Score Normalization for VQ-UBM Based Text-Independent Speaker Verification

Cemal Hanilci[1], and Figen Ertas[1]

[1]Department of Electronic Engineering, Uludag University, Bursa, Turkey
chanilci@uludag.edu.tr, fertas@uludag.edu.tr

Abstract

This paper presents score normalization for recently proposed modeling technique, vector quantization universal background model (VQ-UBM) based speaker verification of cellular data. Test-normalization (TNorm) which is the most widely used score normalization technique, is evaluated for VQ-UBM based speaker verification. Experimental results using NIST 2002 Speaker Recognition Evaluation (SRE) (one-speaker detection task) show that score normalization improves the verification performance and VQ-UBM provides better recognition accuracy than support vector machines generalized linear discri minant sequence kernel (SVM-GLDS), which is one of the state-of-the-art modeling techniques for speaker verification, in terms of both, Equal Error Rate (EER) and Minimun Detection Cost Function (MinDCF).

1. Introduction

Speaker recognition aims to find the identity of a speaker given a speech sample. It is a challenging task with many variable factors which affect the system performance such as transmission channel, length of train/test utterance, session variability, etc. these variabilities can be speaker-dependent, in which the variability comes from the speaker and test-dependent, when the variability comes from the test data [1]. These variabilities have a negative impact on the system performance and compensation techniques can be used to reduce these negative effects.

In the literature different compensation techniques at three different levels have been proposed. Compensation can be in feature level, score level and model level. Feature level methods aim at removing the negative effects of transmission channels from the feature vectors. The most widely used techniques of feature level compensations are cepstral mean subtraction (CMS) [2] and RASTA filtering [3]. CMS is used for compensation of linear channel variations. RASTA is a filtering approach which aims to reduce convolutional noise which exists in the transmission channel. Another feature level compensation technique was proposed by Reynolds which maps feature vectors into a channel independent space to mitigate channel effects [4]. Score level techniques aim to remove score scales and shifts caused by varying transmission channel conditions. These methods have been proposed to deal with score variability and to select more robust and effective speaker-independent threshold. The well-known score domain techniques are Hnorm [5], Tnorm and Znorm [6]. Model level compensation attempts to minimize the effects of varying channels by modifying the verification models such as Speaker Model Synthesis [7].

State-of-the-art speaker recognition systems use Maximum a *Posteriori* (MAP) adapted modeling techniques. MAP adapted Gaussian mixture models with universal background model (GMM-UBM) has been the most popular modeling algorithm [5]. In GMM-UBM method a background model is created from a set of background speakers which are not included in training speakers via conventional expectation maximization (EM) algorithm and then speaker models are constructed from the background model by MAP adaptation. Recently Hautamaki et. al. introduced MAP adaptation of vector quantization (VQ-MAP or VQ-UBM) which adapts the centroid vectors via MAP algorithm and it has been reported that VQ-UBM algorithm provides recognition accuracy as good as GMM-UBM with a significant speed-up [8], [9].

In this paper we focus on score level compensation techniques for VQ-UBM based speaker recognition system. To the best of our knowledge there is no previous study which investigates the score level compensation techniques on this new modeling algorithm. Well known TNorm method is evaluated on VQ-UBM based speaker recognition and compared with another well-known and popular modeling technique, support vector machines-generalized linear discriminant sequence kernel (SVM-GLDS) [10].

The organization of our paper is as follows. In section 2, VQ-UBM method is described. SVM-GLDS method is briefly explained in section 3. Section 4 describes the application of TNorm technique to VQ-UBM and SVM-GLDS. In section 5, speaker verification setup is described and the experimental results are given in section 6. Finally, the conclusions are summarized in section 7.

2. VQ-UBM

The general structure of VQ-UBM based speaker recognition system is shown in Figure 1. As seen from the figure VQ-UBM algorithm first creates a background model (UBM) from the background data which consists of several speech samples from different speakers taken from a database not used in the training/test set of speaker recognition database. UBM model is created using conventional clustering algorithm such as *K*-means. Obviously, the UBM model is a codebook which consists of M centroids (M is called model order of codebook size), $U = \{u_1, u_2, ..., u_M\}$, generated by clustering algorithm. After UBM model is created the speaker is trained using features computed from training speech sample of speaker and UBM model by using MAP adaptation. Output of the training step of the algorithm is the codebook of speaker with M centroids, $C = \{c_1, c_2, ..., c_M\}$, which is defined as *speaker model*. VQ-

UBM algorithm iteratively adapts the centroid vectors, c_i, using UBM centroids, u_j, by MAP algorithm.

In the recognition phase, feature vectors computed from the speech sample of unknown speaker are scored against UBM model, U, and claimed speaker model, C, and if the match score above a pre-defined threshold identity claim is accepted by the system or if the score is below the threshold unknown speaker is rejected. The match score between unknown feature set, $X = \{x_1, x_2, ..., x_T\}$ and UBM model, U, and speaker model, C, is computed by the following formula:

$$Score = MSE(X,U) - MSE(X,C) \quad (1)$$

where, *MSE* is the Mean Square Error and computed as:

$$MSE(X,Y) = \frac{1}{|X|} \sum_{x_i \in X} \min_{y_k \in Y} \|x_i - y_k\|^2 \quad (2)$$

where $|X|$ denotes the number of vectors in the set of X, and $\|x_i - y_k\|^2$ is the squared Euclidean distance between the vectors, x_i and y_k. For more details about the algorithmic description of VQ-UBM method readers are referred to [8], [9].

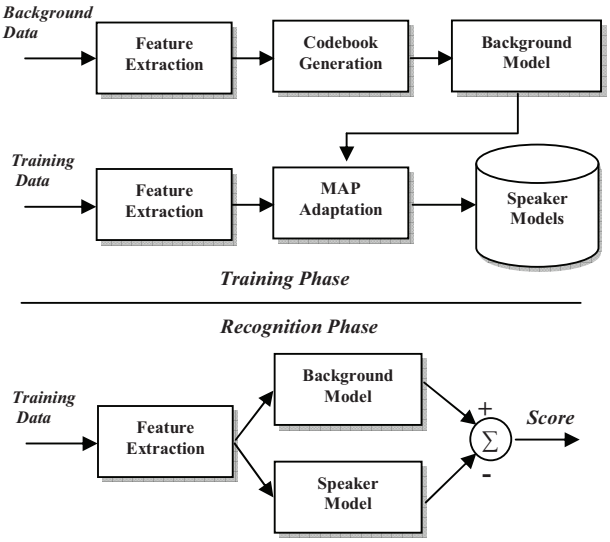

Fig. 1. General structure of VQ-UBM based speaker recognition system

3. Support Vector Machines

Support vector machines (SVM) have become a powerful technique for pattern classification applications. It has gained so much popularity in recent years. SVM is a discriminative classifier which models the boundary between two classes. Since speaker verification is also a two-class problem, SVM has proven to be one of the most powerful techniques in speaker recognition. The most important problem for speech applications using SVM is the amount of training and test data. Since speech is not a static signal, the features are extracted from shifted short-time frames (typical frame duration is 20-30 ms with 10-15 ms frame shift), and this yields a set of feature vectors not a single vector for a speech sample. This results in long training times for SVM approach. There are several studies in literature which focus on this problem [11], [12]. In recent years, use of sequence kernels for application of SVM in speech processing has gained more attention. Sequence kernel basically maps the set of training/test feature vectors into a single characteristic vector and trains the SVM by using this vector and scoring is a simple dot product. One of the most powerful sequence kernel method for speaker and language recognition is the generalized linear discriminant kernel (GLDS) [10]. The GLDS method creates a single vector (also known as supervector) by mapping the set of feature vectors into kernel space using a polynomial expansion [13]. As an example, second-order polynomial expansion of a 2-dimensional vector $x = [x_1, x_2]^t$ is given by $b(x) = [1, x_1, x_2, x_1^2, x_1 x_2, x_2^2]^t$. During training, all the background and speaker features, $X = \{x_1, x_2, ..., x_T\}$, are represented by average expanded feature vectors, which is computed by:

$$b(X) = \frac{1}{T} \sum_{t=1}^{T} b(x_t) \quad (3)$$

The averaged vectors of background data and training speaker data are assigned with the appropriate label for SVM training (+1 for target speaker vectors and -1 for background vectors). The output of the SVM training is a set of support vectors, b_i, their weights, α_i, and a bias term d. These are collapsed into a single model vector by:

$$w = \sum_{i=1}^{N_{sv}} \alpha_i t_i b_i + d \quad (4)$$

where N_{sv} and $t_i \in \{+1, -1\}$ are the the number of support vectors and ideal outputs, respectively.

During the recognition phase of SVM-GLDS, the match score in the GLDS method is computed as an inner product:

$$Score = w_{target}^t b_{test} \quad (5)$$

where w_{target} and b_{test} are the model vector of the target speaker and averaged expanded feature vector of the test sample, respectively. Since all speaker models, train and test data are represented by single vectors, the recognition phase is computationally efficient. The details about the SVM-GLDS can be found in [10], [13].

4. Application of TNorm

The last step in a speaker verification system is the decision making. This is done by comparing the match score with a threshold. If the score is higher than the threshold, identity claim is accepted, otherwise rejected. The score normalization techniques aim to tune decision thresholds. It is known that the scores of speaker verification trials vary a lot. The variability of scores come from different sources such as phonetic content of

speech sample, channel noise, duration of the speech sample, speaker's health, emotion, age, etc. [14]. To avoid these problems score normalization is applied to make it easier to tune speaker-independent decision threshold.

General structure of TNorm method is shown in Figure 2. TNorm works as follows: given a set of feature vectors $X = \{x_1, x_2, ..., x_T\}$ extracted from a test speech sample, and a speaker model λ (C for VQ-UBM and w for SVM-GLDS), a match score, $S(X, \lambda)$, between X and λ is computed. TNorm uses a set of cohort models (also known as impostor models) $\Phi = \{\theta_1, \theta_2, ..., \theta_N\}$ to obtain impostor scores $S_{TNorm} = \{S(X, \theta_1), S(X, \theta_2), ..., S(X, \theta_N)\}$. Then the speaker verification scores are normalized by:

$$S_{TNorm}(X, \lambda) = \frac{S(X, \lambda) - \mu_{TNorm}}{\sigma_{TNorm}} \quad (6)$$

where μ_{TNorm} and σ_{TNorm} are the mean and standard deviation of the impostor scores S_{TNorm}. Obviously, TNorm assumes that scores have a Gaussian distribution and centers the score distribution. More details about the TNorm can be found in [6].

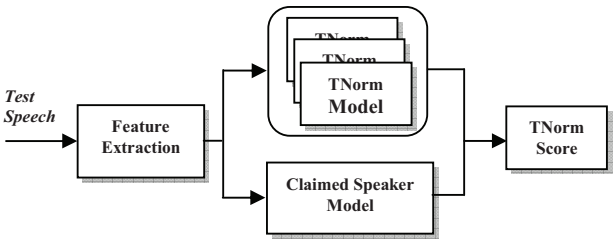

Fig. 2. TNorm score normalization method

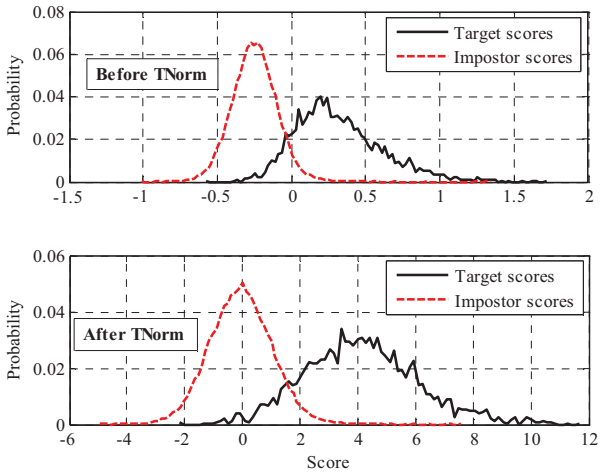

Fig. 3. Effect of TNorm on score distributions.

Figure 3 shows the effect of TNorm on score distribution of a VQ-UBM based speaker recognition system. It can be seen from figure that after TNorm the impostor score distribution has zero mean and unit standard deviation. The intersection of target and impostor score distributions causes the wrong decisions and after TNorm there are less number of scores in this region so the performance is improved. In our application we select a set of speakers which doesn't exist in the background database and speaker recognition database for TNorm models. For the VQ-UBM codebook and for the SVM case the model vector, w for each TNorm speaker corresponds to a TNorm model.

5. Speaker Verification Setup

We use NIST 2002 speaker recognition evaluation (SRE) database for the speaker verification experiments [15]. Database contains speech samples transmitted over different cellular networks. NIST 2002 contains 139 male and 191 female speakers and consists of 2982 target and 36277 impostor trials. For each speaker two minutes of speech data is available for training. The duration of test files varies between 15 seconds to 45 seconds. The gender-dependent background models and TNorm models are created using NIST 2001 SRE.

We use standard MFCC features as speaker-specific features. MFCC extraction algorithm is as follows: discrete Fourier transform (DFT) magnitude spectrum of 30 ms hamming windowed frames is computed in every 15 ms. Magnitude spectrum is smoothed by a 27-channel triangular filterbank. The logharithmic filterbank outputs are converted into MFCCs by discrete Cosine transform (DCT). We use first 12 MFCCs. The first and second order derivatives (also known as delta and double deltas) of the MFCC features are appended to original features which yields 36-dimensional feature vectors. The last step is the cepstral mean and variance normalization (CMVN).

We use two standard metrics as the speaker recognition performance criteria: equal error rate (EER) and minimum detection cost function (MinDCF). EER corresponds to threshold at which the miss rate (P_{miss}) and false alarm rate (P_{fa}) are equal; MinDCF is the minimum value of a weighted cost function given by $0.1 \times P_{miss} + 0.99 \times P_{fa}$. All reported MinDCF values are multiplied by *100* for ease of comparison. These two metrics are the most well-known criteria in speaker recognition community. Additionally, a few selected detection error tradeoff (DET) curves are plotted in order to see the full behavior of the false alarms and miss detections of the proposed systems. For VQ-UBM based system we obtain the results for three different model orders (codebook size), $M \in \{64,512,1024\}$ and for SVM-GLDS we use two different polynomial expansion orders, $m = 2$ and $m = 3$.

6. Results

The speaker verification results for different modeling techniques (VQ-UBM and SVM-GLDS) and with and without TNorm are shown in Table 1. Figure 4 and Figure 5 show DET plots of VQ-UBM (M=512) and SVM-GLDS (m=3) systems and compares the performances with and without TNorm, respectively.

Examining the Table 1, Figures 4 and 5, the following observations are made: VQ-UBM outperforms the SVM-GLDS methods in terms of both, EER and MinDCF. Best speaker recognition performance is obtained for M=512 for VQ-UBM and m=3 for SVM-GLDS. TNorm technique improves the speaker recognition performance for both methods, VQ-UBM and SVM-GLDS. TNorm has a big effect on MinDCF (Table I). It can be seen that from Figures 4 and 5 TNorm reduces the miss

probability errors for a fixed false alarm rate, significantly (left top corners of DET plots).

Table 1. Speaker recognition performance

Method	Without TNorm		With TNorm	
	EER (%)	MinDCF	EER (%)	MinDCF
VQ-UBM (M=64)	9.27	4.28	8.91	3.49
VQ-UBM (M=512)	7.98	3.86	7.69	3.13
VQ-UBM (M=1024)	8.25	3.94	8.26	3.31
SVM-GLDS (m=2)	12.97	6.35	12.45	5.22
SVM-GLDS (m=3)	9.69	4.36	10.09	3.73

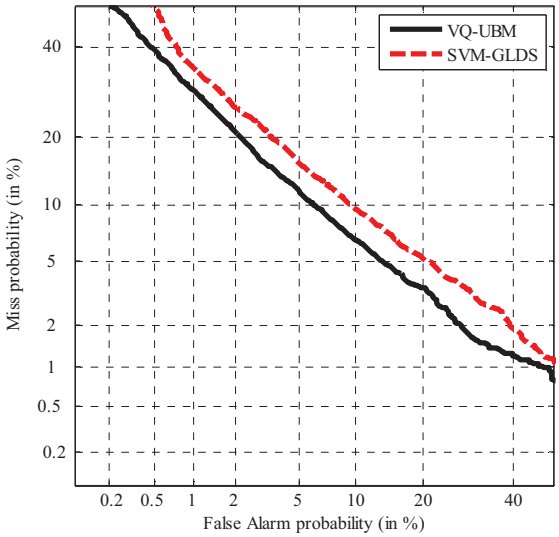

Fig. 4. Comparison of VQ-UBM and SVM-GLDS without TNorm.

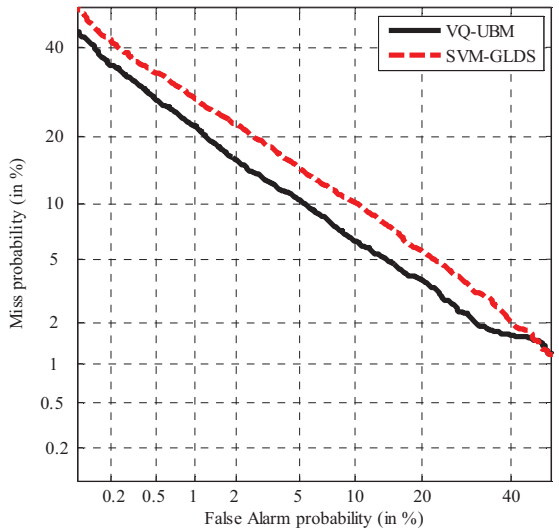

Fig. 5. Comparison of VQ-UBM and SVM-GLDS with TNorm.

.

7. Conclusions

In this paper we evaluated a well-known score normalization technique, TNorm, on VQ-UBM based speaker recognition system and compared the system performance with a state-of the-art modeling method, SVM-GLDS. Our results indicate that TNorm, independent of modeling method (VQ-UBM or SVM-GLDS), model order (codebook size for VQ-UBM, and polynomial expansion order for SVM-GLDS), increase the speaker recognition performances in terms of EER and MinDCF values. In conclusion, TNorm is a simple and easy to implement method and needs to be considered in speaker recognition applications.

8. References

[1] D. R. Castro, J. F. Aguilar, J. G. Rodriguez, and J. O. Garcia, "Speaker verification using speaker-and test-dependent fast score normalization", *Pattern Recognition Letters*, vol. *28,* no. *1,* pp. *90-98,* Jan., 2007.

[2] A. A. Garcia, and R. J. Mammone, "Channel robust speaker identification using modified-mean cepstral mean normalization with frequency warping", in *International Conference on Acoustics, Speech and Signal Processing (ICASSP 99)*, Phoenix, AZ., USA, 1999, pp. *325-328.*

[3] H. Hermansky, and N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech and Audio Proc.*, vol. *2,* no. 4, pp. *578-589,* Oct., 1994.

[4] D. A. Reynolds, "Channel robust speaker verification via feature mapping", in *International Conference on Acoustics, Speech and Signal Processing (ICASSP 03)*, Hong Kong., China, 2003, pp. *53-56.*

[5] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Proc.*, vol.*10,* no. 1-3, pp. *19-41,* Jan., 2000.

[6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Proc.*, vol. *10,* no. 1-3, pp. *42-54,* Jan., 2000.

[7] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition", in *International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000, pp. *495-498.*

[8] V. Hautamaki, T. Kinnunen, I. Karkkainen, M. Tuonen, J. Saastamoinen, and P. Franti, "Maximum a Posteriori estimation of the centroid model for speaker verification", *IEEE Signal Proc. Letters*, vol.*15,* no. 1-3, pp. *162-165,* Jan., 2008.

[9] T. Kinnunen, J. Saastamoinen, V. Hautamaki, M. Vinni, and P. Franti, "Comparative evaluation of maximum a Posteriori vector quantization and Gaussian mixture models in speaker verification", *Pattern Rec. Letters*, vol.*30,* no. 4, pp. *341-347,* March, 2009.

[10] W. Campbell, J. Campbell, D.A. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", *Comp. Speech and Language*, vol.*20,* no. 2-3, pp. *210-229,* April, 2006.

[11] Z. Lei, Y. Yang, and Z. Wu, "Mixture of support vector machines for text-independent speaker recognition", in *Interspeech*, Lisbon, Portugal, 2005, pp. *2041-2044.*

[12] V. Wan, and S. Renals, "SVMSVM:Support vector machine speaker verification methodology", in *International Conference on Acoustics, Speech and Signal*

Processing (ICASSP 03), Hong Kong, China, 2003, pp. *221-224.*

[13] W. Campbell, K. Assaleh, and C. Broun, "Speaker recognition with polynomial classifiers", *IEEE Trans. Speech and Audio Proc*., vol. *10*, no. 4, pp. *205-212,* May, 2002.

[14] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification", *Eurasip Journal of Applied Signal Proc.*, vol. *4*, no. 4, pp. *430-451,* 2004.

[15] NIST 2002 Speaker Recognition Evaluation, webpage, http://www.itl.nist.gov/iad/mig/tests/spk/2002/