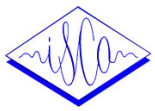

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

DOUBLE GAUSSIAN BASED FEATURE NORMALIZATION FOR ROBUST SPEECH RECOGNITION

Bo Liu, Li-Rong Dai, Jin-Yu Li, Ren-Hua Wang

iFlytek Speech Lab
University of Science and Technology of China, Hefei, Anhui, P. R. China
liubo@ustc.edu, lrdai@ustc.edu.cn, jinyuli@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

In this paper, a new feature normalization approach based on Cumulative Density Function (CDF) matching principle is proposed. Since speech features in noisy environments usually follow bimodal distributions, we fully utilize this characteristic by representing the CDF of the features with a double Gaussian model. Feature normalization process is performed according to the estimated CDF. The experimental results on Aurora2 database show that the performance of our method is much better than that of the conventional Mean and Variance Normalization (MVN) method, and comparable to that of the method combining the spectral subtraction and histogram equalization (HE). Moreover, further improvement has been gained by combining our method with a simple temporal feature smoothing process. This result suggests that our new method has the potential to be integrated with other techniques to provide even better performance.

1. INTRODUCTION

There are many sources that will severely degrade the performance of speech recognition systems in realistic environments. A variety of techniques were proposed to improve the robustness of the system. They can be roughly divided into two categories: Adaptation methods, the first class, usually cope with the specific adverse environments by adjusting the acoustic models. Whereas in the second class, normalization methods, speech features are transformed to reduce the mismatch between training and test environments.

Cepstral Mean Normalization (CMN) is a classical and widely used normalization method, though in principle it can only compensate the shift of the mean of the speech features. Its natural extension, Mean and Variance Normalization (MVN), is proposed to normalize both the mean and the variance [1], so it can improve the system robustness to additive noises, as well as the channel effects. Both methods are based on parametric models.

Non-parametric models can also be used in normalization methods, such as cumulative histogram used in histogram equalization method [1]. HE method is further developed into quantile equalization method [2][3][4], in which less data is required for the estimation of quantiles. The quantile equalization can be combined with many other methods, such as spectral subtraction (SS) [5] and vector Taylor series (VTS) [6]. In [7][8], two independent cumulative histograms are estimated from speech segments and pure noise segments respectively, a VAD algorithm is required to discriminate speech from noise, but usually the discrimination cannot be very precise. Cumulative histogram is also used in a nonlinear unsupervised adaptation technique [9], though the performance of the technique itself is not yet much satisfying if not combined with MLLR.

It is shown that distributions of speech features in noisy environments are usually bimodal [7]. It is an important characteristic that should be fully utilized. However, the present parametric normalization methods, such as MVN, are incapable of representing such structure. Non-parametric normalization methods, such as HE, can express any distributions in principle, but a larger amount of data is needed to estimate the features distributions precisely. Therefore, we propose a double-Gaussian-based feature normalization method. It is a parametric model that is in nature capable of representing the bimodal structure precisely with only a small amount of data. Satisfying results have been achieved in our experiments on Aurora2 database.

The rest of this paper is organized as follows. In section 2, we introduce the CDF matching principle, which is the basis of our new method. Two applications of this principle under specific assumptions, MVN and HE method, are also briefly introduced. In section 3, we describe the principle of our double Gaussian feature normalization method and its implementation details. Experimental results on Aurora2 are shown in section 4. Finally, we present the conclusions in section 5.

2. FEATURE NORMALIZATION

2.1 Cumulative Density Function (CDF) Matching Principle

In order to reduce the mismatch between training and test environments, it is very natural to transform the speech features to make their probability density functions match[1][7]. This operation is equivalent to making the cumulative density function (CDF) match, since CDF is just the integral of probability density function. According to this principle, the feature transformation function can be obtained from the CDF of the data and its inverse function, as follows:

Assume feature transformation function is $x = T[y]$, where y is the feature before transformation, and x is the feature after transformation.

Let $C_X(x)$ be the CDF of x, and $C_Y(y)$ be the CDF of y, then the CDF matching principle can be expressed as

$$C_Y(y) = C_X(x) \quad (1)$$

So the feature transformation function can be obtained as

$$x = T[y] = C_X^{-1}(C_Y(y)) \quad (2)$$

The above transformation is also called feature compensation. However, in practical applications, it is usually much more convenient to implement the feature normalization, in which $C_X(x)$ is always the CDF of a fixed reference distribution. Both training and test feature distributions are normalized into that fixed distribution, therefore the mismatch can be reduced. The most widely used reference distribution is standard Gaussian distribution, which is also adopted in the implementation of our new method.

2.2 Mean and Variance Normalization (MVN)

MVN is a simple but effective robustness method [1]. The principle of this method is to normalize the feature by adjusting both the mean and the variance simultaneously.

MVN can also be viewed as an example of applying CDF matching principle: when feature distribution is exactly Gaussian distribution, the MVN method is equivalent to the CDF matching. However, single Gaussian model cannot express the complicated distribution of the data, such as the bimodal structure.

2.3 Histogram Equalization (HE)

HE is another example of the applying CDF matching principle. It has been widely used in image processing. In recent years, HE is also used in robust speech recognition and some satisfying results have been achieved [5]. A non-parametric model, cumulative histogram, is adopted to estimate the CDF in HE, and that is the primary difference of HE from MVN. However, a larger amount of data is needed to estimate a precise cumulative histogram due to its non-parametric essence.

HE is combined with spectral subtraction method to achieve better performance in [5]. Spectral subtraction is first applied in spectral domain to reduce the effects of additive noises, then HE is used in cepstral domain to compensate the non-linear distortion caused by residual additive noises and channel effects.

3. DOUBLE GAUSSIAN BASED FEATURE NORMALIZATION

It is shown that the speech features are usually bimodal in noisy environments [7]. It is quite reasonable because the statistical characteristic of the features is quite different for speech and noise, therefore the features of speech and noise scatter in different regions.

In order to use a parametric model to express the bimodal structure of CDF of speech features, a double Gaussian model is adopted in our normalization method. EM algorithm [10] is used to estimate the model parameters. Compared with the single Gaussian model used in MVN, double Gaussian model can match the bimodal structure much more precisely.

If the probability density function of feature y is

$$p_Y(y) = \sum_{k=1}^{K} c_k N_k(y; \mu_k, \psi_k) \quad (3)$$

where $K = 2$ if double Gaussian model is used. N_k denotes k th Gaussian mixture, and c_k, μ_k and ψ_k are the weight, mean and variance of it, respectively.

Then the CDF of y is

$$C_Y(y) = \sum_{k=1}^{K} c_k F(\frac{y - \mu_k}{\psi_k}) \quad (4)$$

where $F(t)$ is the CDF of standard Gaussian distribution. Since there is no analytic expression of this function, we implement it using table lookup.

Finally, we normalize both training and test features as

$$x = T[y] = C_X^{-1}(C_Y(y)) = F^{-1}(C_Y(y)) \quad (5)$$

3.1 Model Parameters Estimation

EM algorithm [10] is used in the estimation of the model parameters. If double Gaussian model is as follows

$$\begin{aligned} p(y \mid \phi) &= \sum_{k=1}^{2} c_k p_k(y \mid \phi_k) \\ &= \sum_{k=1}^{2} c_k N_k(y \mid \mu_k, \psi_k) \end{aligned} \quad (6)$$

Given the data sequence $y = (y_1, ..., y_N)$, then

$$\gamma_k^i = [c_k p_k(y_i \mid \phi_k)] / [p(y_i \mid \phi)] \quad (7)$$

$$\gamma_k = \sum_{i=1}^{N} \gamma_k^i \quad (8)$$

After one iteration, the updated model parameters can be computed as follows:

$$c_k^{'} = \gamma_k / N \quad (9)$$

$$\mu_k^{'} = [\sum_{i=1}^{N} \gamma_k^i y_i] / \gamma_k \quad (10)$$

$$\psi_k^{'} = [\sum_{i=1}^{N} \gamma_k^i (y_i - \mu_k)(y_i - \mu_k)^{'}] / \gamma_k \quad (11)$$

In our experiments, five iterations are performed to obtain the reliable model parameters, though we find that the algorithm usually converges after only two or three iterations.

3.2 Approximate Implementation

Neither the CDF of standard Gaussian distribution $F(t)$ nor its inverse function $F^{-1}(t)$ has analytic expression, so we approximately implement these two functions using table lookup.

Each item of the table contains a data pair $[t, F(t)]$, and linear interpolation is used to get a more precise function value from two neighboring data pairs. We also adopt Binary Search algorithm to accelerate the table lookup process.

It is interesting to find that the final performance of our implementation is virtually insensitive to the number of data pairs in the table. In fact, 300 data pairs have shown to be quite enough in our experiments, and this 300-point-table is used in all of our experiments listed in this paper.

3.3 Temporal Smoothing of the Features

In order to find out whether our method can cooperate well with other methods, a simple temporal ARMA filter [11] is included in our experiments.

In current implementation of our method, speech features are normalized on a frame-by-frame basis, so it is less likely for the normalized feature of the adjacent frames to be very continuous. The ARMA filter can smooth out the undesirable spikes owing to its low pass characteristics.

4. RECOGNITION EXPERIMENTS

4.1 Speech Databases and Back-End Configurations

Our experiments are performed on Aurora2 database. This database is a subset of TI digits database distributed by ETSI, with artificially mixed additive noises and channel effects. Two strategies of acoustic model training are defined, one trained with clean speech (Clean condition), and the other trained with both clean and noisy speech (Multi condition). Three test sets are defined for both strategies: set A, noise type is matched for training and test; set B, noise type is mismatched for training and test; set C suffers from channel effects besides additive noises.

The back-end configurations of the recognizer are also defined in Aurora2 [12]. HTK is used to perform both the training and test; 16 emitting states used for each digit model, and 3 Gaussian mixtures per state; silence model has 3 states and 6 Gaussian mixtures per state; short pause model has only 1 state tied with the middle state of silence model. By giving all these definitions, Aurora2 makes it possible to compare the performance of different front-end noise-robust techniques.

4.2 Experimental Setup and Results

First, MVN method is implemented for the comparison with our method. MFCC features (include C0) are used and experiments are done on a sentence-by-sentence basis. Normalization is applied on the static features before the computation of derivatives. The results are shown in Table 2. Compared with the MFCC baseline results (Table 1), relative error rate reduction is 32.76%.

Then, in order to compare our method with HE, the absolute performance of the method in [5] is also cited (Table 3). This method combines the spectral subtraction and HE. It can be worked out that the relative performance is 41.91%.

Our double Gaussian feature normalization method is also implemented on MFCC features (include C0). Each dimension of the static features is normalized independently. Five EM iterations are performed to get the parameters of double Gaussian model. Then the transformation functions are obtained and the features are normalized. The acoustic model is retrained using normalized training features, while normalized test features are sent to the back-end. The experimental results are listed in Table 4. The relative performance is 41.16%.

Finally, the ARMA filtering is combined with our double Gaussian feature normalization method. The results are shown in Table 5. The relative performance is 48.04%. It is a significant improvement.

The obvious superiority of double Gaussian features normalization method to the MVN method is attributed to the utilization of bimodal characteristic in our method. The performance of our method is also comparable to that of the

Table 1: Baseline Results

Aurora2 Absolute Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	87.82	86.27	83.78	86.39
Clean	61.34	55.75	66.14	60.06
Average	74.58	71.01	74.96	73.23

Table 2: Results for MVN

Aurora2 Absolute Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	90.38	90.23	89.73	90.19
Clean	74.32	75.48	75.79	75.08
Average	82.35	82.86	82.76	82.63

Aurora2 Relative Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	21.03%	28.85%	36.72%	27.92%
Clean	33.57%	44.59%	28.51%	37.59%
Average	27.30%	36.72%	32.61%	32.76%

Table 3: Results for the Combination of Spectral Subtraction and HE (Cited from [5])

Aurora2 Absolute Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	90.89	89.80	90.11	90.30
Clean	82.51	82.78	81.87	82.49
Average	86.70	86.29	85.99	86.39

Aurora2 Relative Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	25.21%	25.72%	39.03%	28.18%
Clean	54.76%	61.08%	46.46%	55.63%
Average	39.99%	43.40%	42.75%	41.91%

Table 4: Results for Double Gaussian Feature Normalization

Aurora2 Absolute Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	90.93	91.02	90.52	90.88
Clean	78.68	81.03	79.34	79.75
Average	84.81	86.02	84.93	85.32

Aurora2 Relative Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	25.56%	34.56%	41.59%	33.01%
Clean	44.85%	57.14%	38.97%	49.30%
Average	35.21%	45.85%	40.28%	41.16%

Table 5: Results for
Double Gaussian Feature Normalization + ARMA Filtering

Aurora2 Absolute Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	91.67	91.48	91.53	91.57
Clean	82.47	84.15	82.99	83.24
Average	87.07	87.82	87.26	87.41

Aurora2 Relative Performance				
Training Mode	Set A	Set B	Set C	Overall
Multi	31.60%	37.97%	47.81%	38.04%
Clean	54.64%	64.17%	49.77%	58.04%
Average	43.12%	51.07%	48.79%	48.04%

combination method of spectral subtraction and HE [5]. Furthermore, better performance is obtained when ARMA temporal filtering procedure is combined with our method.

5. CONCLUSIONS

In this paper, we propose a new speech feature normalization algorithm based on CDF matching principle. The double Gaussian model is used in our algorithm as a parametric method to take the advantage of bimodal characteristics of speech features contaminated by noises.

On Aurora2 database, the performance of our double Gaussian based feature normalization method is much better than that of MVN method, and comparable to that of the combination method of spectral subtraction and HE.

As a standalone technique, our approach can already achieve quite satisfying performance that is comparable to the results of sophisticated compound methods. Moreover, significant improvement is observed when a simple ARMA filtering procedure is concatenated with our method. So it is reasonable to expect even better performance by combining our method with other front-end noise-robust techniques, such as spectral subtraction, Wiener filtering or VTS.

6. ACKNOWLEDGEMENT

This research is supported by the National Natural Science Foundation of China (No. 60275038).

7. REFERENCES

[1] A. de la Torre, J.C. Segura, and C. Benitez, etc. “Non-linear Transformations of the Feature Space for Robust Speech Recognition,” *Proc. ICASSP’02*, pp. 401-404, 2002.

[2] F. Hilger, H. Ney, “Quantile Based Histogram Equalization for Noise Robust Speech Recognition,” *Proc. EUROSPEECH’01*, pp. 1135-1138, 2001.

[3] F. Hilger, S. Molau, and H. Ney, “Quantile Based Histogram Equalization for Online Application,” *Proc. ICSLP’02*, pp. 237-240, 2002.

[4] F. Hilger, H. Ney, “Evaluation of Quantile Based Histogram Equalization with Filter Combination on the Aurora 3 and 4 Databases,” *Proc. EUROSPEECH’03*, pp. 341-344, 2003.

[5] J.C. Segura, M.C. Benitez, A. de la Torre, and A.J. Rubio, “Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization for Robust ASR,” *Proc. ICSLP’02*, pp. 225-228, 2002.

[6] J.C. Segura, M.C. Benitez, and A. de la Torre, etc. “VTS Residual Noise Compensation,” *Proc. ICASSP’02*, pp. 409-412, 2002.

[7] S. Molau, F. Hilger, D. Keysers, and H. Ney, “Enhanced Histogram Normalization in the Acoustic Feature Space,” *Proc. ICSLP’02*, pp. 1421-1424, 2002.

[8] S. Molau, F. Hilger, and H. Ney, “Feature Space Normalization in Adverse Acoustic Conditions”, *Proc. ICASSP’03*, pp. 656-659, 2003.

[9] S. Dharanipragada, M. Padmanabhan, “A Nonlinear Unsupervised Adaptation Technique for Speech Recognition,” *Proc. ICSLP’00*, pp. 556-559, 2000.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society*, vol.39, no.1, pp. 1-38. 1977.

[11] C.P. Chen, J. Bilmes, and K. Kirchhoff, “Low-resource Noise-robust Feature Post-processing on AURORA2.0,” *Proc. ICSLP’02*, pp. 2445-2448, 2002.

[12] H.G. Hirsch, D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” *ISCA ITRW ASR 2000*, 2000.