NORMALIZATION OF TIME-DERIVATIVE PARAMETERS USING HISTOGRAM EQUALIZATION

Yasunari Obuchi[1] *and Richard M. Stern*
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{obuchi, rms}@cs.cmu.edu

Abstract

In this paper we describe a new framework of feature compensation for robust speech recognition. We introduce Delta-Cepstrum Normalization (DCN) that normalizes not only cepstral coefficients, but also their time-derivatives. In previous work, the mean and the variance of cepstral coefficients are normalized to reduce the irrelevant information, but such a normalization was not applied to time-derivative parameters because the reduction of the irrelevant information was not enough. However, Histogram Equalization provides better compensation and can be applied even to delta and delta-delta cepstra. We investigate various implementation of DCN, and show that we can achieve the best performance when the normalization of the cepstra and delta cepstra can be mutually interdependent. We evaluate the performance of DCN using speech data recorded by a PDA. DCN provides significant improvements compared to HEQ. We also examine the possibility of combining Vector Taylor Series (VTS) and DCN. Even though some combinations do not improve the performance of VTS, it is shown that the best combination gives better performance than VTS alone. Finally, the advantages of DCN in terms of the computation speed are also discussed.

1. Introduction

Speech signals are a mixture of various information. A part of the speech signal is produced according to the speaker's intention, while another part is introduced by the environment and not helpful to the understanding of the content of speech. If the former part is dominant, speech recognition is easy. If not, robust speech recognition techniques must be introduced to reduce the effects of irrelevant information. Since the separation of relevant and irrelevant information is not apparent, those techniques need to use prior knowledge or assumptions about the speech and the acoustical environment.

Cepstral Mean Normalization (CMN) [1] is a well-known method to reduce the environmental distortion. It assumes that the mean of the cepstral coefficients is invariant for various utterances. Therefore, there is no relevant information in the mean, and subtracting it reduces only irrelevant information. In CMN, the irrelevant information is assumed to be convolutional channel noise or spectral tilt. In some cases, such a strong assumption may cause a loss of relevant information, but the greater reduction of irrelevant information in adverse conditions generally results in better performance. A natural extension of CMN is Mean and Variance Normalization (MVN) [2,3], where the assumption is still stronger. In MVN the mean and the variance of the cepstral coefficients of clean speech are assumed to be invariant. Therefore, removing mean and variance is assumed to reduce only irrelevant information, no matter what that information may be. A third technique, Histogram Equalization (HEQ) [4,5] uses the stronger assumption that the shape of the entire distribution of cepstral coefficients is invariant. In HEQ, any detail of the cepstral distribution is regarded irrelevant and to be removed.

From this perspective, we can say that any normalization can be applied to any parameter if the invariance assumption is valid. That is, the motivation of our work, in which we try to apply normalization techniques not only to cepstral parameters, but also to their time-derivatives. Although it is true that the cepstral mean can be interpreted as spectral tilt, we do not pay much attention to the origin of the irrelevant information in applying the other methods. Instead, we focus only on finding transformations that preserve the relevant information in speech. This paper compares and discusses such simple models and transformations, and shows that speech recognition performance can be improved by their use. More importantly, these transformations are extended to the delta cepstra.

The remainder of this paper is organized as follows. In the next section, we describe the concept of HEQ and our implementation of it. In section 3, various versions of Delta-Cepstrum Normalization (DCN) are introduced. Section 4 presents experimental results, and conclusions are given in the last section.

2. Histogram Equalization

Histogram Equalization is a procedure that is commonly used in image processing. Balchandran and Mammone [6] first applied it to the amplitudes of speech signals, and Dharanipragada and Padmanabhan [7] applied it to cepstral features as an adaptation method. Some more recent papers (*e.g.* [4,5]) applied feature normalization methods for robust speech recognition.

The basic idea of HEQ is that the distribution of cepstral coefficients in the test data should be identical to that of the training data. In the case where we can treat each dimension of the cepstral vector as independent, finding the transformation is easy by using the cumulative density function (CDF), the integral of the probability density function (PDF). Since the CDF is a monotonic increasing function between 0 and 1, the inverse function can be defined. Thus, the transformation of HEQ is defined as follows:

[1] Currently at Advanced Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo 185-8601, Japan

$$x_i = HEQ(y_i) = C_X^{-1}(C_Y(y_i)) \quad (1)$$

where C_X is the CDF estimated from training data and C_Y is the CDF of the test data, y_i is a cepstral coefficient of the i^{th} frame, and x_i is the corresponding transformed cepstral coefficient. Since HEQ is applied to each cepstral dimension independently, we omit the other subscript for the cepstral dimension in this paper.

Usually there is a huge number of samples in the training data, and we can get an almost continuous curve of the CDF from the precise histogram. The number of samples in a test utterance is small, but we can define the CDF at sample points simply by sorting the cepstral parameters and obtaining their relative ranks, because the CDF is a function of the number of frames that have smaller values than the current point. After sorting, we calculate $C_X^{-1}(t/N)$ for t = 0, 1, 2, ... N (where N is the number of frames) by interpolation using the pre-stored numeric table of C_X^{-1}.

There are some issues in the implementation of HEQ. In [4], the CDF obtained from the Gaussian PDF was used as the reference. Even though the distribution of cepstral coefficients tends to be Gaussian in some cases, we made the reference CDF according to (1) to make it more precise. Another issue is whether MVN should be applied to the training data before obtaining the reference CDF. We thought that HEQ should be a natural extension of MVN, so we applied MVN to the training data before developing the CDF. There is also a concern about the domain of HEQ. In [5], it is said that applying HEQ in the Mel-filterbank domain is better than applying it in the cepstral domain. However, our preliminary experiments showed the opposite results, so we decided to apply it in the cepstral domain.

3. Delta-Cepstrum Normalization

It is well known that the use of time-derivative parameters such as delta and delta-delta cepstra improves recognition accuracy. However, there have been few previous studies that attempt to normalize these features. The RASTA method [8] and other filtering approaches make use of inter-frame information, but they do not use the entire distribution of delta parameters. Mean subtraction of delta parameters does not help because the mean of delta parameters is always zero by definition. The variance of delta parameters can be non-zero, but it was reported in [3] that MVN does not need to be applied to delta and delta-delta cepstra. It is possible that the improvement obtained using MVN is smaller than the loss of relevant information. However, if compensation using HEQ provides more gain than loss, we could have different results.

Since the delta and delta-delta cepstra are not independent from the cepstrum, there are several ways with which these features could be compensated. Figure 1 shows three types of Delta-Cepstrum Normalization. The simplest option is called Independent DCN, where the delta and delta-delta cepstra are calculated from the original cepstrum, and then HEQ is applied to the cepstrum, the delta cepstrum and the delta-delta cepstrum independently. The second option is called Sequential DCN, where HEQ is applied to the original cepstrum, then time-derivative operation is carried out using the normalized cepstrum, and finally HEQ is applied to the delta and delta-delta cepstra.

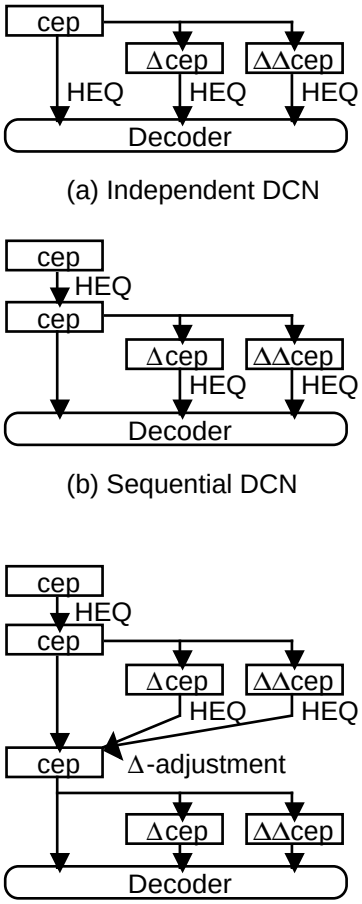

Fig. 1. *Schematic diagram of DCN. (a) Independent DCN. (b) Sequential DCN. (c) Feedback DCN*

In this method, the delta and delta-delta cepstra part can take advantage of the normalization of the cepstrum. The third option is called Feedback DCN, where the output of Sequential DCN is fed back to the cepstrum part, and "Δ-adjustment" is executed. Δ-adjustment is a procedure described in more detail below that reduces the mismatch between the normalized cepstrum and the normalized delta and delta-delta cepstra. By introducing Δ-adjustment, the cepstral normalization can take advantage of the normalization of the delta and delta-delta cepstra. However, even though both delta and delta-delta cepstra are expected to be helpful, we perform Δ-adjustment using the delta-cepstrum only, because it is difficult to define an appropriate Δ-adjustment procedure that makes use of both delta and delta-delta. A more detailed description of Feedback DCN including Δ-adjustment follows.

In Feedback DCN we describe the observed cepstral coefficients by y_i. After applying HEQ, we obtain normalized cepstral coefficients z_i.

Table 1. *Recognition results for real data*

	WER (%)
Baseline (CMN)	41.5
MVN	33.5
HEQ	30.2
Independent DCN	27.5
Sequential DCN	27.0
Feedback DCN	25.6
Close-talk	16.4

Table 2. *Recognition results for real data using combinations with VTS*

	WER (%)
VTS (CMN)	23.3
VTS + MVN	25.4
VTS + HEQ	27.3
VTS + Independent DCN	23.4
VTS + Sequential DCN	23.6
VTS + Feedback DCN	22.7

$$z_i = HEQ(y_i) \quad (2)$$

Delta-cepstral coefficients are defined as follows.

$$\Delta z_i = \frac{1}{2}(z_{i+1} - z_{i-1}) \quad (3)$$

The error function is then defined to be the difference between the original delta cepstrum and the normalized delta cepstrum:

$$e_i = HEQ(\Delta z_i) - \Delta z_i \quad (4)$$

Finally, the cepstrum is modified so that the error function decreases.

$$x_i = z_i - \alpha(e_{i+1} - e_{i-1}) \quad (5)$$

α is a weight parameter, whose value was set to 1 empirically. Using these values of x_i, the delta and delta-delta cepstra are recalculated, and the resulting parameters are fed into the decoder.

4. Experimental Results

The proposed algorithms were evaluated in a series of recognition experiments. Triphone HMMs with 2000 tied states (8 Gaussians/state) were trained using the 5000-word LDC Wall Street Journal database (WSJ0). The Sphinx-III decoder developed by CMU was used for decoding, with a trigram language model. Speech input was sampled by 11.025kHz, and 13 MFCCs were computed every 10ms.

We recorded 330 utterances from eight speakers simultaneously using two microphones: the built-in microphone of the PDA (Compaq iPAQ PocketPC Model 3630) and a close-talk microphone (Optimus Nova 80). Each speaker uttered 40 to 43 sentences chosen from the WSJ0 database. Using these recordings, we prepared two test sets. The first set was the real data recorded by the PDA microphone. The SNR of the first set was estimated as 18dB using NIST's stnr tool. The data are corrupted by both additive noise from computer fans in the recording room and the spectral tilt of the PDA microphone. A second set of artificial data were obtained by digitally adding the relatively clean speech data recorded using the close-talk microphone to noise recorded by the PDA microphone with varying SNR from 0dB to 25dB. The spectral tilt of the close-talk microphone is small, and the additive noise is the same as the first set except that the amplitude is adjusted to each SNR value.

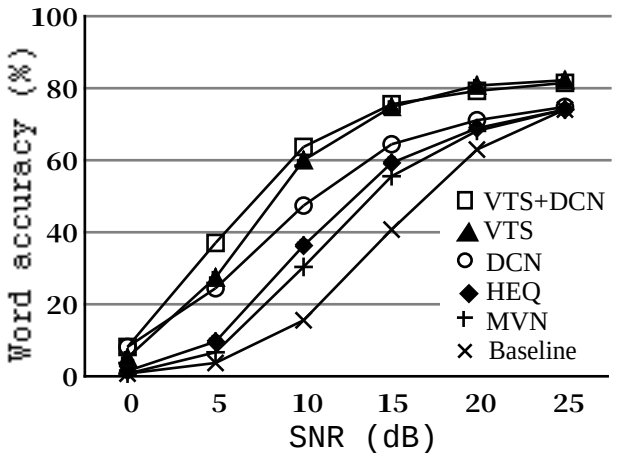

Fig. 2. *Recognition results for artificial data*

4.1 Experiments Using Real Data

Table 1 shows the word error rates (WER) obtained by various methods using the real data set. MVN and HEQ improve the accuracy as expected. Independent DCN provides an improvement over HEQ, that is 9% relative WER reduction. Sequential DCN works slightly better than Independent DCN, providing a relative improvement in WER of 11%. Finally, Feedback DCN results in the best performance, providing a relative reduction in WER of about 15% compared to HEQ. As the reference, the WER using the close-talk microphone recording with no additional noise is 16.4%, and that is regarded as the lower limit in WER to be obtained by any similar compensation method.

4.2 Combination with VTS

VTS [9] is known as one of the most powerful compensation algorithms developed for quasi-stationary additive noise and linear filtering. In [10], it is reported that HEQ reduces

Table 3. *Execution time for 1 second speech*

	time (sec)
MVN	0.0001
HEQ	0.0012
Independent DCN	0.0033
Sequential DCN	0.0033
Feedback DCN	0.0019
VTS	2.8395

the residual noise of VTS, so one can achieve better results by applying HEQ after VTS. To verify this result and check its extensibility to DCN, we performed some additional experiments using VTS.

Table 2 shows the word error rates obtained using VTS as well as various combination of VTS and other methods. Although VTS by itself works better than even the best form of DCN, the WER becomes greater when we apply HEQ after VTS, that is opposite to the result described in [10]. Independent DCN and Sequential DCN are better than HEQ, but they are still worse than VTS alone. However, if we apply Feedback DCN after VTS, we obtain a relative improvement in WER of about 3% compared to VTS alone.

4.3 Experiments Using Artificial Data

Figure 2 shows the recognition accuracy obtained using the artificial data using various SNRs. Since Feedback DCN was the best among three types of DCN in the previous experiments, we used Feedback DCN only. VTS in combination with DCN was also tested.

As shown in the figure, the use of DCN does not improve recognition accuracy over the results obtained with VTS for SNRs above about 10dB. DCN is more helpful when used with VTS at lower SNRs, and is even better by itself than VTS at 0dB.

4.4 Computational Complexity

One of the advantages of HEQ is fast execution owing the possibility of being implemented via table lookup. On the other hand, EM-based algorithms such as VTS are usually very slow. To confirm the same advantage for DCN, we measured the time consumed by the CPU to compensate 330 utterances of the real data set, and calculated the average time to compensate one second of speech. The experiment was carried out with an Intel Celeron 2.0GHz processor and 256MB memory running on the Linux operating system. Execution times for the various algorithms are shown in Table 3. In Independent and Sequential DCN, there are three equalization operations for the cepstrum, the delta cepstrum, and the delta-delta cepstrum. That is why it takes approximately three times as much time as HEQ. In Feedback DCN, we did not apply HEQ to the delta-delta cepstrum, so the execution time is about twice that of HEQ. Apart from those small differences, all of three DCN algorithms ran in less than 1% of real time. In contrast, VTS requires much more than real time due to its time-consuming EM iterations.

5. Conclusions

In this paper, we have introduced a new feature normalization algorithm that is based on the normalization of time-derivative parameters. This procedure, referred to as Delta-Cepstrum Normalization (DCN) is quite simple to implement and provides greater recognition accuracy than either Cepstral Mean Normalization (CMN) or Histogram Equalization (HEQ). The performance of DCN approached that of Vector Taylor Series (VTS) and with only of a small fraction of the computational cost of VTS. We investigated three implementations of DCN, Independent, Sequential, and Feedback Delta-Cepstrum Normalization. The best implementation, Feedback DCN, provide a relative improvement of 15% compared to standard HEQ using real data recorded by the built-in microphone of an iPAQ. We also showed that Feedback-DCN can reduce recognition error rate when it is applied after VTS.

Fast run times for the HEQ and DCN algorithms are observed when the algorithms are implemented using table lookup. As a result, these algorithms are attractive for small devices used in noisy conditions, such as PDAs and in-vehicle systems.

References

[1] B. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," Journal of the Acoustical Society of America, vol. 55, pp.1304-1312, 1974

[2] J. Openshaw and J. Mason, "On the Limitations of Cepstral Features in Noise," Proc. of ICASSP 1994

[3] P. Jain and H. Hermansky, "Improved Mean and Variance Normalization for Robust Speech Recognition," Proc. of ICASSP 2001

[4] A. de la Torre, J. Segura, C. Benitez, A. Peinado, and A. Rubio, "Non-linear Transformation of the Feature Space for Robust Speech Recognition," Proc. of ICASSP 2002

[5] S. Molau, M. Pitz, and H. Ney, "Histogram Based Normalization in the Acoustic Feature Space," Proc. of ASRU 2001

[6] R. Balchandran and R. Mammone, "Non-parametric Estimation and Correction of Non-linear Distortion in Speech systems," Proc. of ICASSP 1998

[7] S. Dharanipragada and M. Padmanabhan, "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition," Proc. of ICSLP 2000

[8] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP Speech Analysis," ICSI Technical Report TR-91-069, UC Berkeley, 1991

[9] P. Moreno, B. Raj, and R. Stern, "A Vector Taylor Series Approach For Environment Independent Speech Recognition," Proc. of ICASSP 1996

[10] J. Segura, M. Benitez, A. de la Torre, S. Duponi, and A. Rubio, "VTS Residual Noise Compensation," Proc. of ICASSP 2002