A Quality Measure Method Using Gaussian Mixture Models and Divergence Measure for Speaker Identification*

Rong Zheng, Shuwu Zhang, Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
{rzheng,swzhang,xubo}@hitic.ia.ac.cn

Abstract

Previous work has demonstrated the promise of frame-level quality measure methods to robust speaker recognition. This paper explores the issues involved in applying soft estimates to quality measures as weighting factors in score computation. A quality measure algorithm using Gaussian mixture density and Jensen divergence measure is presented for traditional GMM-UBM scoring mechanism. Derivation and validation of the quality measurement are reported in this paper. We investigate the usefulness of different feature processing, different GMM-based quality models and incoporation of divergence measure for quality estimation. Comparison experiments performed on the NIST1999 SRE corpus show the effectiveness of the proposed method.

Index Terms: speaker identification, quality measure, Gaussian mixture density, divergence measure

1. Introduction

In recent years, Gaussian Mixture Models (GMMs) have become the dominant approach to text-independent speaker identification systems. The Gaussian Mixture Model-Universal Background Model (GMM-UBM) method has reported high performance in several NIST evaluations and is introduced for speaker identification in this paper [1-2].

The conventional GMM-based speaker identification uses the average log-likelihood scores to make a decision with respect to the whole test utterance. Previous work in speaker recognition has shown the promise of incorporating quality measures into the recognition process [3-5]. We are concerned with quality-based score computation here [5]. The motivation in the use of quality measures is to automatically determine the weighting factors for each feature vector that contributes to identify a speaker. There are mainly two kinds of quality measures: hard decisions and soft decisions.

Hard decisions based methods, e.g. frame pruning technique, were investigated to prune out some error-prone frames that have lower cohort-normalized likelihood scores [3-4]. The final decision is based on a subset of active frames. However, quality measures incorporating soft estimates replace the discrete decisions with an estimate of the probability that the feature vectors are reliable. The probability calculation is then used as weighting factors for each frame. In [5], a generic framework incorporating quality measure has been proposed and a frame-level quality measure meeting a goodness criterion based on deviation from the fundamental frequency was used.

In this paper, we propose a quality measure algorithm using soft estimates to determine the degree of frame reliability based on GMMs. Generally, each individual component Gaussian is interpreted to represent some broad acoustic classes (underlying broad phonetic sounds) [6], which are used to describe potentially phonetic-specific speaker characteristics here. Good features are selected and GMM reference models for quality estimation are then created. In order to automatically extract from the training speech signal the part that best contributes to estimate the frame-level quality, a procedure based on Jensen divergence measure are also applied during the training and the identification process.

The remainder of this paper is organized as follows. In section 2, a brief review of the conventional GMM-UBM scoring based speaker identification is provided for notation and definition. Section 3 describes how different quality measures are incorporated into the score computation. In section 4, the considered quality measure methods are detailed. The experimental results are shown in Section 5. Finally, some conclusions are given in Section 6.

2. Conventional GMM-UBM based speaker identification

In GMM-UBM based speaker identification, a UBM is trained using speech utterances from a large group of speakers to represent the characteristics of all different speakers. Each speaker model is derived from the UBM by employing maximum a posteriori (MAP) adaptation using speaker-specific training speech [1]. The UBM and speaker models are modeled by GMM that is a weighted sum of multivariate Gaussian probability distributions. A GMM with M Gaussian components is parameterized mathematically by the notation $\lambda = \{w_m, \mu_m, \Sigma_m\}, m = 1, 2, ..., M$. For a sequence of input feature vectors $X = \{x_1, x_2, ..., x_T\}$, the average log-likelihood value of speaker model $\lambda_i, i = 1, 2, ..., I$ is given by

$$\log p(X / \lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t / \lambda_i) \qquad (1)$$

For closed-set recognition, the speaker corresponding to the most likely GMM is hypothesized as the speaker of the test utterance.

3. Review of integrating quality measure into score computation

The traditional scoring mechanism has the drawback of regarding all the preserved information as equal in terms of

* This work was partially supported by NSFC under the grant No. 60475014 and National Hi-tech Research Plan under the grant No. 2005AA114130.

importance. Promising results have been reported by incorporating quality measures into the recognition process. Two kinds of quality measures have been introduced into quality-based score computation: hard decisions and soft decisions. We will give a brief description of them in the following subsections.

3.1. Hard decisions based quality measure

The underlying idea in the quality-based score computation utilizing hard decisions suggests feature vector mask. The mask consists of 0's and 1's, with 0 meaning the feature vector is eliminated and with 1 indicating the vector is desirable.

In previous work, frame pruning has been proposed and demonstrated the promise of the elimination of some frames from speaker recognition process [4]. These frames should be the parts of the speech utterance lacking of speaker specific information. Assuming that all non-target frame scores will be close to the UBM, fixed-rate pruning (i.e., a predefined ratio of the frame vectors) and adaptive-rate pruning (i.e., the frames whose scores do not exceed some preset pruning threshold) have been presented. In order to obtain a meaningful comparison between different frames [4], frame pruning is based on the distance between the hypothesized model's likelihood score and the UBM's likelihood score in GMM-UBM speaker recognition system.

Although frame pruning technique is effective to some extent, it will obviously increase the likelihood of other speaker's attacks in the model set. Pruning of these frames is a random process ideally [4].

3.2. Soft decisions based quality measure

Instead of forcing hard decisions, Garcia-Remero et al. [5] replace discrete decisions with soft estimates of the quality measure as weighting factors in the score computation process. The motivation is to use intermediate values to indicate the degree of confidence whether or not the feature vector is masked. A frame-level quality measure based on unimodal deviation from the fundamental frequency was presented.

4. Description of the proposed method

The Gaussian components can be considered to model some underlying broad phonetic sounds which characterize a person's voice [6]. Our basic assumption is that integrating speaker-specific phonetic variability could potentially contribute to identify a speaker and so improve the performance.

In our quality measure design, we are considering the variability introduced by phonetic classes to be preserved. In the training phase, a speaker's speech is used to produce a quality reference model. This is accomplished by modeling the probability distributions of the employed features for each speaker with a GMM. Features extracted from the test signal are assessed using the quality models, by calculating a similarity measure with respect to each quality GMM. The similarity values can be viewed as indicators of speech quality and are utilized as weighting factor in score computation.

Fig.1 shows a general block diagram of the proposed quality measure method. The details of the algorithm's functional blocks and related issues to obtain the quality value are discussed as follows.

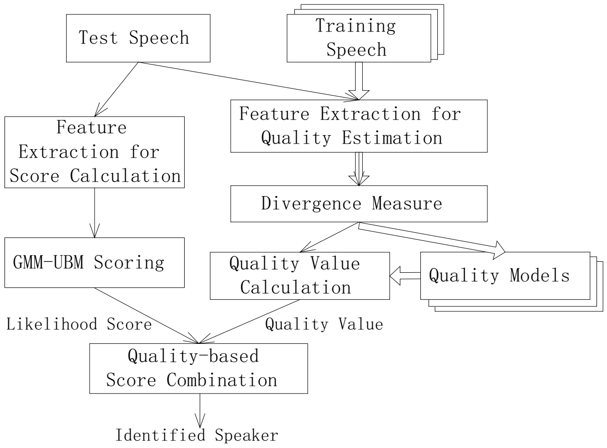
Test Speech
Training Speech
Feature Extraction for Quality Estimation
Feature Extraction for Score Calculation
Divergence Measure
GMM-UBM Scoring
Quality Value Calculation
Quality Models
Likelihood Score
Quality Value
Quality-based Score Combination
Identified Speaker

Figure 1 *Block diagram of the proposed quality measure method.*

4.1. Quality value calculation and score combination

In order to design a quality estimation criterion and produce a quality value, the training speech of each speaker is used for the modeling of a speaker-dependent Gaussian mixture model which will be utilized to assess the features in recognition.

For each dimension of individual Gaussian component, instead of calculating the likelihood score, a quality value is computed using equation 2, which are the sum of shadow areas (left plots) or lengths of vertical lines (right plots) depicted in Figure 2. A similar method was used for quality estimation based on unimodal Gaussian model of pitch in [5].

$$q_n(y_t^d) = p(| y_t^d - \mu_{Y_n^d} | < | Y_n^d - \mu_{Y_n^d} |) \tag{2}$$

where $Y_n^d \sim N(\mu_{Y_n^d}, \sigma_{Y_n^d})$ is the d -th dimension's normal feature distribution of n -th Gaussian component. $y_t^d, d = 1, 2, ..., D$ is the feature value for quality estimation at time instant t, and p denotes the probability. For each test file, the final frame-level quality value is estimated as a weighted combination of Gaussian component quality signal that is the average of multidimensional quality value from the lowest feature order to the highest order as in equation 3 (i is omitted for concision in the right-hand side of (3)),

$$Q_t^i = \sum_{n=1}^{N} w_n \{\frac{1}{D} \sum_{d=1}^{D} q_n(y_t^d)\} \tag{3}$$

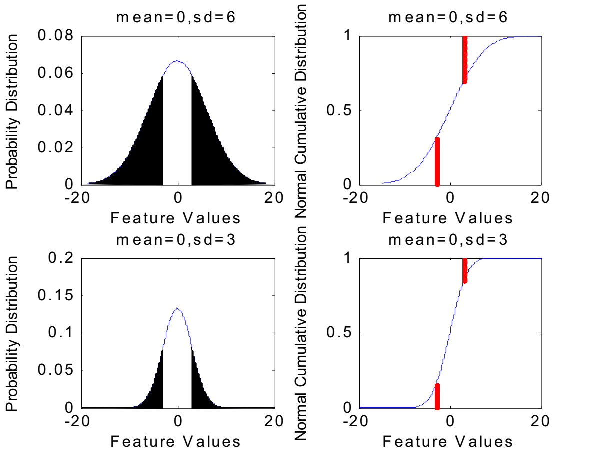

Figure 2 *Illustration of dimensional quality calculation and changing the variance of the normal feature distribution.*

where $w_n, n = 1,...N$, is the mixture weight of a particular quality model $i, i = 1,2,...,I$. The resulting value lies in the closed interval from 0 to 1, which gives the confidence degree of acoustic unit's reliability. The quality-based likelihood scores are computed as follows to find the identified speaker.

$$\hat{i} = \arg\max_{1\le i\le I} \frac{1}{\sum_{t=1}^{T} Q_t^i} \sum_{t=1}^{T} Q_t^i * \log p(x_t / \lambda_i) \quad (4)$$

4.2. Feature extraction

In order to obtain an effective quality measure algorithm, speaker-dependent information with discriminative capabilities among speakers should be used to train the quality criterion.

Although other discriminative features can be adopted, the acoustic front-end is fixed to MFCC processing in this study. As we are interested in measuring quality variation due to the speakers, higher order spectral features determine the amount of detail in the speech spectrum and tend to contain more useful speaker information. Because experimental evaluations are performed on telephone speech, 19 Mel filters are utilized in useful telephone bandwidth and the features, containing 16 static coefficients with zeroth coefficient excluded, are mean-subtracted and evaluated for quality estimation.

Fig.2 illustrates graphically the problem that the dimensional quality value changes when the mean of the feature distribution stays constant and the standard deviation changes. For the same feature value case, a smaller quality value occurs with smaller variance.

So identification results are compared for two kinds of cepstral coefficients processing for deriving features with different variance scales. 1) mean-subtracted and bandpass liftering in cepstral domain (BLift) [7]; 2) mean-subtracted and short-time cepstrum mean and variance normalization (CMVN)[8]. As 16 static MFCCs are extracted, we also compare the experiments conducted on the lower 8-order (L8) static features and the higher 8-order (H8) static features separately. The results are provided in Section 5.2.

4.3. GMM and GMM-UBM for quality modeling

In this study, two different quality modeling techniques, GMM and GMM-UBM, are implemented respectively.

For the GMM-UBM modeling, the quality models for individual speakers are created by adapting a quality UBM. During recognition, the quality values of the test speech signal are assessed for each quality GMM. In this implementation, only 5 components chosen from the most probable mixtures in the quality UBM are used for the quality value calculation.

For the conventional GMM framework, the features are used to model the variability introduced by acoustic classes to describe the speaker identity. All the mixture components are computed to obtain the quality signal in recognition.

4.4. Divergence measure

In this subsection, a divergence measure is introduced to extract the most "important" part of speech signal to represent the particular speaker during the quality modeling procedure and to calculate the more correct quality signal in identification.

Jensen divergence measure was used to automatically extract from the input speech signal the part that best contributes to identify a speaker [3].

Here, we use this divergence measure to select the most representative speech frames and remove some confusion-prone ones for quality estimation. Only the vectors with higher Jensen divergence measure are modeled and scored. We quantify the contribution of each feature vector using the Jensen difference (JD) (cf. [3] for details).To conform the GMM-UBM baseline system, a vector of dimension I (i.e., the number of target speakers, 230 and 309 in our cases separately) is evaluated for each input speech frame. If the value of Jensen difference is smaller than the preset threshold, the input vector is considered to be no contribution to the quality modeling and identification process. A unitary preset threshold is determined experimentally.

5. Experimental results

5.1. Corpus and GMM-UBM baseline system

The NIST1999 speaker recognition evaluation (SRE) one-speaker corpora are derived from the Switchboard 2 corpus and consist of variable length test utterances (with a primary focus on segments with 15s~45s) extracted from conversational telephone speech. This corpus consists of 230 male and 309 female speakers with 2 minutes of training speech from two telephone calls (each training segment is 1 minute long). 1-speaker test segments from the same telephone line with mixed handsets (i.e., the training and recognition conditions match closely) and all of the telephone lines are used for identification separately. A detailed description of the corpus can be found in [9]. In section 5.2 we selected the male subset containing 230 speakers to find the optimal configuration. Then the fine-tuning parameters are validated on the female subset with 309 speakers in section 5.3. No cross-gender experiments are performed.

Speech utterances are divided into 24-ms frames with 50% overlap, ignoring about 10%~15% low-energy frames. 16 MFCCs and 16 delta coefficients are calculated. CMVN is applied to mitigate channel effects [8]. The UBM with 1024 mixture components is trained on the gender-dependent partition of the NIST2000 SRE [9]. All speaker models are created by MAP adaptation of the UBM with the mixture weight and the covariance matrix unchanged. The top five mixture components are used for the calculation of the likelihood value during the mixture scoring procedure [1].

5.2. Performance of speaker identification systems by integrating quality measure into the baseline system

Comparison experiments are conducted to obtain the most effective combination of feature processing and mixture model configuration using the same line condition of the male subset.

Table 1 shows the performance of the baseline system and the comparative results when integrating quality measure into the baseline system. The identification rate of the baseline system compares favorably with those reported by Kinnunen et al. in [10] (16.9% identification error rate for GMM method). In Table1, the experimental results incorporating quality estimation are performed with 16-component diagonal GMMs and 1024-mixture GMM-UBM quality models, respectively. We note that the recognition rates of GMM and GMM-UBM quality modeling are quite similar. However, bandpass liftering processing gives a slight reduction from the baseline. An experiment was also done to see the effectiveness when incorporating dynamic features to static features for quality measure. However, we have found the benefits of dynamic features to be negligible compared to static features only. The main cause might be the fact that too detailed speaker-specific acoustic representation for quality estimation seems undesirable for text-independent speaker identification. Moreover, fewer feature dimensions allow a computational savings when modeling and scoring the speech. So only the experimental results performed on static features are reported.

Furthermore, we investigate the impact of combination of CMVN and Jensen divergence measure, which are listed in Table 2. Under the same line condition, the best system results in a 31.9% relative reduction in error compared to the baseline. It is also observed that the removal of high-order cepstral coefficients reduces identification accuracy. Based on these observations it appears reasonable to extract more representative part of speech and preserve high quefrencies information when improving performance in future work.

The identification rates using the best configuration for the same line condition and all line condition are summarized in Table 3. Compared to the conventional GMM-UBM scoring mechanism, the proposed method is more effective, which yields 13.4% relative reduction in error for all line condition of the male partition. The increase of absolute identification rate under different line condition is higher than that under the same line condition as indicated in Table 3.

In Table 3, the proposed method slightly outperforms frame pruning or Jensen difference, even though the difference is not large. The criterion based on deviation from the fundamental frequency gives worse identification rates which are not provided here. These results may suggest limited utility of single-frame spectral representation in quality estimation. However, the benefits of the proposed algorithm are obvious. A quality measure algorithm based on GMMs is presented and alternative features preserving salient speaker information can be included and modeled according to this framework. One would hope that there are further gains by combining the baseline system and improved GMM-based quality estimation.

5.3. Validation

The fine-tuning parameters, which led to the best results, are validated on the female partition of the NIST1999 SRE corpus. A comparison of the identification rates for different experimental systems is also given in Table 3. The comparable performances show the interest of the proposed algorithm for speaker identification.

6. Conclusions

A quality measure algorithm using Gaussian mixture density and Jensen divergence measure in score computation has been provided in this paper. It has the advantage of estimating quality by extracting the best part of speech to potentially utilize broad phonetic-specific speaker characteristics by GMM modeling. Even though the recognition time is increased, experimental results demonstrated that the proposed algorithm is feasible and can significantly improve the performance compared to the conventional GMM-UBM scoring. The proposed method also outperforms the hard decisions based techniques, e.g. frame pruning. Future work should include the choice of good speaker-dependent features, e.g. long-term higher-level features, for quality estimation and the incorporation of fast recognition procedure to reduce the computational load [10].

Table 1. *Comparison of quality measure methods with different feature processing and modeling approaches on data from the male partition of the NIST1999 SRE corpus (same line).*

Experiments	Identification rates (%)
Baseline	85.6
Baseline+BLift GMM/UBM	85.4
Baseline+BLift GMM	84.1
Baseline+CMVN GMM/UBM	**87.1**
Baseline+CMVN GMM	**86.7**

Table 2. *Comparison of quality measure methods with different number of feature dimensions and integration of Jensen difference measure on data from the male partition of the NIST1999 SRE corpus (same line).*

Experiments	Identification rates (%)
Baseline+CMVN GMM	86.7
Baseline+L8 CMVN GMM+JD	89.0
Baseline+H8 CMVN GMM+JD	89.6
Baseline+CMVN GMM+JD	**90.2**
Baseline+CMVN GMM/UBM+JD	89.0

Table 3. *Summary of the identificaiton rates for different recognition conditions on the male partition of the NIST1999 SRE corpus and validation of the proposed algorithm on data from the female partition (all line=same line+different line).*

Experiments	Male		Female	
	Same line	All line	Same line	All line
Baseline	85.6	57.6	74.1	54.9
Baseline+Pruning	89.5	62.8	77.7	58.9
Baseline+JD	88.2	61.8	75.2	57.0
Proposed	**90.2**	**63.3**	**78.1**	**59.1**

7. References

[1] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, Vol.10 pp. 19-41, 2000.

[2] Reynolds, D.A., Campbell, W. et al.: The 2004 MIT lincoln laboratory speaker recognition system. Proc. ICASSP, Vol.1 pp. 177-180, 2005.

[3] Vergin, R., O'Shaughnessy, D.: A double Gaussian mixture modeling approach to speaker recognition. Proc. Eurospeech, 1997

[4] Besacier, L., Bonastre, J.F., Fredouille, C.: Localization and selection of speaker-specific information with statistical modeling. Speech Communication. Vol.31 pp.89-96, 2000

[5] Garcia-Romero, D., Fierrez-Aguilar, J. et al.: On the use of quality measures for text-independent speaker recognition. Proc. Speaker Odyssey, 2004

[6] Reynolds, D.A., Rose, R.C.: Robust Text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. SAP, pp.72-83, 1995

[7] Zhen, B., Wu, X. et al: On the use of bandpass liftering in speaker recognition. Proc. ICSLP, Vol.2 pp. 933-936, 2000

[8] Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, Vol.25 pp.133-147, 1998

[9] NIST. The 1999 and 2000 speaker recognition evaluation plans. Available: http://www.nist.gov/speech/tests/spk/index.htm

[10] Kinnunen, T., Karpov, E., Franti, P.: Real-time speaker identification and verification. IEEE Trans. SAP, pp.277-288, 2006