CEPSTRAL SHAPE NORMALIZATION (CSN) FOR ROBUST SPEECH RECOGNITION

Jun Du, Ren-Hua Wang

University of Science and Technology of China, Hefei, P. R. China, 230027

unuedjwj@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

In this paper, we propose a new feature normalization approach for robust speech recognition. It is found that the shape of speech feature distributions is changed in noisy environments compared with that in the clean condition. So cepstral shape normalization (CSN) which normalizes the shape of feature distributions is performed by exploiting an exponential factor. This method has been proven effective in noisy environments, especially under low SNRs. Experimental results show that the proposed method yields relative word error rate reductions of 38% and 25% on aurora2 and aurora3 databases, respectively, in comparing with those of the conventional mean and variance normalization (MVN). It is also shown CSN consistently outperforms other traditional methods, such as histogram equalization (HEQ) and higher order cepstral moment normalization (HOCMN).

Index Terms: robust speech recognition, shape normalization

1. INTRODUCTION

With the progress of automatic speech recognition (ASR), the noise robustness of speech recognizers attracts more and more attentions for practical recognition systems. Various noise robust technologies can be constructed either in the feature domain or the model domain [1]. In this paper, we focus on the feature domain. Quite several well-known normalization methods for feature domain have been developed. Cepstral mean normalization (CMN) is a simple but effective way to remove the time-invariant distortions introduced by the transmission channel. A natural extension of CMN is mean and variance normalization (MVN) [2] which normalizes both the mean and variance. So it can improve the robustness to additive noises, as well as the channel effects. Higher order cepstral moment normalization (HOCMN) [3] can be considered as the extension of CMN and MVN, where the mean and variance are related with the first and second moments, respectively. Double Gaussian normalization (DGN) [4] uses cumulative density functions (CDFs) matching principle under the assumption that the distributions of speech features in noisy environments are usually bimodal.

The above methods are based on parametric models. Non-parametric models can also be used in feature normalization, such as cumulative histogram used in histogram equalization (HEQ) [5]. The non-linear transformation of HEQ has shown its superiority over the linear compensation approaches, such as CMN and MVN. Several extensions, for example, quantile HEQ [6], progressive HEQ [7] and polynomial-fit HEQ [8], are also proposed recently. The advantage of HEQ methods is that they not only attempt to match speech feature mean and variance, but also completely match the feature distribution of the training and testing data.

The motivation of our cepstral shape normalization (CSN) method is based on the following two points. First, as reported in [9, 10, 11], modelling of speech feature distributions is discussed. Also our preliminary analysis shows that speech feature distributions of each dimension can be well approximated by employing generalized Gaussian density (GGD) model in noisy environments. Second, compared with traditional normalization methods, CSN has more directly physical meanings and the stronger pertinency. In HEQ and DGN, a larger mount of data is needed to accurately estimate the feature distribution. In HOCMN, the solution of odd order is approximative and not easy. In CSN, we only need to estimate the shape factor. Also the solution of the shape factor is accurate and simple.

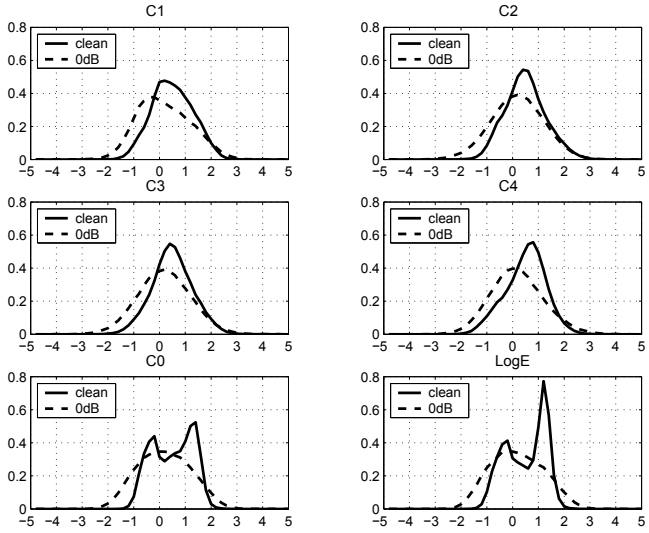

Fig. 1. *Distributions for MFCC C0-C4 and LogE compared between clean condition and noisy environments.*

The rest of this paper is organized as follows. In section 2, we propose the CSN method including its analysis and formulation. Experimental setup and results are discussed in section 3 and 4. Finally in section 5, we give our conclusions.

2. CEPSTRAL SHAPE NORMALIZATION (CSN)

2.1. Analysis for speech distribution

Before our algorithm is described, first we give some preliminary analysis of speech distribution. As shown in Figure 1, the distributions of different feature dimensions, are compared between clean condition and noisy environments. Here features are processed by MVN because we only focus on the shape of distributions. In the clean condition, the distributions of C0 and logE are bimodal. In

the case of other dimensions only one peak is observed. In noisy environments where SNR is 0dB, the shape of each dimensional distributions is changed. All the distributions, even for the C0 the logE, are Gaussian-like. Only the main difference among these distributions is the shape or skewness.

Generalized Gaussian density (GGD) model, which is introduced in [11] to approximate the distribution of speech features, is used here for speech modelling in noisy environments. For the signal x, with zero mean and unit variance, the PDF of GGD is defined as:

$$p_x(x|\nu) = \frac{\nu \mathrm{A}(\nu)}{2\Gamma(1/\nu)} \exp(-[\mathrm{A}(\nu)|x|]^{\nu}) \quad (1)$$

in which

$$\mathrm{A}(\nu) = \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}} \quad (2)$$

where $\Gamma(\cdot)$ defines the Gamma function given by:

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx, \quad z > 0 \quad (3)$$

$\mathrm{A}(\nu)$ defines the *dispersion* and *scale* of the distribution, while parameter ν describes the exponential rate of decay and, in general, the *shape* or *skewness* of the distribution $p_x(x|\nu)$. A well-known special case of the GGD function is a standard Gaussian or normal distribution($\nu = 2$). In effect, smaller values of the shape parameter ν correspond to heavier tails and therefore to more peaked distributions.

2.2. Algorithm description

Based on the analysis of section 2.1, Cepstral Shape Normalization (CSN) is summarized to the following two steps, which are applied to both training and testing data.

Step1: Cepstral parameters are processed by MVN.

$$y(n,k) = \frac{x(n,k) - \mu(k)}{\sigma(k)} \quad (4)$$

where $x(n,k)$ is the kth dimensional component of the original cepstral vector at frame time n; $\mu(k)$ and $\sigma(k)$ represent the mean and standard deviation of the current utterance for the kth dimension,respectively.

Step2: Shape normalization is performed by exploiting an exponential factor.

$$z(n,k) = [y(n,k)]^{\alpha(k)} \quad (5)$$

where $\alpha(k)$ denotes the shape factor for the kth dimension which is similar to the shape parameter ν in GGD. Our goal of transformation in Eq. (5) is to make the final processed features satisfy a fixed reference distribution, which is represented by GGD with a shape parameter ν_0.

In order to obtain the solution of shape factor $\alpha(k)$, we adopt *moment matching estimator*(MME) [11]. The rth-order absolute central moment for a GGD with a shape parameter ν_0 is given by:

$$E[|X|^r] = \int_{-\infty}^{+\infty} |x|^r p_x(x|\nu_0) dx \quad (6)$$

where $E[\cdot]$ represents the expectation operator. Substituting Eq. (1) into Eq. (6), we can further show that the rth-order moment is essentially defined as:

$$m_r = E[|X|^r] = \mathrm{A}^{-r}(\nu_0) \frac{\Gamma((r+1)/\nu_0)}{\Gamma(1/\nu_0)}, \quad \nu_0 > 0 \quad (7)$$

Then the *generalized Gaussian ratio function*, is used:

$$\mathcal{M}(\nu_0, r) = \frac{m_{2r}}{m_r^2} = \frac{\Gamma((2r+1)/\nu_0)\Gamma(1/\nu_0)}{\Gamma^2((r+1)/\nu_0)} \quad (8)$$

Based on Eq. (8), we define the following equation:

$$F(\alpha(k)) = \overline{\mathcal{M}}(\alpha(k), r) - \frac{\Gamma((2r+1)/\nu_0)\Gamma(1/\nu_0)}{\Gamma^2((r+1)/\nu_0)} = 0 \quad (9)$$

where $\overline{\mathcal{M}}(\alpha(k), r)$ can be estimated as follows:

$$\overline{\mathcal{M}}(\alpha(k), r) = \frac{\frac{1}{N}\sum_{n=1}^{N} |z(n,k)|^{2r}}{(\frac{1}{N}\sum_{n=1}^{N} |z(n,k)|^r)^2} \quad (10)$$

N is the frame number of the current utterance.

Obviously, $\alpha(k)$ is the root of Eq. (9). Although there is no close-form solution, $F(\alpha(k))$ is a monotonically increasing function of $\alpha(k)$. So a numerical solution for $\alpha(k)$ can be obtained by exploiting the *secant method*, which is a fast iterative method.

Finally, two free parameters, the shape parameter ν_0 and the moment order r, should be set. Some preliminary experiments show that the configuration of $\nu_0 = 2$ and $r = 2$, which is used in all experiments, achieves the best performance. $\nu_0 = 2$ represents the standard gaussian distribution, which is widely used as a fixed reference distribution in other normalization methods. The physical meaning of Eq. (8) ($r = 2$) is referred to as the *kurtosis*, which is the degree of peakedness of a distribution.

2.3. Temporal smoothing of the features

Though the above CSN approach is very effective in matching the global feature statistics of the testing (or noisy) speech to those of the training (or reference) speech, the undesired sharp peaks or valleys of the feature vector component sequence of a noisy speech utterance, which are caused by some non-stationary noise, can not be restored well to that of the original clean speech utterance. Therefore, a simple temporal Mth-order ARMA filter [12] is included in our experiments.

3. EXPERIMENTAL SETUP

Our experiments are performed on both aurora2 and aurora3 databases. The aurora2 task consists of English digits in the presence of additive noise and linear convolutional distortion. These distortions have been synthetically introduced to clean TIDigits data. Two training conditions (clean-condition/multi-condition) and three testing sets (sets A/B/C) are defined by aurora2. Only clean training condition is considered here because it represents a more serious mismatch situation and requires more robust speech features.

The aurora3 task consists of Danish, German, Spanish, and Finnish digits in realistic automobile environments. Three experiments are defined for the evaluation: well-matched, high-mismatch, and mid-mismatch. The experiment names refer to the relationship between the testing and training data.

The speech features, including 14 cepstral coefficients (MFCCs, C0-C12 plus the log-energy), are produced by the reference WI007 front-end. All the normalization methods are applied to these static features. Only one of C0 and log-energy will be preserved in the model training. Then the first and second derivatives are computed through the processed features. HMMs are trained in the manner prescribed by the scripts included with the aurora task. The details of the two databases, baseline front-end and back-end can be found in [13, 14].

Method	Word Error Rate (%) of Clean Condition Training				Relative Error Rate Reduction
	Set A	Set B	Set C	Avg.	
MVN	29.82	29.23	33.63	30.35	-
DGN	21.73	20.46	21.92	21.26	30.0%
HEQ	20.14	19.19	19.57	19.65	35.3%
HOCMN	19.75	18.76	20.87	19.58	35.5%
CSN	19.13	18.35	19.16	18.82	38.0%

Table 1. *Performance comparison of CSN with several normalization methods for different testing sets in clean condition training on aurora2.*

Method	20dB	15dB	10dB	5dB	0dB	-5dB
MVN	4.10	8.79	20.47	43.94	74.43	89.41
DGN	3.34	6.07	12.61	27.48	56.80	84.03
HEQ	3.74	6.29	12.12	24.71	51.37	82.62
HOCMN	2.97	5.20	10.80	24.22	54.71	84.50
CSN	3.67	6.07	11.88	24.22	48.27	76.74

Table 2. *Performance (Word Error Rate) comparison of CSN with several normalization methods for different SNRs in clean condition training on aurora2.*

4. EXPERIMENTAL RESULTS

4.1. Comparison of CSN with other normalization methods

In this section, LogE(log-energy) is used for all experiments. We compare the performances of five feature normalization methods(MVN, DGN, HEQ, HOCMN, CSN). MVN can be considered as the baseline. For HOCMN, the odd and even order are set to 3 and 4, respectively.

As shown in Table 1, CSN consistently achieves the best performance among these methods for different testing sets on aurora2. Compared with MVN, our method yields relative WER(word error rate) reduction of 38.0% for the average of all testing sets.

From the viewpoint of SNRs, the performances of different methods are also compared in Table 2. It can be found that for high SNRs, CSN is comparable to other methods such as HOCMN and HEQ. But when the SNR is below 5dB, The performance of CSN is much better than the others.

Moreover, the effectiveness of different normalization approaches can also be observed with the following average distance measure:

$$d = E\left[\frac{||\bar{\mathbf{y}} - \bar{\mathbf{x}}||}{||\bar{\mathbf{x}}||}\right] \quad (11)$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the 13-dimensional vectors of MFCC parameters for clean and noisy speech processed by a certain normalization method, respectively. $||\cdot||$ is the Euclidean distance, and the average $E[\cdot]$ is performed over all testing utterances on aurora2, including all different noise types but separated for different SNRs. The distance measure d reflects how the normalized feature vectors are "individually" matched to their clean speech versions. The results of the distance measure d are listed in Table 3. The conclusion is that the measure d for CSN-processed features is consistently smaller than others under low SNRs($<$5dB), which is almost the same as that of the WER. But there are two main differences between distance measure and WER. First, the relative reduction of distance measure is much smaller than that of WER. Second, the reduction of WER do not necessarily result in the reduction of distance measure, which can be observed by comparing the rows of MVN and HEQ in Table 2 and 3 respectively.

Method	20dB	15dB	10dB	5dB	0dB	-5dB
MVN	0.821	0.908	1.000	1.098	1.203	1.306
DGN	0.841	0.926	1.014	1.107	1.205	1.303
HEQ	0.855	0.939	1.025	1.115	1.211	1.307
HOCMN	0.818	0.901	0.988	1.081	1.180	1.279
CSN	0.826	0.907	0.990	1.081	1.178	1.277

Table 3. *Distance measure comparison of CSN with several normalization methods for different SNRs on aurora2.*

Aurora 3 Word Error Rate (MVN)					
	Finnish	Spanish	German	Danish	Average
Well (x40%	6.69%	5.47%	6.45%	10.28%	**7.22%**
Mid (x35%)	13.89%	9.20%	14.28%	23.59%	**15.24%**
High (x25%	53.11%	27.70%	18.36%	43.85%	**35.76%**
Overall	**20.82%**	**12.33%**	**12.17%**	**23.33%**	**17.16%**

Aurora 3 Word Error Rate (DGN)					
	Finnish	Spanish	German	Danish	Average
Well (x40%	6.00%	4.73%	6.11%	9.44%	**6.57%**
Mid (x35%)	12.45%	8.83%	12.96%	22.60%	**14.21%**
High (x25%	36.29%	17.35%	14.66%	38.40%	**26.68%**
Overall	**15.83%**	**9.32%**	**10.65%**	**21.29%**	**14.27%**

Aurora 3 Word Error Rate (HEQ)					
	Finnish	Spanish	German	Danish	Average
Well (x40%	6.50%	4.84%	5.85%	9.71%	**6.72%**
Mid (x35%)	13.20%	8.56%	12.15%	23.73%	**14.41%**
High (x25%	27.63%	15.73%	13.64%	35.11%	**23.03%**
Overall	**14.13%**	**8.86%**	**10.00%**	**20.97%**	**13.49%**

Aurora 3 Word Error Rate (HOCMN)					
	Finnish	Spanish	German	Danish	Average
Well (x40%	6.70%	4.49%	6.41%	9.83%	**6.86%**
Mid (x35%)	13.06%	8.52%	14.20%	24.29%	**15.02%**
High (x25%	36.61%	21.11%	15.73%	42.08%	**28.88%**
Overall	**16.40%**	**10.06%**	**11.47%**	**22.95%**	**15.22%**

Aurora 3 Word Error Rate (CSN)					
	Finnish	Spanish	German	Danish	Average
Well (x40%	6.33%	4.58%	5.99%	9.35%	**6.56%**
Mid (x35%)	13.41%	8.45%	13.10%	21.75%	**14.18%**
High (x25%	24.20%	20.18%	11.56%	28.64%	**21.15%**
Overall	**13.28%**	**9.83%**	**9.87%**	**18.51%**	**12.87%**

Table 4. *Performance comparison of CSN with several normalization methods on aurora3*

Table 4 presents WER results for different methods on aurora3. For all three experiments including well-matched, mid-mismatch and high-mismatch, the average performances of CSN are consistently better than those of other methods, especially for high-mismatch case. And a 25% relative reduction in overall WER is achieved when compared with MVN. Also consistent improvements of CSN for each language are obtained except for the case of Spanish language in DGN and HEQ. Furthermore, the rank of overall performances of DGN, HEQ and HOCMN is not consistent on aurora2 and aurora3 as shown in Table 1 and 3. So our CSN method, which achieves the best performance on both aurora2 and aurora3, is more stable.

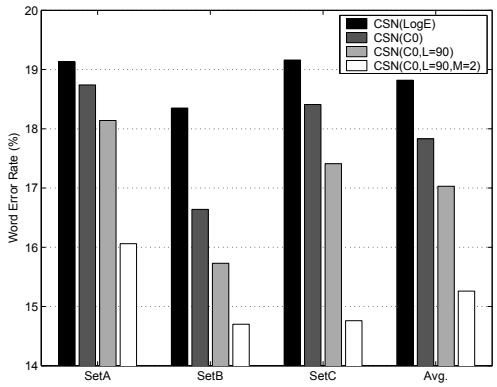

Fig. 2. *Performance comparison of CSN with several techniques applied on aurora2*

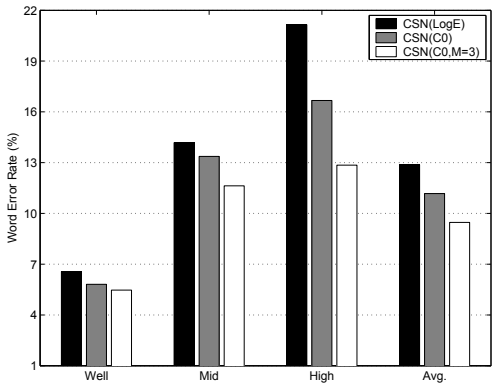

Fig. 3. *Performance comparison of CSN with several techniques applied on aurora3*

4.2. Further improvements with CSN

To achieve further improvements with CSN, several techniques are applied as follows. 1) LogE is replaced by C0, which is considered more robust in noisy conditions. 2) The implementation for CSN can be in the manner of full-utterances or segments. Our experiments in section 4.1 exploits the former. In the latter, or by segments with length L, the summation in Eq. (10) is performed by a segment of frames including the preceding $\frac{L}{2}$ frames and following $\frac{L}{2}$ frames. $L = 90$ is properly used for aurora2 experiments. For aurora3, the effectiveness of segmental implementation is not obvious, which is not reported here. 3) the ARMA filtering mentioned in section 2.3 is combined with CSN. The order M is set to 2 and 3 for aurora2 and auror3, respectivley.

The results are listed in Figure 2 and 3. With the above techniques, significant improvements are achieved for both aurora2 and aurora3, especially in mismatch case. We obtain relative WER reductions of 18.9% and 26.4%, which are the best average performances on aurora2 and aurora3, respectively, compared with the original CSN without using any techniques.

5. CONCLUSIONS

The CSN algorithm, which normalizes the shape of feature distributions as described in this paper, is an efficient feature normalization method for robust speech recognition. The performances of our method on aurora2 and aurora3 are significantly improved compared with the MVN method, and also consistently better than those of other traditional normalization methods. Further improvements of performance are achieved by several techniques such as concatenating a simple ARMA filtering procedure. In our future work, it is reasonable to expect even better performance by combining our method with other noise-robust techniques.

6. REFERENCES

[1] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, Vol. 16, No. 3, pp. 261-291, 1995.

[2] O. Viikki and K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Communication*, Vol. 25, No. 1, pp. 133-147, 1998.

[3] C.-W. Hsu and L.-S. Lee, "Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition," *Proc. of ICASSP*, I-197-I-200, 2004.

[4] B. Liu, L.-R. Dai, J.-Y. Li and R.-H. Wang, "Double Gaussian Based Feature Normalization for Robust Speech Recognition," *Proc. of ISCSLP*, pp. 253-256, 2004.

[5] *Á*. de la Torre, J.C. Segura, C. Benitez, A.M. Peinado and A.J. Rubio, "Non-linear Transformations of the Feature Space for Robust Speech Recognition," *Proc. of ICASSP*, I-401-I-404, 2002.

[6] F. Hilger and H. Ney, "Quantile Based Histogram Equalization for Noise Robust Speech Recognition," *Proc. of EUROSPEECH*, pp. 1135-1138, 2001.

[7] S.-N. Tsai and L.-S. Lee, "A New Feature Extraction Front-End for Robust Speech Recognition using Progressive Histogram Equalization and Multi-Eigenvector Temporal Filtering," *Proc. of ICSLP*, pp. 165-168, 2004.

[8] S.-H. Lin, Y.-M. Yeh and B. Chen, "Exploiting Polynomial-fit Histogram Equalization and Temporal Average for Robust Speech Recognition," *Proc. of ICSLP*, pp. 2522-2525, 2006.

[9] S. Gazor and W. Zhang, "Speech Probability Distribution," *IEEE Signal Processing Letters*, Vol. 10, No. 7, pp. 204-207, 2003.

[10] J.W. Shin, J.-H. Chang and N.S. Kim, "Statistical Modeling of Speech Signals Based on Generalized Gamma Distribution," *IEEE Signal Processing Letters*, Vol. 12, No. 3, pp. 258-261, 2005.

[11] K. Kokkinakis and A.K. Nandi, "Speech Modelling Based on Generalized Gaussian Probability Density Functions," *Proc. of ICASSP*, I-381-I-384, 2005.

[12] C.-P. Chen, J. Bilmes and K. Kirchhoff, "Low-Resource Noise-robust Feature Post-processing on Aurora 2.0," *Proc. of ICSLP*, pp. 2445-2448, 2002.

[13] H.G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *Proc. of ISCA ITRW ASR*, pp. 181-188, 2000.

[14] A. Moreno, et al., "SpeechDat-Car: A Large Speech Database. for Automotive Environments," *Proc. of LREC*, pp. 373-378, 2000.