Exploiting GMM-based Quality Measure for SVM Speaker Verification

Rong Zheng, Hongchen Jiang, Shuwu Zhang, Bo Xu

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

Abstract. In this paper, we examine the problem of quality measurement for speaker verification using support vector machines (SVMs). An efficient Gaussian mixture models (GMMs) based quality estimation algorithm is proposed to potentially utilize speaker-specific broad acoustic-class characteristics. Some verification strategies are also considered in the test phase. We perform clustering-based vector pre-quantization to reduce the computational load and the redundancy in speech signal. Quality estimation is also integrated into test-length normalization. We then apply it to a text-independent speaker verification task using the NIST 2002 speaker recognition evaluation (SRE) database. Experimental results show that the proposed method can produce good classification accuracy.

Keywords: Speaker verification, quality measure, vector pre-quantization, support vector machines, Gaussian mixture models.

1 Introduction

In the case of speaker verification, the most common approach are based on Gaussian mixture models (GMMs), which has been proven highly successful due to their good performance, especially with the use of Maximum A Posterior (MAP) adaptation [1]. Support Vector Machines (SVMs), as proposed by [2], are more and more often used in machine learning applications. Several recent studies consider the application of SVMs to speaker recognition [3][4]. A sequence kernel-based learning algorithm, named Generalized Linear Discriminant Sequence (GLDS), is to proven to be the one of the most powerful approaches [3]. It is based on training on one sequence and testing on another one. The key principle of this approach is to explicitly map sequences to a high-dimensional feature space using an average polynomial expansion.

The system called GLDS uses a very simple averaging strategy to make yes/no decision. A more intelligent approach would yield better results. An exciting area of recent work in speaker recognition is the use of quality measure methods for conventional scoring mechanism [5][6]. The quality values can be used to characterize the consistency between the observation and the model. The basic principle of quality measure is to automatically determine the weighting factors for feature vectors that contribute to identify a speaker.
In this paper, we combine the GLDS SVM framework with GMM-based quality measure concept. We develop a quality measure algorithm using soft estimates to determine the degree of acoustic reliability based on Gaussian mixture density. Generally, each individual component Gaussian is interpreted to represent some broad acoustic classes [7], which are used to describe potentially phonetic-specific speaker characteristics here. In order to improve the computational efficiency of SVMs and the redundancy in speech signal when incorporating quality measure strategy, clustering-based representative vectors are used to represent similarly acoustic characteristics during the test phase. Furthermore, a test-length score normalization (Tnorm) incorporating GMM-based quality measure is also employed during the verification process [8].

The organization of this paper is as follows. In section 2, SVM speaker verification involved in the experiments and their implementation are presented. Following is a brief description of quality measurements. In section 4, integrating GMM-based quality measure into SVM speaker verification is presented and some verification strategies are also considered. Experimental settings and classifier performances are presented in section 5. Conclusions are reported in section 6.

2 SVM Speaker Verification

Speaker verification requires the system to perform binary decision. A SVM classifier is ideally suited for such a task. Even if the speaker verification task can be seen as a two-class classification problem, SVMs can not be applied directly, because the examples are sequence and classical SVMs can only work with fixed size vectors.

A strategy to construct sequence kernels, recently developed by Campbell in [3], is based upon training on one sequence and testing on another one. Campbell amounts to mapping explicitly sequences to a fixed dimension space (called feature space) using a polynomial expansion, and to perform a dot product in this space. A new kernel called GLDS of the form was proposed:

\[
K(X_i, X_j) = \Phi(X_j)\Gamma^{-1}\Phi(X_j)
\]

(1)

where \(X_i\) and \(X_j\) are two input sequences, and \(\Gamma\) is a matrix derived by the metric of the feature space induced by \(\Phi()\). This matrix is usually a diagonal approximation \(\Upsilon\) of the covariance matrix computed over all concatenated impostor sequences. Campbell proposed a method to normalize each expansion coefficient using \(\Upsilon\). Further expression was defined as follows:

\[
\Phi(X_i) = \frac{1}{T_i} \sum_{t_i=1}^{T_i} \phi(x_{t_i}), \quad \phi_{\text{norm}}(x_{t_i}) = \frac{\phi(x_{t_i})}{\sqrt{\Upsilon}}.
\]

(2)

where \(x_{t_i}\) is a frame of sequence \(X_i\) at discrete time \(t_i\), and \(\phi_{\text{norm}}()\) is the normalized version of \(\phi()\). So equation 1 can thus rewrite as:
\begin{equation}
K(X_i, X_j) = \frac{1}{T_j} \sum_{t_i=1}^{T_i} \phi^{\text{norm}}(x_{t_i}) \cdot \frac{1}{T_j} \sum_{t_j=1}^{T_j} \phi^{\text{norm}}(x_{t_j}).
\end{equation}

Once all vectors are computed and normalized, they can be used as input to a linear SVM. Instead of making a Gaussian assumption, the optimization condition of SVM relies upon a maximum margin concept. SVM is a purely discriminative approach where the boundaries are directly learnt from the data.

SVMs are trained to separate the target from the background samples. Each conversation side generates a point in hyper-space, scores are computed as a distance from the hyper-plane to the point. The probability of the entire sequence can be computed as follows.

\begin{equation}
P(X_i, \ldots, X_j | w) = \frac{1}{T_j} \sum_{t_i=1}^{T_i} w_i \cdot \phi^{\text{norm}}(X_{t_i}) = w \cdot \left( \frac{1}{T_j} \sum_{t_i=1}^{T_i} \phi^{\text{norm}}(X_{t_i}) \right).
\end{equation}

where \(w\) is the target model that is compacted by all the support vectors from the training set, and \(t\) denotes matrix transpose. Scores for each target speaker are an inner product between the speaker model and the average expansion.

### 3 Quality Measures

The traditional scoring mechanism has the drawback of regarding all the preserved information as equal in terms of importance. Promising results have been reported by incorporating quality measures into the recognition process. There are mainly two kinds of quality measurements that have been introduced into quality-based score computation: hard decisions and soft decisions. Some comparative experiments of quality measure methods for GMM-UBM speaker identification were reported in [9].

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 for recognition. In previous work, frame pruning has been proposed and demonstrated the promise of the elimination of some frames from speaker recognition process [5]. These frames should be the parts of the speech utterance lacking of speaker specific information. 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 which are based on the distance between the hypothesized model’s likelihood score and the world model’s likelihood score.

Instead of forcing hard decisions, Garcia-Remero et al. [6] 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 Integrating Quality Measure into SVM Speaker Verification

In this paper, a better technique to determine the score for a speaker is studied. Speech quality is one of the most important factors to be considered in evaluating speaker verification system. In order to illustrate the extent to which data quality can affect the performance of verification system, we consider their impact on classification by assigning different quality values according to data characteristics. We aim at obtaining some parts of the evaluation set containing ‘good’ and ‘bad’ samples and give low quality values to the ‘bad’ ones since they are likely to be poorly classified.

The components of the proposed SVM speaker verification system integrating GMM-based quality measure are shown in Figure 1. Feature extraction is performed for the silence-removed speech frames. The feature vectors are per-quantized to a smaller number of the centroid (mean vector) of the clusters using k-means clustering. On one hand, the match scores for each speaker are calculated, on the other hand, quality-based Tnorm scores are also computed. Finally, we utilize the test-length normalized output score to obtain verification results.

Fig. 1. Block diagram of SVM speaker verification integrating GMM-based quality measure.

4.1 Integrating Quality Measure into Verification Scores

In our quality measure design, we are considering the consistency between the observation and the speaker model. In the training phase, GMM quality reference models are trained for each target speaker. 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 will be utilized as weighting factors in score computation.

In this subsection, we describe the proposed quality measure method. Figure 2 shows an example how the quality value develops with feature dimension.
For each dimension of individual Gaussian component, instead of calculating the likelihood score, a quality value is computed using equation 5, which are the sum of shadow areas (left plots) or lengths of vertical lines (right plots) depicted in Figure 1. A similar method was used for quality estimation based on unimodal Gaussian model of pitch in [6].

$$q_m(y^d_{t_j}) = P\left(\left| y^d_{t_j} - \mu_{Y_m^d} \right| < \left| Y_m^d - \mu_{Y_m^d} \right| \right).$$  \hspace{1cm} (5)

where $Y_m^d \sim N(\mu_{Y_m^d}, \sigma_{Y_m^d})$ is the $d$-th dimension’s normal feature distribution of $m$-th Gaussian component. $y^d_{t_j}, d=1,2,\ldots,D$ is the feature value for quality estimation at time instant $t_j$, 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 multi-dimensional quality value from the lowest feature order to the highest order as in equation 6,

$$Q_{t_j} = \sum_{m=1}^{M} \omega_m \left\{ \frac{1}{D} \sum_{d=1}^{D} q_m(y^d_{t_j}) \right\}.$$ \hspace{1cm} (6)

where $\omega_m, m=1,\ldots,M$, is the mixture weight of a particular quality model. 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 score is computed as follows, where $w$ is the target speaker’s SVM model.

$$\hat{P}(X_1,\ldots,X_{t_j} | w) = \frac{1}{\sum_{t_j=1}^{T_j} Q_{t_j}^w} \left( \sum_{t_j=1}^{T_j} Q_{t_j}^w \cdot \phi_{\text{norm}}(X_{t_j}) \right).$$ \hspace{1cm} (7)
4.2 Clustering-based Vector Pre-quantization

As indicated in equation 7, if we estimate a quality value for each input vector, the computation time is very expensive. In practice, the adjacent feature vectors are close to each other in the acoustic space because of the gradual movements of the articulators, so there is large redundancy in speech vectors. Some of speech features are redundant or dependent on other features, which is not useful for classification. Non-discriminative frame-level quality values are given for the similarity of acoustic characteristics. So clustering-based pre-quantization is performed before the scoring procedure for two reasons. The first one is to improve the verification efficiency of SVM scoring and reduce the computational load. The second one is to achieve a more concise data representation and to deal with the redundancy in speech frames.

In clustering-based pre-quantization, we partition the sequence $X$ into $C_N$ clusters using the k-means clustering algorithm (the number of clusters is set a priori according to the total number of speech frames). In averaging, the representative vectors are the centroid (mean vector) of the clusters.

So we can rewrite equation 7 as:

$$
\hat{P}(X_1, \ldots, X_{C_n} | w) = \frac{1}{\sum_{c_n=1}^{C_n} \text{Count}(c_n) * Q_{c_n}^w} \cdot \left\{ \sum_{c_n=1}^{C_n} (\text{Count}(c_n) * Q_{c_n}^w) \cdot \left( w' \cdot \phi_{\text{norm}}(X_{c_n}) \right) \right\}
$$

where $Q_{c_n}^w$ is the quality value of representative vector of class $c_n$, against a particular GMM-based quality model associated with $w$.

4.3 Quality-based Tnorm

Tnorm was proposed to compensate for bias due to test length and more [8]. Tnorm parameters are estimated from scores of each test segment $X$ against a set of imposter speaker models at test time. Then the mean and standard deviation of the imposter scores are used to transform the target speaker’s score.

After pre-quantization, we calculate quality-based impostor scores so as to adjust the speaker’s score more conformingly. We not only train Tnorm impostor models within the framework of SVMs, but create Tnorm models with GMM approach for quality-based Tnorm parameter calculation using the similar expression as in equation 8 with the exception for impostor quality models.
5 Experiments and Results

The described speaker verification system has been developed and evaluated using data from NIST2001 and NIST2002 speaker recognition evaluations (SRE) [10]. In this section, we will describe the corpus, the baseline system and comparative performances of speaker verification integrating GMM-based quality measure.

5.1 Corpus

The NIST2002 corpus includes cellular data extracted from the Switchboard Cellular part 2. This corpus consists of 139 male and 191 female speakers with 2 minutes of training speech from a single cellular phone call. There are 3570 test segments (1442 males and 2128 females) which range from few seconds to a minute (with a primary focus on segments with 15s~45s). All of the verification utterances have to be scored against gender-matching impostors and true speaker. A detailed description of the evaluation corpus can be found in [10]. No cross-gender experiments are performed.

NIST2001 SRE is used to collect the background set and the needed speech for detection score normalization for SVMs. This corpus includes 60 development speakers (38 males and 22 females), 174 test speakers (74 males and 100 females) and a total of 2038 verification speech segments (850 males and 1188 females). Gender-dependent Tnorm scores are computed using 100 males and 100 females from the NIST2001 SRE.

5.2 Baseline System

In order to validate the proposed quality measure method, we compare its performance to state-of-the-art system using the GLDS kernel (cf. [11] for details). It consists of an average polynomial expansion. All experiments are implemented using the LIBSVM package [12]. For fair comparison, exactly the same development and evaluation data are used for two systems.

Feature extraction is performed as follows. Speech utterances are divided into 24ms frames with a shift of 12ms, ignoring about 10%~15% low energy frames based on a simple energy detector. 16 Mel-Frequency Cepstral Coefficients (MFCCs) and 16 delta coefficients are calculated. Cepstral mean subtraction and variance normalization is applied to mitigate linear channel effects [13]. The maximal polynomial order is set to 3 and the size of the mapping is 6545. Tnorm is performed to transform output scores. Details on Tnorm for GLDS SVM classifier can be found in [14].

5.3 Performance Assessment of Comparison Experiments

We illustrate an added benefit of considering data characterization within the scoring procedure, namely the effect of the quality values of speech samples in classification. 16-component diagonal GMMs for quality estimation are trained using expec-
tation-maximization (EM) algorithm. The performance assessment of both the baseline GLDS SVM speaker verification system and the system combining the GMM-based quality measure with GLDS SVM are evaluated using Detection Error Trade-off (DET) curves [15] and Detection Cost Function (DCF) [10]. The DCF is defined to weight two types of errors, that is, miss detections and false alarms.

For both male and female partitions, the experiments of single-frame quality estimation using equation 7, namely not including clustering-based pre-quantization, report similar results compared to the baseline. A large amount of time is needed to obtain the verification results. So the performances of DET curves employing single-frame quality estimation are omitted here. The results show that possibly relying on acoustic characteristics produce an increase in the confusion or do not contribute to discrimination for frame-level quality estimation, due to the redundancy in the feature vectors and data variability caused by channel noise and acoustical environment. During the training phase we do not make any selection among the feature vectors intended for reference creation, it means the quality models may be created with environment noise. Because there are essential difference among sessions, recognition data affects on the quality estimation performance and does not provide recognition improvements.

The DET curves for comparison experiments using clustering-based quality measurement are shown in Figure 3 and Figure 4, for male and female partitions, respectively. Table 1 gives the corresponding EER and the minimal DCF for the above described corpus. Some interesting results are that the proposed algorithm satisfactorily improves the performance at low false alarm rates. The presented verification approach brings 9.0% relative reduction of EER and 10.0% relative reduction of the minimal DCF compared to the baseline system for male data of the NIST2002 SRE, and 6.0% relative reduction of EER and 9.1% relative reduction of the minimal DCF for female partition. This may be explained by the fact that the clustering-based feature space reduces the huge variation of frame-level acoustic features, while cluster representatives are preserved to describe acoustic data similarity. As we can see in Figure 2, the closer the obtained feature vector is to the means of Gaussian components, the larger the quality value is. For example, if the feature vectors are distributed within the vicinity of the means of quality models, the centroid gives the largest quality estimation. The experimental results demonstrate the promise of the proposed strategy and the verification performance is well improved by integrating clustering-based quality measure into the GLDS SVM speaker verification system.
Fig. 3. Verification performance on data from male partition of the NIST 2002 SRE. The minimal DCF operating point is indicated with a circle.

Fig. 4. Verification performance on data from female partition of the NIST 2002 SRE.

Table 1. EER and minimal DCF comparisons for male and female partitions.

<table>
<thead>
<tr>
<th>Partition</th>
<th>EER (%)</th>
<th>Min. DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>8.85</td>
<td>0.0319</td>
</tr>
<tr>
<td>Female</td>
<td>8.63</td>
<td>0.0363</td>
</tr>
</tbody>
</table>
6 Conclusions

In this paper, we have compared quality measure for speaker verification over cellular data. We investigate the usefulness of GMM-based quality measure as weighting factors in score computation. Clustering-based pre-quantization is presented to cope with the time-consuming computational load and the redundancy in speech frames. Quality-based test normalization is also used to compensate the variability of output score distribution. The performance of speaker verification system has been well improved by combining quality estimation with the GLDS SVM classifier in a NIST speaker recognition evaluation task.

References