CEPSTRAL GAIN NORMALIZATION FOR NOISE ROBUST SPEECH RECOGNITION

Shingo YOSHIZAWA, Noboru HAYASAKA, Naoya WADA and Yoshikazu MIYANAGA

Graduate School of Engineering, Hokkaido University, Sapporo, Japan

ABSTRACT

This report describes a robust speech recognition technique which normalizes cepstral gains in order to remove effects of additive noise. We assume that the effects can be expressed by an approximate model which consists of gain and DC components in log-spectrum. Accordingly, we propose cepstral gain normalization (CGN) which normalizes the gains by means of calculating maximum and minimum values of cepstral coefficients in speech frames. The proposed method can extract noise robust features without a prior knowledge and environmental adaptation because it is applied to both training and testing data. We have evaluated recognition performance under noisy environments using Noisex-92 database and a 100 Japanese city names task. The CGN provides improvements of recognition accuracy at various SNRs comparing with combinations of conventional methods.

1. INTRODUCTION

In recent years, speech recognition technologies have considerably progressed and achieved high accuracy in clean environments. However, since the recognition performance declines when speech is corrupted by noise, an improvement of noise robust techniques is required. Noise robustness approaches can be classified into several categories, i.e., feature or model adaptation[1,2] noise robust feature extraction[3] and noise suppression[4,5]. In the noise suppression, spectral subtraction (SS)[4] is effective for background noise suppression and widely used in speech recognition, speech enhancement and speaker verification. This method estimates noise spectra from non-speech intervals and subtracts them from noisy speech spectra. Also an advanced spectral subtraction method is proposed[6].

However, SS has two problems in case of utilizing in speech recognition. In the first, speech/non-speech separation is a non-continuous and non-linear operation in frame time. It causes distortion of original speech and recognition performance declines under high SNR environments. In the second, the efficacy of the noise subtraction weakens in log-spectrum domain. It is known that log-spectrum is very sensitive to noise since spectral valleys are more affected by noise than spectral peaks. When residual noise spectrum exists after SS processing, the noise spectrum is considerably enhanced in non-speech and pause intervals. Root-cepstrum analysis[7] has been proposed in order to suppress the sensibility of noise, however, recognition performance falls for clean speech or speech-like noise.

In this report, we consider speech log-spectrum affected by the additive noise and propose an approximate model. The model assumes that log-spectrum of the additive noise is expressed by changes of gain and DC components. This approximate model can be applied to cepstrum domain. And then we propose cepstral gain normalization (CGN) which normalizes cepstral gains by means of calculating maximum and minimum values of cepstral coefficients in speech frames. The CGN does not distort original speech because of simple linear operations. Moreover, it has an advantage that a priori knowledge and adaptation are not required under any environments on account of executing same processing in both training and testing data.

2. APPROXIMATE MODEL IN ADDITIVE NOISE

When speech is corrupted by unknown additive stationary noise and unknown multiplicative distortion, we describe a model of noisy environment given by the following equation:

$$X(n,\omega) = S(n,\omega)H(\omega) + A(\omega) \tag{1}$$

where $X(n,\omega)$ is a power spectrum of noisy speech at frequency ω and frame time n, $S(n,\omega)$ is a power spectrum of clean speech, $H(\omega)$ is the multiplicative distortion and $A(\omega)$ is a power spectrum of the additive noise. The additive noise is assumed to be uncorrelated with speech signal. We express $E(n,\omega)=S(n,\omega)H(\omega)$ to simplify. Log-transformation of Eq. 1 can be expressed as

$$\log X(n,\omega) = \log(E(n,\omega) + A(\omega))\,. \tag{2}$$

Figure 1 shows the second channels of mel-scale filterbank in power and log spectrum domain respectively, where the value of $A(\omega)$ is set to *0.002*. The noise spectrum $A(\omega)$ is artificially added to the spectrum $E(n,\omega)$ in power spectrum domain. The speech sample is extracted from utterance 'hachinohe', analyzed by 512-points short-time Fourier transforms and compressed by 40 mel-scaled filterbanks. The waveform of log-spectra in noisy speech $log(E(n,\omega)+A(\omega))$ decreases in gain (distance between maximum and minimum values) and increases in DC level, comparing with that of clean speech. We take notice of the following points in order to formulate these changes.

1. The power spectrum of clean speech $E(n,\omega)$ has a minimum value in each frequency. The value must be non-zero.

2. The power spectrum of additive noise $A(\omega)$ is much larger than the minimum values $min(E(n,\omega))$, that is $A(\omega)>>min(E(n,\omega))$.

Ideally, $min(E(n,\omega))$ is a zero value in clean environment. However, the zero value is converted to infinity in log-spectrum domain. It prevents speech recognition training or testing on computer. This constraint is inevitable and the minimum value necessarily becomes an important factor determining the gain in the waveform of log-spectra.

The change of gains $G(\omega)$ can be expressed as

$$G(\omega)=\frac{G_N(\omega)}{G_C(\omega)} \tag{3}$$

$$G_N(\omega)=\log\left(\frac{\max(E(n,\omega))+A(\omega)}{\min(E(n,\omega))+A(\omega)}\right) \tag{4}$$

$$G_C(\omega)=\log\left(\frac{\max(E(n,\omega))}{\min(E(n,\omega))}\right) \tag{5}$$

where *$G_N(\omega)$* and *$G_C(\omega)$* are gain factors in noisy and clean environments respectively. The gain of clean speech is mostly determined by the minimum value *$min(E(n,\omega))$*. On the other hand, the gain of noisy speech is determined by the noise value *$A(\omega)$* and becomes much smaller than that of clean speech on account of *$A(\omega)>>min(E(n,\omega))$*.

The DC offsets of clean and noisy speech can be expressed as

Noisy:

$$D_N(\omega)=\frac{\sum_{n=1}^{L}\log(E(n,\omega)+A(\omega))}{L} \tag{6}$$

Clean:

$$D_C(\omega)=\frac{\sum_{n=1}^{L}\log E(n,\omega)}{L} \tag{7}$$

where *$D_N(\omega)$* and *$D_C(\omega)$* are DC offsets of noisy and clean speech respectively and *L* is the number of speech frames. Consequently, by adjusting the gains and DC offsets, the log-spectrum of noisy speech can be expressed as the following approximate equation:

$$\log X(n,\omega)\approx G(\omega)[\log E(n,\omega)-D_C(\omega)]+D_N(\omega)\,. \tag{8}$$

Furthermore,

$$\log X(n,\omega)\approx G(\omega)\log E(n,\omega)+B(\omega) \tag{9}$$

where *$B(\omega)$* is a DC bias given by

$$B(\omega)=D_N(\omega)-G(\omega)D_C(\omega)\,. \tag{10}$$

3. EFFECTIVENESS OF APPROXIMATE MODEL

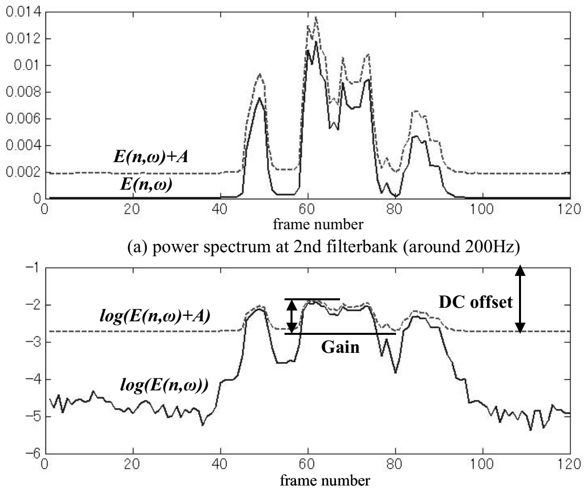
The effects of additive noise are expressed by changes of gain and DC components in log-spectrum domain. In this section, we discuss effectiveness of the approximate model. Equation (9) can be generalized as

$$\log(f(x)+A)\approx\gamma\log f(x)+\beta\,. \tag{11}$$

Exponential-transformation of Eq. (11) can be described as

$$f(x)+A\approx e^{\beta}(f(x))^{\gamma} \tag{12}$$

where *$f(x)>0$*. The representation of Eq. (12) enlarges approximation errors if *$f(x)$* is a strictly increasing function.



(a) power spectrum at 2nd filterbank (around 200Hz)

(b) log-spectrum at 2nd filterbank (around 200Hz)

Figure 1: Power spectrum and log-spectrum of clean and noisy environments.

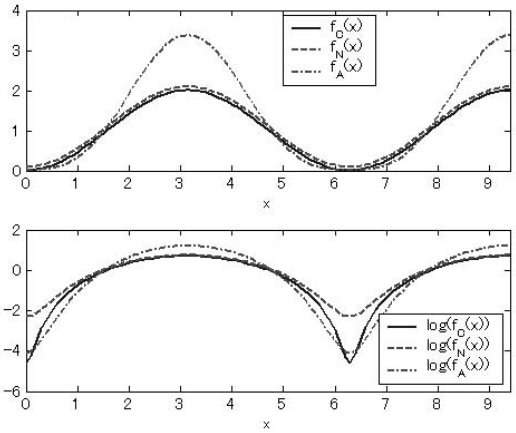


Figure 2: Comparison of linear and logarithm functions.

However, a time trajectory of power spectra in speech is similar not to an increasing function but to a periodic function. We utilize a simple periodic function given by *$f(x)=-cos(x)+1$*. Model functions in clean and noisy environments are expressed as

Clean: $$f_C(x)=-\cos(x)+1+\delta \tag{13}$$

Noisy: $$f_N(x)=-\cos(x)+1+A \tag{14}$$

where δ is a threshold assumed in clean environment and A is additive noise assumed in noisy environment. SNR values are determined by these parameters. When δ is set to 0.01, a SNR value becomes 40dB. When A is set to 0.1, it produces noisy environment at SNR 20dB. According to the approximate model, *$f_N(x)$* can be approached to *$f_C(x)$* by canceling effects of β and γ, given by

$$f_A(x)=e^{-\beta}(f_N(x))^{\frac{1}{\gamma}} \tag{15}$$

where *$f_A(x)$* is an approximate function. The values of β and γ are calculated in the same way as the previous section.

Figure 2 shows a comparison of three functions in linear and logarithm domain *(δ=0.01, A=0.1)*. Although the distance between $f_A(x)$ and $f_C(x)$ enlarges at the spectral peak(x=π, 3π), it becomes compressed after logarithm transformation. The distance between $logf_A(x)$ and $logf_C(x)$ becomes much smaller than that of between $logf_N(x)$ and $logf_C(x)$ at the spectral valley($x=0, 2\pi$). In Eq. (15), the root function $(.)^{1/\gamma}$ brings the minimum value of noise spectra close to that of clean spectra at the spectral valley. The exponent function *exp(–β)* can suppress errors at the spectral peak which is produced by the root operation.

Focusing on logarithm transformation, the transformation of Eq. (15) can generate robust feature parameters which are less affected by noise than no robust processing at spectral valleys. The approximate model suppresses a trade-off that logarithm transformation is sensitive to small changes of values at spectral valleys and insensitive to large changes at spectral peaks.

4. CEPSTRAL GAIN NORMALIZAION

When recognition training is executed in clean environment and recognition testing is evaluated in noisy environment, a difference of log-spectra between training and testing environments can be removed by adjusting the gain and the DC offsets. The adjusted log-spectrum is obtained by the following equation:

$$\log S'(n,\omega) = \log E(n,\omega) - D_C(\omega)\,. \qquad (16)$$

In the noisy environment, *logS'(n,ω)* can be obtained by canceling the gain *G(ω)* and the DC offset $D_N(\omega)$ according to Eq. (8). The DC offset $D_N(\omega)$ can be calculated from an average of log-spectra. The gain *G(ω)* can be eliminated by normalizing both clean and noisy log-spectra gain $G_N(\omega)=1$, $G_C(\omega)=1$. These operations can be applied into cepstral parameters approximately. A series of procedures are summarized to the following two steps, which are applied to both training and testing data.

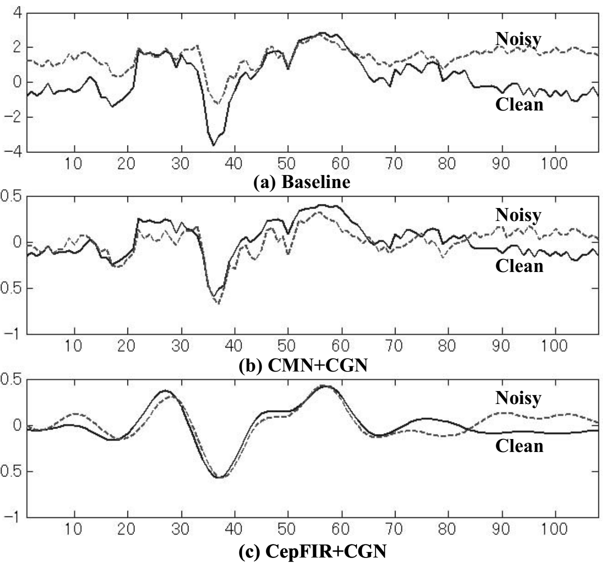
Step1: Subtract an average of cepstral coefficients. The operation is known as cepstral mean normalization (CMN)[5].

$$C'(n,k) = C(n,k) - \left(\sum_{n=1}^{L} C(n,k)\right) / L \quad \text{for } 1 \le k \le M \qquad (17)$$

Step2: Normalize gains by calculating the maximum and the minimum values of cepstral coefficients. It is called as cepstral gain normalization (CGN) in this report.

$$C''(n,k) = C'(n,k) / (\max_{1\le n\le L} C'(n,k) - \min_{1\le n\le L} C'(n,k)) \quad \text{for } 1 \le k \le M \qquad (18)$$

where *k* is quefrency in *M*-order cepstrum. These steps are applied to delta cepstrum and delta-delta cepstrum as well. CGN is similar to cepstral variance normalization (CVN). While CGN is based on the approximate model, CVN is not sufficient to clarify the effects of additive noise. As for additive noise, the normalization using cepstral gain by CGN realizes better noise robust performance.



(a) Baseline

(b) CMN+CGN

(c) CepFIR+CGN

Figure 3: Cepstral coefficients of clean and noisy speech (2nd MFCC).

5. TEMPORAL FILTERING

Although we have assumed stationary additive noise in the above discussion, non-stationary noise should be considered under real environments. If the non-stationary noise produces sharp peaks in cepstrum, the peaks may prevent estimating an exact gain of noisy cepstrum. Smoothing cepstral coefficients is required to remove such turbulences. Therefore we use temporal filtering which is proposed by RASTA filtering[8] and FIR band-pass filtering[9]. These processing remove DC components and higher modulation frequency components of cepstral coefficients. The removal of DC components substitutes for CMN. Speech intelligibility mostly exists between 4Hz and 16Hz in modulation frequency. The band-pass filtering reduces noise components existing in the other frequency bands.

Figure 3 shows the second MFCC in clean and noisy environments. The noisy speech is generated artificially from speech babble noise at SNR 10dB. In this example, it turns out that the combination of CGN and CepFIR (which means the FIR band-pass filtering) succeeds in fitting cepstral gains between clean and noisy speech.

6. EXPERIMENTS

6.1. EXPERIMENTAL CONDITIONS

Noise robust performance has been evaluated in an isolated word speech recognition task. We have utilized NOISEX-92 noise database[10] and 100 words Japanese city names of JEIDA (Japanese Electronic Industry Development Association) database. A speech data is sampled at 11.025KHz and 16bit quantization. In speech analysis, MFCC features are extracted after pre-emphasis and Hanning windowing, and converted to 38 dimensional feature vectors. Frame length and shift are 23.2ms and 11.6ms respectively. The feature vectors consist of 12 MFCC, 12 delta MFCC, 12 delta-delta MFCC, delta log energy and delta-delta log energy. In training, we have created 100 word-level HMMs from 40 males speech data. These models have 32 states and 1 mixture per states based on a continuous density function. In testing, clean speech is artificially added with noise at various SNRs. Recognition accuracy has been measured from 10 speaker-independent test sets in recognition experiments.

6.2. EXPERIMENTAL RESULTS

We have evaluated various combinations in order to find optimum techniques suited with CGN. Table 1 shows average recognition accuracy in F-16 noise, factory noise, and speech babble noise. At CepFIR, we have implemented a 240-tap FIR filter which passes between 1Hz and 10Hz. CMN and CepFIR are computed on a whole speech utterance. Without CGN or CVN, CepFIR has pointed the best recognition performance of four methods. The recognition rates of CGN based methods are higher than those of CVN. CVN based methods distort recognition performance in clean environment (over SNR 30dB) comparing with usage of only CMN. At last, the combination of CepFIR and CGN has provided the highest recognition performance.

Table 2 shows average recognition accuracy in all fifteen types of noisy environments in NOISEX-92. We have compared the combination of CepFIR and CGN with other combinations of conventional methods. SS parameters are set to over-estimation $\alpha=1.5$ and flooring $\beta=0.1$. The noisy spectrum of SS is calculated from an average of 7 frames in non-speech intervals. While SS method drops recognition performance under clean environment, the recognition rate of CGN is higher than Baseline by 0.5%. These results prove that CGN does not distort original speech. In the combination of SS, CepFIR and CGN, its improvement is intangible. As a result, the combination of CepFIR and CGN not only improves recognition accuracy comparing with combinations of conventional method under noisy environments and but also preserves high recognition accuracy under clean environment.

7. CONCLUSIONS

In this paper, we have described the approximated model of additive noise in log-spectrum and cepstral mean normalization (CGN) which normalizes cepstral gains. The CGN improves not only noise robust performance under any noisy environments and but also does not without distort original speech. Since the proposed algorithm is quite simple and has low computation cost, it can be embedded into speech analysis front-end. However, CGN processing should start after speech endpoint detection in the same way as CMN computed on a whole speech utterance. Our research future goals are to modify for frame-wise processing and continuous speech recognition and to combine other noise robust techniques which are represented by model compensation.

8. ACKNOWLEDGEMENTS

The authors would like to thank Research and Development Headquarters, Yamatake Corporation for fruitful discussions. This study is supported in parts by Semiconductor Technology Academic Research Center (STARC), Program 112 and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B2) (15300010), 2003.

Table 1: Average recognition rates in F-16 noise, factory noise, and speech babble noise.

	Baseline	CMN	RASTA	CepFIR
0dB	4.6	4.8	4.6	9.8
10dB	63.3	72.8	70.1	75.1
20dB	95.0	97.9	96.2	97.3
Clean	99.1	99.3	99.4	99.3
Average	**65.5**	**68.7**	**67.6**	**70.4**

	CMN+CVN	CMN+CGN	CepFIR+CVN	CepFIR+CGN
0dB	27.4	26.0	36.8	43.2
10dB	81.9	83.3	81.5	85.9
20dB	96.1	97.8	95.2	97.3
Clean	99.1	99.3	99.0	99.6
Average	**76.1**	**76.6**	**78.1**	**81.5**

Table 2: Average recognition rates in all fifteen types of noisy environments.

	0dB	10dB	20dB	Clean
Baseline	20.6	65.3	95.8	99.3
SS	32.9	82.5	94.7	98.6
SS+CMN	43.6	82.0	92.8	98.1
SS+RASTA	44.9	85.9	95.5	98.5
SS+CepFIR	47.9	85.8	95.0	98.2
CepFIR+CGN	53.3	88.7	97.6	99.6
SS+CepFIR+CGN	49.9	88.8	97.8	99.4

9. REFERENCES

[1] A. Acero and R. M. Stern, “Environmental robustness in automatic speech recognition,” Proc. ICASSP90, pp.849-852, 1990.

[2] P. J. Moreno et al., “A vector Taylor series approach for environment-independent speech recognition,” Proc. ICASSP96, pp.849-852, 1996.

[3] H. Hermansky, “RASTA processing of speech,” IEEE Trans. Speech Audio Processing, vol.2, pp.578-589, 1994.

[4] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoust., Speech, Signal Processing, ASSP-33,vol.27, pp. 113-120, 1979.

[5] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am. vol.55, pp.1304-1312, 1974.

[6] N. Flores, A. and Young, S. J., “Continuous speech recognition in noise using spectral subtraction and HMM adaptation,” Proc. ICASSP94, pp.409-412, 1994.

[7] P. Alexandre and P. Lockwood, “Root cepstral analysis: A unified view. Application to speech processing in car noise environments,” Speech Comm. vol.12, pp.277-288, 1993.

[8] H. Hermansky, “Should recognizers have ears?,” ESCA Tutorial and Research Workshop on Robust speech recognition for unknown communication channels, pp. 1-10, ESCA-NATO, 1997.

[9] T. Arai, M. Pavel, H. Hermansky and C. Avendano, “Intelligibility of speech with filtered time trajectories of spectral envelopes,” Proc ICSLP96, pp.2490-2493, 1996.

[10] A. Varga and H. J.M. Steenken, “Assessment for automatic speech recognition : II. NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems,” Speech Comm., vol.12, no.3, pp.247-251, 1993.