Modified SNR-Normalization Technique for Robust Speech Recognition

*Hoi-In Jung, **Kab-Jong Shim, and *Hyung-Soon Kim

Abstract

One of the major problems in speech recognition is the mismatch between training and testing environments. Recently, SNR normalization technique, which normalizes the dynamic range of frequency channels in mel-scaled filterbank, was proposed[1]. While it showed improved robustness against additive noise, it requires a reliable speech detection mechanism and several adaptation parameters to be optimized.

In this paper, we propose a modified SNR normalization technique. In this technique, we take simply the maximum of filterbank output and predetermined masking constant for each frequency band. According to the speaker-independent isolated word recognition in car noise environments, proposed modification yields better recognition performance than the original SNR normalization method, with rather reduced complexity.

Ⅰ. Introduction

The main problem for robust speech recognition is the mismatch between training and testing environments. This problem is mainly caused by a different recording channel, a different speaker, and a variability of ambient noise. Several approaches have been considered for robust speech recognition. These approaches are classified into three categories: preprocessing speech enhancement techniques, robust feature extraction and distance measure techniques, and model compensation techniques[2].

Recently, SNR normalization technique, which normalizes the dynamic range of frequency channels in mel-scaled filterbank, was proposed and showed improved robustness against additive noise[1]. In this technique, adaptive masking constant is added to the outputs of a mel-scaled triangular filterbank for both training and test data. The goal is to normalize the SNR in each frequency band by adapting the masking constant depending on measured SNR or dynamic range in each band[1]. The idea of this technique is that the HMM models are trained in similar conditions to the test noise environments. But conventional SNR-normalization requires a reliable speech detection mechanism to measure the dynamic range, and needs several adaptation parameters which should be optimized during recognition experiment. These problems sometimes yield improper adaptation of the masking constant, thereby limiting the amount of performance improvement.

To alleviate these problems, we proposed a modified SNR-normalization technique. The proposed modification take simply the maximum of filterbank output and predetermined masking constant for each frequency band, and does not involve the speech detection mechanism and several adaptation parameters. The energy fluctuations in non-speech interval which is one of the main sources of environmental mismatch are minimized. Experimental results showed that the proposed technique yielded, better performance than conventional SNR-normalization in noisy environments.

Ⅱ. Conventional SNR-normalization

Fig.1 shows the influence of car noise on the output of 11-th mel-scaled filterbank(its center frequency is about 820Hz.) in log energy spectral domain. The mismatch between two signals concerns mainly the low energy portions. Reducing mismatch for these portions can be achieved by normalizing the dynamic range in each frequency band.

SNR-normalization technique employs the Mel-frequency cepstral coefficients(MFCC) and its general scheme is given in Fig.2. This technique goes into operation for each critical-band filterbank output.

The output of each mel-scaled triangular critical band is normalized to the target SNR by adding a adaptive masking constant. The algorithm contains the following steps:

- Initialized the masking constants a[i] by calculating

* Department of Electronics Engineering, Pusan National University, Korea
** Passenger Car E&R Center II, Hyundai Motor Company, Korea

Manuscript Received: September 23, 1997.

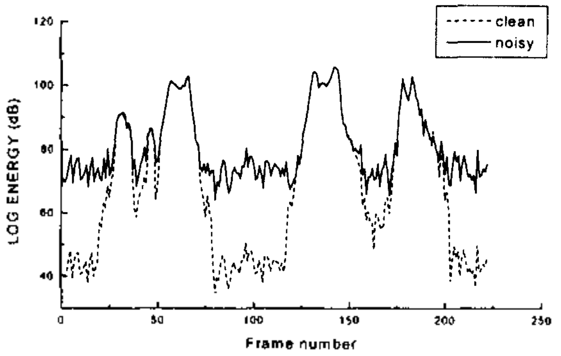


Figure 1. Effect of car noise on a frequency band

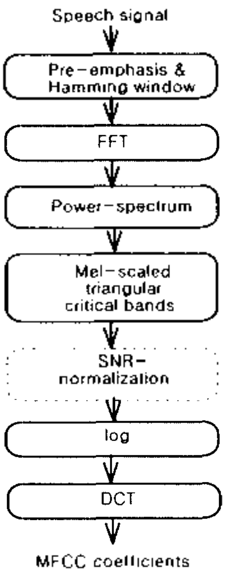


Figure 2. Basic signal processing front-end

the dynamic range of an initialization part of the data.

- Main loop (for each frame):
  1. Mask the filterbank outputs x[i]:
     $y[i] = x[i] + a[i]$
  2. Measure the instantaneous dynamic range $\mathrm{SNR}_{inst}[i]$ of the masked signal y[i].
  3. Adapt the masking constants depending on a fixed target SNR:
     – increase a[i] if $\mathrm{SNR}_{inst}[i] > \mathrm{SNR}_{target}$
     – decrease a[i] if $\mathrm{SNR}_{inst}[i] < \mathrm{SNR}_{target}$

The estimation of the instantaneous dynamic range, $\mathrm{SNR}_{inst}[i]$, is based on the measured maximum, max[i], and minimum, min[i], for the i-th filterbank outputs. The maxima and minima follow respectively sudden peaks and valleys immediately. In the speech parts where the signal is smaller than the measured maximum, max[i] is decreased with a certain time constant. During the noise/silence intervals where the signal is larger than the measured minimum, min[i] is increased with a different time constant. The detailed algorithm is given in [1].

Fig.3 represents the result of the filterbank output after the SNR-normalization. It can be seen that the mismatch between clean and noisy signal was reduced after SNR-normalization.

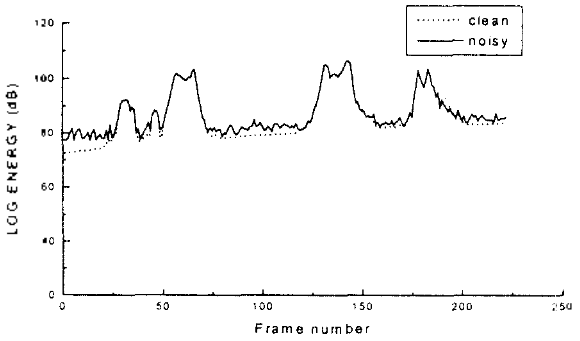


Figure 3. Result of SNR-normalization for the frequency band shown in Fig. 1.

However, the above-mentioned SNR-normalization requires a reliable speech detection mechanism to measure the dynamic range, and needs several adaptation parameters to be optimized during recognition experiments. These problems sometimes yield improper adaptation of the masking constant, thereby limiting the amount of performance improvement.

Ⅲ. Modified SNR-normalization

Modified SNR-normalization proposed in this paper is a simple and straightforward technique. The idea is to normalize the dynamic range by setting the log energy levels of low energy portions to the predetermined value, which is related to the noise level for the target SNR. Thus, the proposed method takes a maximum value between the filterbank output energy and the predetermined target noise level, as follows:

$$y[i] = \max(x[i], TH)$$

where TH is the target noise level.

The practical issue with this technique is how to select the target noise level. In this paper, we determined the target noise level to maximize the recognition accuracy for the various noise environments to be considered. Fig. 4 shows the result of the modified SNR normalization for the same filterbank output shown in Figs. 1 and 3. It can be seen that the modified SNR normalization yields good match between the clean and noisy speech. In particular, time-varying fluctuations for the non-speech or noise-only periods are minimized by the proposed method. This fact is the major difference between the proposed method and the somewhat similar approach presented previously[3], which can be represented as

$$y[i]=x[i]+TH.$$

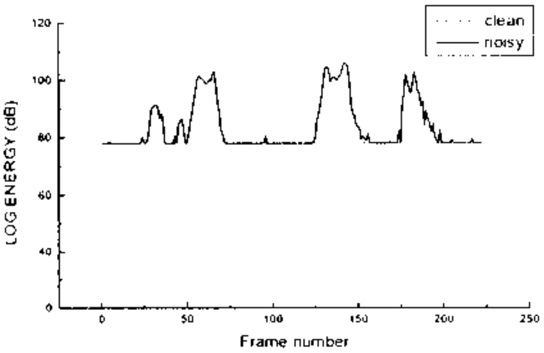
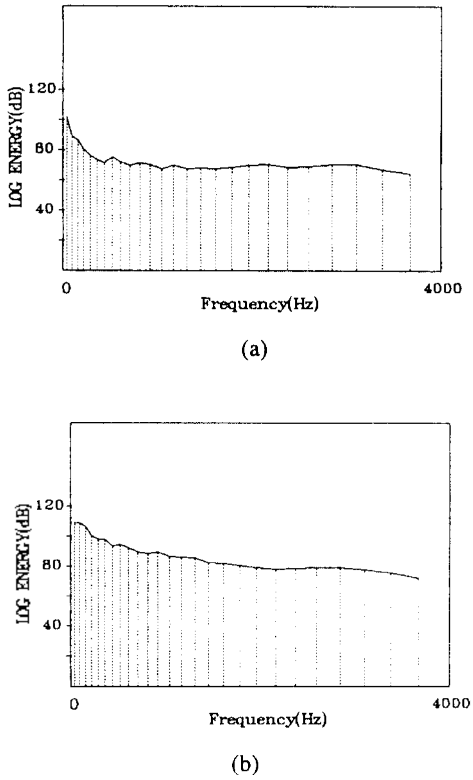


Figure 4. Result of modified SNR-normalization for the frequency band shown in Fig. 1.

Another issue with the proposed method is the frequency dependent optimization of the target noise level. In general, the environmental noise characteristics cannot be known in advance, and keeping the target noise level constant for all filterbanks gives unexpectedly good performance as to be shown in next section. However, if broad spectral characteristics of additive noise can be known a priori, frequency-dependent target noise level based on the prior knowledge of noise characteristics may yield better performance. For example, the power spectra of car noise, which has periodical characteristics, is strongly concentrated at lower frequencies. Fig. 5 shows the typical average spectral characteristics of car noise at idle and 100km/h condition. In this figure, the dashed vertical lines indicate the center frequencies of filterbanks. Even after preemphasis that acts as a high-pass filter, domination of low frequency components is apparent.

For the car noise environment, it is reasonable to have higher values of the target noise level for the lower frequency bands than those for the higher frequency bands. Thus we made a further modification for the car noise environment as follows:



(a)

(b)

Figure 5. The spectral characteristics of car noise(after preemphasis). (a) idle condition, (b) the speed of 100km/h.

$$y[i]=\begin{cases} max(x[i], THL) & for \quad 1\leq i\leq L \\ max(x[i], TH) & for \quad L+1\leq i\leq N \end{cases}$$

where N is the number of filterbanks for MFCC computation and L is the number of lower frequency bands. In this paper, we used the values of N = 26 and L = 4 for car noise environment. THL and TH are the target noise level for the lower frequency bands and the remaining bands, respectively, and $THL>TH$.

IV. Experimental Results

We tested the proposed techniques in an isolated word recognition problem. The database used for this recognition task was 50 words for command and control in car environment. Speech data were collected from 59 male speakers and sampled at 8 kHz. We use 49 persons for training and 10 persons for test. We constructed discrete HMM for each word using HTK V2.0[4], and number of states for HMM was assigned as three times of the number of phones in each word. The feature parameters used were 12 MFCC and the size of codebook was 64. To make noisy speech data, car noise at the speed of 100km/h was added to clean speech data.(The estimated SNR is about 0dB.)

We performed two experiments to evaluate the proposed techniques. In the first experiment, we compared conventional SNR-normalization and modified SNR-normalization. Secondly, we experimented filterbank dependent SNR-normalization which uses frequency-dependent target noise levels.

Table 1 shows the baseline test which is the recognition experiment without any processing.

Table 2 shows the performance comparison of conventional SNR-normalization and modified SNR-normalization with various target SNR and target noise level. Table 2(a) and (b) represent conventional SNR-normalization and modified SNR-normalization, respectively. For clean speech test, conventional SNR-normalization and modified SNR-normalization show a similar performance to baseline test. For noisy speech test, however, both of two techniques yield better performance than baseline test and modified SNR-normalization outperforms conventional SNR-normalization. In modified SNR-normalization experiment, best recognition accuracy was achieved when the target noise level was set to 72dB.

Table 1. Recognition accuracy of baseline test.

	Clean speech	Noisy speech
Baseline	96.4%	80.2%

Table 2. Recognition accuracy of original and modified SNR-normalization.

(a) Original SNR-normalization

Target SNR	Clean speech	Noisy speech
1dB	92.0%	84.2%
2dB	94.8%	88.8%
3dB	95.6%	90.8%
4dB	96.8%	88.4%
5dB	96.0%	87.8%

(b) Modified SNR-normalization

Target noise level	Clean speech	Noisy speech
77dB	94.8%	94.4%
76dB	95.4%	95.0%
74dB	95.2%	93.6%
72dB	**96.2%**	**93.6%**
69dB	95.0%	93.0%

Table 3. Recognition accuracy of filterbank-dependent modified SNR-normalization.

	Clean	Noisy
FB-DEP Mod-SNR(12dB, 18dB)	96.8%	93.8%
FB-DEP Mod-SNR(13dB, 18dB)	95.4%	91.8%
FB-DEP Mod-SNR(14dB, 18dB)	**96.0%**	**95.0%**
FB-DEP Mod-SNR(16dB, 18dB)	95.0%	94.2%

Table 3 shows the recognition accuracy of filterbank-dependent modified SNR-normalization(FB-DEP Mod-SNR). In this table, two values in parenthesis indicate the target noise level for the lower frequency bands and the rest of frequency bands, respectively. Filterbank dependent modified SNR-normalization represents different threshold for lower frequency bands(1st to 4th filterbanks) to reflect the spectral characteristic of car noise. Several threshold values for these bands in filterbank dependent modified SNR-normalization were examined and threshold for the rest of filterbanks was set to the target noise level of 72dB, which yields good performance in the first experiment.

With appropriately chosen target noise level, filterbank dependent modified SNR-normalization gives an improved result on noisy speech. Especially when the target noise level for the lower frequency band and that for the rest of filterbanks are 76dB and 72dB, respectively, we get the best performance on noisy speech data with near-the-best performance on clean speech data. The recognition accuracy in noisy test is improved about 25% compared with that of baseline test. Therefore, the filterbank dependent modified SNR-normalization better working than conventional SNR-normalization and modified SNR-normalization.

V. Conclusions

In this paper, we presented a modified SNR-normalization technique to reduce the mismatch between the training and test environments. The proposed technique is easy to implement and requires little additional computational loads. The recognition results showed that the performance of the proposed modification is superior to that of the original SNR-normalization in car noise environments. We also introduce the frequency-dependent target noise level for the case that there is a prior knowledge on the spectral characteristics of additive noise, like car noise. Currently, recognition experiments for various noise environments is under consideration, and we have a plan to study a hybrid approach using the modified SNR-normalization and other preprocessing speech enhancement.

References

1. T. Claes and D. Van Compernolle, "SNR-normalization for robust speech recognition," *Proc. ICASSP*, 1996.
2. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Publishers, 1996.
3. J. P. Openshaw and J. S. Mason, "On the limitations of cepstral features in noise," *Int. Conf. Acoust. Speech and Signal Processing*, pp. II-49-52, 1994.
4. S. Young, *HTK: Hidden Markov Model toolkit V2.0*, Eng. Dept. Speech Group, Cambridge Univ., Cambridge UK, Tech. Rep., 1992.

▲Hoi In Jung

Hoi In Jung received the B.S. degree in electronic engineering from Pusan National University. From March 1996 to present, he is in M.S. course in electronic engineering at Pusan National University. His research area is speech recognition in noisy environment and his research interests include speech recognition, speech synthesis and speech enhancement.

▲Kab Jong Shim

Kab Jong Shim received his B.S. and M.S. degree in electronics engineering from Kang Won National University, Korea, in 1989 and 1991, respectively. He has been working for Research & Development Division in Hyundai Motor Company since 1991 and was engaged in the research and development of automotive electronic systems. His current research interests include speech recognition, synthesis and intelligent human interface systems in vehicle.

▲Hyung Soon Kim

Hyung Soon Kim received the B.S. degree in electronic engineering from Seoul National University in 1983, and the Ph.D. degree in electrical and electronic engineering from the Korea Advanced Institute of Science and Technology(KAIST) in 1989.

From 1987 to 1992, he was with Digicom Institute of Telematics, where he was Technical manager of the Speech Communication Division. Since 1992, he has been with the faculty of the Department of Electronics Engineering at Pusan National University, and is an Assistant Professor. His research interests include digital signal processing, speech recognition and speech synthesis.