KURTOSIS NORMALIZATION IN FEATURE SPACE FOR ROBUST SPEAKER VERIFICATION

Yanlu Xie, Beiqian Dai, Zhiqiang Yao, Minghui Liu

MOE-Microsoft Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China
(xieyl@mail.ustc.edu.cn bqdai@ustc.edu.cn zqyao@ustc.edu liumh@ustc.edu).

ABSTRACT

The acoustic mismatch between the training and test environments will lead to the difference of the statistical characteristics of speech parameters. Since the statistical characteristics of the kurtosis can measure the non-Gaussianity of a random variable, kurtosis normalization will make the training and test speech parameters match the standard normal distribution in some sense. In this paper, a kurtosis normalization method using sigmoid functions (logit functions) in feature space is presented for GMM-UBM based text-independent speaker verification system. Experimental results on the 2004 NIST SRE database show that with the new method significant improvement can be achieved in not only equal error rate but also minimum detection cost compared with baseline system (more than 33% relative reduction for long speech).

1. INTRODUCTION

The acoustic mismatch between the training and test data leads to the performance degradation of speaker verification. Feature derived from spectrum of speech may be affected by many factors such as the microphones, the acoustic environments, the transmission channels and so on. Many of the approaches used to solve the problems concentrate on the normalization of speech features.

In feature space, resent research has proved the mismatch can be lessened when the cumulative distribution functions (CDFs) of training and test data are both transformed to match that of the standard normal or other uniform distribution. The transformations include histogram equalization [1][2], feature warping [3][4], short-time Gaussianization [5] and so on.

Histogram equalization and short-time Gaussianization have achieved much improvement. However, these methods especially concern the short-time distribution of speech cepstral and don't normalize the statistical characteristics such as means, variance and higher order cepstral moment directly. Nevertheless transforming the characteristic of all the speech parameters can also be used to reduce the mismatch [6].

This paper presents a statistical characteristic based normalization method. Since the standard kurtosis is able to measure the peakness or non-Gaussianity of a random variable, the method can lessen the non-Gaussianity by optimizing the kurtosis of each dimension of the feature parameters with sigmoid functions (logit functions). In the method the parameters of sigmoid functions are optimized until the kurtosises reach approximate zero (the same as the normal distribution). Because of the normalization on both the training and test speech, the mismatch between them is decreased and the performance of system is improved.

In addition, short-time feature space normalization methods such as short-time Gaussianization have to use adjacent frames of parameters to estimate the transform. Thus the first and last several frames of parameters are wasted, which is sub-optimal when the length of speech is short. On the other hand, the one-parameter sigmoid function used in this paper is trained offline with other speech data. None of the speech is wasted. So the method is more suitable for speaker verification with short speech. The scheme of NIST'04 8conversations-1conversation and 10seconds-10seconds tasks is used to evaluate the performance of the proposed normalization method.

The paper is organized as follows: in section 2, the short-time feature space Gaussianization is presented. In section 3, we present our kurtosis normalization method. And the experiment results are showed in section 4.

2. FEATURE SPACE GAUSSIANIZATION TRANSFORM

For a random variable $\mathbf{X} \in R^D$, the definition of its Gaussianization transformation is to make the transformed variable $T(\mathbf{X})$ follow the standard normal distribution:

$$T(\mathbf{X}) \sim N(0, \mathbf{I})$$

the transform function $T(\mathbf{X})$ is invertible, differential and existent [7].

It is difficult to calculate the Gaussianization transformation for high dimensional data. In order to

employ one dimensional Gaussianization directly, B. Xiang applied a linear transformation to the feature to make the assumption of independence of feature vector components less strong [5]. The linear transformation estimated by the Expectation Maximization (EM) algorithm can also make the transformed feature more suited to diagonal covariance Gaussian mixture models (GMMs). One-dimensional Gaussianization transformation is deduced that

$$T(x) = \Phi^{-1}(F_X(x)) \tag{1}$$

where $F_X(x)$ and $\Phi(x)$ are respectively the CDFs of variable x to be transformed and the standard normal's.

If the $F_X(x)$ has been known for each value of x, the transform can be determined by looking up standard normal table. Two approaches have been presented to estimate $F_X(x)$, i.e. a piece-wise constant function approximation and Gaussian mixture models approximation. The former method is much simpler and easier than the latter. With the former method, we can get that

$$F(x_i) = {rank(x_i)}/{N} \tag{2}$$

where $rank(x_i)$ is the rank of x_i in sorted list of samples. Combining (1) with (2) yields the final local feature space Gaussianization transformion

$$T(x_i) = \Phi^{-1}({rank(x_i)}/{N}),\ 1 \le i \le N \tag{3}$$

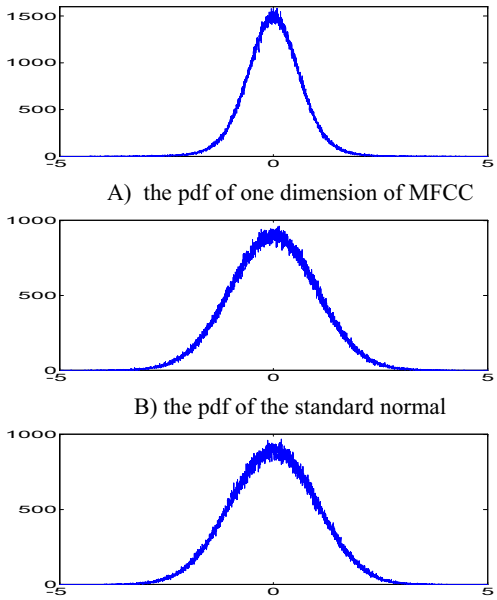
The function (2) suggests that the evaluation of $F_X(x)$ uses the adjacent N frames parameters, so that the transformation takes into account the distribution of a short segmental speech. At the same time $F_X(x)$ of the first and last $N/2$ frames can't be computed exactly. B. Xiang has shown the best result is determined when N=300 [5]. If the parameter vectors are computed every 10 ms, about 3s of speech is wasted. When the training and test speech is short it becomes sub-optimal. The short-time Gaussianization method especially concerns the short-time distribution of speech cepstral. But it doesn't normalize the statistical characteristic directly. In fact, some statistical characteristics can perfectly measure the non-Gaussianity, and the normalization of them can also take the effect of Gaussianization. Thus we would like to propose a kurtosis normalization method here.

3. KURTOSIS NORMALIZATION BASED ON SIGMOID FUNCTIONS

As we know the standard kurtosis can measure the peakness or non-Gaussianity of a random variable. The kurtosis of a random variable x is defined as

$$K(x) = ({E(x^4)}/{E(x^2)^2}) - 3 \tag{4}$$

Kurtosis is a scale independent dimensionless parameter. A normal random variable has a kurtosis of zero. If a random variable has a kurtosis less than zero, it is termed platykurtic i.e. sub-Gauss. If it has kurtosis greater than zero, it is termed leptokurtic i.e. super-Gauss. Speech signals are generally leptokurtic, so are speech cepstral parameters. A) and B) of Figure1 compare the probability density function (pdf) of one dimension of the MFCC with that of the standard normal. In the experiment, 1130089 samples are used in both of the two distributions.



A) the pdf of one dimension of MFCC

B) the pdf of the standard normal

C) the pdf of one dimension of MFCC after kurtosis normalization

Figure 1: *The comparison of pdfs*

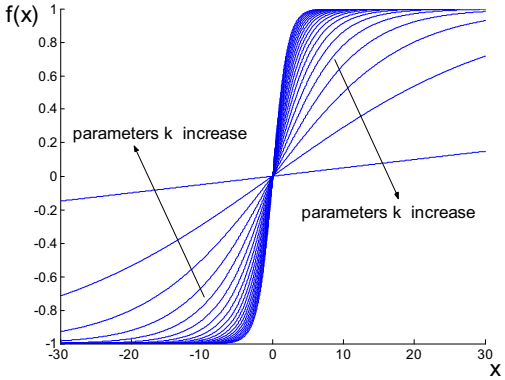


Figure 2: *the sigmoid functions (the logit functions)*

The sigmoid function (the logit function) can be expressed as

$$f(x,k) = \frac{a}{1+\exp(-kx)} - b \tag{5}$$

where a and b are constant coefficients, k>0. In order to keep the means of speech parameters invariable when the sigmoid function is used, coefficients a and b are chosen to be 2 and 1 respectively.

From figure 2, it can be found that the sigmoid function changes greatly when the variable x is small. When x becomes larger, the variety of $f(x)$ will be lessened. The smaller the parameter k in the sigmoid function becomes, the more obvious the trend tends to be. Because of the difference between MFCC and the standard normal, substituting the sigmoid function for (3) can also take effect of Gaussianization transformation. We have proved mathematically that the optimization of kurtosis can be achieved by selecting the transforming factor k of sigmoid functions. How to optimize k is shown as follows.

Conceptually, the transforming factor k represents the deviatability between one dimension of MFCC and the normal random variable. In the work described here, the transforming factor k is chosen to minimize the absolute value of the kurtosis of the transformed parameter, i.e.

$$\hat{k} = \arg\min_{k} \left| K(f(x,k)) \right| = \arg\min_{k} \left| \frac{E(f(x,k)^4)}{E(f(x,k)^2)^2} - 3 \right| \quad (6)$$

A closed-form solution for $\hat{k}$ from (6) may be difficult to obtain. In fact, the experiments show that the optimum transforming factor is obtained by searching over a grid spaced between $0.050 \leq k \leq 1.070$. Computing each factor of the grid can get the optimized value of factor k**.**

After kurtosis normalization, cepstral variance normalization (CVN) can also be used. It is because the CVN technique only multiplies a coefficient without changing the kurtosis.

In fact, the non-Gaussianity of each dimension of the MFCCs is different. The experiment shows that their kurtosises vary from 0.205 to 2.516. Thus training different k parameters for each dimension is necessary. From B) and C) of figure1, we find the MFCC approximates the normal distribution when its kurtosis is decreased to zero.

It seems that each dimension of separate speaker using different sigmoid functions may be more reasonable. But experiments show that normalization kurtosis for each speaker will lead to the worst result, nearly entirely error. Thus kurtosis normalization is performed globally over a development database.

4. EXPERIMENTS AND RESULTS

4.1. Database

The performance of the described normalization methods is evaluated on NIST'04 8conversations-1conversation (8c-1c) and 10seconds-10seconds (10s-10s) Speaker Recognition Evaluation (SRE) tasks. Only male speakers are used here. This database is a subset of the Switchboard (SWB) cellular telephone corpus. The 8c-1c task uses eight single channel conversational sides of one speaker for training and then tests on one single channel conversation side. Each conversation side is about 5 minute including silence duration. The 10s-10s task uses approximately 10 seconds of estimated speech in training and test. There are 170 male speakers along with a total of 8088 verification trials in the 8c-1c task and 246 male speakers along with 9375 trials in the 10s-10s task.

4.2. Evaluation Measure

The evaluation of the speaker verification system is based on Detection Error Tradeoff (DET) curves which show the tradeoff between the two types of detection errors (false alarm, and false rejection). On the DET curve typically two specific operating points may be of interest. One of them is the Equal-Error Rate (EER) where the FA rate equals the FR rate, which is used as a summary performance measure for comparing systems. The other is the point having the lowest detection cost. Detection cost function (DCF) is defined for the NIST evaluation.

$$DCF = C_{FA} P_{FA|N} P_N + C_{FR} P_{FR|T} P_T$$

where P_N and P_T are the priori probability of the specified nontarget and target speakers with P_N=0.99 and P_T=0.01. The specific cost factors are C_{FA}=1 and C_{FR}=10, which shifts the point of interest toward low FA rates [8][9].

4.3. System Description

The baseline system is essentially a GMM-UBM based text-independent speaker verification system [10]. The feature vectors are composed of 16 mel-cepstal coefficients (MFCCs) and their 16 corresponding deltacepstra coefficients without the zeroth one. The vectors are computed every 10 ms with a 20 ms Hamming window. An energy based silence removal technique is used to discard silence frames in both training and test vectors. Cepstral mean subtraction (CMS) and RASTA processing are also used to normalize the linear channel.

With the data from NIST'01 male speakers, expectation Maximization (EM) estimation is used to train a universal background model (UBM) with 2048 mixtures for the 8c-1c task and a UBM with 512 mixtures for the 10s-10s task respectively. Given the UBM, the target speaker models are then derived using Bayesian adaptation. For the systems used in the experiment, only the means of the mixture components are adapted.

4.4. Experimental Results

Table 1 and figure 3 show the results of normalization methods described in Section 2 and 3. In the short-time Gaussianization method, the number of the frames used to estimate the distribution is 300. The kurtosis normalization

method is based on sigmoid functions whose parameters are trained with the same data as that use to train UBM. The two normalization methods are both applied after the MFCC are processed by CMS and RASTA technique.

Table 1: *the comparison of the DCF and EER*

Task	Method	Min. DCF	EER (%)
10s-10s	baseline	0.0928	29.89
	Short-time Gaussianization	0.0898	29.70
	Kurtosis normalization	**0.0881**	**28.25**
8c-1c	baseline	0.0541	11.31
	Short-time Gaussianization	0.0375	7.60
	Kurtosis normalization	**0.0354**	**7.49**

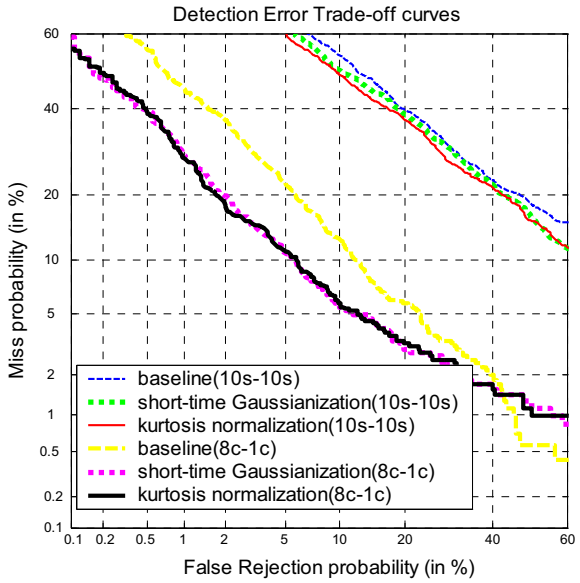


Figure 3: *the comparison of the DET curves*

From table 1 and figure 3, it is found that the performance of the two tasks is improved with both the two normalization methods.

As to the 8c-1c task, the performance is improved much with kurtosis normalization compared with baseline system (34.6% relative reduction for the minimum DCF and 33.8% for the EER). The performance of Kurtosis normalization is comparable with short-time Gaussianization, a little better in the 10s-10s task. It is because with the short-time Gaussianization method the first and last 150 frames (about 3s of speech) can't be computed exactly and it is sub-optimal when the training and test speech is only 10s.

5. CONCLUSIONS

In this paper we presented a kurtosis normalization method for text-independent speaker verification. While the traditional short-time Gaussianization normalizes short-time distribution. Our method based on sigmoid functions normalizes the training and test data concerning the statistical characteristic. Experiments on the NIST'04 SRE database have proved kurtosis normalization is comparable with short-time Gaussianization, when speech is short the former performs a little better. In the future, we will test the effective of the technique on larger database.

6. ACKNOWLEDGEMENTS

This work has been supported by the Science Research Fund of MOE-Microsoft Key Laboratory of Multimedia Computing and Communication (No.05071810) and the National Science Foundation of China (No.60272039).

7. REFERENCES

[1] S. Molau, M. Pitz, H. Ney. "Histogram Based Normalization in the Acoustic Feature Space", *Proc. ASRU 2001*, Trento, Italy, 2001.

[2] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition", *in EUROSPEECH*, 2001, pp.1135–1138.

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

[4] Dharanipragada, Satya, Padmanabhan, Mukund, "A nonlinear unsupervised adaptation technique for speech recognition", *In ICSLP-2000*, vol.4, 556-559.

[5] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification", *ICASSP'02*, 2002, vol. 1, pp. 681–684.

[6] Chang-wen Hsu, Lin-shan Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition", *ICASSP '04*, Volume: 1, 17-21 May 2004 Pages:I - 197-200 vol.1

[7] Ramesh Gopinath "Gaussianization", *IMA Workshop*: Mathematical Foundations of Speech Processing and Recognition, http://www.ima.umn.edu/talks/workshops/9-18-22.2000/gopinath/talk.pdf September 18-22, 2000

[8] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation-overview, methodology, systems, results, perspective", *Speech Communication,* vol. 31, pp.225–254, 2000.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. "The DET curve in assessment of detection task performance", *In Proc. Of the EUROSPEECH*, pages 1895–8, Rhodes, Greece, September 1997.

[10] Reynolds, Douglas A., Quatieri, Thomas F. and Dunn, Robert B. "Speaker Verification Using Adapted Gaussian Mixture Models" *Digital Signal Processing*, Volume 10, Issue 1-3, January, 2000, Pages 19-41