SPEAKER MODELING DISTANCE NORMALIZATION TECHNIQUE IN MULTILINGUAL SPEAKER VERIFICATION

KSHIROD SARMAH & UTPAL BHATTACHARJEE

Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India

ABSTRACT

Speaker modeling distance normalization (D-Norm) is one of the important score normalization techniques in speaker verification (SV) system. For D-Norm implementation, it doesn’t need any additional speech data or external speaker population, that being one of the essential advantages of D-Norm. But traditional D-Norm has still disadvantage of its time complexity due to the Monte-Carlo based Kullback –Leibler distance estimation approach. In this paper we used a simplified an innovative D-Norm technique, which is based on the upper bound of the KL divergence between two statistical models namely GMM-UBM adapted target speaker model and the UBM model. This approach proved better performance. From the experimental point of view we can conclude that the performance of the multilingual SV system has been degrades approximately 3.0% due to language mismatching of the training and testing of the system. Also by applying D-Norm, the performance of the baseline system has been improved approximately 2.5% by reducing the EER of 8.91% to 6.06% for language matching condition and 11.66% to 10.60% for language mismatching conditions respectively. Finally, it is also observed that the performance of SV system after applying D-Norm is better in language matching condition than that of mismatching condition.

KEYWORDS: GMM-UBM, Kullback-Leibler Distance, D-Norm, Speaker Verification

INTRODUCTION

Automatic Speaker Verification (ASV) is the use of a machine to verify a person’s claimed identity from his or her voice [1]. In multilingual environment, the speaker can be determined with the utterance spoken in different languages that of the training system. The performance of the multilingual SV system degrades due to mismatching of phonetic contents due to language variability of training and testing process. So, for robustness of multilingual activities lots of normalization techniques have been already applied. To deal with this problem, various compensation techniques for channel effects and mismatching between training and testing languages have been proposed.

The pattern-matching problems can be formulated as measuring the likelihood of an observation given speaker model using stochastic model [1]. Classification methods for speaker recognition in recent years have centered on statistical approaches [3]. For this works, the Gaussian Mixture Model (GMM)-Universal Background Model (UBM) classifier has become the most dominant approach over the past decades and achieves state-of-art system performance. Also Support Vector Machine (SVM) has been proved to be an effective method for speaker recognition in the recent years. Above all, several other classifiers are developed recently. The most important and successful one including SVM using GMM supper vector (GSV-SVM) which concatenate the GMM mean vectors as the input for SVM training and testing. The Joint Factor Analysis (JFA) introduced by Kenny (Kenny et al.2007) that jointly models the channel subspace and the speaker subspace. Although other methods achieve their good performance, GMM-UBM is still the dominant basis for their developments.
In generally there are two categories of normalization techniques, namely feature level normalization, and score level normalization. In feature level normalization it aims at removing the channel effects, phonetic context mismatching effects from the feature vectors prior to train the speaker models. Cepstral Mean Substraction (CMS), Cepstral Variance Normalization (CVN), Relative Spectral (RASTA), etc. are some common example of feature normalization approaches. In score domain normalization compensation it attempts to remove model score scales and shifts caused by varying input channel condition. For example Z-Norm, T-Norm, H-Norm, HT-Norm etc. In speaker modeling domain, the main aim is to modify verification model and to minimize the effects of varying channels, phonemes mismatching effects etc. D-Norm can be categorized into this Speaker Modeling Distance Normalization. Speaker model-domain compensation involves modifying the speaker model parameters instead of the feature vectors [2].

Score normalization techniques have been mainly introduced by Li and Porter in their studies [5].The main purpose of the score normalization techniques is to transform scores from different speakers into a similar range in order to get a common speaker independent verification threshold [2]. The score variability comes from different sources, for example nature of enrollment data can vary among the speakers, mismatching phonetic contents, the duration, the environmental noises as well as the quality of the speaker modeling training techniques [4].For various effects that cause the score variability, it can be reduced by corresponding score normalization technique. For example, T-Norm is used to reduce the divergence between different testing utterances. H-Norm is applied to handle the reduced handset effects in the speech utterance. Z-Norm is an approach applied to normalize the speaker dependent scores to a uniform distribution. In the similar way handset-dependent T-Norm (HT-Norm) has been used for robust speaker verification application. Recently, a speaker-dependent cohort approach has been proposed in the context of test-normalization, namely adaptive Tnorm or ATnorm (Sturim and Reynolds, 2005).

Speaker model distance normalization (D-Norm) is another special approach of score normalization technique which doesn’t need any additional speech data or external speaker population for its implementation in reality which is its advantage. The traditional D-Norm is estimated by the Kullback-Leibler (KL) divergence with the help of Monte-Carlo method, which need lots of time and computation, this being its another disadvantage. So, a modified and simplified version of D-Norm based on the upper bound of the KL divergence between two models (Target speaker model and UBM) has been proposed by Dong Yuan [6], which achieve similar performance gains as traditional, but reducing the time and computation complexity. In this work we have used the simplified and modified version of D-Norm in multilingual speaker verification system.

The rest of the paper is organized as follows: Section 2 describes the GMM-UBM as speaker modeling techniques. Traditional D-Norm as well as principle of modified D-Norm has been explained in Section 3. Experimental Setup and Results are given in Section 4. Section 5 concludes the paper with a Summary.

GMM-UBM AS SPEAKER MODELING TECHNIQUE

Current state-of-the-art approaches for text-independent speaker verification. Gaussian Mixture Model-Universal Background Model (GMM-UBM) has become the basis of the top performing systems in the NIST SREs for better performance and better robustness [2,4,7]. The UBM is a large GMM background model trained to represent the speaker-independent distribution of features which has been collected from the set of imposters. UBM model is generated using the Expectation-Maximization (EM) algorithm. The target speaker model is derived from the UBM using Maximum a Posteriori (MAP) adaptation algorithm with the corresponding training data from the speaker-set. For better performance, only the mean parameter is adapted [7]. In this case all the models are diagonal covariance GMMs.
Normally, SV systems use mel-frequency cepstral coefficients (MFCCs) as a feature vector and the speaker model $\lambda_s$ is parameterized by the set $\{w_i, \mu_i, \Sigma_i\}$ where $w_i$ are the weights, $\mu_i$ are the mean vectors, and $\Sigma_i$ are the covariance matrices. In the testing phase, for a given $T$ sequence of feature vectors $X = \{x_1, x_2, x_3, \ldots, x_T\}$ are extracted from a test signal. A log-likelihood ratio $\Lambda(X)$ is computed by scoring the test feature vectors against the claimant model and the UBM.

$$\Lambda(X) = \log p(X | \lambda_s) - \log p(X | \lambda_{UBM})$$

(1)

Where $\lambda_s$ and $\lambda_{UBM}$ represent models for the hypothesized speaker and imposter respectively.

The claimant speaker is accepted if $\Lambda(X) \geq \theta$ or else rejected. The observations are assumed statistically independent, therefore the log-likelihoods of the observation sequence the hypothesized speaker model and the UBM (imposter) model are given by,

$$\log p(X | \lambda_s) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t | \lambda_s)$$

(2)

$$\log p(X | \lambda_{UBM}) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t | \lambda_{UBM})$$

(3)

The important problem in speaker verification (SV) is to find a decision threshold $\theta$ for the decision making [8,9]. The uncertainty in $\theta$ is mainly due to score variability between the trials.

**D-NORM BASED ON KULLBACK-LEIBLER (KL) DIVERGENCE**

The KL-divergence is commonly used in statistics as a measure of similarity between two density distributions. We can define the KL-divergence for any two given probability density function $f(x)$ and $h(x)$ as follows:

$$D(f \| h) = \int f(x) \log \frac{f(x)}{h(x)} \, dx$$

(4)

KL-divergence is not distance in a strict sense because it usually doesn’t verify the symmetry condition and triangle inequality.

Similarly, we have

$$D(h \| f) = \int h(x) \log \frac{h(x)}{f(x)} \, dx$$

(5)

Hence KL distance between $f(x)$ and $h(x)$ can be computed as

$$KL_d(f \| h) = D(f \| h) + D(h \| f)$$

(6)

Direct computation of KL distance is not possible for most complex statistical distributions of $D(f \| h)$ and $D(h \| f)$, such as GMM or HMM. Instead, the Monte Carlo simulation method is normally employed.

With the expression of the model distance, then D-Norm can be performed by

$$S_{D-Norm} = \frac{LLR}{KL_d(X)}$$

(7)

Where $LLR(X)$ is the log-likelihood ratio of the test data $X$ to target model $g$ and the UBM model $h$ can be expressed as

$$LLR(X) = LLR(X \| g) - LLR(X \| h)$$

(8)
And $KL_{d}(X_d)$ is the symmetrical KL distance corresponding to the speaker $X_d$.

The disadvantage of the traditional D-Norm is that it is time consuming and computational cost. To solve this problem the modified D-Norm has been developed based on the upper bound of KL divergence distance.

**Modified D-Norm**

A simplified approach of D-Norm [6, 10] of estimating the KL divergence based on the upper bound approach, in which authors proved that KL divergence between two GMMs is upper bounded.

$$KL(f||h) \leq KL(w_f||w_h) + \sum_{i=1}^{N} w_i^f KL(N(\cdot, \mu^f_i, \Sigma^f_i)||N(\cdot, \mu^h_i, \Sigma^h_i))$$

(9)

In GMM-UBM speaker modeling system, only mean of speaker model is adapted by MAP. So, in this case we have $w_f = w_h$ and $\Sigma^f_i = \Sigma^h_i$ for $i=1,2,\ldots,M$ Gaussian components.

Thus (9) can be written as

$$KL_{d}(f||h) \leq \sum_{i=1}^{N} w_i (\mu^f_i - \mu^h_i) \Sigma^{-1} (\mu^f_i - \mu^h_i)^T = KL_{UB}(\mu^f, \mu^h)$$

(10)

Where $KL_{UB}(\mu^f, \mu^h) = \sum_{i=1}^{N} w_i (\mu^f_i - \mu^h_i) \Sigma^{-1} (\mu^f_i - \mu^h_i)^T$

(11)

(11) Shows that the upper bound of KL divergence $KL_{UB}$ is actually a weighted version of the Mahalanobios distance between two GMM supervector $\mu^f$ and $\mu^h$, which satisfied the symmetry property.

In this case we can used the upper bound $KL_{UB}$ as the distance measure of two speaker models, then the modified D-Norm can be defined by

$$S_{D-Norm} = \frac{LLR}{KL_{UB}(\mu^f, \mu^h)}$$

(12)

**EXPERIMENTAL SETUP AND RESULTS**

In this section, we are reporting experimental setup of the baseline system of Speaker Verification System using GMM-UBM with D-Norm as back ended classifier and MFCC with Prosodic as in the front-ended feature vectors. Subsection 4.1 gives dataset presentation with speech database. Brief overview of Performance evaluation criteria gives in the Subsection 4.2 and Finally Subsection 4.3 explains the results and conclusions.

**Data-Set Presentation with Speech Database**

In these works, the speaker verification in multilingual experiments were conducted on the multilingual speech corpus namely ALS-DB (Arunchali Linguistic Speech Database) which was described in details in [11,12]. We carried out the experiments using data set of single linguistic group of the Device-2 (Headset Microphone) only, in which three languages speech utterances has been collected namely English, Hindi and Local (Galo) languages.

**Baseline System**

In these works, the baseline system of the speaker verification system was developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. For feature extraction, 14-dimensional MFCC feature vectors were calculated from the silence removed using energy based VAD signal analysis from the speech signal every 10 ms frame rate and a 20 ms frame size with Hamming window. Cepstral features were processed with CMN as well as CVN at feature level normalization to eliminate the signal and channel distortions. Delta as well as acceleration
were then computed over frames span and appended to each feature vectors which resulted in dimensionality 39. The 0th cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame [13]. The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter $H(z)=1-0.97z^{-1}$ has been applied before framing. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 22 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by DCT.

In this experiment, we also combine the prosodic features consisting of three components of formants, one pitch and one energy component with its first and second derivatives with MFCC features. So all together we have 9 dimensional prosodic features adding a total of 9 dimensional prosodic features with the 39 dimensional MFCC features. So, finally we got a 48-dimensional feature vectors. Combination of acoustic features (MFCC) with prosodic features improves the performance of the SV system with comparing to the singleton MFCC features.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with 50 male and 50 female speaker’s data about 5 hours of data from the ALS-DB database. A 512 Gaussian components each male and female model with Expectation Maximization (EM) algorithm. Finally gender-independent UBM model is created by pooling the both male and female models of total 1024 Gaussian components. The speaker models were created for 100 speakers (54 male and 46 female) by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data. In this experiment we have used 1000 true trials and 1500 false trails.

**PERFORMANCE EVALUATION CRITERIA**

Results are presented using detection error tradeoff (DET) plots. Along with equal error rate (EER), the minimum detection cost function (MinDCF) value as defined by NIST [14] that was also used as an overall performance measure.

According to the NIST Detection Cost Function (DCF) can be defined as

$$C_{\text{DET}} = (C_{\text{Miss}} \cdot P_{\text{Miss|Target}} \cdot P_{\text{Target}}) + (C_{\text{FalseAlarm}} \cdot P_{\text{FalseAlarm|NonTarget}} \cdot P_{\text{NonTarget}})$$  \hfill (13)

Where $P_{\text{Miss|Target}}$ and $P_{\text{FalseAlarm|NonTarget}}$ are the miss (false rejection) probability and the false alarm (false acceptance) probability respectively. And

Parameter Values are

- Cost of miss $C_{\text{Miss}} = 10$
- Cost of a false alarm $C_{\text{FalseAlarm}} = 1$
- Probability of a target $P_{\text{Target}} = 0.01$
- Probability of a non-target $P_{\text{NonTarget}} = 1 - P_{\text{Target}} = 0.99$

Minimum DCF (MinDCF), defined as the DCF value at the threshold for which $C_{\text{DET}}$ value is smallest, is the optimum cost.

**RESULTS AND CONCLUSIONS**

**Experiment**

In this experiment single language has been considered for training the system and each language has been considered separately for testing the system. Training sample of length 120 seconds from first two sessions have been
considered for training the system and the other two sessions have been considered for testing the system. Testing sample of length 15 seconds, 30 seconds and 45 seconds have been extracted from the speech sample of length 120 seconds. The result of the experiments has been summarized in table 1. Figure 1 shows the DET curves obtained for the language matching and mismatching conditions separately. In this work, the term matching is referred for the system that trained and tested using same language utterances of the speaker set and mismatching means trained the system with a language but tested with the other language of the same speaker set.

### Table 1: Results for GMM-UBM Baseline with D-Norm for Language Matching and Mismatching Conditions

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>MinDCF (x 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-UBM for language mismatching condition</td>
<td>11.66</td>
<td>2.15</td>
</tr>
<tr>
<td>GMM-UBM+D-Norm for language mismatching condition</td>
<td>10.60</td>
<td>1.97</td>
</tr>
<tr>
<td>GMM-UBM for language matching condition</td>
<td>8.91</td>
<td>1.63</td>
</tr>
<tr>
<td>GMM-UBM+D-Norm for language matching condition</td>
<td>6.06</td>
<td>1.14</td>
</tr>
</tbody>
</table>

![DET Curve for the Speaker Verification System for Language Matching and Mismatching Conditions](image)

**Figure 1:** DET Curve for the Speaker Verification System for Language Matching and Mismatching Conditions

**CONCLUSIONS**

From the experimental point of view we have come to conclude that the performance of the multilingual SV system has been degraded approximately 3.0% due to language mismatching of the training and testing of the system. Also by applying D-Norm, the performance of the baseline system has been improved approximately 2.5% by reducing the EER of 8.91% to 6.06% for language matching condition and 11.66% to 10.60% for language mismatching conditions respectively. Finally, it is also observed that the performance of SV system after applying D-Norm is better in language matching condition than that of mismatching condition.

**ACKNOWLEDGEMENTS**

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.
REFERENCES
