:291–298, 1994.

[6] T. Kinnunen, E. Karpov, and P. Fränti. Efficient online cohort selection method for speaker verification. In *Proceeding of 8th Int. Conf. on Spoken Language Processing*, volume 3, pages 2401–2402, 2004.

[7] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proceeding of A Speaker Odyssey, The Speaker Recognition Workshop*, pages 213–218, 2001.

[8] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. 10:19–41, 2000.

[9] P. Sivakumaran, J. Fortuna, and A. Ariyaeeinia. Score normalization applied to open-set, text-independent speaker identification. In *Eurospeech*, page 2669–2672, 2003.

[10] Y. Zigel and A. Cohen. On cohort selection for speaker verification. In *Eurospeech*, pages 2977–2980, 2003.