SHORT-TIME GAUSSIANIZATION FOR ROBUST SPEAKER VERIFICATION

Bing Xiang[1,2], *Upendra V. Chaudhari*[1], *Jiří Navrátil*[1], *Ganesh N. Ramaswamy*[1], *Ramesh A. Gopinath*[1]

[1]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598
[2]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853
Email: bxiang@ece.cornell.edu,{uvc,jiri,ganeshr,rameshg}@us.ibm.com

ABSTRACT

In this paper, a novel approach for robust speaker verification, namely short-time Gaussianization, is proposed. Short-time Gaussianization is initiated by a global linear transformation of the features, followed by a short-time windowed cumulative distribution function(CDF) matching. First, the linear transformation in the feature space leads to local independence or decorrelation. Then the CDF matching is applied to segments of speech localized in time and tries to warp a given feature so that its CDF matches normal distribution. It is shown that one of the recent techniques used for speaker recognition, feature warping[1] can be formulated within the framework of Gaussianization. Compared to the baseline system with cepstral mean subtraction(CMS), around 20% relative improvement in both equal error rate(EER) and minimum detection cost function(DCF) is obtained on NIST 2001 cellular phone data evaluation.

1. INTRODUCTION

Robust speaker recognition has been an active area of research for a long time because performance degradation due to mismatched conditions has been a significant barrier for deployment of speaker recognition technologies. Previous approaches to channel compensation and adaptation fall into three categories: (1) feature-based, (2) model-based, and (3) score-based.

Cepstral mean subtraction(CMS)[2] and RASTA[3] are two of the standard feature-based approaches. But channel and handset mismatch can still cause lots of errors after CMS or RASTA. Recently, some new approaches were proposed, such as discriminative feature design with neural networks [4]. The most recent technique is feature warping[1] which transforms the distribution of a cepstral coefficient feature stream to a standard distribution over a specified time interval based on cumulative distribution function (CDF) matching. This technique brought significant improvements for speaker verification compared to standard techniques. As explained later, this method can be formulated within the framework of Gaussianization [5], a technique originally proposed for high dimensional density estimation.

One of the recently proposed model-based methods is the speaker model synthesis (SMS) system [6], which applies a speaker-independent model transformation to synthesize speaker models for those channels with no speaker data available during enrollment. Score-based approaches include ZNORM or HNORM [7], which normalizes the distribution of the imposter scores to be zero mean and unit variance.

In this paper, a novel approach for robust speaker verification, namely short-time Gaussianization, is proposed. It is initiated by a global linear transformation of the features, followed by a short-time windowed CDF matching. The experiment results show that the use of short-time Gaussianization makes the speaker verification system more robust compared to our baseline system with CMS. Around 20% relative improvement in both equal error rate (EER) and minimum detection cost function (DCF) is obtained on the NIST 2001 cellular phone data evaluation.

This paper is organized as follows. Section 2 briefly reviews the method of feature warping. In section 3, short-time Gaussianization is introduced. The experimental results for short-time Gaussianization are shown in section 4. Section 5 includes the conclusion.

2. FEATURE WARPING

Telephony speech is characterized by channel distortions that affect the distribution of features. For example, linear channel effects will shift the mean of Mel-frequency cepstral coefficients(MFCC), and additive noise will tend to modify the variance [1]. Mapping the raw features to an ideal distribution, such as the standard normal distribution, appears to be a good way to make the features more robust to different channel and noise effects. This can be done via CDF matching, which warps a given feature so that its CDF matches a desired distribution [1], e.g. $N(0,1)$. The warping can be viewed as a nonlinear transformation T from the

original feature $\mathbf{X}$ to a warped feature $\hat{\mathbf{X}}$, i.e.,

$$\hat{\mathbf{X}} = T(\mathbf{X}). \tag{1}$$

The method assumes that the components of the MFCC vector are independent (this assumption will be made weaker in next section). Each component is processed as a separate stream. CDF matching is performed over a sliding window, the size of which is set to be N. Only the central frame of the window is warped based on CDF matching. The features in a given window of the utterance are sorted in ascending order. Suppose the central frame has a rank r (between 1 and N). Its corresponding CDF value is approximated as

$$\Phi = (r - 1/2)/N. \tag{2}$$

Then the warped value $\hat{x}$ should satisfy

$$\Phi = \int_{-\infty}^{\hat{x}} f(z)dz, \tag{3}$$

where $f(z)$ is the probability density function (PDF) of standard normal distribution, i.e.

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}). \tag{4}$$

$\hat{x}$ can be quickly found by lookup in a standard normal CDF table.

3. SHORT-TIME GAUSSIANIZATION

Gaussianization[5] was originally proposed for high dimensional density estimation and exploited the independence structures in the data to alleviate the curse of dimensionality. This is achieved by an iterative scheme. In each iteration, the feature space is first transformed by a linear transformation. Then, it is followed by marginal Gaussianization, which is global CDF matching with the normal distribution as the target. As mentioned before, the feature warping method described in section 2 can be formulated in the framework of Gaussianization, with some modifications. The global marginal Gaussianization step is replaced by short-time windowed feature warping. Thus, this new approach is called short-time Gaussianization. It is applied to speaker verification as a new way to compensate channel and handset variability.

Suppose the original feature set is $\mathbf{X}$, with dimension D. And there is a linear transformation $A_{D\times D}$ such that the transformed features are

$$\mathbf{Y} = A\mathbf{X}. \tag{5}$$

The probability density of Y is modeled by a so-called compound Gaussian model (CGM) [8] as

$$p(\mathbf{y};\theta) = \sum_{k=1}^{K} \rho_k |A| \prod_{d=1}^{D} \sum_{i=1}^{I_{k,d}} \pi_{k,d,i} \phi(y_d, \mu_{k,d,i}, \sigma^2_{k,d,i}), \tag{6}$$

where K is the number of compound Gaussian components, D is the dimension of features and $I_{k,d}$ is the number of Gaussians for the k-th CGM component and d-th dimension. ρ_k is the prior for the k-th CGM component, $\pi_{k,d,i}$ is the prior of the i-th Gaussian of the k-th CGM component for dimension d, and ϕ is the PDF of a univariate Gaussian distribution, i.e.

$$\phi(y_d, \mu_{k,d,i}, \sigma^2_{k,d,i}) = \frac{1}{\sqrt{2\pi}\sigma_{k,d,i}} \exp(-\frac{(y_d - \mu_{k,d,i})^2}{2\sigma^2_{k,d,i}}), \tag{7}$$

with mean $\mu_{k,d,i}$ and variance $\sigma^2_{k,d,i}$. The full parameter set is $\theta = \{A, \rho_k, \pi_{k,d,i}, \mu_{k,d,i}, \sigma_{k,d,i}\}$.

Since we have

$$y_d = \mathbf{a}_d \mathbf{x}, \tag{8}$$

where $\mathbf{a}_d$ is the d-th row of matrix A, the probability of $\mathbf{X}$ can be expressed as

$$p(\mathbf{x};\theta) = \sum_{k=1}^{K} \rho_k |A| \prod_{d=1}^{D} \sum_{i=1}^{I_{k,d}} \pi_{k,d,i} \phi(\mathbf{a}_d \mathbf{x}, \mu_{k,d,i}, \sigma^2_{k,d,i}), \tag{9}$$

The CGM actually assumes that the transformation could make the new feature space achieve local decorrelation or independence, depending on the number of Gaussian mixtures $I_{k,d}$ for the k-th compound model and d-th dimension. When $I_{k,d} = 1$ for all k and d, the feature transformation achieves local decorrelation since it is modeled as a mixture of diagonal covariance Gaussians. However, when $I_{k,d} > 1$, the transformation achieves local independence instead. An EM algorithm is derived in [8] for estimating the parameter set θ. It is based on maximizing the likelihood of $\mathbf{x}$ and θ. Generally five EM iterations are sufficient for training.

Among those parameters in θ, only the linear transformation matrix A is used for short-time Gaussianization as compared to marginal Gaussianization which uses the parameters in θ for global CDF matching. It can be implemented in two steps, similar to general Gaussianization.

(1) Linearly transform the data:

$$\mathbf{Y} = A\mathbf{X}; \tag{10}$$

(2) Nonlinearly transform the data by marginal short-time Gaussianization:

$$\hat{\mathbf{X}} = T(\mathbf{Y}), \tag{11}$$

where T is the short-time windowed feature warping mentioned in section 2.

One of the differences from general Gaussianization is that no iterative scheme is used here. Only one pass is applied for each utterance. A global transformation matrix A is estimated on speaker-independent data and applied to the features from all speakers. Another difference is that CDF

matching in short-time Gaussianization is implemented in a non-parametric scheme, instead of parametric scheme.

It has been mentioned that feature warping is based on the assumption of independence of feature vector components. Since the linear transformation A is introduced before the warping, this assumption is less strong than that without such a transform.

4. EXPERIMENTS

4.1. Database

The short-time Gaussianization techniques are evaluated on the cellular telephone speech, used in the NIST speaker recognition evaluation for 2001. 2 hours of speech from 38 male and 22 female speakers, with 2 minutes each speaker, are used for training the background model and also the CGM. There are 74 male and 100 female target speakers. Each speaker has 2 minutes of speech for training. 20380 gender-matched verification trials form the test set. The duration of each test segment varies from a few seconds to one minute, with the majority of tests falling into a range between 15 to 45 seconds. The ratio between target and impostor trials is roughly 1:10.

4.2. Evaluation measure

The evaluation of the speaker verification system is based on Detection Error Tradeoff (DET) curves, which show the tradeoff between false alarm (FA) and false rejection (FR) errors. Besides the equal error rate (EER), there is also a detection cost function (DCF) defined for the NIST evaluation:

$$DCF = C_{FA}Pr(FA|N)Pr(N) + C_{FR}Pr(FR|T)Pr(T) \quad (12)$$

where $Pr(N)$ and $Pr(T)$ are the a priori probability of non-target and target tests with $Pr(N) = 0.99$ and $Pr(T) = 0.01$. And the specific cost factors $C_{FA} = 1$ and $C_{FR} = 10$. So the point of interest is shifted towards low FA rates.

4.3. Speaker verification system

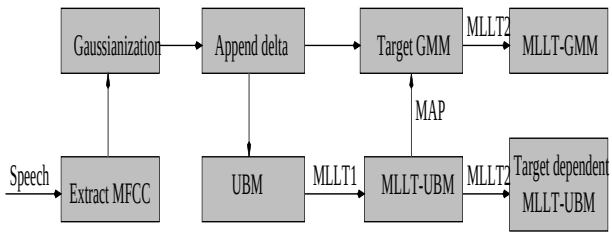

Fig. 1. Speaker verification system

As shown in Fig. 1, short-time Gaussianization is applied on the extracted 19 dimensional MFCC first. The frame rate is set to 10ms. Then delta coefficients are calculated based on the warped features for each dimension. These are appended to the warped MFCCs to form a 38-dimension feature vector used in all the experiments. The speaker verification system is based on Gaussian mixture models (GMMs). During the training session, a universal background model (UBM) is trained from the speaker-independent data and enhanced by a maximum likelihood linear transformation (MLLT) [9], i.e. MLLT1 is applied to generate MLLT-UBM. Then, for each enrolled speaker, a target GMM is created by adapting MLLT-UBM with MAP adaptation. After applying speaker-specific MLLT (MLLT2) on both the target GMM and the MLLT-UBM, we get the MLLT-GMM and target-dependent MLLT-UBM for each speaker [10]. 512 Gaussian components are used for each GMM and UBM. During testing, only the best component as determined in the MLLT-UBM is counted for each frame. And the score of each trial is obtained from the average of frame log likelihood ratios between the claimed target MLLT-GMM and its corresponding MLLT-UBM.

4.4. Experimental results

4.4.1. Variable parameters for CGM

For short-time Gaussianization, variable parameters for the CGM were evaluated. Window size for CDF matching was set as 300 frames in these experiments. Tab. 1 shows that CGM with 32-component compound model and 1 Gaussian for each dimension performs the best in this work. As mentioned above, with this configuration, CGM is trying to achieve local decorrelation in feature space. With the same number of Gaussian mixtures, it is better than using a 1-component compound model with 32 Gaussians for each dimension as well as using 8-component compound model with 4 Gaussians for each dimension.

K	$I_{k,d}$	DCF(10^{-3})	EER(%)
1	32	46.5	11.2
8	4	45.7	11.0
32	1	44.0	10.8

Table 1. Variable parameters for CGM

4.4.2. Variable window size

Tab. 2 shows the results of short-time Gaussianization with variable window sizes during CDF matching. Among the selected window size, 300(3 seconds) gives the best results in terms of the minimum DCF. When the window size is increased to the length of whole utterance (up to one minute),

the performance obviously becomes worse. So focusing on the local distribution is more beneficial than on the global distribution.

window size(frame)	DCF(10^{-3})	EER(%)
100	47.9	11.2
300	44.0	10.8
500	44.6	11.1
utterance	48.2	12.4

Table 2. Variable window size

4.4.3. Compared to baseline and feature warping

Fig. 2 compares the performance of the baseline system with CMS, feature warping and short-time Gaussianization. Both short-time Gaussianization and feature warping improve the performance. Short-time Gaussianization brought 23% improvement compared to the baseline in terms of minimum DCF and 19% improvement in terms of EER. And it also shows advantages over feature warping, especially in low FA rates, and stronger capability of preserving speaker-specific information.

4.4.4. Gender-dependent short-time Gaussianization

Since the evaluation of NIST is gender-matched, gender-dependent UBMs can be used for background models. Different CGMs are trained for male and female speakers first. Then gender-dependent UBMs are trained after the gender-dependent short-time Gaussianization. Another improvement is obtained for DCF, as shown in Fig. 2.

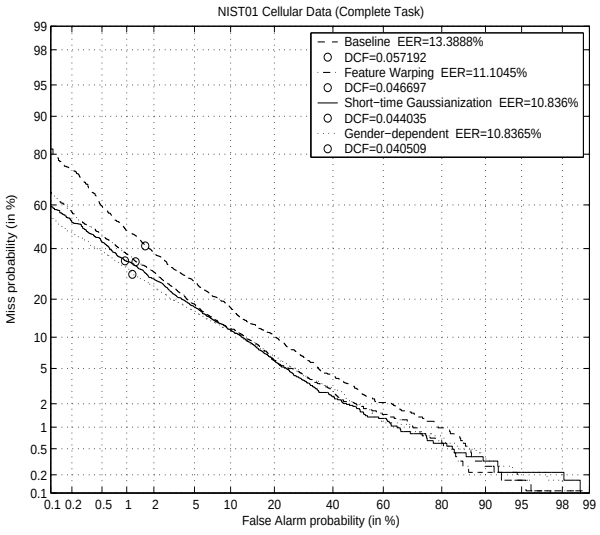

Fig. 2. Comparison between different systems

5. CONCLUSIONS

Our experimental results show that short-time Gaussianization significantly improves the verification performance. It brings around 20% improvement compared to the baseline for both minimum DCF and EER. And it also shows advantages over feature warping. Thus short-time Gaussianization is a robust approach to speaker verification. Such techniques may also be used for speech recognition. More experiments are needed in this area in the future.

Acknowledgements The authors would like to thank Sabine Deligne and Ran Zilca in IBM, and Prof. Toby Berger in Cornell for their help and discussion.

6. REFERENCES

[1] J.Pelecanos and S.Sridharan, "Feature warping for robust speaker verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[2] S.Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust. Speech Signal Processing, vol.ASSP-29, pp.254-272, Apr.1981.

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

[4] L.P.Heck, Y.Konig, M.K.Sonmez and M.Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design", Speech Communication, vol.31, pp.181-192, 2000.

[5] S.Chen and R.A.Gopinath, "Gaussianization", Proc. NIPS 2000, Denver Colorado.

[6] R.Teunen, B.Shahshahani and L.P.Heck, "A model-based transformational approach to robust speaker recognition", Proc. ICSLP, 2000.

[7] D.A.Reynolds, T.F.Quatieri and R.B.Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, 10(1-3), pp.19-41, 2000.

[8] http://www.research.ibm.com/people/r/rameshg/chen-gaussianization-neuralcomputation2000.ps.

[9] U.V.Chaudhari, J.Navrátil, S.H.Maes and R.Gopinath, "Transformation enhanced multi-grained modeling for text-independent speaker recognition", ICSLP, 2000.

[10] J.Navrátil, U.V.Chaudhari and G.N.Ramaswamy, "Speaker verification using target and background dependent linear transforms and multi-system fusion", Eurospeech, 2001.