Studies on Model Distance Normalization Approach in Text-independent Speaker Verification

DONG Yuan[1] LU Liang[1] ZHAO Xian-Yu[2] ZHAO Jian[1]

Abstract Model distance normalization (D-Norm) is one of the useful score normalization approaches in automatic speaker verification (ASV) systems. The main advantage of D-Norm lies in that it does not need any additional speech data or external speaker population, as opposed to the other state-of-the-art score normalization approaches. But still, it has some drawbacks, e.g., the Monte-Carlo based Kullback-Leibler distance estimation approach in the conventional D-Norm approach is a time consuming and computation costly task. In this paper, D-Norm was investigated and its principles were explored from a perspective different from the original one. In addition, this paper also proposed a simplified approach to perform D-Norm, which used the upper bound of the KL divergence between two statistical speaker models as the measure of model distance. Experiments on NIST 2006 SRE corpus showed that the simplified approach of D-Norm achieves similar system performance as the conventional one while the computational complexity is greatly reduced.

Key words Gaussian mixture model (GMM), Kullback-Leibler distance, model distance normalization, speaker recognition, speaker verification

Speaker verification is a process of determining whether an utterance is spoken by a claimant or not. For this task, the Gaussian mixture model (GMM)-universal background mode (UBM)[1] framework has become the dominant approach over the past decade and achieves state-of-the-art system performance, whereas support vector machine (SVM) also has been proved to be an effective method for speaker recognition in recent years[2−4]. In classical GMM-UBMs, a speaker model is adapted by maximum a posteriori (MAP) from universal background model (UBM), and in the test phase, decision making is performed by the log-likelihood ratio (LLR) detection, in which the LLR is compared with the pre-established global threshold to determine whether to accept the claimed identity or not. However, due to the score variability caused by various factors (environment noise, mismatch between testing and training data, inter- and intra-speaker divergence, etc), an appropriate speaker independent global threshold is usually difficult to set for decision making.

To cope with this problem, various compensation techniques for channel effects and mismatch between training and testing data have been proposed, which can be generally divided into three categories, namely, feature domain, channel domain, and score domain compensation. Feature domain compensation is aimed at removing the channel effects from the feature vectors prior to model training or verification, e.g., cepstral mean subtraction (CMS), RelAtive SpecTral (RASTA), etc. In model domain, the aim is to modify verification models to minimize the effects of varying channels, e.g., speaker model synthesis (SMS), eigenchannels[5] for GMM-UBM based system or nuisance attribute projection (NAP)[6] for SVM based system, etc. And finally, score domain compensation attempts to remove model score scales and shifts caused by varying input channel conditions. Examples of this includes widely used score normalization approaches such as H-Norm[7], T-Norm[8], Z-Norm[9], etc. For different effects that cause the score variability can be reduced by corresponding score normalization approach, for example, H-Norm is proposed to handle the variability caused by different handset types, and T-Norm is mean to reduce the divergence between different testing utterances. Z-Norm is an approach to normalize the speaker dependent scores to a uniform distribution. Score normalization is an effective approach for speaker verification and can improve the system performance significantly in most cases. However, the fatal drawback of the above score normalization approaches is that they bring heavy computational burden to the verification system and needs large amount of additional development speech data, which, in some cases, is difficult to obtain.

Another interesting score normalization approach, namely, model distance normalization (D-Norm)[10], is a special one compared with those above, which does not need any additional speech data or external speaker population, but still can achieve promising system performance gains in most cases. This advantage of D-Norm makes it more practical in reality. In original D-Norm, the model distance is estimated by the Kullback-Leibler (KL) divergence through a Monte-Carlo method, which, however, needs much additional time and computation. In this paper, D-Norm was re-investigated systematically and its principles were presented in another perspective, which is more logically reasonable. In addition, a simplified approach to perform D-Norm was also proposed, which used the upper bound of the KL divergence between two models as the measure of model distance. Experiments on NIST 2006 SRE corpus showed that the simplified approach of D-Norm could achieve similar system performance gains as the traditional one while the computational complexity was greatly reduced.

The rest of the paper is organized as follows: in Section 1, some preliminary knowledge is presented, and in Section 2, principles of model distance normalization are discussed and a new simplification approach for performing D-Norm is proposed. Some experimental results on 2006 NIST SRE corpus are given in Section 3. Section 4 concludes the paper with a summary.

1 Preliminaries

1.1 GMM-UBM

GMM-UBM is the predominant approach used in speaker recognition systems, particularly for text-independent task[1]. According to this approach, a UBM is trained using the EM (Expectation-maximization) algorithm on a larger quantity of exclusive speech, and the

Received March 24, 2008; in revised form October 12, 2008
Supported by the Key Project of the Ministry of Education of China (108012)
1. Department of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China
2. France Telecom Research and Development Center (Beijing), Beijing 100080, P. R. China
DOI: 10.3724/SP.J.1004.2009.00556

target speaker model is adapted from the gender specific UBM using MAP estimation. For a D-dimensional feature vector $\boldsymbol{x}$, the probability density of $\boldsymbol{x}$ given a speaker model λ, which has M Gaussian mixtures is defined as

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\boldsymbol{x}) \tag{1}$$

The density is a weighted linear combination of M unimodal Gaussian densities $p_i(\boldsymbol{x})$, which is parameterized by a mean vector $\boldsymbol{\mu}_i$ and a covariance matrix Σ_i. The speaker model can be characterized by $\lambda = (w_i, \boldsymbol{\mu}_i, \Sigma_i), i = 1, \cdots, M$.

Given a GMM model λ, the average log-likelihood of the test utterance $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ is computed by

$$LL(X|\lambda) = \frac{1}{T} \sum_{t} \log p(\boldsymbol{x}_t|\lambda) \tag{2}$$

And the log-likelihood ratio (LLR) used for detection is defined as

$$LLR(X) = LL(X|\lambda_{spk}) - LL(X|\lambda_{ubm}) \tag{3}$$

where λ_{ubm} is the UBM model and λ_{spk} indicates the speaker model. $LLR(X)$ is finally compared with the predefined threshold θ to decide whether the utterance presented is from the claimant or not.

1.2 Kullback-Leibler (KL) divergence

The KL-divergence[11], also known as the relative entropy in the information theory, is commonly used in statistics as a measure of similarity between two density distributions. For two probability density functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, the KL-divergence is defined as

$$D(f\|g) = \int f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \tag{4}$$

KL-divergence is not distance in a strict sense because it usually does not verify the symmetry condition and triangle inequality. However, the divergence satisfies the following three properties:

1) Self-similarity: $D(f\|f) = 0$;
2) Self-identification: $D(f\|g) = 0$ only if $f = g$;
3) Positivity: $D(f\|g) \geq 0$ for all f, g.

The KL divergence is used in many aspects of speech and image recognition as a kind of similarity or distance measurement. For two d-dimensional Gaussians, the KL divergence has a closed form expression:

$$D(N(\cdot;\boldsymbol{\mu}, C)\|N(\cdot;\tilde{\boldsymbol{\mu}}, \tilde{C})) = \frac{1}{2}\left[\log\frac{\det\tilde{C}}{\det C} - d + \mathrm{tr}(\tilde{C}^{-1}C) + (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}}\tilde{C}^{-1}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\right] \tag{5}$$

whereas for two GMMs, no such closed form expression exists.

2 Model distance normalization

2.1 Principles of model distance normalization

As discussed in the introduction, the purpose of score normalization is to alleviate the variability caused by numerous reasons, and currently, most normalization approaches are achieved by rescaling the impostor score distribution of each speaker to a normal distribution (zero mean and unit variance), just as

$$LLR_{Norm} = \frac{LLR(X) - \mu_{Norm}}{\sigma_{Norm}} \tag{6}$$

where μ_{Norm} and σ_{Norm} are the mean and standard deviation of the impostor scores estimated through statistical approach, respectively by different estimation approaches of the two parameters μ_{Norm} and σ_{Norm} , corresponding score normalization approaches will be obtained, e.g. Z-Norm, T-Norm, etc.

Reference [10] proposed the model distance normalization (D-Norm) based on some experimental observations, in which the impostor score had a correlation with $KL(\lambda_{spk}\|\lambda_{ubm})$ (It is the KL divergence between two GMMs defined as (4)). However, in this paper, it is proposed based on the following proposition (before being validated, it is only a hypothesis): the distance between the speaker model λ_{spk} and background model λ_{ubm} also causes the variability of the final log-likelihood score. More concretely, the final log-likelihood score $LLR(X)$ in (3) will also depend on the value of $D(\lambda_{spk}, \lambda_{ubm})$, at least to some extent (for many other factors it will also affect the final score), in which $D(\cdot,\cdot)$ is a kind of distance measure between two GMMs. Moreover, it is believed that a small value of $D(\lambda_{spk}, \lambda_{ubm})$, which means the speaker model is more similar to the UBM, more likely results in a small range of LLR for a set of test utterances. Contrarily, a large distance between λ_{spk} and λ_{ubm} will also tend to enlarge the range of LLR. Thus, if the above proposition is valid, then it is straightforward that the LLR can be rescaled by the model distance $D(\lambda_{spk}, \lambda_{ubm})$ in order to eliminate the variability caused by this kind of distance divergence, which is what D-Norm exactly does.

Thus, before the introduction of the technical approach about D-Norm, it is necessary to validate the presupposition discussed above. In addition, an appropriate distance measure approach is also needed to estimate the value of $D(\lambda_{spk}, \lambda_{ubm})$. Since λ_{spk} and λ_{ubm} are two statistical models, just as the conventional approach, the commonly used Kullback-Leibler (KL) divergence, was used in this paper to estimate $D(\lambda_{spk}, \lambda_{ubm})$. In the following subsections, we will first demonstrate the reasonability of D-Norm approach (theoretically and by some experiments) and then introduce the technical approach of performing D-Norm.

2.2 Relationship between *LLR* and KL divergence

To explain the principles of model distance normalization, it is meaningful to investigate some relationship between the KL divergence and the log-likelihood ratio. For Gaussian mixtures, a closed form expression for KL divergence only exists when the number of Gaussian mixtures is 1. Thus, we will only use one Gaussian mixture to perform our deduction in the next subsection. And for simplicity, the formulas are based on mono-dimensional input data, as the extension to multi-dimensional data is straightforward.

Let $g = N(\cdot;\mu_g, \Sigma_g)$ be the Gaussian distribution for the target, and $f = N(\cdot;\mu_f, \Sigma_f)$ be the Gaussian distribution of the UBM. Give a test utterance, $X = o_1, \cdots, o_n$, which is also supposed to obey a Gaussian distribution $h = N(\cdot;\mu_h, \Sigma_h)$. The expected log-likelihood $LL(X|g)$ of X given the Gaussian distribution g is obtained as

$$LL(X|g) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{(2\pi\Sigma_g)^{\frac{1}{2}}} \exp(-\frac{(o_i - \mu_g)^2}{2\Sigma_g})\right]$$

By developing the square term, it will be of the following form:

$$LL(X|g) = -\frac{1}{2}\Bigg[\log(2\pi\Sigma_g) + \frac{1}{\Sigma_g}\Bigg(\frac{1}{n}\sum_{i=1}^{n} o_i^2 -$$

$$2\mu_g\frac{1}{n}\sum_{i=1}^{n}o_i+\mu_g^2\Big)\Big]=-\frac{1}{2}\Big[\log(2\pi\Sigma_g)+\frac{1}{\Sigma_g}(\Sigma_h+\mu_h^2-2\mu_h\mu_g+\mu_g^2)\Big]=-\frac{1}{2}\left[\log(2\pi\Sigma_g)+\frac{\Sigma_h}{\Sigma_g}+\frac{(\mu_h-\mu_g)^2}{\Sigma_g}\right] \quad (7)$$

Similarly, for the background model f, the average log-likelihood will be

$$LL(X|f)=-\frac{1}{2}\left[\log(2\pi\Sigma_f)+\frac{\Sigma_h}{\Sigma_f}+\frac{(\mu_h-\mu_f)^2}{\Sigma_f}\right] \quad (8)$$

Thus, the log-likelihood ratio of the test data X to target model g and background model f can be expressed as

$$LLR(X)=LL(X|g)-LL(X|f)=\frac{1}{2}\left[\log\left(\frac{\Sigma_f}{\Sigma_g}\right)+\frac{\Sigma_h}{\Sigma_f}-\frac{\Sigma_h}{\Sigma_g}+\frac{(\mu_h-\mu_f)^2}{\Sigma_f}-\frac{(\mu_h-\mu_g)^2}{\Sigma_g}\right] \quad (9)$$

Note that for two Gaussians, the KL divergence has the closed form expression in the form of (5). Having a little change in (9), we will get the following expression:

$$\begin{aligned}LLR(X)=&\frac{1}{2}\left[\log\left(\frac{\Sigma_f}{\Sigma_h}\frac{\Sigma_h}{\Sigma_g}\right)+1-1+\frac{\Sigma_h}{\Sigma_f}-\frac{\Sigma_h}{\Sigma_g}+\frac{(\mu_h-\mu_f)^2}{\Sigma_f}-\frac{(\mu_h-\mu_g)^2}{\Sigma_g}\right]=\\&\frac{1}{2}\left[\log\left(\frac{\Sigma_f}{\Sigma_h}\right)-1+\frac{\Sigma_h}{\Sigma_g}+\frac{(\mu_h-\mu_f)^2}{\Sigma_f}\right]-\\&\frac{1}{2}\left[\log\left(\frac{\Sigma_g}{\Sigma_h}\right)-1+\frac{\Sigma_h}{\Sigma_f}+\frac{(\mu_h-\mu_f)^2}{\Sigma_g}\right]=\\&D(h\|f)-D(h\|g)\end{aligned} \quad (10)$$

Unfortunately, since the KL divergence does not satisfy the property of triangle inequality, we cannot further deduct the following inequality from (10)

$$LLR(X)=D(h\|f)-D(h\|g)\le D(f\|g)$$

Moreover, the KL divergence of two GMMs is more complicated than that of two Gaussian distributions, and does not have such closed form expression, but from the deduction above, we can consider that the range which will be affected by the KL divergence between two GMMs is a reasonable hypothesis. In the next subsection, we will present the D-Norm approach, and through the experimental results further demonstrate the validation of the proposition.

2.3 D-Norm approach description

D-Norm was first proposed by Ben[10], in which the model distance was described by a simply symmetrized version of KL divergence between two statistical models, just as the following expression shows:

$$KL2(p_a\|p_b)=KL(p_a\|p_b)+KL(p_b\|p_a) \quad (11)$$

where p_a and p_b was two probabilistic models corresponding to two speakers, respectively. And $KL(\cdot\|\cdot)$ denotes the KL divergence as (4). Obviously, $KL2$ satisfies the symmetry condition, but not the triangle inequality. With the expression of the model distance, then D-Norm can be performed by

$$LLR_{D-Norm}=\frac{LLR}{KL2(X_l)} \quad (12)$$

where $KL2(X_l)$ is the symmetrized KL distance corresponding to the speaker X_l. The D-Norm approach looks tidy in expression, and it also has the advantage as discussed above, that is, it does not need the additional development speaker data or external speaker population. However, to estimate the value of $KL2(X_l)$ is not an easy task, since the KL divergence does not have a closed form expression for two GMMs. Reference [10] gave a Monte-Carlo method to estimate the KL distances by synthesizing data that follow the statistical laws of the speaker and UBM model, which is described as

$$D_{MC}(f\|g)=\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(\boldsymbol{x}_i)}{g(\boldsymbol{x}_i)}\to D(f\|g)$$

where $\{\boldsymbol{x}_i\}_{i=1}^n$ are n samples drawn from the prabability density function (PDF) f. As $n\to\infty$, the Monte-Carlo method can yield convergence, and it also satisfies the similarity property discussed in Section 1. But it has the drawback that it is time consuming and computational costly for a large amount of synthesized data is needed. Additionally, it cannot hold the property of positivity absolutely, especially when the amount of synthesized data is small. This is a relatively serious problem in practice. In the following subsection, we will introduce another approach to estimate the value of the KL divergence, which avoids the above problems.

2.4 D-Norm based on the upper bounds of KL

As for the drawbacks of the Monte-Carlo method based estimation of the KL distance, in this subsection, a simplified method of estimating the KL divergence, namely, the upper bound approach, is introduced as an alternative. It is based on the work in [12], in which the authors proved that the KL divergence between two GMMs is upper bounded by

$$KL(p_a\|p_b)\le KL(w_a\|w_b)+\sum_{i=1}^{N}w_i^aKL(N(\cdot;\boldsymbol{\mu}_i^a,\Sigma_i^a)\|N(\cdot;\boldsymbol{\mu}_i^b,\Sigma_i^b)) \quad (13)$$

In most of classical GMM-UBM systems, only means of the speaker model are adapted by MAP, and in this case, there will be $w^a=w^b$ and $\Sigma_i^a=\Sigma_i^b, i=1,\cdots,N$. Thus, (13) can be rewritten as

$$KL2(p_a\|p_b)l\le\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_i^a-\boldsymbol{\mu}_i^b)\Sigma_i^{-1}(\boldsymbol{\mu}_i^a-\boldsymbol{\mu}_i^b)^{\mathrm{T}}=D_{UB}^2(\boldsymbol{\mu}^a,\boldsymbol{\mu}^b) \quad (14)$$

where

$$D_{UB}^2(\boldsymbol{\mu}^a,\boldsymbol{\mu}^b)=\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_i^a-\boldsymbol{\mu}_i^b)\Sigma_i^{-1}(\boldsymbol{\mu}_i^a-\boldsymbol{\mu}_i^b)^{\mathrm{T}} \quad (15)$$

In the case of state-of-the-art speaker verification systems, the covariance matrices used are diagonal. If we hypothesize that $\Sigma_i=\mathrm{diag}\{\sigma_{i1},\cdots,\sigma_{id}\}$, then (15) can be expressed as

$$D_{UB}^2(\boldsymbol{\mu}^a,\boldsymbol{\mu}^b)=\sum_{i=1}^{N}w_i\sum_{j=1}^{d}\frac{(\mu_{ij}^a-\mu_{ij}^b)^2}{\sigma_{ij}^2} \quad (16)$$

(16) shows that the upper bound of KL divergence D_{UB}^2 is actually a weighted version of the Mahalanobios distance between two GMM supervectors $\boldsymbol{\mu}^a$ and $\boldsymbol{\mu}^b$. It is obvious that comparing with the Monte-Carlo based KL divergence estimation approach, D_{UB}^2 satisfies the symmetry property

as well as those showed in Subsection 1.2, namely, the self-similarity, self-identification, and positivity. Moreover, the computational cost of D^2_{UB} is relatively small. In this paper, we will not discuss and compare the two methods in depth, for more information about this, please refer to [13].

In this case, if we use the upper bond D^2_{UB} as the distance measure of two speaker models, then the D-Norm can be performed by

$$LLR_{D-Norm} = \frac{LLR}{D^2_{UB}(\boldsymbol{\mu}^{spk}, \boldsymbol{\mu}^{UBM})} \tag{17}$$

In the following section, some experiments are performed to examine the performance of the D-Norm approach discussed above, and a comparison with the conventional one will also be presented. In addition, the proposition for D-Norm is also investigated from some experiment observations, and the result is very convincing.

3 Experimental results

In this section, we will report experimental results on GMM based speaker verification system using the model distance normalization discussed above. Subsection 3.1 presents the datasets used in our experiments. Subsection 3.2 gives a brief introduction of the evaluation criteria. The result of these experiments are discussed in Subsections 3.3 and 3.4.

3.1 Database description

For cepstral feature extraction, 13-dimensional PLP vectors were calculated from the silence removed speech signal every 10 ms using a 25 ms Hamming window. Band-limiting was performed by only retaining the filterbank outputs from the frequency range 300 Hz $\sim$ 3 400 Hz. Cepstral features were processed with RASTA filtering to eliminate channel distortion. Delta, acceleration, and triple-delta coefficients were then computed over frames span and appended to each feature vector, which resulted in dimensionality 52. Feature mapping and histogram equalization (HEQ) were performed to improve channel and noise robustness. Heteroscedastic linear discriminant analysis (HLDA) was then used to decorrelate the features and reduce the dimensionality from 52 to 51 (1 dimension was left out as nuisance). Speaker verification experiments were conducted on the 2006 NIST SRE corpus[14]. We focused on male part of the single-side 1 conversation train and single-side 1 conversation task, which contains 1 570 true trails and 20 561 false trails. A gender independent UBM with 2 048 Gaussians was used in all the experiments, which was trained using about 40 hours of data from the Switchboard II corpora (phases 1 and 2). The speaker GMM models were taken from UBM by MAP adaptation with the relevance fact set to be 16 (only the means were used).

3.2 Detection cost function (DCF)

Results are presented using detection error tradeoff (DET) plots. Along with the equal error rate (EER), the minimum detection cost function (DCF) value, as defined by NIST[14], was also used as an overall performance measure. The DCF defined for the NIST evaluation is of the following expression:

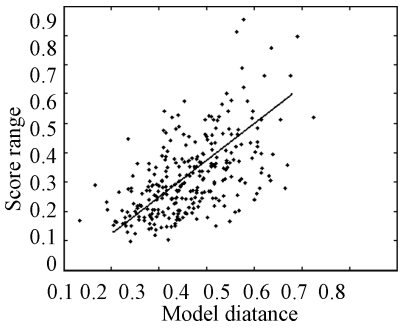
$$\begin{aligned} C_{Det} = & C_{Miss} \times P_{Miss|Target} + C_{FalseAlarm} \times \\ & P_{FalseAlarm|NonTarget} \times (1 - P_{Target}) \end{aligned} \tag{18}$$

The parameters of this cost function are the relative costs of detection errors, $C_{Miss} = 10$ and $C_{FalseAlarm} = 1$, and the a priori probability of the specified target speaker $P_{Target} = 0.01$.

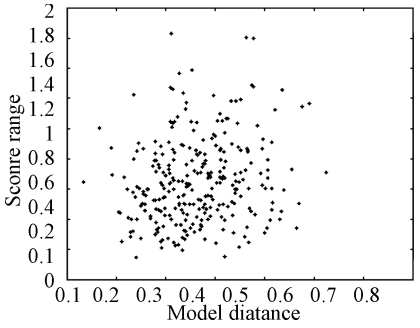
3.3 Correlation between D^2_{UB} and LLR

In this subsection, the presupposition brought forward in Section 2 will be validated from some experimental results. Fig. 1 presents the distribution of the points $(\Delta s, D^2_{UB})$, in which $\Delta s = \bar{s}_{tgt} - \bar{s}_{imp}$ describes the score range for a particular speaker ($\bar{s}_{tgt}$ and $\bar{s}_{imp}$ are the mean values of the target score and impostor score, respectively), and D^2_{UB} is the model distance between this speaker model and UBM.

From the picture above, we can find that an approximately linear correlation between Δs and D^2_{UB} indeed exists, although it is not very prominent. This shows the presupposition in Section 2 is reasonable, and after D-Norm, the correlation between LLR and model distance is removed, just as Fig. 1 (b) shows.



(a) Before D-Norm ($\Delta(s)$ has some correlation with D^2_{UB}, which means D^2_{UB} can affect the range of LLR.)



(b) After D-Norm (The correlation is removed.)

Fig. 1 Distributions of points $(\Delta(s), D^2_{UB})$

3.4 Performance of D-Norm

In this part, we give the experimental results of D-Norm and compare the performance of the simplified approach in this paper with the conventional one. Fig. 2 shows the DET curves of GMM baseline system and GMM with model distance normalization in 1conv4w-1conv4w task of the 2006 NIST SRE, and Table 1 gives the results of the systems in terms of both minimum DCF and EER, in which "MC-DNorm" denotes the conventional Monte-Carlo based D-Norm approach while "UB-DNorm" denotes the upper-bound based approach proposed in this paper.

Table 1 Results for GMM baseline and GMM with two kinds of D-Norms

System	EER (%)	MinDCF (× 100)
GMM baseline	7.64	4.61
GMM with MC-DNorm	7.21	3.63
GMM with UB-DNorm	7.26	3.70

In Monte-Carlo based D-Norm approach, the KL distance was estimated through about twenty thousands synthetic acoustic vectors, which were equivalent to approxi-

mately $2 \sim 3$ minutes long utterances. The results showed that since the relationship between the range of LLR and model distance was not very prominent, as Fig. 1 indicated, the D-Norm did not improve the system performance significantly, but it is still promising, especially when considering its low cost. Moreover, the simplified approach proposed in this paper achieved the performance similar to the traditional one, while reducing the complexity and computing cost greatly. Additionally, it is worthwhile to try the combination between D-Norm and T-Norm (just as ZT-Norm) to further improve the system performance.

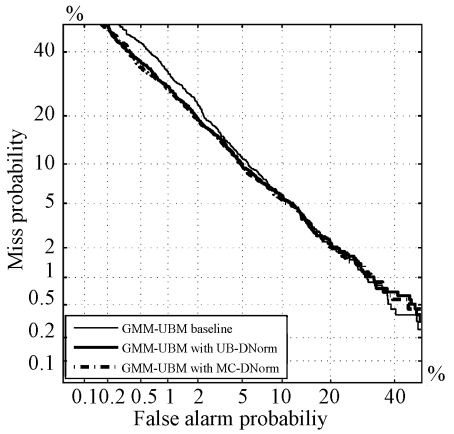


Fig. 2 DET curves for GMM baseline and GMM with D-Norm in the 1conv4w-1conv4w task of the 2006 NIST SRE

4 Conclusion

This study investigated the conventional model D-Norm approach for GMM-UBM based speaker verification systems. Based on some previous works, this paper illustrated the principles of D-Norm in a new perspective which is more logically reasonable. In addition, a simplified approach for performing D-Norm, namely, the upper bound model distance estimation, was also presented. Compared with the original Monte-Carlo based D-Norm, the simplified one in this paper reduced the computational complexity in a great deal, while can still achieve a similar performance, as was shown in the experiments on 2006 NIST SRE corpus.

References

1 Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, **10**(1-3): 19−41

2 Wan V, Renals S. Speaker verification using sequence discriminant support vector machines. *IEEE Transactions on Speech and Audio Processing*, 2005, **13**(2): 203−210

3 Campbell W M. Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA: IEEE, 2002. 161−164

4 Campbell W M, Sturim D E, Reynolds D A, Solomonoff A. SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Washington D. C., USA: IEEE, 2006. 97−100

5 Kenny P, Mihoubi M, Dumoucheln P. New MAP estimators for speaker recognition [Online], available: http://www.crim.ca/perso/patrick.kenny/, November 19, 2008

6 Solomonoff, Campbell W M, Boardman I. Advances in channel compensation for SVM speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, USA: IEEE, 2005. 629−632

7 Heck L, Weintraub M. Handset dependent background models for robust text-independent speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Washington D. C., USA: IEEE, 1997. 1071

8 Auckenthaler R, Carey M, Thomas H L. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 2000, **10**(1-3): 42−54

9 Li K P, Porter J E. Normalizations and selection of speech segments for speaker recognition scoring. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. New York, USA: IEEE, 1988. 595−598

10 Ben M, Blouet R, Bimbot F. A Monte Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA: IEEE, 2002. 689−692

11 Kullback S. *Information Theory and Statistics*. New York: Dover Publications, 1968

12 Do M N. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 2003, **10**(4): 115−118

13 Hershey J R, Olsen P A. Approximating the Kullback-Leibler divergence between Gaussian mixture models. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Washington D. C., USA: IEEE, 2007. 317−320

14 Information Access Division. The NIST 2006 speaker recognition evaluation plan [Online], available: http://www.nist.gov/speech/tests/spk/spk/2006/, Octobor 24, 2007

DONG Yuan Received his Ph. D. degree on telecommunication from Shanghai Jiao Tong University in 1999. Then, from 1999 to 2001, he was with Nokia Research Center as a research and development scientist, working on voice recognition on Nokia mobile phone. From 2001 to 2003, he worked as a post doctoral research staff at Cambridge University, UK, working on European speech recognition project − CORETEX. Since 2003, he has worked as an associate professor at Beijing University of Posts and Telecommunications. His research interest covers speaker recognition, audio indexing, speech recognition, and speech synthesis. Corresponding author of this paper.
E-mail: yuandong@bupt.edu.cn

LU Liang Master student at the School of Information Engineering, Beijing University of Posts and Telecommunications. His research inerest covers speaker verification and speech recognition.
E-mail: luliang07@gmail.com

ZHAO Xian-Yu Received his bachelor and master degrees in electronic engineering from Harbin Institute of Technology in 1997 and 1999, respectively, and Ph. D. degree in electrical engineering from Tsinghua University in 2005.
E-mail: xianyu.zhao@orange-ftgroup.com

ZHAO Jian Master student at the School of Information Engineering, Beijing University of Posts and Telecommunications. His research interest covers speaker verification and speech recognition.
E-mail: michaeljianzhao@gmail.com