Statistical Model-Based Noise Reduction Approach for Car Interior Applications to Speech Recognition

Sung Joo Lee, Byung Ok Kang, Ho-Young Jung, Yunkeun Lee, and Hyung Soon Kim

This paper presents a statistical model-based noise suppression approach for voice recognition in a car environment. In order to alleviate the spectral whitening and signal distortion problem in the traditional decision-directed Wiener filter, we combine a decision-directed method with an original spectrum reconstruction method and develop a new two-stage noise reduction filter estimation scheme. When a tradeoff between the performance and computational efficiency under resource-constrained automotive devices is considered, ETSI standard advance distributed speech recognition font-end (ETSI-AFE) can be an effective solution, and ETSI-AFE is also based on the decision-directed Wiener filter. Thus, a series of voice recognition and computational complexity tests are conducted by comparing the proposed approach with ETSI-AFE. The experimental results show that the proposed approach is superior to the conventional method in terms of speech recognition accuracy, while the computational cost and frame latency are significantly reduced.

Keywords: Speech enhancement, ETSI standard Aurora advanced front-end, two-stage mel-warped Wiener filter, clean spectrum reconstruction, Gaussian mixture model, speech recognition.

Manuscript received Feb. 19, 2010; revised July 14, 2010; accepted July 26, 2010.

This work was supported by the Industrial Strategic technology development program, 10035252, development of dialog based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Rep. of Korea).

Sung Joo Lee (phone: +82 42 860 5732, email: lee1862@etri.re.kr), Byung Ok Kang (email: bokang@etri.re.kr), Ho-Young Jung (email: hjung@etri.re.kr), and Yunkeun Lee (email: yklee@etri.re.kr) are with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Hyung Soon Kim (email: kimhs@pusan.ac.kr) is with the Department of Electronics Engineering, Pusan National University, Busan, Rep. of Korea.

doi:10.4218/etrij.10.1510.0024

I. Introduction

The enhancement of noisy speech has attracted a great deal of research interest for many years and has received attention concerning widespread applications such as voice communications, speech recognition, and hearing aids. Speech enhancement in the past decades has focused on the suppression of additive background noises since speech quality and intelligibility are dependent on short-term spectral amplitude and insensitive to spectral phase. The traditional decision-directed Wiener filter also exploits those characteristics of speech. The goal of speech enhancement for automatic speech recognition is to eliminate noise components while preserving the characteristics of original speech, and it plays an important role in maintaining an average recognition rate, especially in the presence of ambient noise. The past progress of speech enhancement technology for speech recognition allows a driver to control in-car installations via voice. For example, a driver is able to select a song or make a phone call using a voice command while driving. Recently, the use of speech recognition has increased due to its contributions to traffic safety as well as user convenience. However, despite the efforts in voice recognition research, speech recognition performance is often degraded in adverse noisy conditions [1], [2]. In a car interior environment, the amount of background noise depends on the car speed and also on further boundary conditions, such as opened or closed windows. In general, a relatively good signal-to-noise ratio (SNR) is achieved while at a stand-still or slow speed, and a poor SNR is obtained at medium or high speed over 60 km/h. Although the employment of speech enhancement technology enables robust voice recognition, it is still difficult to maintain the average voice recognition rate in a fast moving car. According

to human-computer interface research dealing with state-of-the-art speech recognition technologies, an additive noise consisting of engine, wind, and tire noise is regarded as a major obstacle to prevent high-accuracy recognition.

There are many publications that have reported speech enhancement methods such as the Wiener filter [3], ETSI-AFE [4]-[7], spectral subtraction [8], Ephraim and Malah MMSE [9], log spectral amplitude estimation [10], the autoregressive model-based Kalman filter [11], [12], the auditory perceptual criteria-based method [13], the hidden Markov model (HMM)-based methods [14]-[16], and the model-based Wiener filter [17]. Among the traditional methods, the two-stage mel-warped Wiener filter in ETSI-AFE became popular as a signal preprocessor for voice recognition due to the noise-robustness, relative algorithm simplicity, and online frameworks [4], [7]. Although ETSI-AFE shows relatively robust recognition accuracy when compared with the traditional methods in [3]-[13], it has been reported that the traditional methods based on pre-trained speech and noise knowledge in [14]-[17] outperform ETSI-AFE. However, the traditional model-based methods in [14]-[16] are too complex and computationally demanding to be equipped into modern automotive devices. Therefore, traditional model-based methods are not suitable for the purpose of this work. The goal of this work is to develop a robust end-efficient speech enhancement algorithm for voice recognition which can be employed into a commercial automotive device. The detailed benefits and drawbacks of the traditional model-based methods are found in [17]. The model-based method in [14] is relatively simple when compared with the other model-based methods in [14]-[16]. However, it still requires additional computational costs such as the discrete cosine transform (DCT) and inverse DCT (IDCT) processes. Besides the computational cost, the HMM-based methods in [14]-[16] are unsuitable for online signal processing since the best hidden state sequence needs to be found before the optimal estimate on noise suppression filter gain. Therefore, we assume that ETSI-AFE is an appropriate technology when the hardware specifications of modern automotive devices are considered.

The traditional two-stage mel-warped Wiener filter in ETSI-AFE consists of two filtering stages. The first stage whitens colored noise while preserving speech components, and the second stage removes any residual noise [4]-[6]. Therefore, the cooperation of the two filtering stages overcomes the spectral whitening drawback in the traditional Wiener filter. According to ETSI-AFE, a posteriori SNR is directly obtained from the input spectra, and a decision-directed method is employed to estimate a priori SNR. The major weakness of this decision-directed approach is output speech distortion, which is caused by contaminated a posteriori SNR estimation in the presence of ambient noise. In this paper, we propose a new noise suppression approach to alleviate the spectral whitening and signal distortion problem at the same time. In the proposed method, the two filtering stages are compressed into a single filter by integrating a decision-directed method and an original spectrum reconstruction method. Consequently, the two filtering stages in ETSI-AFE are substituted by the proposed two-stage noise reduction filter estimation scheme. Nevertheless, the two filtering stage cooperation concept to cope with the spectral whitening defect still remains in the proposed approach. The aim of spectrum reconstruction is to restore the original spectrum with the noise-corrupted observation. In general, the reconstruction performance can be guaranteed when the noise-impaired frequency region is limited in a narrow area. We assume that this condition is relatively well matched to a car environment since the major vehicle-noise components are concentrated in the low frequency region. In this work, two statistical model-based spectrum reconstruction methods are adopted to restore the original spectrum from the noisy observation. One is based on a single density Gaussian mixture model (GMM) which is built of clean speech spectra, and the other is a joint density GMM which is established of joint samples (clean and noise-corrupted spectrum pairs). In the proposed method, the de-noised spectrum is obtained by the decision-directed method, and the original clean spectrum is estimated from the de-noised observation. Then, a posteriori SNR is derived from the reconstructed speech spectrum. Therefore, a posteriori SNR with statistical precision is attained, and the signal distortion problem is alleviated by exploiting pre-trained knowledge of speech. Since the secondary filtering stage is physically removed in the proposed method, both the computational load and frame latency can be reduced when it is compared with ETSI-AFE.

The remainder of this paper is organized as follows. After a brief review of the conventional two-stage mel-warped Wiener filter in section II, the proposed approaches are explained in section III. In section IV, the performance evaluation is described before conclusions are given in section V.

II. Conventional Two-Stage Mel-Warped Wiener Filter

The goal of speech enhancement in signal processing is to find the optimal estimate on an original clean speech, given a noisy observation. The Wiener filter was originally proposed by Norbert Wiener during the 1940s and was published in 1949 [18]. It is based on a statistical approach. Typical filters are designed for desired frequency response. However, the design of Wiener filter is from a different angle. Recent popular examples of the Wiener filter can be found in [3] and [4].

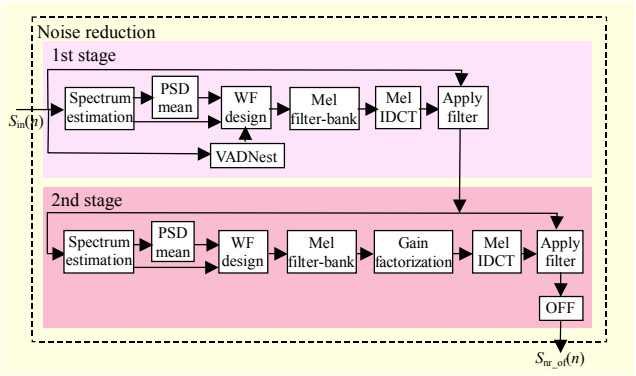

Fig. 1. Block diagram of the conventional two-stage mel-warped Wiener filter.

According to the conventional two-stage mel-warped Wiener filter in Fig. 1, it attempts to remove additive noise throughout the two filtering stages. The first stage coarsely reduces noise and whitens residual noise. Then, the second stage removes any residual noise [4], [5]. Since there is a latency of 2 frames (20 ms) in each stage, the total frame latency of the traditional two-stage mel-warped Wiener filter is 4 frames (40 ms) [4]. Noise reduction is conducted on a frame-by-frame basis. Therefore, the frameworks in the two-stage mel-warped Wiener filter are suitable for online signal processing. The advantage of this online framework is to minimize the whole response time of the voice recognition system.

As shown in Fig. 1, the linear spectrum of each frame is estimated after framing the input signal. The power spectral density (PSD) of the signal spectrum is obtained by smoothing the spectra along the time indexes. Then, the frequency response of the Wiener filter is calculated from the speech and noise PSD. Linear Wiener filter coefficients are further smoothed along the frequency axis by mel-frequency filter-banks, resulting in a mel-warped frequency domain Wiener filter. The impulse response of this mel-warped Wiener filter is achieved by applying a mel-warped IDCT. Finally, the enhanced signal is achieved by applying the Wiener filter to the input signal.

According to ETSI-AFE, the frequency response for noise reduction is derived from a priori SNR. The a priori SNR is estimated from a posteriori SNR according to a decision-directed approach [4], [9]. The a priori SNR is derived according to (1) through (4). The a posteriori SNR is obtained by

$$\xi(bin,t)=\frac{P_{\text{in\_PSD}}^{1/2}(bin,t)}{P_{\text{noise}}^{1/2}(bin,t)}-1\ , \tag{1}$$

where $P_{\text{in\_PSD}}^{1/2}(bin,t)$ and $P_{\text{noise}}^{1/2}(bin,t)$ denote the magnitude of the PSD of an input signal and background noise on the frequency index, *bin*, at time *t*.

$$\eta(bin,t)=\beta\times\frac{P_{\text{den3}}^{1/2}(bin,t-1)}{P_{\text{noise}}^{1/2}(bin,t)}+(1-\beta)\times\max\left[\xi(bin,t),0\right], \tag{2}$$

$$P_{\text{den2}}^{1/2}(bin,t)=\eta(bin,t)P_{\text{in\_PSD}}^{1/2}(bin,t), \tag{3}$$

$$\eta_2(bin,t)=\max\left[\frac{P_{\text{den2}}^{1/2}(bin,t)}{P_{\text{noise}}^{1/2}(bin,t)},\eta_{TH}\right]. \tag{4}$$

The transfer function of the Wiener filter is then obtained using the a priori SNR by

$$H(bin,t)=\frac{\eta_2(bin,t)}{1+\eta_2(bin,t)}. \tag{5}$$

The spectral amplitude of a noiseless signal is calculated as

$$P_{\text{den3}}^{1/2}(bin,t)=H(bin,t)P_{\text{in}}^{1/2}(bin,t), \tag{6}$$

where $P_{\text{in}}^{1/2}(bin,t)$ denotes the amplitude spectrum of the input signal. As shown in (1), the a posteriori SNR is obtained directly from the input spectra even in the presence of adverse noise. In a fast moving car, the input spectra are already contaminated by ambient noise. Therefore, the output signal distortion of ETSI-AFE is inevitable due to this direct estimate on a posteriori SNR from input observations.

III. Proposed Approaches

The simple motivation of this work is that a posteriori SNR estimation with statistical precision is helpful for enhancing the speech quality of the conventional decision-directed Wiener filter. To realize this idea, we integrate a decision-directed method with a model-based speech spectrum reconstruction method into a new noise reduction filter estimation stage. That is, the de-noised signal spectra are attained by using the conventional decision-directed method, and the original speech spectra are restored from the de-noised spectra. Then, a posteriori SNR with statistical precision is achieved from the reconstructed speech spectra. In this work, two spectrum reconstruction methods are adopted to restore an original clean spectrum. One is based on a single density GMM, which is established with clean signals, and the other involves a joint density GMM pre-trained with joint samples (clean and noisy signal pairs). In the proposed method, the original spectrum is restored by two steps. The first step is a clean amplitude spectrum estimation based on GMM. The mapping function to transform a noisy spectral vector into a clean spectral vector is a least-squares regression estimate [19], [20]. The second step is a generation of the spectral amplitude of a clean signal.

Let $x\in R^n$ be the spectral vector of a noise-corrupted signal, and $y\in R^n$ be the spectral vector of a clean signal.

The goal of spectral reconstruction is to find a mapping function, F, that minimizes the mean square error,

$$\varepsilon_{\text{mse}} = E\left[\left\|y - F(x)\right\|^2\right], \tag{7}$$

where $E[\cdot]$ denotes the expectation, and $F(x)$ is the reconstructed clean spectral vector. Details of the mapping functions are found in [19] and [20]. In our approaches, the clean spectral amplitude is restored in a logarithmic spectrum domain in order to mimic human auditory characteristics.

1. Mapping Function Based on GMM

The distribution density of x is modeled as a Gaussian mixture density, that is, a mixture of Q component densities given by

$$\begin{aligned} p(x \mid \lambda) &= \sum_{i=1}^{Q} \alpha_i b_i(x), \sum_{i=1}^{Q} \alpha_i = 1, \alpha_i \geq 0, \\ b_i(x) &= \frac{\alpha_i}{(2\pi)^n \left|C_i\right|^{1/2}} \exp\left[-\frac{1}{2}\left(x-\mu_i^x\right)^T C_i^{-1}\left(x-\mu_i^x\right)\right], \end{aligned} \tag{8}$$

where λ denotes the GMM of clean speech, and α_i, μ_i, and C_i indicate a priori probability, mean vector, and covariance matrix of class I, respectively. The mapping function that minimizes the mean square error is obtained as follows:

$$F(x) = \sum_{i=1}^{Q} h_i(x) \times \mu_i^x,\ h_i(x) = \frac{b_i(x)}{\sum_{j=1}^{Q} b_j(x)}. \tag{9}$$

The weighting function $h_i(x)$ denotes the a posteriori probability of the i-th Gaussian component generated by vector x.

2. Mapping Function Based on Joint Density GMM

Let $z = (y, x)^T$ be a joint vector between the spectral vector of a clean signal and the spectral vector of a noisy signal. If the joint density of the vectors is modeled as a mixture of Q $2n$-variate Gaussian functions, the mapping function based on the joint density is obtained as follows:

$$F(x) = E[y \mid x] = \sum_{i=1}^{Q} h_i(x)\left[\mu_i^y + C_i^{yx} C_i^{xx-1}\left(x-\mu_i^x\right)\right], \tag{10}$$

where

$$h_i(x) = \frac{\dfrac{\alpha_j}{(2\pi)^n \left|C_i^{xx}\right|^{1/2}} \exp\left[-\dfrac{1}{2}\left(x-\mu_i^x\right)^T C_i^{xx-1}\left(x-\mu_i^x\right)\right]}{\displaystyle\sum_{j=1}^{Q} \frac{\alpha_j}{(2\pi)^n \left|C_j^{xx}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(x-\mu_j^x\right)^T C_j^{xx-1}\left(x-\mu_j^x\right)\right]}$$

with $C_i = \begin{bmatrix} C_i^{xx} & C_i^{xy} \\ C_i^{yx} & C_i^{yy} \end{bmatrix}$ and $\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$.

3. Spectral Compensation Based on GMM

The logarithmic spectral amplitude of the de-noised signal in (6), which is obtained by the tradition decision-directed approach, is used for original spectrum reconstruction. By doing so, the variability of the input spectra is alleviated in the presence of severe noise, and the reconstruction performance becomes reliable. The amplitude spectrum of a clean signal is restored as follows:

$$\begin{aligned} P_{\text{den4}}^{1/2}(bin,t) = {} & \alpha(bin,t) P_{\text{den3}}^{1/2}(bin,t) + \\ & \left(1-\alpha(bin,t)\right) \exp\left[F\left\{\log\left(P_{\text{den3}}^{1/2}(bin,t)\right)\right\}\right]. \end{aligned} \tag{11}$$

The optimized updating factor α can be derived by minimizing the mean square error as follows,

$$\varepsilon_{\text{mse}} = E\left[\left\|P_{\text{noiseless}}^{1/2}(bin,t) - P_{\text{den4}}^{1/2}(bin,t)\right\|^2\right], \tag{12}$$

$$\alpha(bin,t) = \frac{SNR(bin,t) + \overline{SNR}(bin,t)}{1 + SNR(bin,t) + \overline{SNR}(bin,t)}, \tag{13}$$

where

$$SNR(bin,t) = \frac{P_{\text{noiseless}}^{1/2}(bin,t)}{P_{\text{noise}}^{1/2}(bin,t)},\ \overline{SNR}(bin,t) = \frac{\exp\left[F\left\{\log\left(P_{\text{den3}}^{1/2}(bin,t)\right)\right\}\right]}{P_{\text{noise}}^{1/2}(bin,t)}.$$

In this work, the noiseless amplitude spectrum in (13) is substituted by the de-noised spectrum in (6). For the robustness of the algorithm, a time-frequency smoothing step is recommended as

$$\overline{\alpha}(bin,t) = \sum_{i=-L_f}^{i=L_f} \sum_{j=L_t}^{j=L_t} w_f(i) w_t(j) \alpha(bin+i, t+j), \tag{14}$$

where

$$w_f(i) = w_f(-i) = \frac{\left(1-w_f(0)\right)\left(L_f+1-i\right)}{L_f\left(L_f+1\right)},\ 1 \leq i \leq L_f,$$

$$w_t(j) = w_t(-j) = \frac{\left(1-w_t(0)\right)\left(L_t+1-j\right)}{L_t\left(L_t+1\right)},\ 1 \leq j \leq L_t.$$

The transfer function for noise suppression is then obtained according to (15) through (17). An enhanced a posteriori SNR is calculated as

$$\xi_m(bin,t) = \frac{P_{\text{den4}}^{1/2}(bin,t)}{P_{\text{noise}}^{1/2}(bin,t)}. \tag{15}$$

An enhanced a priori SNR is obtained as

$$\eta_3(bin,t) = \beta \times \frac{P_{\text{den5}}^{1/2}(bin,t-1)}{P_{\text{noise}}^{1/2}(bin,t)} + (1-\beta) \times \max\left[\xi_m(bin,t), 0\right]. \quad (16)$$

Then, an enhanced transfer function is obtained as

$$H_2(bin,t) = \frac{\eta_3(bin,t)}{1+\eta_3(bin,t)}. \quad (17)$$

The amplitude spectrum of a clean signal is then obtained according to (18):

$$P_{\text{den5}}^{1/2}(bin,t) = H_2(bin,t) P_{\text{den4}}^{1/2}(bin,t). \quad (18)$$

4. Proposed Model-Based Noise Reduction Approach

Although the two filtering stages are substituted by a single reduction filter in the proposed approach, the spectral whitening and signal distortion drawback are still alleviated by the proposed two-stage filter estimation scheme. The proposed noise reduction filter with statistical precision is achieved as follows. In the first stage, the de-noised spectrum is derived from the input by the conventional decision-directed method. The original spectrum reconstruction is followed in the second stage. Then, an enhanced transfer function is obtained according to (11) through (17). Figure 2 shows a block diagram of the proposed model-based noise reduction approach.

The green blocks in Fig. 2 indicate the proposed second noise reduction filter estimation stage, and the proposed method does not require an actual second filtering stage. Therefore, the frame latency is reduced by half when compared with the two filtering stages of ETSI-AFE.

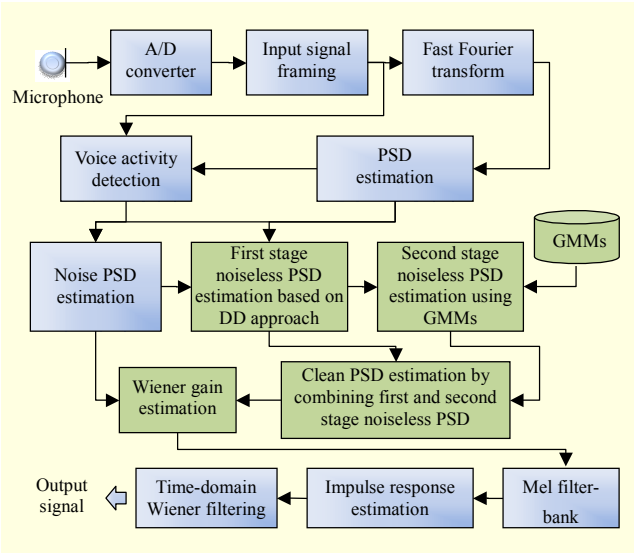

Fig. 2. Block diagram of the proposed model-based noise reduction approach.

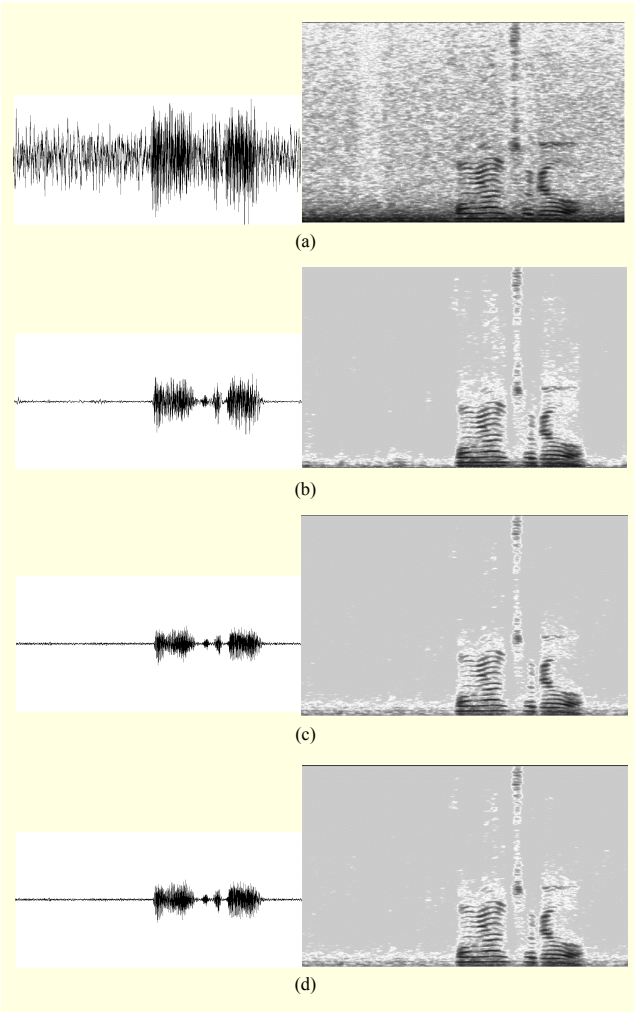

Fig. 3. (a) Original noisy signal and spectrogram and (b) enhanced signal and spectrogram after 2-stage Wiener filter (c) after the proposed approach with a single density GMM and (d) after the proposed approach with a joint density GMM.

Figure 3(a) shows a noisy signal sample and its corresponding spectrogram recorded in a moving car at medium speed of 60 km/h to 80 km/h. The acoustic transducer is located on the center-fascia of a D segment sedan, where after-market navigation is usually installed. The acoustic waveform is digitalized into 16 bit pulse-code modulation (PCM) format at 16 kHz sampling. Since ETSI-AFE supports 8 kHz sampling rate, the original algorithm is modified to process 16 kHz samples. As shown in Fig. 3(a), main vehicle noise components are spread in the low-frequency region. Although the conventional ETSI-AFE removes additive noise components twice, car-noise components are still found in Fig. 3(b), particularly in the speech portions. Although a single noise suppression filter is employed in the proposed approach, Figs. 3(c) and 3(d) show that the car-noise components are efficiently suppressed even in the speech portions.

IV. Experimental Results

The aim of this work is to develop a robust and efficient signal pre-processor for speech recognition in a car environment. Since the resources of an automotive device are constrained, a tradeoff between computational efficiency and performance should be considered in developing a speech enhancement algorithm. Under the hardware specifications of modern commercial automotive devices, it is assumed that the noise suppression approach in ETSI-AFE is a relatively good solution. Therefore, our research is focused on improving the performance and computational efficiency of the traditional two-stage mel-warped Wiener filter while reducing the computational complexity. The proposed model-based approach can be implemented in two different branches. One is based on a GMM (a single density GMM-based method) and the other is using a joint density model (a joint density GMM-based method). The performance evaluation of the proposed approach is also performed by comparing it with ETSI-AFE. The performance evaluation on speech enhancement methods can be performed in terms of SNR or segmental SNR improvement, a signal distortion measure, or a subjective listening test such as mean opinion score test. In this experiment, the evaluation is conducted in terms of average isolated word recognition (IWR) rate since the proposed method is targeted at voice recognition. The simple assessment on the computational cost is also conducted by measuring the program completion time.

1. Speech Recognizer Preparation and GMM building

A HMM-based speech recognition system was prepared to demonstrate the performance of the proposed approach in simulative and real car environments. To establish initial acoustic models (AMs), a number of phonetically optimized utterances were recorded from 1,500 persons. The speech waveforms were digitalized into 16 bit PCM at 16 kHz sampling rate. Since the initial speech signals are collected in an office room, these initial AMs do not match the car environment. Therefore, an AM adaptation is necessary for environmental adjustment. In this paper, the initial AMs are adjusted using the discriminative AM adaptation method in [21], [22]. For this AM adaptation procedure, 8,000 utterances were collected from 80 speakers in moving cars at a medium speed of 60 km/h to 100 km/h. As speech recognition features, mel-frequency cepstral coefficients (13 MFCCs including C0) and first and second derivatives are extracted. The voice recognizer is able to identify 46,000 words, and the vocabulary word list is composed of destination entries for a car navigation system. Some examples of the Korean points of interest and their corresponding lexicons are as follows:

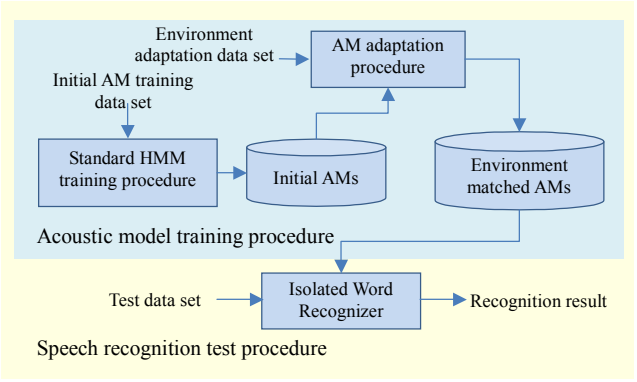

Fig. 4. Block diagram of environment matched AM training procedure.

포곡점현대그린서비스 p o g o xg z v xm h jv xn d E g U r i xn s v b i s U
포구 p o g u
포대포 p o d E p o
포도나무미술학원 p o d o n a m u m i s u xl h a g wv xn
포도마을사거리 p o d o m a U xl s a g v r i
포도원식당 p o d o wv xn s i xg D a N
포동 p o d o N
포동부동산 p o d o N b u d o N s a xn
포드 p o d U
포드서비스센터 p o d U S v b i S U S e xn t v

Figure 4 represents the AM training procedure to obtain an environment matched AM for a speech recognition system. As shown in Fig. 4, the car environment matched AM is established by adjusting the initial models.

In the proposed model-based noise reduction approaches, two kinds of GMMs are required. One is a single density GMM, and the other is a joint density GMM. In order to build a single density GMM which represents the pre-trained knowledge of speech for original spectrum reconstruction, 4,000 utterances are collected from 100 persons (50 males and 50 females) inside an idling car. Since it is difficult to record clean and noisy signals at the same time, the immunity learning scheme is adopted to obtain joint samples (clean and noisy signal pairs). Moving car sounds at medium and high speeds of 60 km/h to 120 km/h are recorded for an hour while driving on a highway including asphalt and concrete road surface. The car noise signal is artificially added to the previous utterances in a random manner. Thus, the 4,000 joint samples (clean and artificially noise-added signal pairs) are obtained, and these artificial joint samples are used to estimate a joint density GMM. A Hyundai Verna, a C segment sedan, was used for this car noise acquisition. The speaker lists involved in the HMM and GMM buildings are not overlapped. In order to find the optimal mixture size of GMMs, several speech recognition

tests are conducted by changing the mixture size (from 8 to 512). The performance difference depending on the mixture size can be ignored due to the time-frequency spectrum smoothing and mel-frequency warping scheme in Fig. 2. Therefore, the GMMs (a single density GMM and a joint density GMM) of 8 mixture size are selected for the computational efficiency.

2. Isolated Word Recognition Tests

The performance of the proposed method is demonstrated in simulative and real car environments. For the simulative test, 269 utterances are collected from 20 speakers (8 males and 12 females) in an idling car. A sports utility vehicle (SUV) manufactured by GM-Daewoo Motors is used for this speech database (DB) acquisition, and car noise signal on an asphalt road is also recorded for artificial noise addition. Since the SNR is estimated in the SNR range between 15 dB and 20 dB in an idling car, car noise is artificially added in the range between –5 dB and 10 dB.

Figure 5(a) shows a signal sample and its corresponding spectrogram recorded in an idling car. Their artificial noisy signal and spectrum pair is shown in Fig. 5(b). Since the SUV sounds are recorded at medium speed of 60 km/h to 80 km/h on a smooth asphalt road, it is seen that car noise components

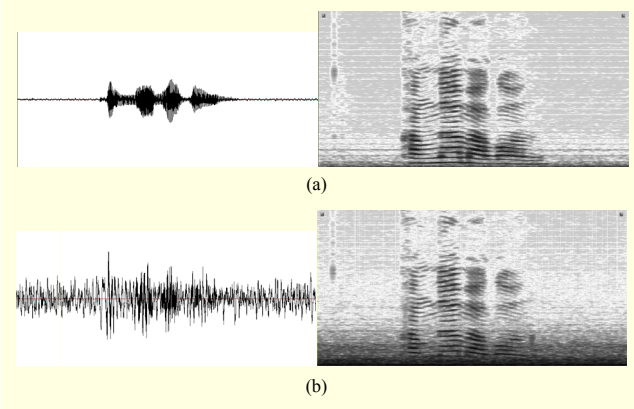

Fig. 5. (a) Signal and spectrogram sample in an idling car and (b) artificial noisy signal and spectrogram sample in –5 dB.

Table 1. IWR results for simulative data.

	Original	AFE noise reduction	Single density GMM	Joint density GMM
Idling	92.19%	97.03%	97.4%	97.77%
SNR 10	88.1%	95.91%	96.28%	96.65%
SNR 5	85.13%	95.91%	96.28%	96.65%
SNR 0	71.38%	92.24%	93.31%	93.68%
SNR –5	46.47%	86.99%	89.96%	90.71%

are concentrated in the low and narrow frequency region. It is known that this kind of colored noise causes only small degradation of voice recognition performance.

Table 1 shows voice recognition test results in the simulative condition. As shown in Table 1, the proposed single density GMM-based method and joint density GMM-based method show a competitive performance with ETSI-AFE. However, the performance improvement of the proposed methods is not noticeable since the characteristics of additive noise are not so complex, and the word lengths of the speech samples are relatively long having more than 4 syllables. Furthermore, the noise only positions in the front and back of the utterance are relatively short at around 500 ms.

For the real environment test, 1,252 utterances are recorded from 30 speakers (15 males and 15 females) in various car environments (idling to high speeds over 100 km/h). Various vehicle types, such as D and E segment sedans and SUVs, served for this real test data collection. The SNR range of the real speech DB is measured from –10 dB to 20 dB. This real test DB represents real driving situations well, including car-engine noise, street noise, wind noise from an air-conditioner or opened window, burst noise caused by a coarse road surface, and so on. Therefore, the voice recognition accuracy without noise reduction is very low (IWR accuracy of 9.5%).

As shown in Fig. 6, the characteristics of real car noise are different from the simulative car noise. Since used cars are served for real speech acquisition and the road surface is not smooth (including a concrete road), the frequency components of real car noise are relatively widespread while the main components are still concentrated in low frequency region. Also, the noise only positions in the front and back of the utterance are relatively long at more than 1 s. Therefore, it is found that the speech recognizer has difficulty identifying the

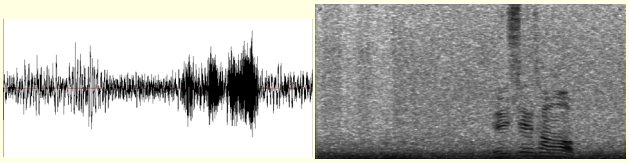

Fig. 6. Signal sample and spectrogram at medium speed of 60 km/h to 80 km/h.

Table 2. IWR results for real data.

	AFE noise reduction	Single density GMM	Joint density GMM
ASR accuracy	76.42%	85.06	86.74%
ERR	0%	36.64%	43.77%

test utterances.

Table 2 shows the average speech recognition rate and the corresponding error reduction ratios (ERRs). As shown in Table 2, the proposed method is superior to ETSI-AFE. Since the pre-trained knowledge on the original and the noise-impaired signals is available in a joint density GMM, the proposed method based on joint density GMM shows a significant performance improvement.

3. Computational Complexity Tests

The computational complexity of the proposed method based on a single density GMM is compared with ETSI-AFE. For this computational cost comparison, the program completion time for several noisy signal samples is measured on a personal computer (PC) equipped with a 2.2 GHz CPU and 2G memory. The operating system of the PC is based on Linux. For the algorithm integration into small automotive devices such as car navigation, fixed-point arithmetic conversion is inevitable since a commercial automotive device does not have a floating-point arithmetic processor for price competiveness. Therefore, ETSI-AFE and the proposed single density GMM-based method are implemented using fixed-point arithmetic in C program-language. A GNU program compiler is adopted to execute both the noise reduction algorithms. The results of the computational complexity tests indicate that the computational load of the proposed method is mitigated by average 40% when compared with ETSI-AFE. The computational load difference between the proposed approaches is caused by the alternative spectral mean estimation procedure in the mapping function. Since a diagonal covariance matrix is used in the proposed joint density GMM-based method for computational efficiency, the additional computational burden can be ignored. Therefore, it is presumed that the proposed joint density GMM-based method is also computationally efficient.

V. Conclusion

In this paper, a statistical model-based noise reduction approach for a speech recognition system in a car environment was proposed. The proposed approach is motivated by the simple idea that the enhancement qualities of the traditional decision-directed Wiener filter might be improved if a more precise estimate on a posteriori SNR is possible. In order to implement this idea, the two filtering concepts in ETSI-AFE are compressed into a single filter in order to alleviate the spectral whitening and signal distortion problem in the traditional decision-directed Wiener filter. In this work, we combine a decision-directed method with an original spectrum reconstruction method and develop a new two-stage noise suppression filter estimation scheme. The proposed two-stage noise suppression filter estimation scheme was demonstrated in terms of average voice recognition rate in the simulative and real car environments. The experimental results showed that the proposed approach is superior to ETSI-AFE because the proposed approach efficiently alleviates the spectral whitening and signal distortion artifacts. Since the two filtering stages in ETSI-AFE are compressed into a single filter in the proposed approach, a secondary filtering stage is not required. Therefore, the computational load of the proposed method is mitigated by an average of 40% while reducing the frame latency by half when compared with ETSI-AFE. The proposed method will be adapted to a huge vocabulary voice recognition system for commercial car navigation [23].

References

[1] Y. Gong, "Speech Recognition in Noisy Environments: a Survey," *Speech Commun.*, vol. 16, no. 3, Apr. 1995, pp. 261-291.

[2] Y. Suh and H. Kim, "Feature Compensation Combining SNR-Dependent Feature Reconstruction and Class Histogram Equalization," *ETRI J.*, vol. 30, no. 5, Oct. 2008, pp. 753-755.

[3] J. Lim and A. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. IEEE*, vol. 67, no. 12, Dec. 1979, pp. 1586-1604.

[4] ETSI Std. Document, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithm," ETSI ES 202 050 V1.1.1 (2002-10).

[5] A. Agarwal and Y. Cheng, "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition," *Proc. IEEE-ASRU Workshop*, 1999, pp. 12-15.

[6] M. Cheng et al., "A Robust Front-End Algorithm for Distributed Speech Recognition," *Proc. EUROSPEECH*, 2001, pp. 425-428.

[7] D. Macho et al., "Evaluation of a Noise-Robust DSR Front-End on Aurora Databases," *Proc. ICSLP*, Sept. 2002, pp. 17-20.

[8] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans., Acoustics, Speech, Signal Process.*, vol. 27, no. 2, Apr. 1979, pp. 113-120.

[9] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 32, no. 6, Dec. 1984, pp. 1109-1121.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 33, no. 2, Apr. 1985, pp. 443-445.

[11] W. Wu and P. Chen, "Subband Kalman Filtering for Speech Enhancement," *IEEE Trans. Circuits Syst. II: Analog Digit.*

Signal Process., vol. 45, no. 8, Aug. 1998, pp. 1072-1083.
[12] J. Gibson, B. Koo, and S. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, Aug. 1991, pp. 1732-1742.
[13] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, Mar. 1999, pp. 126-137.
[14] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proc. IEEE*, vol. 80, no. 10, Oct. 1992, pp. 1526-1555.
[15] H. Sameti et al., "HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise," *IEEE Trans. Speech Audio Process.*, vol. 6, Sept. 1998, pp. 445-455.
[16] J. Wu et al., "A Noise-Robust ASR Front-End Using Wiener Filter Constructed from MMSE Estimation of Clean Speech and Noise," *Proc. IEEE-ASRU Workshop*, 2003, pp. 321-326.
[17] T. Arakawa, M. Tsujikawa, and R. Isotani, "Model-Based Wiener Filter for Noise Robust Speech Recognition," *Proc. ICASSP*, 2006, pp. 537-540.
[18] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley: NY, 1949.
[19] A. Kain and M. Macon, "Spectral Voice Conversion for Text-To-Speech Synthesis," *Proc. ICASSP*, 1998, pp. 285-288.
[20] K. Park and H.S. Kim, "Narrowband to Wideband Conversion of Speech using GMM based Transformation," *Proc. ICASSP*, vol. 3, June 2000, pp. 1843-1846.
[21] B. Kang, H. Jung, and Y. Lee, "Discriminative Noise Adaptive Training Approach for an Environment Migration," *Proc. INTERSPEECH*, Aug. 2007, pp. 2085-2089.
[22] H. Jung, B. Kang, and Y. Lee, "Model Adaptation using Discriminative Noise Adaptive Approach for New Environments," *ETRI J.*, vol. 30, no. 6, Dec. 2008, pp. 865-867.
[23] S. Lee et al., "A Commercial Car Navigation System Using Korean Large Vocabulary Automatic Speech Recognizer," *Proc. APSIPA ASC*, Oct. 2009, pp. 286-289.

Sung Joo Lee received his BS and MS in electronic engineering from Pusan National University, Busan, Korea, in 1996 and 1998, respectively. After graduation, he joined Hyundai Electronics Multimedia Research Center for voice/audio codec applications. Since 2000, he has been with ETRI, Daejeon, Korea, and is a senior researcher in the Speech/Language Information Research Center. His research interests include environment-robust speech signal processing and speech recognition.

Byung Ok Kang received his BS and MS in electrical and electronics engineering from the POSTECH, Korea, in 1997 and 1999, respectively. After graduation, he joined Samsung Electronics S/W Center for mobile phone application. Since 2002, he has been a researcher at the Speech/Language Information Research Center of ETRI, Korea, and has continued his research on speech signal processing and speech recognition applications.

Ho-Young Jung received his MS and PhD in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1995 and 1999, respectively. His PhD dissertation was on robust speech recognition. He joined ETRI, Daejeon, Korea, in 1999 as a senior researcher and has belonged to the Speech/Language Information Research department as a principle researcher. His current research interests include speech recognition, noise-robust processing, blind signal separation, and machine learning. He has published or presented 35 papers in speech recognition.

Yunkeun Lee received the BS and MS in electronic engineering from Seoul National University and KAIST in 1986 and 1988, respectively. He received the PhD in information and communication engineering from KAIST, Seoul, Rep. of Korea, in 1998. Currently he is in charge of the Spoken Language Processing Team, ETRI, Daejeon, Rep. of Korea. His research interests include speech recognition, speech synthesis, and speech enhancement.

Hyung Soon Kim received the BS in electronics engineering from Seoul National University, Korea, in 1983, and the PhD in electrical and electronics engineering from the KAIST, Korea, in 1989. From 1987 to 1992, he was with the DigiCom Institute of Telematics, Korea, where he was a technical manager of the Speech Communication Division. Since 1992, he has been with Pusan National University, Korea. He is a professor in the School of Electrical Engineering. From 1999 to 2000, and from 2006 to 2007, he was a visiting scholar at Carnegie Mellon University and Oregon Health and Science University, USA, respectively. His research interests include speech recognition, speaker adaptation, and noise-robust speech signal processing.