MEAN NORMALIZATION OF POWER FUNCTION BASED CEPSTRAL COEFFICIENTS FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENT

Soonho Baek [1], *Hong-Goo Kang* [2]

School of Electrical and Electronic Engineering
Yonsei University, Seoul, Korea
[1]bestboybsh@dsp.yonsei.ac.kr, [2]hgkang@yonsei.ac.kr

ABSTRACT

This paper presents the effect of mean normalization to various types of cepstral coefficients for robust speech recognition in noisy environments. Although the cepstral mean normalization (CMN) technique was originally designed to compensate channel distortion, it has also been proved that the CMN also improves recognition accuracy in additive noisy environment. However, no one has yet considered the interaction of CMN with spectral mapping functions required for extracting cepstral features. This paper investigates the impact of CMN to the speech recognition system depending on the types of spectral mapping function by mathematically analyzing the amount of spectral distortion between clean and noisy conditions. The analytic result is also confirmed by comparing the type of recognition error patterns in automatic speech recognition experiment with Aurora 2 database. Experimental results show that the performance improvement by adopting CMN becomes significant if the logarithmic function is replaced with the appropriate setting of fractional power mapping function. Especially, the deletion errors are dramatically reduced.

Index Terms— Robust speech recognition, CMN

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems severely degrades in noise environment mainly because of the mismatching between training and test conditions. Many algorithms have been proposed to minimize the mismatch by taking signal domain, feature domain, or model domain approaches [1]-[13]. Speech enhancement techniques such as Wiener filter and spectra subtraction are typical examples for the signal domain approach [5][6]. Introducing a new feature or a compensation processing to features such as power normalized cepstral coefficients (PNCCs) and relative spectral perceptual linear predictive (RASTA-PLP) is an example of the feature domain approach [1][2][7]. Adapting statistical model parameters to match the ASR system with distorted environment such as using a vector Taylor series (VTS) belongs to the model domain approach [9][10].

Although recent works on robust speech recognition are somewhat biased to model domain approaches, both signal and feature domain approaches are still important because the unified or merged approach could be more beneficial to overcome the ASR problems in noisy environment. This paper focuses on the feature domain algorithm, especially tries to analyze the impact of mean normalization to various types of cepstral coefficients.

The mel-frequency cepstral coefficient (MFCC) has been popularly used for ASR systems thanks to its good performance in various environments [15]. Although many researchers claimed that their features were better than MFCC in certain conditions, they could not be used for wide applications. As the important of robustness in noisy environments becomes increased, however, a slightly modified version of MFCC has been applied to real applications nowadays. The PNCC that utilizes the fractional power function instead of the natural logarithm function used for MFCC is a typical example [1][2]. It proved that the recognition accuracy could be improved in noisy conditions compared to conventional MFCC and PLP features. By excluding the additional noise reduction module that the PNCC extraction process requires, it was verified that the recognition accuracy increased only if the CMN process was included [12]. In other words, only replacing the spectral non-linear mapping function with the fractional power function was not meaningful in noisy conditions. From the result, it is clear that there is a close interaction between the type of spectral mapping function used and the CMN technique. However, the interaction between CMN and spectral mapping functions has never been considered in earlier studies.

This paper further investigates the impact of CMN to the ASR system depending on the types of the spectral mapping function. At first, the spectral mapping function is extended to the form of generalized logarithmic function to make the mathematical analysis be easier. Then, the spectral distortion between noisy and clean speech is measured. As we can expect from the characteristic of generalized logarithmic function, the amount of spectral distortion varies depending on the type of spectral mapping functions and SNR conditions. In

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

addition, the distribution of recognition error types is investigated when CMN is combined or not. To verify the analysis results, an ASR system is designed and tested with Aurora 2 database in various noisy environments. Experimental results confirm that the amount of performance improvement varies depending on the type of spectral mapping function. The best performance can be obtained by properly setting the type of non-linear spectral mapping function with including CMN, of which conclusion is in the same line with the previous study that the power normalized cepstral coefficients (PNCCs) show better performance than MFCCs only if the CMN processing is combined.

The layout of the paper is as follows. Section 2 shows the performance comparison of MFCC and PNCC. In Sec. 3, the fractional power function based ASR system is represented. Section 4 analyzes the spectral distortion between the noisy speech and the clean speech. The recognition error patterns are described in Sec. 5. Finally, conclusion is included in Section 6.

2. COMPARISON BETWEEN MFCCS AND PNCCS

The procedure for extracting PNCCs includes the three main modules [1][2]: the noise reduction, the replacement of spectral mapping function, and the CMN. The impact of each module to word accuracy in noisy environment is compared to MFCC based ASR system in Fig.1. It indicates that the ASR performance is significantly improved by applying the noise reduction (denoted NR in figure) or CMN technique. However, when the CMN is not included, there is no big differences between the MFCC based system and the PNCC based system. This means that the replacement of spectral mapping function does not improve the ASR performance when CMN is not included.

Next section investigate the impact of mean normalization to the power spectral mapping function based ASR system by measuring the spectral distortion between the noisy speech and the clean speech. In addition, the error patterns are also analyzed in Sec.5.

3. POWER FUNCTION BASED ASR SYSTEM

As a spectral mapping function, the generalized logarithmic function that is a general form of the logarithmic function is adopted. In this section, the generalized logarithmic function is described, and its impact to the spectrum in noisy environments is analyzed.

3.1. Generalized Logarithmic Function

In [14], the generalized logarithmic function is defined by

$$f_\gamma(x) = \frac{1}{\gamma}\left(x^\gamma - 1\right), \quad \gamma \neq 0, \tag{1}$$

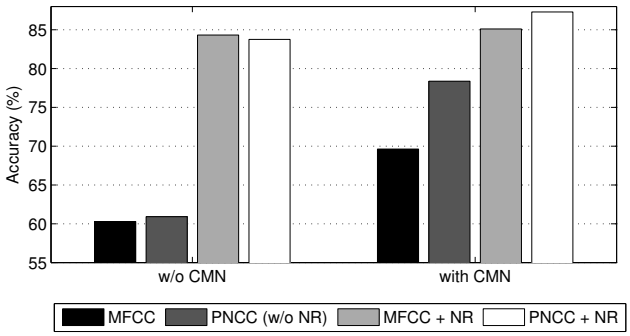

Fig. 1. Comparing the ASR performance.

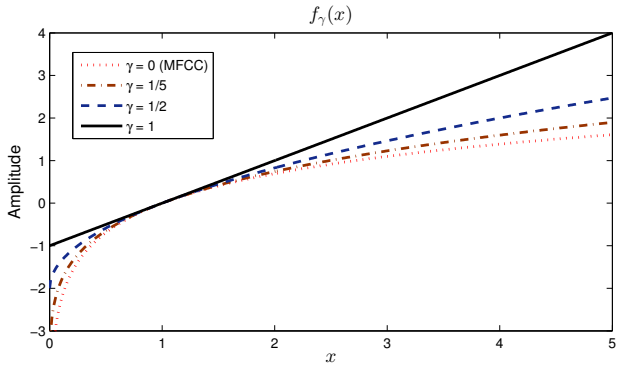

Fig. 2. Generalized logarithmic function for various value of γ.

where γ is a real value of $|\gamma| \leq 1$. if γ approaches 0, the generalized logarithmic function $f_\gamma(x)$ is equivalent to the natural logarithmic function, which was mathematically proved in [14]:

$$\lim_{\gamma \to 0} f_\gamma(x) = \log x. \tag{2}$$

Fig.2 shows the curves of a generalized logarithmic function $f_\gamma(x)$ for several values of γ. It is observed that if γ is close to zero, the corresponding curve is close to the natural logarithmic function. In this paper, γ is set to $1/15$, which is the value of maximizing the performance of PNCC in [2].

3.2. Generalized logarithmic spectrum of noisy speech

Given the spectrum of the clean speech $X(l)$, the additive noise $D(l)$, and the channel noise $H(l)$, the spectrum of observed signal $Y(l)$ at l^{th} frame is represented by

$$Y(l) = X(l)H(l) + D(l), \tag{3}$$

where l denotes the frame index. For simplicity, the frequency bin index is omitted. If the noise reduction process is in-

cluded, the spectrum of enhanced speech is represented by

$$\begin{aligned} Y_{NR}(l) &= X(l)H(l) + D(l) - N(l) \\ &= X(l)H(l) + e(l), \end{aligned} \tag{4}$$

where $N(l)$ and $e(l)$ are the estimated noise components and the residual noise components, respectively. Since the residual noise is assumed as an additive term, the signal model of including the noise reduction module is equivalent to Eq.(3).

After taking the generalized logarithmic function and some manipulation, the generalized log-spectrum in noisy environments is obtained by

$$\begin{aligned} Y^{\gamma}(l) &= f_{\gamma}(Y(l)) \\ &= X^{\gamma}(l) + N_a^{\gamma}(l) + N_c^{\gamma}(l), \end{aligned} \tag{5}$$

where

$$N_a^{\gamma}(l) = (\gamma D^{\gamma}(l) + 1)\, g_{\gamma}(X(l)H(l)/D(l)), \tag{6}$$

$$N_c^{\gamma}(l) = (\gamma X^{\gamma}(l) + 1)\, H^{\gamma}(l), \tag{7}$$

and

$$g_{\gamma}(x) = f_{\gamma}(x+1) - f_{\gamma}(x). \tag{8}$$

The distortion caused by additive noise in the generalized log-spectrum is represented by $N_a^{\gamma}(l)$ that depends on signal-to-noise ratio (SNR). $N_c^{\gamma}(l)$ is the distortion caused by channel noise, which depends on the speech components and the channel noise components. In addition, their dependency is determined by the value of γ.

When CMN is combined, the generalized logarithmic spectrum is expressed by

$$\bar{Y}^{\gamma}(l) = Y^{\gamma}(l) - E\{Y^{\gamma}(l)\}. \tag{9}$$

Actually, Eq.(9) does not take into account the influence of liftering. However, since the liftering is equivalent to just spectral smoothing process and DCT is linear, it is reasonable to analyze the impact of CMN in the generalized log-spectral domain instead of the quefrency domain.

4. SPECTRAL DISTORTION

To analyze the distortion quantitatively, the distance between the generalized cepstral coefficients of noisy speech and clean speech is measured. Given the value of γ, the distortion in the generalized log-spectral domain at the l^{th} analysis frame is defined as

$$\bar{e}^{\gamma}(l) = \bar{Y}^{\gamma}(l) - \bar{X}^{\gamma}(l). \tag{10}$$

Substituting Eq.(5) and Eq.(9) into Eq.(10), we obtain

$$\bar{e}^{\gamma}(l) = \bar{e}_a^{\gamma}(l) + \bar{e}_c^{\gamma}(l), \tag{11}$$

where

$$\bar{e}_a^{\gamma}(l) = N_a^{\gamma}(l) - E\{N_a^{\gamma}(l)\}, \tag{12}$$

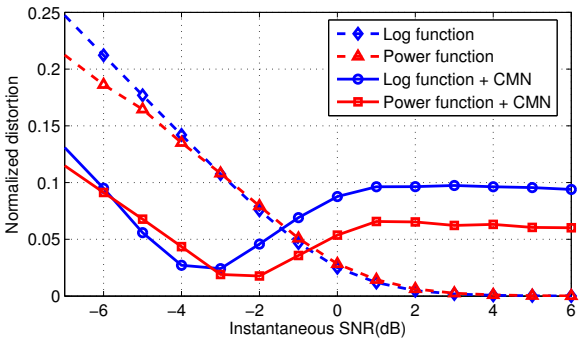

Fig. 3. Normalized distortion in the generalized spectral coefficients for perturbed values of SNRs.

and

$$\bar{e}_c^{\gamma}(l) = N_c^{\gamma}(l) - E\{N_c^{\gamma}(l)\}. \tag{13}$$

In this paper, it is assumed that the length of the channel noise is much shorter than a analysis window. Thus, we only focus on analyzing the influence of additive noise.

Fig. 3 compares the distortion $e_a^{\gamma}(l)$ in the case of the logarithmic function based system ($\gamma = 0$) and the power function based system ($\gamma = 1/15$) when CMN is included. The distortion is computed by measuring Euclidean distance between clean speech and noisy speech in the generalized log-spectral domain, and it is normalized by the dynamic range of the generalized log-spectral coefficients of the corresponding clean speech.

In Eq.(6) and Eq.(12), it indicates that $\bar{e}_a^{\gamma}(l)$ depends on SNRs, i.e., inversely proportional to SNRs at low SNRs. Note that when CMN is not used, there is no considerable difference between the logarithmic function based system and the power function based system. After applying the CMN, however, the distortion at low SNR region is reduced, but the distortion at high SNR region is even increased. It may be interpreted that the CMN processing introduces oversubtraction at the speech dominant region. However, in the case of power function based system, the distortion due to oversubtraction is smaller than the one with MFCC. Accordingly, it concludes that the generalized cepstral coefficients reduce the oversubtraction caused by adopting the mean normalization process. In the next subsection, we investigate the impact of distortion variation to the recognition error patterns depending on the type of spectral nonlinear mapping function used.

5. RECOGNITION ERROR PATTERN

To investigate the impact of the distortion to ASR systems addressed in the previous subsection, we count the recognition errors by dividing them into three types: deletion error, substitution error, and insertion error. A deletion error occurs when a word is recognized as its neighboring word or as a non-speech segment. An insertion error occurs when a word

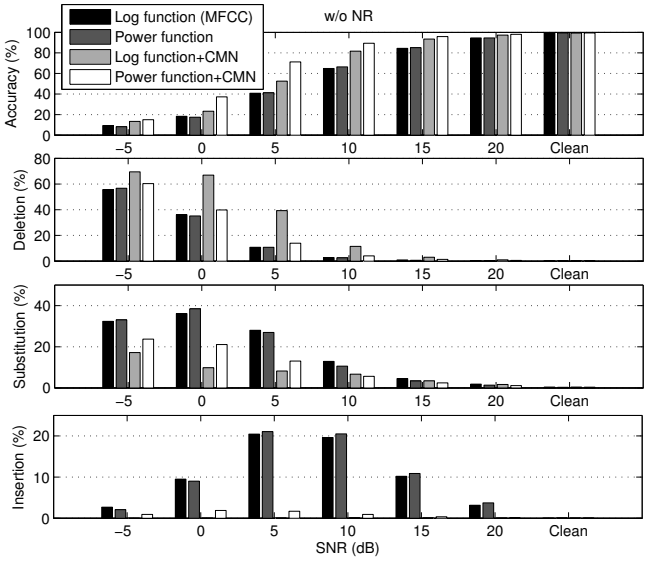

Fig. 4. Word accuracy and error patterns with and without using CMN and with either a natural logarithmic function (MFCC) or a fractional power function (PNCC).

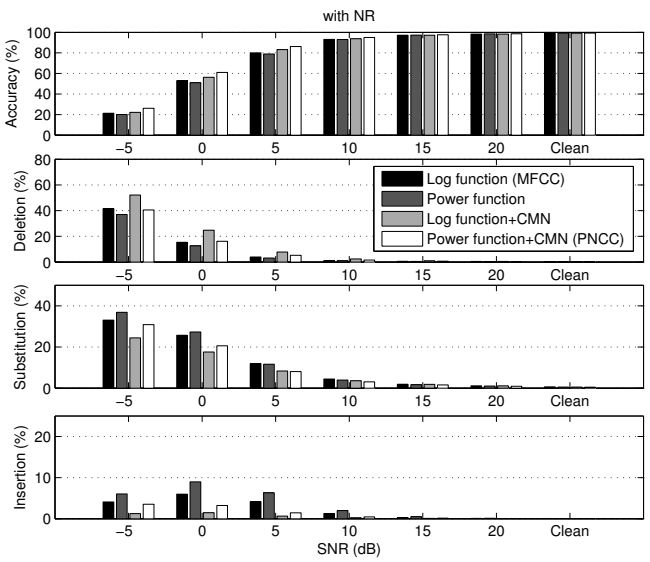

Fig. 5. Word accuracy and error patterns when noise reduction is included.

is recognized as multiple words or the non-speech segment is recognized as words.

The recognition experiments are conducted with Aurora 2 database that is a noisy speech database distributed by European telecommunications standards institute (ETSI) for the purpose of defining distributed speech recognition standard [16]. Its source speech is obtained by downsampling the database of TIDIGITS consisting of English connected digit strings [16]. The different types of noise signals are artificially added to clean speech to generate signals having various SNRs.

The simulation task is done by the method introduced in [16] with HTK toolkit v3.4. The recognizer has eleven whole word HMMs with 16 states and 3 Gaussian mixtures. Two pause models, sil and sp, are defined. The sil model consists of 3 states and each state consists of 6 Gaussian mixtures. The sp model consists of a single state which is tied with the middle state of the sil model.

Fig.4 shows the error patterns for the case of combining CMN or not. The results are obtained by taking an average value for all types of noise. Note that in case CMN is not combined, there is no big difference between the logarithmic function based ASR system and the fractional power function based ASR system. In addition, the number of insertion errors is significantly large, because the noise segments in the non-speech region are recognized as words. By combining CMN technique, the insertion errors are reduced. However, reducing insertion errors leads to the increment of deletion errors, which results from the distortion caused by oversubtraction as shown in Fig.3. It can be intuitively understood that there are many cases where the speech dominant regions are misrecognized as the non-speech region due to oversubtraction. The distortion caused by oversubtraction can be reduced by adopting a fractional power mapping function as shown at the previous section. It results in the fact that the increment of deletion errors due to the oversubtraction is significantly reduced in contrast to MFCCs, which is shown in Fig.4. While the substitution errors and the insertion errors are slightly increased, the overall word accuracy is still significantly increased if the fractional power function is used.

Fig.5 shows the error patterns when the noise reduction process based on asymmetric filtering used in [2] is included in the procedure of feature extraction. The impact of CMN to the fractional power function based ASR system are same to the case where the noise reduction is not adopted. Since the noise reduction process considerably suppresses the noise components, the performance improvement by replacing the spectral mapping function is slightly reduced.

6. CONCLUSION

In this work, we presented the impact of the mean normalization to the power mapping function based ASR system in noisy environment. Only replacing the natural logarithm function with the power function did not improve the word accuracy. Its impact was became significant when the mean normalization technique is combined. The analysis on the spectral distortion and error pattern has demonstrated convincingly that the fractional power function based ASR system is very effective to minimize the deletion errors which is unavoidable if CMN is adopted.

7. REFERENCES

[1] Chanwoo Kim and Richard M Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction.," in *INTERSPEECH*, 2009, pp. 28–31.

[2] Chanwoo Kim and Richard M Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4101–4104.

[3] F-H Liu, Alejandro Acero, and Richard M Stern, "Efficient joint compensation of speech for the effects of additive noise and linear filtering," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 257–260.

[4] Fu-Hua Liu, Richard M Stern, Xuedong Huang, and Alejandro Acero, "Efficient cepstral normalization for robust speech recognition," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1993, pp. 69–74.

[5] Anshu Agarwal and Yan Ming Cheng, "Two-stage mel-warped wiener filter for robust speech recognition," in *Proc. ASRU*, 1999, vol. 99, pp. 67–70.

[6] ETSI Standard, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050, pp. v1, 2007.

[7] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[8] Hynek Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[9] Pedro J Moreno, Bhiksha Raj, and Richard M Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 733–736.

[10] Alex Acero, Li Deng, Trausti T Kristjansson, and Jerry Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition.," in *INTERSPEECH*, 2000, pp. 869–872.

[11] Pedro J Moreno, Bhiksha Raj, and Richard M Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 733–736.

[12] Jinkyu Lee, Soonho Baek, and Hong-Goo Kang, "Signal and feature domain enhancement approaches for robust speech recognition," in *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*. IEEE, 2011, pp. 1–4.

[13] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[14] Takao Kobayashi and S Imai, "Spectral analysis using generalized cepstrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 5, pp. 1087–1089, 1984.

[15] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[16] Hans-Günter Hirsch and David Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.