Speech Recognition in Car Noise Environments Using Multiple Models Based on a Hybrid Method of Spectral Subtraction and Residual Noise Masking

Myung Gyu Song*, Hoi In Jung*, Kab-Jong Shim**, Hyung Soon Kim*

Abstract

In speech recognition for real-world applications, the performance degradation due to the mismatch introduced between training and testing environments should be overcome. In this paper, to reduce this mismatch, we provide a hybrid method of spectral subtraction and residual noise masking. We also employ multiple model approach to obtain improved robustness over various noise environments. In this approach, multiple model sets are made according to several noise masking levels and then a model set appropriate for the estimated noise level is selected automatically in recognition phase. According to speaker independent isolated word recognition experiments in car noise environments, the proposed method using model sets with only two masking levels reduced average word error rate by 60% in comparison with spectral subtraction method.

Ⅰ. Introduction

In the presence of noise or channel distortions, performance of automatic speech recognizers is greatly reduced because of the mismatch introduced between training and testing environments. It is mostly due to the distortions induced by different channels, the intra- and inter-speaker variability, and environmental noise. Especially, for speech recognition in a car, the mismatch due to various environmental noise is to blame for the performance degradation. Approaches dealing with these problems have been classified into two categories[1]:

- speech enhancement which focuses on the restoration of either the waveform or the parameters of clean speech embedded in noise
- model compensation which aims at compensating and adapting the parameters of recognition models instead of extracting the optimal estimate of speech from a noisy speech signal

In this paper, we employ spectral subtraction method, one of speech enhancement techniques, to reduce the effect of added noise in speech. It estimates the magnitude spectrum of clean speech by explicitly subtracting noise magnitude spectrum from noisy magnitude spectrum. However, even after spectral subtraction, undesirable variations due to the residual noise still remain. To cope with this residual noise, we introduce noise masking technique which raises spectral subtracted magnitude value lower than a pre-defined masking level up to the masking level. In this technique, how the masking level is determined is an important issue. If the masking level is too low, the masking on noise becomes less effective. On the other hand, if the masking level is too high, speech signal itself becomes distorted. To alleviate the problem of selecting one appropriate masking level for various noise environments, we employ multiple model approach, where multiple model sets are made according to several noise masking levels and then a model set appropriate for the estimated noise level is selected automatically in recognition phase. Our experimental results indicate that the proposed method achieves excellent performance over various noisy environments.

This paper is organized as follows. In Section 2 we present a hybrid method of spectral subtraction and residual noise masking. In Section 3 we then describe recognition system using multiple model sets according to noise masking levels. In Section 4 we present experimental results. Finally we summarize our major findings and outline our future work.

*Dept. of Electronics Eng., Pusan National Univ., Korea
**Passenger Car E&R Center II, Hyundai Motor Company, Korea
Manuscript Received : November 9, 1998

Ⅱ. Spectral Subtraction and Residual Noise Masking

If noisy speech signal $x(t)$ is given by

$$x(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ is clean speech signal and $n(t)$ noise signal uncorrelated with $s(t)$, then the Fourier transform of i-th windowed signal is given by

$$X_i(\omega) = S_i(\omega) + N_i(\omega) \tag{2}$$

Spectral subtraction is a typical speech enhancement technique which reduces the effects of added noise in speech[2]. The estimate of the magnitude of clean speech can be obtained by subtracting the noise magnitude spectrum from the noisy speech magnitude spectrum as follows[3].

$$|\hat{S}_i(\omega)| = \begin{cases} |X_i(\omega)| - \alpha |\overline{N(\omega)}|, \\ \quad if\ |X_i(\omega)| - \alpha |\overline{N(\omega)}| > \beta |\overline{N(\omega)}| \\ \beta |N(\omega)|, otherwise \end{cases} \tag{3}$$

where α is an overestimation factor and β a spectral flooring factor. $|\overline{N(\omega)}|$ is the estimated noise magnitude spectrum determined by averaging several frames of noise spectrum typically during the initial non-speech intervals as follows.

$$|\overline{N(\omega)}| = \frac{1}{M}\sum_{i=1}^{M} |N_i(\omega)| \tag{4}$$

Even after spectral subtraction, undesirable variations due to the residual noise still remain. This is clearly shown in Figure 1. Figure 1 shows mel-scaled filter bank energy contours at a specific frequency band with center frequency of 820 Hz. In this figure, (a) is the filter bank energy contour of original clean and noisy speech, and (b) is the one after spectral subtraction. We can see from Figure 1(b) that frame-by-frame fluctuations still remain mainly in non-speech intervals, and these are major source of mismatch between clean and noisy speech. To reduce the mismatch due to the residual noise, we introduce noise masking technique. This approach has already been proposed in [7]. Noise masking is a psychological phenomenon which reduces the perception of speech sounds in the presence of noise[1]. A speech recognition system utilizing this effect can reduce the contribution of low energy regions in the discrimination. In other words, speech spectra can be modified to immune the system from variations on background noise[4][5][6]. A simple noise masking such as Klatt's technique, which replaces the model mean or the observation with composite mask in case that either the model mean or the observation is below the composite mask, provided robust performance for speaker dependent digit recognition with an SNR as low as 3dB[6].

In this paper, by raising the spectral components lower than the pre-defined masking level up to the masking level, we remove the residual noise which still remains after spectral subtraction and thereby reducing the mismatch between two different environments, especially in non-speech intervals as shown in Figure 1(c). Thus noise masking technique in this paper is implemented by taking larger value as the resultant spectrum between spectral subtracted magnitude spectrum and a predefined masking level(ML).

$$Y(\omega) = \max(|\hat{S}(\omega)|, ML) \tag{5}$$

where $|\hat{S}(\omega)|$ is the spectral subtracted magnitude spectrum and $Y(\omega)$ is the resultant spectrum of spectral subtraction and residual noise masking. The practical problem of this approach is how the masking level is determined.

Ⅲ. Recognition System using Multiple Models according to Masking Levels

Although the hybrid method of spectral subtraction and residual noise masking mentioned above is effective in reducing mismatch between two different environments, it requires an appropriate choice of the masking level. If the masking level is too low, the masking on noise becomes less effective. On the other hand, if the masking level is too high, speech signal itself becomes distorted. Therefore, it seems not feasible to treat various noise environments effectively with only one masking level. In order to alleviate this problem, we introduce multiple model approach based on several representatives masking levels. In this approach, multiple model sets are made according to several noise masking levels, and then a model set appropriate for the estimated noise level of input speech is selected automatically in recognition phase.

Recognition procedure using the proposed multiple models is as follows. Assuming that several frames of

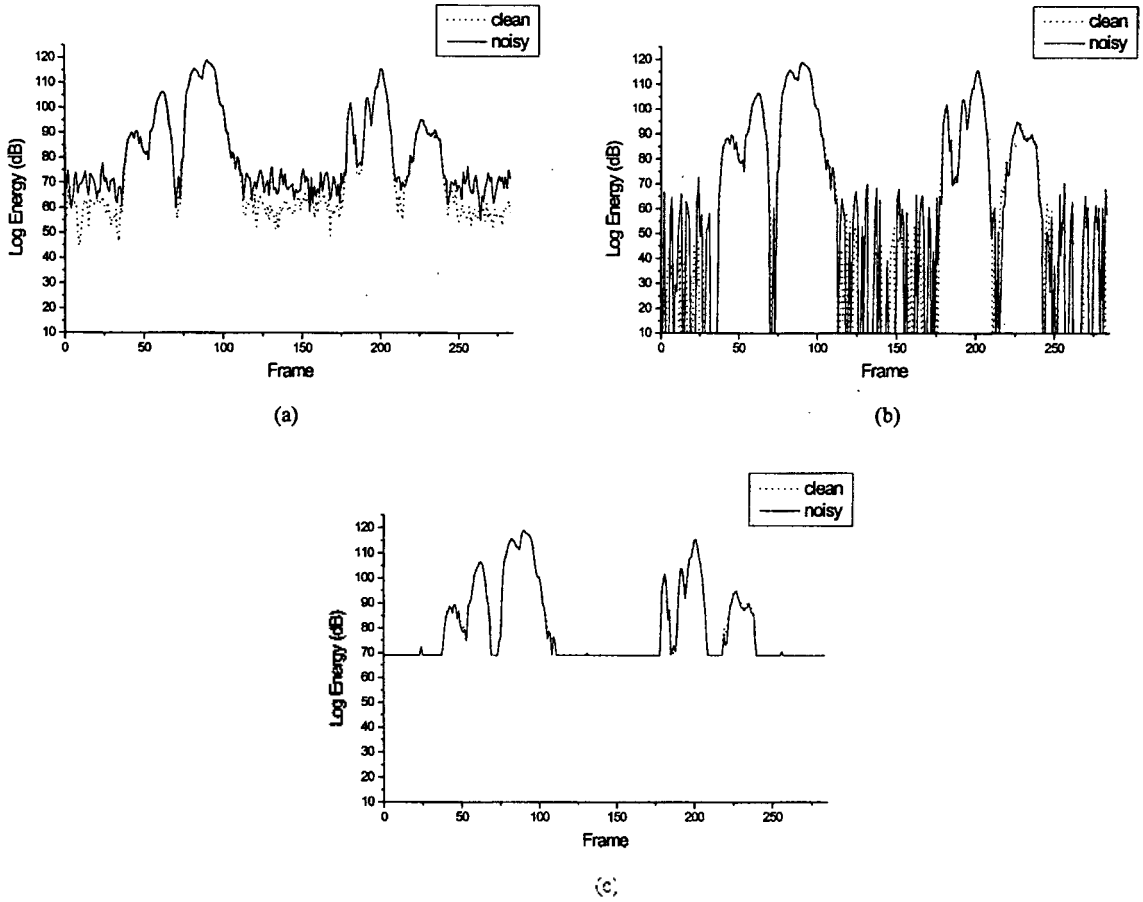

Figure 1. Mel-scaled filter bank energy contour of clean and noisy speech for the utterance "Window down, headlight" at specific frequency band with center frequency of 820 Hz : (a) filter bank energy contour of original speech, (b) filter bank energy contour after spectral subtraction, (c) filter bank energy contour after both spectral subtraction and residual noise masking in case of *ML*=69dB.

initial input speech signal are non-speech intervals, noise characteristics such as mean and standard deviation of mel-scaled filter bank energy are extracted from those frames. Spectral subtraction is performed by regarding the mean as the noise estimate for each filter bank output. After that we determine the noise masking level using the measured standard deviation of noise frames, based on the idea that the amount of fluctuations due to residual noise after removing the noise mean from input speech signal by spectral subtraction is related to the noise standard deviation. In this paper, masking level is determined by

$$ML_{dB} = 20\log_{10}(\sigma \times \gamma) \qquad (6)$$

where σ is the measured noise standard deviation and γ is the weighting factor for σ. The above masking level is then quantized into one of several representative masking levels so as to reduce the number of required models, or consequently the memory requirement of the system. Figure 2 illustrates our approach of masking level quantization in case of the system with three model sets according to three representative masking levels *ML*1, *ML*2 and *ML*3.

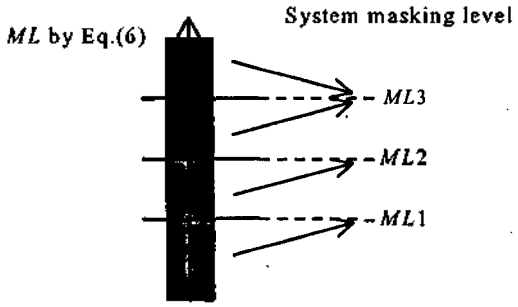

Figure 2. Example showing the quantization of masking level into one of the representative masking levels.

According to the quantized masking level, residual noise masking is performed on the spectral subtracted magnitude spectrum by Eq.(5). Final feature parameters are computed by taking logarithm of the resultant spectrum and discrete cosine transform(DCT). In recognition module, the recognizer takes over the information of the quantized masking level and performs recognition using the model set with that masking level. This technique can improve the robustness of system, because the recognition system can consider various noise environments by recognizing with an appropriate model set which is selected based on the automatic noise environment analysis.

Ⅳ. Experiments and Results

4.1. Database and recognition system

The speech data used in our experiments are 50 isolated words for command and control in car. These were recorded at 8kHz sampling rate, from 59 male speakers. We used the data from 49 speakers for training, those from the rest 10 speakers for test. The car noise data were recorded from five different environments in car: stopping with motor off (env1), stopping with motor on(env2), driving in a low traffic road(env3), driving in a heavy traffic road(env4) and driving in an express highway(env5). The average SNR of these five noise environments are 21dB, 10dB, 2dB, 0dB and -5dB, respectively. The noisy speech was simulated by adding the noise data of each environment to speech signal so as to have the same SNR as the real environments.

Discrete HMM for each isolated word was modeled using HTK. The size of codebook used is 64, and the number of states of each word model is determined as three times the number of phonemes in the word. Feature parameters used are mel-frequency cepstral coefficients(MFCCs), which is known as relatively robust to noise. 12 MFCCs are computed every 10ms with an analysis window of 20ms, from 26 triangular filter bank outputs.

4.2. Evaluations and results

At first, baseline test without any processing and the test with standard spectral subtraction were performed. The results are shown in Table 1, and it can be seen that the average recognition rate using spectral subtraction is slightly higher than that of baseline. Relatively lower performance improvement of spectral subtraction seems to be due to the non-stationary characteristics of car noise in our experiments.

Next, we performed the test using the hybrid method of spectral subtraction and residual noise masking. Table 2 shows the recognition rates of hybrid method with single masking level according to various pre-defined masking levels from 69dB to 86dB.

Table 1. Recognition rate of baseline and spectral subtraction (%).

	Clean speech	Noisy speech					Average
		env1	env2	env3	env4	env5	
Baseline	96.4	93.0	90.8	76.6	64.6	22.4	74.0
Spectral Subtraction	95.6	93.6	92.6	76.2	66.0	35.4	76.6

Table 2. Recognition rate of hybrid method using spectral subtraction and residual noise masking (%).

ML	Clean speech	Noisy speech					Average
		env1	env2	env3	env4	env5	
69dB	**96.2**	**95.8**	**95.6**	93.2	**90.2**	61.8	**88.8**
72dB	93.2	93.8	94.0	92.8	87.2	59.6	86.8
74dB	94.4	93.4	94.0	**93.4**	90.0	65.2	88.4
76dB	94.4	93.0	93.4	90.8	88.8	66.6	87.8
77dB	93.4	92.2	92.0	90.2	88.0	69.0	87.5
78dB	93.0	93.6	93.0	91.2	87.4	71.4	88.3
80dB	91.8	90.8	90.6	88.6	86.2	**74.4**	87.1
84dB	89.6	89.6	89.6	87.6	85.0	72.8	85.7
86dB	86.0	84.2	84.4	82.6	79.6	69.8	81.1

From above tables, the hybrid method for the case of *ML* = 69dB reduced recognition errors by 52% and 57% in comparison with spectral subtraction and baseline, respectively. For clean speech test, as masking level goes higher, recognition rate is reduced because of the increased distortions in speech.

Recognition experiments with multiple model sets are performed as follows. Noise mean and standard deviation of mel-scaled filter bank energy are extracted from initial 8 frames. Spectral subtraction is performed by regarding the mean as noise estimate for each filter bank output. After that we determine the noise masking level using the measured standard deviation of noise frames. Weighting factor γ for noise standard deviation as in Eq.(6) was examined from 0.2 to 2.5. Multiple model sets were built variously according to masking levels, but we presented typical two cases, one with three masking levels and the other with two masking levels, in this paper. Table 3(a)

shows the recognition rate according to γ and Table 3(b) the distribution of test utterances according to γ and masking level, when multiple model sets have three masking levels(69dB, 74dB, 80dB). For example, when γ is set to 0.5, the average recognition rate over 6 different environments including clean speech is 90.9% and the recognition rate for the environment of driving in a low traffic road(env3) is 93.4%. Among total 500 test utterances in the environment of driving in a low traffic road, 414 utterances corresponded to masking level of 69dB, 64 utterances to that of 74dB, and the rest 22 utterances to that of 80dB. Those of multiple model sets with only two masking levels(69dB, 80dB) were shown in Table 4.

We note that the average recognition rate of multiple model set with three masking level is 90.9%, which is the same as the theoretical upper limit of recognition rate, given by averaging the highest recognition rate in each environment from Table 2. We also note that that of multiple model set with only two masking level is 90.7%, which is very close to the theoretical upper limit. This confirms the validity of our masking level selection method based on the standard deviation of non-speech intervals. From the experimental results, we knew that multiple model sets with only two masking levels gave enough performance, considering memory requirements. Figure 3 summarizes the results of Table1 and 4. From this figure, one can see that while the optimal recognition rate is given at γ = 0.3, the variations of the recognition rate according to γ are not significant.

Table 3. Recognition experiments with three multiple model sets
(a) The recognition rate according to weighting factor γ (%)
(b) The distribution of test utterances according to γ and ML.

(a)

γ	Clean speech	Noisy speech					Average
		env1	env2	env3	env4	env5	
0.2	96.2	95.8	95.6	93.0	90.4	63.4	89.1
0.5	96.2	95.8	95.6	93.4	90.6	73.8	**90.9**
1.5	96.2	94.8	95.6	89.0	87.4	74.4	89.6

(b)

γ	ML	Clean speech	Noisy speech				
			env1	env2	env3	env4	env5
0.2	69dB	500	500	500	492	490	391
	74dB	0	0	0	8	6	109
	80dB	0	0	0	0	4	0
0.5	69dB	500	493	500	414	315	0
	74dB	0	7	0	64	143	12
	80dB	0	0	0	22	42	488
1.5	69dB	500	336	489	1	29	0
	74dB	0	144	10	88	19	0
	80dB	0	20	1	411	452	500

V. Conclusions

In this paper, we proposed a pre-processing technique for robust speech recognition in car noise environments. The proposed method is based on the hybrid method of spectral subtraction and the residual noise masking. We also employ multiple model approach, where multiple model sets are made according to several noise masking levels and then a model set appropriate for the estimated noise level is selected automatically in recognition phase. Our experimental results indicated that the proposed method with only two model sets yield about 60% decrease in error rate over various noisy environments. Since we used only clean data for training in our

Table 4. Recognition experiments with two multiple model sets
(a) The recognition rate according to weighting factor γ (%)
(b) The distribution of test utterances according to γ and ML.

(a)

γ	Clean speech	Noisy speech					Average
		env1	env2	env3	env4	env5	
0.3	96.2	95.8	95.6	92.8	89.6	74.4	**90.7**
0.7	96.2	95.6	95.6	91.4	87.0	74.4	90.0
1.5	96.2	93.8	95.6	88.6	87.2	74.4	89.3

(b)

γ	ML	Clean speech	Noisy speech				
			env1	env2	env3	env4	env5
0.3	69dB	500	500	500	478	458	12
	80dB	0	0	0	22	42	488
0.7	69dB	500	488	500	252	95	0
	80dB	0	12	0	248	405	500
1.5	69dB	500	336	489	1	29	0
	80dB	0	164	11	499	471	500

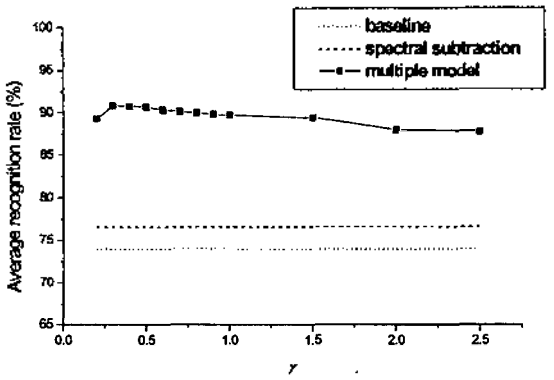

Figure 3. Recognition rate according to (with multiple model sets having with only two masking levels.

experiments, additional performance improvement is expected by including various noisy speech data for training. Our further works also include introduction of frequency-dependent masking level and applying multiple model method in this case.

References

1. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Publishers, 1996.
2. S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 28, pp.752-755, 1980.
3. M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *In Proc. ICASSP*, pp.208-211, 1979.
4. D. H. Klatt, "A digital filter bank for spectral matching," In *Proc. ICASSP*, pp.573-576, 1976.
5. J. N. Holmes and N. C. Sedwick., "Noise compensation for speech recognition using probabilistic models," In *Proc. ICASSP*, pp.741-744, 1986.
6. A. Varga and K. Ponting, "Control experiments on noise compensation in hidden Markov model based continuous word recognition," In *Proc. EUROSPEECH*, pp.167-170, 1989.
7. D. V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, pp.151-167, 1989.

▲Myung Gyu Song

Myung Gyu Song was born in Pusan, Korea. He received the B.S. and M.S. degree in Electronics Engineering from Pusan National University, in 1996 and 1998, respectively. He is a Ph.D student in Electronics Engineering Department at Pusan National University. His current research interests include speech recognition, speech synthesis and digital signal processing.

▲Hoi In Jung

Hoi In Jung received the B.S. and M.S. degree in electronics engineering from Pusan National University, in 1996 and 1998, respectively. From March 1998 to present, he works for Naval Weapon System R&D Center in Agency for Defense Development(ADD). His research area is signal processing and speech recognition.

▲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 Associate Professor. His research interests include digital signal processing, speech recognition and speech synthesis.