A Robust Adaptive Speech Enhancement System for Vehicular Applications

Jwu-Sheng Hu, *Member*, IEEE, Chieh-Cheng Cheng, Wei-Han Liu, and Chia-Hsing Yang

Abstract —*This work proposes and implements a novel and robust adaptive speech enhancement system, which contains both time domain and frequency domain beamformers using H_∞ filtering approach to provide a clean and undisturbed speech waveform and improve the speech recognition rate in vehicle environments. A microphone array data acquisition hardware is also designed and implemented for the proposed speech enhancement system. Mutually matched microphones are needed for traditional multidimensional noise reduction methods, but this requirement is not practical for consumer applications from the cost standpoint. To overcome this issue, the proposed system adapts the mismatch dynamics to maintain the theoretical performance allowing unmatched microphones to be used in an array. Furthermore, to achieve a high speech recognition performance, the speech recognizer is usually required to be retrained for different vehicle environments due to different noise characteristics and channel effects. The proposed system using the H_∞ filtering approach, which makes no assumptions about noise and disturbance, is robust to the modeling error in a channel recovery process. Consequently, the real vehicular experimental results show that the proposed frequency domain beamformer provides a satisfactory speech recognition performance without the need to retrain the speech recognizer.*[1]

Index Terms —Human Machine Interaction, Speech Enhancement, Automatic Speech Recognition, H_∞ filtering.

I. INTRODUCTION

Electronic systems in vehicles are increasingly popular. Given concerns over driving safety and convenience, these in-vehicle electronic devices such as global positioning system (GPS), CD or VCD player, air conditioner, etc. should not be accessed by hands while driving. Therefore, intelligent human-computer interaction interfaces with speech recognition have recently been proposed [1]-[3] to control these in-vehicle devices through voice. However, the poor speech quality, acoustic echo of the far-end speech, and environmental noises degrade the speech recognition performance, resulting in a low acceptance of the speech recognition technology by consumers. Therefore, speech enhancement techniques such as single channel [4]-[5] and multi-channel [6]-[13] noise suppression approaches have been introduced to overcome these issues. Although using a single channel approach can reduce the hardware complexity, the performance degrades due to various problems, such as musical tones [12].

The work in [12] demonstrates that applying a high-pass filter with the cut off frequency of 240Hz can significantly improve the speech recognition rate, when dual channel microphones are used. This is because the vehicular noise caused by driving is dominated by low frequency components, i.e., 50-800 Hz from the engine, air flow, tire noise, road noise, and so on. However, the low frequency components contain useful information about speech characteristics. Consequently, applying the high-pass filter may cause speech distortion and requires extra training processes in a car for the speech recognizer to obtain a good recognition rate. Moreover, when the in-car audio system is turned on, the spectrum of the environmental noise exhibits a wide band behavior instead of a lower frequency one. Consequently, the dual channel noise suppression method with a high-pass filter cannot provide a satisfactory performance.

To overcome this limitation, microphone array based noise suppression approaches, such as Frost beamformer [6], generalized sidelobe canceller (GSC) [7], [9], and robust adaptive beamformer [8] have been proposed to achieve better performance than the single and dual channel cases. However, these methods also have limitations. For example, the microphones must to be mutually matched and have no coherent interference signal. Dahl et al. [10] proposed a finite impulse response (FIR) based normalized least-mean-square (NLMS) filtering approach to perform indirect microphone calibration and to minimize the speech distortion due to the channel effect by using pre-recorded speech signals and a desired signal which are acquired when the environment is quiet. Because the variation between pre-recorded speech signals and the desired signal contain useful information about the dynamics of channel, electronic equipments uncertainties, and microphones' characteristics, the method in [10] outperforms other un-calibrated algorithms in real applications [11]. However, the FIR filter using a finite number of taps is unlikely to characterize the full channel dynamics between sound sources and microphone arrays in an enclosure [14]. Moreover, the NLMS based formulation assumes that the disturbance is uncorrelated to the source, zero mean and

[1] Jwu-Sheng Hu is with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: jshu@cn.nctu.edu.tw).

Chieh-Cheng Cheng is with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: canson.ece89g@nctu.edu.tw).

Wei-Han Liu is with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: lukeliu.ece89g@nctu.edu.tw).

Chia-Hsing Yang is with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: chyang.ece92g@nctu.edu.tw).

Contributed Paper
Manuscript received July 10, 2006

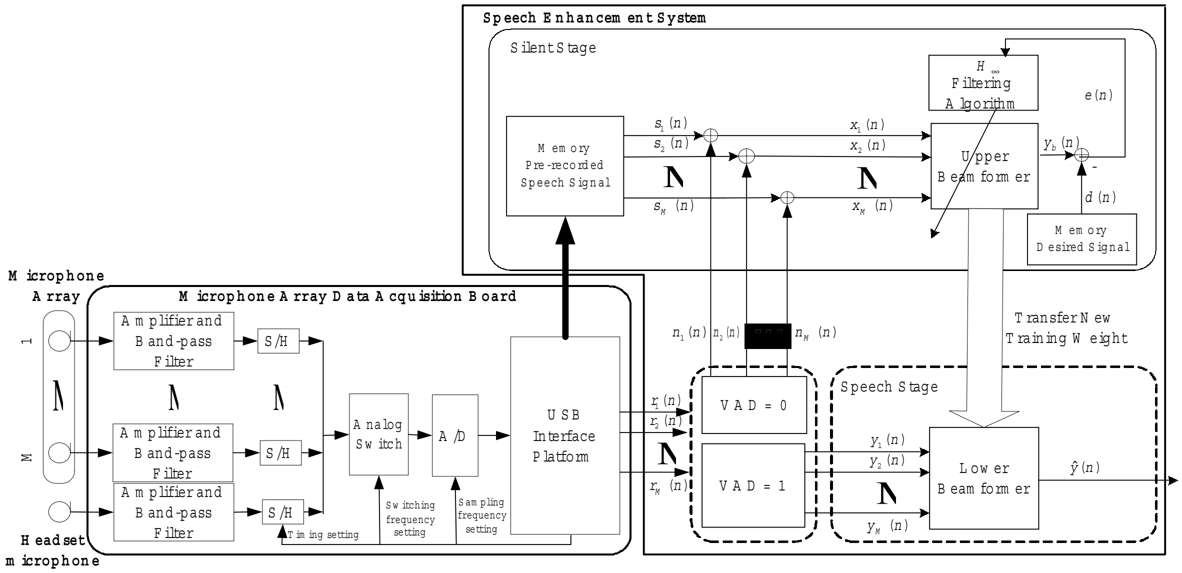


Fig. 1. Overall System Architecture

Gaussian distributed. These assumptions limit speech enhancement performance.

This work presents two beamformers using H_∞ filtering approach to suppress environmental noises and decrease the speech distortion caused by the channel effect. The two beamformers using H_∞ filtering approach are robust to the modeling error caused by the finite tap length of FIR filters and make no assumptions regarding the characteristics of environmental noises [15], [16]. Furthermore, the pre-recorded speech signals and the desired signal can be adopted to suppress the gain from noises to the output especially in low frequencies and the characteristics of the received multi-channel signals can be automatically adjusted to those of the desired signal. Consequently, high-pass filtering and extra retraining processes for a speech recognizer in vehicles are not needed in this work. In this work, a time domain beamformer using H_∞ filtering approach is proposed to produce a clean and undisturbed speech waveform. On the other hand, for speech recognition applications, a frequency domain beamformer using H_∞ filtering approach is proposed to reduce the effect of uncertainty in signal transformation between the time domain and the frequency domain by treating several frames as a single block. The proposed beamformers using two microphones outperform dual channel delay-and-sum beamformer with a high-pass filter introduced in [12] especially when the in-car audio system plays music which acts like a broadband disturbance. This work compares the performance of different numbers of the microphones and the experimental results show that using more microphones improves the performance.

The remainder of this work is organized as follows. The proposed speech enhancement system and the designed microphone array data acquisition hardware are introduced in section 2. Section 3 presents the two proposed beamformers using H_∞ filtering approach in both the time and frequency domains. Section 4 presents several representative experiments in a real vehicle and discusses the experimental results. Conclusions are finally drawn in the last section.

II. Speech Enhancement System and Microphone Array Data Acquisition Hardware Implementation

The overall system architecture can be illustrated as Fig. 1 and can be divided into two sub-systems. The first sub-system consists of a microphone array whose geometry can be flexibly arranged and a data acquisition electronics prototype designed by this work. The main feature of this design is its ability to digitalize the received sound signals and transmit them in real-time via a USB interface. The second sub-system is the speech enhancement system.

A. Microphone Array and Microphone Array Data Acquisition Board

The microphone array consists of M omni-directional condenser microphones and a headset microphone. The frequency response of the microphone ranges from 50 Hz to 16 kHz. The microphone array acquisition board comprising a four-layer board can be divided into three parts. The first part includes microphone amplifiers and filters. The second part is digitization. The third part contains control and communication.

In the first part, the microphone signals are amplified and filtered by six amplifiers and six band-pass filters designed by taking the microphone sensitivity and anti-aliasing into the consideration. The gains of the M microphones and the headset microphone are respectively set to 60dB and 20dB. The second part comprises six sample-and-hold circuits (S/H), one analog switch, and one analog-to-digital (A/D) converter. By acquiring the six channel sound signals with one analog switch and one A/D converter, the system can significantly reduce the power consumption requirement and improve the flexibility for portable consumer applications.

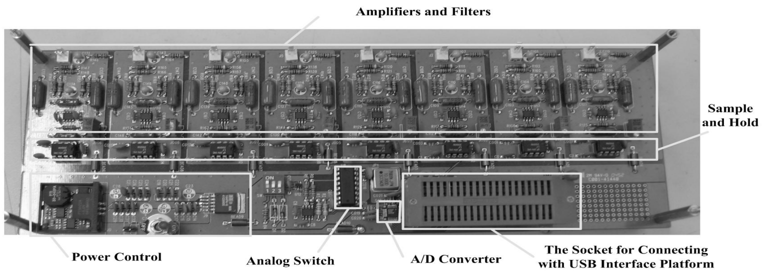


Fig. 2. Microphone array data acquisition board

The third part contains the control and data transmission lines, which are controlled by the USB interface. The USB interface platform can control the timing of the sample-and-hold circuits, analog switch, and A/D converter through the control line. The switching frequency and the timing of the system can be flexibly selected, and the sampling frequency is set to 8 kHz in this work. The converted 16-bit digital data are transmitted in real-time through the USB interface with timing controlled by a micro-controller. The picture of the microphone array data acquisition board is shown in Fig. 2. Fig. 3 shows the installation of the array inside the vehicle. Note that the headset microphone is used only to collect the desired signal, i.e., the user does not need the headset microphone during the online applications.

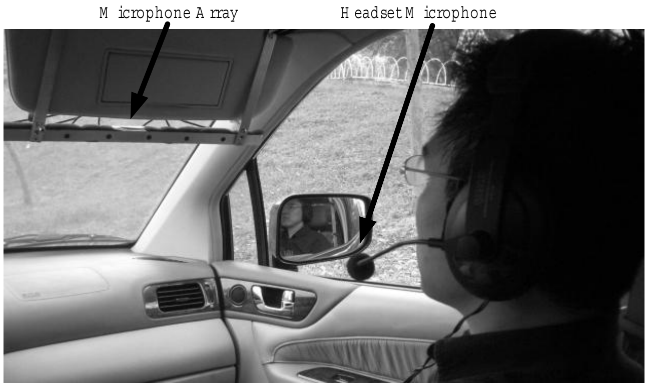


Fig. 3. Installation of the array inside the vehicle

B. Speech Enhancement System

The Speech Enhancement system is separated into two stages, namely the silent and speech stages, by a voice activity detector (VAD) that identifies speech in the received signals. The voice activity detection algorithm can be found in [17] and [18]. If the result of the VAD equals to zero, which means that no speech exists, the system will be run in the silent stage. The system can be switched to the speech stage when the result of the VAD equals to one.

The pre-recorded speech signals shown in the silent stage in Fig. 1 are collected when the environment is quiet and the speaker is at the desired location. The pre-recorded speech signals contain both the characteristics of microphones and the acoustical characteristics of the desired location. The desired signal, $d(n)$, is derived from a headset microphone when the pre-recoded speech signals are collected. Since the headset is close to the mouth, the desired signal contains little channel distortion. The desired signal only needs to be collected when the desired location varies, so the headset microphone is not needed during the online applications. In the silent stage, the environmental noises without speech signals are recorded online. The environmental noises are assumed to be additive, so the signals received when a speaker is talking in a noisy environment can be expressed as a linear combination of the speech signals and the environmental noises. Therefore, in this stage, the system combines the online recorded environmental noises, $n_1(n),\Lambda ,n_M(n)$, with the pre-recorded speech signals, $s_1(n),\Lambda ,s_M(n)$, to construct the training signals, $x_1(n),\Lambda ,x_M(n)$. The weighting vector is derived from the training signals using H_∞ based adaptive filtering approach. In the speech stage, the trained weighting vector is passed to the lower beamformer to purify and recover the noisy received signals, $y_1(n),\Lambda ,y_M(n)$.

C. Voice Activity Detection (VAD) Algorithm

The VAD algorithm [17] in this work can adjust itself according to current environmental noise based on the estimation of the long-term spectral envelope (LTSE). The LTSE tracks the spectral envelope using long-term rectangular speech window information. Assume that $r_i(n)$, the received signal of the ith microphone, is utilized to detect the speech. Then, the J-order LTSE can be defined as:

$$\mathrm{LTSE_J}(b,k)=\max\{R_i(b,k+j)\}_{j=-J}^{j=+J} \tag{1}$$

where $R_i(b,k)$ denotes the amplitude spectrum of $r_i(n)$ at frame k and frequency band b. In addition, the decision rule is formulated in terms of the long-term spectral divergence (LTSD). The J-order LTSD is defined as:

$$\mathrm{LTSD_J}(k)=10\log_{10}\left(\frac{1}{B}\sum_{b=0}^{B-1}\frac{LTSE_J{}^2(b,k)}{N_P^2(b,k-1)}\right) \tag{2}$$

where B means the number of frequency bands. $N_P^2(b,k-1)$ is the noise spectrum for the band b .Then, the J-order LTSD is used to compare with an adaptive threshold γ to determine the existence of speech signals. The threshold γ is adapted depending on the value of noise energy E:

$$\gamma=\begin{cases}\gamma_0 & E\le E_0\\ \gamma_0+\dfrac{\gamma_1-\gamma_0}{E_1-E_0}(E-E_0) & E_0<E<E_1\\ \gamma_1 & E\ge E_1\end{cases} \tag{3}$$

with $E=\sum_{b=0}^{B-1}N_P^2(b,k-1)$

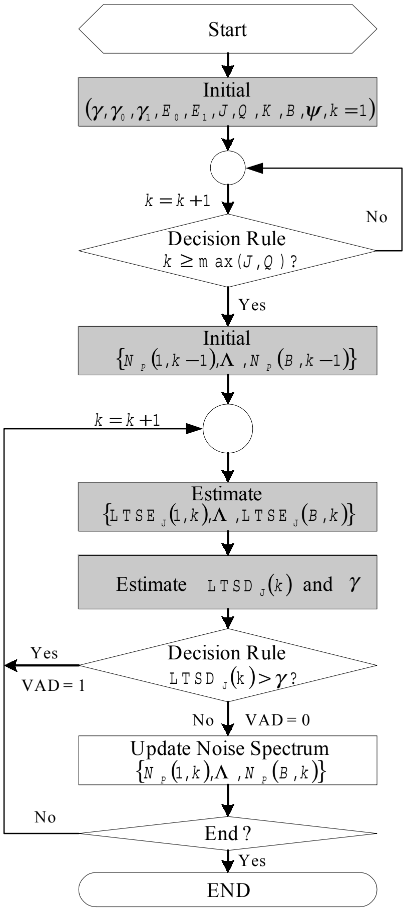
where γ_0 and γ_1 are thresholds when the system is working in the quietest and noisiest conditions respectively. E_0 and E_1 are the noise energies. The result of VAD is set to 1 (i.e., speech signal is detected) when the value of LTSD is larger than the value of γ. If the result of VAD equals to zero (i.e., no speech signal is detected), then the noise spectrum in (2) is updated:

$$N_P(b,k)=\begin{cases}\psi N_P(b,k-1)+(1-\psi)N_Q(b,k) & if\ VAD=0\\ N_P(b,k-1) & if\ VAD=1\end{cases} \quad (4)$$

where ψ is a weights and $N_Q(b,k)$ is the average noise spectrum magnitude at frequency band b over $(2Q+1)$ frame:

$$N_Q(b,k)=\frac{1}{2Q+1}\sum_{q=-Q}^{q=Q}R_i(b,k+q) \quad (5)$$

The flowchart of the VAD algorithm is illustrated as Fig. 4.



Start
Initial
$(\gamma,\gamma_0,\gamma_1,E_0,E_1,J,Q,K,B,\psi,k=1)$
$k=k+1$
No
Decision Rule
$k\geq \max(J,Q)$?
Yes
Initial
$\{N_P(1,k-1),\Lambda,N_P(B,k-1)\}$
$k=k+1$
Estimate
$\{LTSE_J(1,k),\Lambda,LTSE_J(B,k)\}$
Estimate $LTSD_J(k)$ and γ
Yes
VAD = 1
Decision Rule
$LTSD_J(k)>\gamma$?
No
VAD = 0
Update Noise Spectrum
$\{N_P(1,k),\Lambda,N_P(B,k)\}$
No
End ?
Yes
END

Fig. 4. Flowchart of the VAD algorithm

III. Proposed Speech Enhancement Approaches

A. Time Domain Beamformer Using H_∞ Filtering Approach

Based on the system architecture shown in Fig. 1, the formulation of microphone array speech enhancement system can be expressed as the following linear model:

$$d(n)=\boldsymbol{x}^T(n)\boldsymbol{w}+e(n) \quad (6)$$

where M denotes the number of microphones, P denotes the filter order of each microphone, and the superscript T denotes the transpose operation. $d(n)$ is the desired signal and $\boldsymbol{x}(n)=[\boldsymbol{x}_1(n)\ \ \Lambda\ \ \boldsymbol{x}_M(n)]^T$ is a $MP\times 1$ training signal vector. $\boldsymbol{x}_i(n)=[x_i(n)\ \ \Lambda\ \ x_i(n-P+1)]$ is a $1\times P$ training signal vector and each component in the silent stage is constructed from the linear combination of the pre-recorded speech signals and the online recorded environmental noise as $x_i(n)=s_i(n)+n_i(n)$. In addition, $\boldsymbol{w}=[w_{11}\ \ \Lambda\ \ w_{1P}\ \ \Lambda\ \ w_{M1}\ \ \Lambda\ \ w_{MP}]^T$ is the $MP\times 1$ unknown filter coefficient vector of the time domain beamformer that we intent to estimate. $e(n)$ is the unknown estimation disturbance, which may also include modeling error. In this work, italics fonts represent scalars, bold italics fonts represent vectors, and bold upright fonts represent matrices.

To apply the adaptive H∞ filtering approach, the linear model, as in (6), is transformed into its state space form:

$$\begin{aligned}&\boldsymbol{w}(n+1)=\boldsymbol{w}(n)\\&d(n)=\boldsymbol{x}^T(n)\boldsymbol{w}(n)+e(n)\\&\text{with } \boldsymbol{w}(n)=\boldsymbol{w}\end{aligned} \quad (7)$$

To find the optimal estimation, the criterion in the sense of H∞ based filtering is:

$$\min_{\hat{\boldsymbol{w}}(n)}\ \max_{(e(n),\hat{\boldsymbol{w}}(0))} J=-\frac{1}{2}\xi^2\mu_0^{-1}\left|\boldsymbol{w}-\hat{\boldsymbol{w}}(0)\right|^2+\frac{1}{2}\sum_{n=0}^{N}\left[\left|\boldsymbol{w}-\hat{\boldsymbol{w}}(n)\right|^2-\xi^2\left|e(n)\right|^2\right] \quad (8)$$

where μ_0 is a weighting parameter and $\hat{\boldsymbol{w}}(n)$ is the $MP\times 1$ estimated filter coefficient vector. $|\cdot|^2$ denotes the square of the 2-norm. According to [19], the solution of $\hat{\boldsymbol{w}}(n)$ can be approximated by the iteration:

$$\mathbf{M}^{-1}(n+1)=\mathbf{M}^{-1}(n)+\boldsymbol{x}(n)\boldsymbol{x}^T(n)-\xi^{-2}\mathbf{I} \quad (9)$$

$$\hat{\boldsymbol{w}}(n+1)=\hat{\boldsymbol{w}}(n)+\mathbf{M}(n)\boldsymbol{x}(n)\frac{(d(n)-\boldsymbol{x}^T(n)\hat{\boldsymbol{w}}(n))}{(1+\boldsymbol{x}^T(n)\mathbf{M}(n)\boldsymbol{x}(n))} \quad (10)$$

$$\hat{\boldsymbol{w}}(0)=\boldsymbol{0},\ \ \mathbf{M}^{-1}(0)=(\mu_0^{-1}-\xi^{-2})\mathbf{I} \quad (11)$$

where $\mathbf{M}(n)$ is an $MP\times MP$ matrix and $(\cdot)^{-1}$ denotes the matrix inverse operation. In order to ensure that $\mathbf{M}(n)$ remains positive definite, ξ should be chosen such that $\mathbf{M}^{-1}(n)+\boldsymbol{x}(n)\boldsymbol{x}^T(n)-\xi^{-2}\mathbf{I}>0$. For this reason, ξ is selected as $\delta\sqrt{\text{eig}(\mathbf{M}^{-1}(n)+\boldsymbol{x}(n)\boldsymbol{x}^T(n))^{-1}}$ during the iteration, where $eig(\mathbf{z})$ denotes the maximum eigenvalue of $\mathbf{z}$ and $\delta>1$ in order to keep ξ greater than the minimum value.

The adaptation of the filter coefficient vector is performed in the silent stage. When the system is switched to speech stage, the adaptation stops and the filter coefficient vector is passed to lower beamformer. The output of the speech purification system can be calculated by

$$\hat{y}(n)=\boldsymbol{y}^T(n)\hat{\boldsymbol{w}}(n) \quad (12)$$

where $\hat{y}(n)$ is the purified result, and $\boldsymbol{y}(n)=[\boldsymbol{y}_1(n) \quad \Lambda \quad \boldsymbol{y}_M(n)]^T$ is the $MP\times 1$ online recorded polluted speech signal vector acquired by the microphone array, where $\boldsymbol{y}_i(n)=[y_i(n) \quad \Lambda \quad y_i(n-P+1)]$.

B. Frequency Domain Beamformer Using H_∞ Filtering Approach

For the automatic speech recognition (ASR) application, the purified spectrum data should be computed directly to save computation effort, since most speech recognition algorithms are performed in the frequency domain. In this case, the parameters can be updated on a block of data. Hence, the problem is transformed into the frequency domain by using short time Fourier transform (STFT). Consequently, the linear convolution is implemented by padding the unknown estimation disturbance shown in (7) with zeros to make it twice as long as the window length. The unknown estimation disturbance at frame k and frequency ω can be written as:

$$\begin{aligned} E(\omega,k) &= D(\omega,k)-\boldsymbol{W}^H(\omega,k)\boldsymbol{X}(\omega,k) \\ &= D(\omega,k)-\boldsymbol{W}^H(\omega,k)\big(\boldsymbol{S}(\omega,k)+\boldsymbol{N}(\omega,k)\big) \end{aligned} \qquad (13)$$

with $\boldsymbol{W}(\omega,k)=\boldsymbol{W}(\omega)$

where $D(\omega,k)$ is the desired signal in the frequency domain and $\boldsymbol{W}(\omega)$ denotes the $M\times 1$ unknown filter coefficient vector at frequency ω. The superscript H denotes Hermitian operation. $\boldsymbol{X}(\omega,k)=[X_1(\omega,k) \quad \Lambda \quad X_M(\omega,k)]^T$, $\boldsymbol{N}(\omega,k)=[N_1(\omega,k) \quad \Lambda \quad N_M(\omega,k)]^T$ and $\boldsymbol{S}(\omega,k)=[S_1(\omega,k) \quad \Lambda \quad S_M(\omega,k)]^T$ represent the frequency domain training signal vector, the online recorded environmental noise vector, and the pre-recorded speech signal vector, respectively.

In conjunction with the spectrum-based ASR, the window size in the STFT has to equal to that in ASR in order to obtain a more accurate result. However, the window size may be too small to capture the acoustic channel response. For this reason, a previous work [11] proposed an approach called soft penalty frequency domain block beamformer (SPFDBB). SPFDBB takes the frame average over several frames as a block improving the approximation of the channel response. The number of frames in a block is denoted as the frame number L. However, the NLMS algorithm used in [11] limits its performance, because it contains inherent assumptions on the disturbances and channel dynamics. Consequently, the H$_\infty$ based filtering approach is adopted to improve the performance further. The performance index of the frequency domain beamformer using H$_\infty$ filtering approach can be formulated as:

$$\begin{aligned} &\min_{\boldsymbol{W}(\omega,k)} \max -\frac{1}{2}\xi^2\mu_0^{-1}\left|\boldsymbol{W}(\omega)-\hat{\boldsymbol{W}}(\omega,0)\right|^2 \\ &+\frac{1}{2}\sum_{k=0}^{K}\left\{\left|\boldsymbol{W}(\omega)-\hat{\boldsymbol{W}}(\omega,k)\right|^2-\xi^2\begin{bmatrix}\boldsymbol{V}(\omega,k)\\ \boldsymbol{U}(\omega,k)\end{bmatrix}^H \boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{V}(\omega,k)\\ \boldsymbol{U}(\omega,k)\end{bmatrix}\right\} \end{aligned} \qquad (14)$$

where μ_0 is a weighting parameter, $\hat{\boldsymbol{W}}(\omega,k)$ is the $M\times 1$ estimated filter coefficient vector, and

$$\begin{aligned} \boldsymbol{V}(\omega,k) = [&D(\omega,k)-\boldsymbol{W}^H(\omega)\boldsymbol{S}(\omega,k) \quad \Lambda \\ &D(\omega,k+L-1)-\boldsymbol{W}^H(\omega)\boldsymbol{S}(\omega,k+L-1)] \end{aligned}$$

$$\boldsymbol{U}(\omega,k)=[\boldsymbol{W}^H(\omega)\boldsymbol{N}(\omega,k) \quad \Lambda \quad \boldsymbol{W}^H(\omega)\boldsymbol{N}(\omega,k+L-1)]^T$$

$\boldsymbol{\Lambda}=\begin{bmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2\\ \boldsymbol{\Lambda}_3 & \boldsymbol{\Lambda}_4\end{bmatrix}$ is a $2L\times 2L$ matrix.

$\boldsymbol{\Lambda}_1=\begin{bmatrix}1+\lambda & \lambda & \Lambda & \lambda\\ \lambda & 1+\lambda & \mathrm{O} & \mathrm{M}\\ \mathrm{M} & \mathrm{O} & \mathrm{O} & \lambda\\ \lambda & \Lambda & \lambda & 1+\lambda\end{bmatrix}$, $\boldsymbol{\Lambda}_4=\mathbf{I_L}$, and $\boldsymbol{\Lambda}_2=\boldsymbol{\Lambda}_3=-\mathbf{I_L}$

where $\mathbf{I_L}$ is an identity matrix with dimension $L\times L$.
The H$_\infty$ iterative solutions can be shown as:

$$\hat{\boldsymbol{W}}(\omega,k+1)=\hat{\boldsymbol{W}}(\omega,k)+\mathbf{K}(\omega,k)[\boldsymbol{D}_L(\omega,k)-\mathbf{H}(\omega,k)\hat{\boldsymbol{W}}(\omega,k)] \qquad (15)$$

$$\mathbf{K}(\omega,k)=\mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\big(\mathbf{I}+\mathbf{H}(\omega,k)\mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\big)^{-1} \qquad (16)$$

$$\mathbf{P}^{-1}(\omega,k+1)=\mathbf{P}^{-1}(\omega,k)+\mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k)-\xi^{-2}\mathbf{I_M} \qquad (17)$$

$$\mathbf{H}(\omega,k)=\left[\boldsymbol{X}(\omega,k) \quad \Lambda \quad \boldsymbol{X}(\omega,k+L-1) \quad \mu^{\frac{1}{2}}\sum_{\boldsymbol{j=k}}^{k+L-1}\boldsymbol{S}(\omega,j)\right]^H \qquad (18)$$

$$\boldsymbol{D}_L(\omega,k)=\left[D(\omega,k) \quad \Lambda \quad D(\omega,k+L-1) \quad \mu^{\frac{1}{2}}\sum_{j=k}^{k+L-1}D(\omega,j)\right]^T \qquad (19)$$

$$\mathbf{P}^{-1}(\omega,1)=\mu_0\mathbf{I_M} \text{ and } \hat{\boldsymbol{W}}(\omega,0)=[0 \quad 0 \quad \Lambda \quad 0]^T \qquad (20)$$

where the superscript * denotes the complex conjugate. $\mathbf{H}(\omega,k)$ is a $(L+1)\times M$ dimensional matrix at kth block. $\mathbf{I_M}$ is an identity matrix with dimension $M\times M$. The value of ξ during the iteration is chosen as $\delta eig\big(\mathbf{P}^{-1}(\omega,k)+\mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k)\big)$ where $eig(\mathbf{z})$ denotes the minimum eigenvalue of $\mathbf{z}$. δ is a positive constant and lower than one to ensure that (17) is positive definite.Consequently, the purified output signal at kth block can be obtained by the following equation:

$$\hat{\boldsymbol{Y}}(\omega,k)=\hat{\boldsymbol{W}}^H(\omega,k)\mathbf{Y}(\omega,k) \qquad (21)$$

where $\hat{\boldsymbol{Y}}(\omega,k)=[\hat{Y}(\omega,k) \quad \Lambda \quad \hat{Y}(\omega,k+L-1)]$ is the purified result and $\mathbf{Y}(\omega,k)$ is the $M\times L$ online recorded polluted speech signal matrix.

TABLE I
TEN EXPERIMENTAL CONDITIONS AND ISOLATED AVERAGE SNRS

Condition Number	Speed	Power of In-car Audio System	Average SNR (dB)	Condition Number	Speed	Power of In-car Audio System	Average SNR (dB)
C1	20 km/h	Off	4.20	C6	20 km/h	On	-0.08
C2	40 km/h	Off	2.84	C7	40 km/h	On	-2.19
C3	60 km/h	Off	2.72	C8	60 km/h	On	-2.28
C4	80 km/h	Off	-1.90	C9	80 km/h	On	-4.75
C5	100 km/h	Off	-3.04	C10	100 km/h	On	-5.40

The step k is chosen as $0, L, 2L, 3L, \Lambda$ to perform the adaptation process every L frames.

IV. EXPERIMENTAL RESULTS

A. Experimental Conditions and Parameters

An experiment is performed on passenger seat of a mini-van vehicle instead of the driver's seat due to the driving safety consideration. A uniform linear microphone array of five un-calibrated microphones with 0.07 m spacing is mounted in front of the passenger seat. Additionally, the distance between the microphone array and the mouth of the speaker who sits in the passenger seat is about 0.62 m. The performance of the proposed approaches is demonstrated by 341 pairs of the vehicle identification numbers and ten conditions (C1-C10 of Table I). Table I shows the average SNRs in the ten conditions. A music piece containing vocal sound is played repeatedly from six build-in loudspeakers when the in-car audio system is turned on. The desired signal utilized in this experiment is derived from the headset microphone with the lowest channel distortion. The first and second microphones are utilized for the dual microphone case ($M = 2$) and the first, second, and third microphones are used when $M = 3$ and so on. The experimental results are compared with those of a delay-and-sum beamformer with a high-pass filter (DS+HP) introduced in [12]. The related VAD parameters are listed in Table II.

TABLE II
PARAMETERS OF THE VAD ALGORITHM

Parameter	Value	Parameter	Value
J	6	E_1	225
Q	3	γ	36
B	512	γ_0	20
E_0	190	γ_1	160
ψ	0.95		

B. Time Domain Performance Evaluation

Two performance indices, signal recover ratio (SRR) and noise power ratio (NPR), are defined and adopted instead of the signal to noise ratio (SNR) to evaluate the degree of signal distortion and noise suppression, because a higher SNR does not necessarily imply lower signal distortion. SRR is defined as:

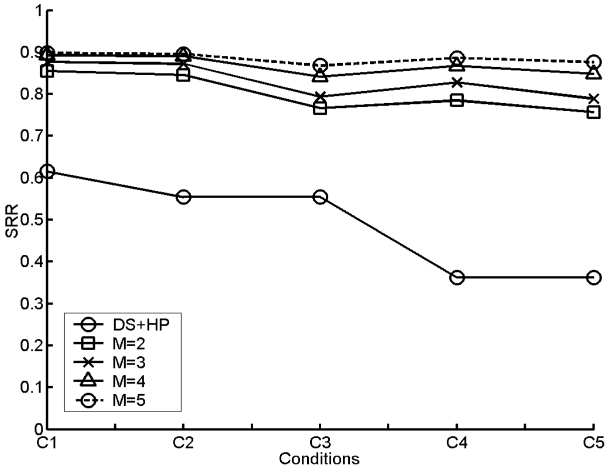
$$SRR(n) = \frac{\mathrm{cov}\left(\left(d(n), \left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right)\right)\right)}{\sqrt{\mathrm{cov}\left(d(n), d(n)\right) \times \mathrm{cov}\left(\left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right), \left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right)\right)}} \tag{22}$$

where $\mathrm{cov}(\cdot)$ is the iacovarnce operation. Further, NPR is defined as:

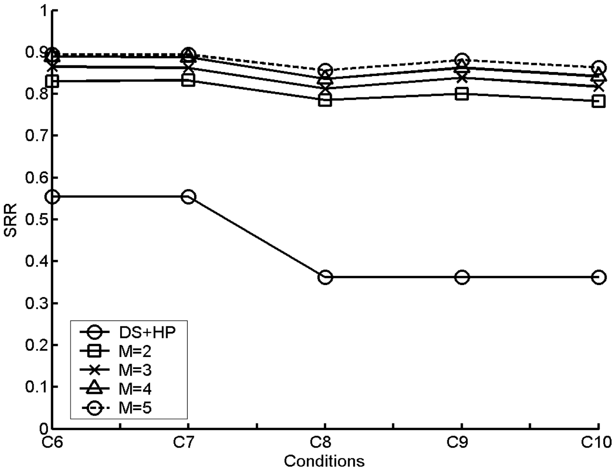
$$NPR(n) = \sum_{n=1}^{V} \left\| \boldsymbol{w}(n)^T \boldsymbol{n}(n) \right\|_2 \Big/ \sum_{n=1}^{V} \left\| n_1(n) \right\|_2 \tag{23}$$

where V in (25) denotes the length of the desired signal, $d(n)$. SRR is defined as the correlation coefficient between the desired signal, $d(n)$, and the recovered signal, $\boldsymbol{w}(n)^T \boldsymbol{s}(n)$. Consequently, a higher value of SRR indicates better speech recovery. NPR represents the ratio of the noise power after beamformer processing ($\boldsymbol{w}(n)^T \boldsymbol{n}(n)$) to the noise power measured at the silent stage (using microphone 1). A smaller value of NPR represents a cleaner speech signal.

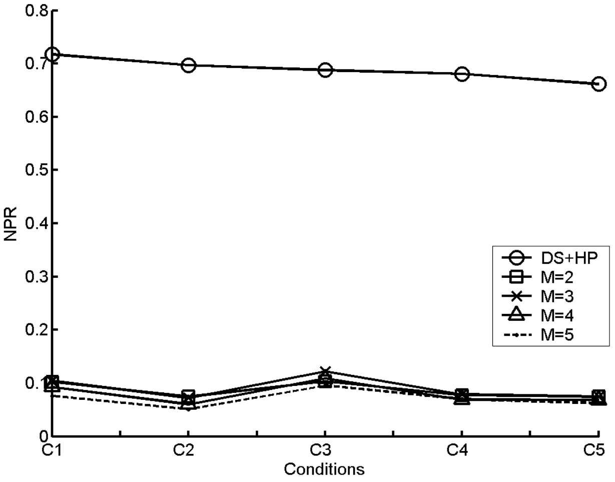
The order of the time domain beamformer using H_∞ filtering approach is set to 128, and μ_0 and ξ are set to 0.9 and 0.95 respectively. The values of SRR and NPR after the processes of the DS+HP and time domain beamformer using H_∞ filtering approach for the ten testing conditions are illustrated in Fig. 5. As shown in Figs. 5 (a) and 5 (b), the SRRs of the proposed approach are higher than those of the DS+HP when adopting two microphones including the cases when the in-car audio system is turned on (conditions C6 to C10). This is because the proposed system can recover the channel distortion and is robust to modeling errors. Moreover, the high-pass filter in the DS+HP suppresses the magnitude of low frequency components of the speech signal, which may decrease the SRR further. The values of NPR of the proposed method are also better than those of the DS+HP in conditions C1 to C10. The values of NPR in C6 to C10 are larger than those in C1 to C5 because switching the in-car audio system on raises the noise complexity. The improvement in SRR and NPR is consistent with the number of microphones used, which indicates that a larger number of microphones can improve speech quality.



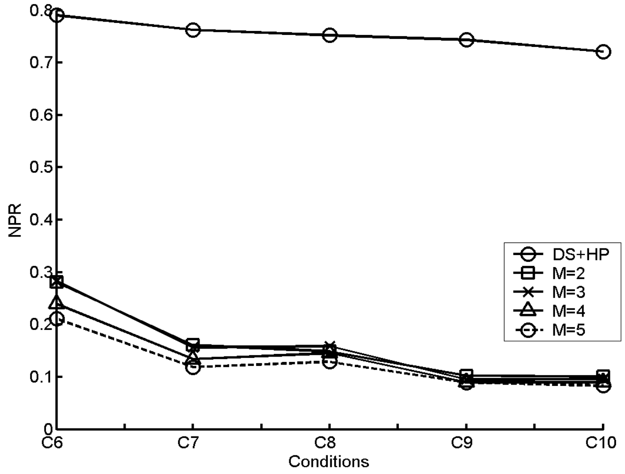
(a) SRRs of conditions C1 to C5.



(b) SRRs of conditions C6 to C10.



(c) NPRs of conditions C1 to C5



(d) NPRs of conditions C6 to C10

Fig. 5. SRRs and NPRs of conditions C1 to C10

C. Frequency Domain Performance Evaluation

The results of the frequency domain beamformer using H_∞ filtering approach are directly delivered to a benchmark speech recognizer, HMM toolkit (HTK) [20]. The related parameters of HTK are shown in Table III. In the experiments, μ_0 and ξ are set to 0.9 and 0.95 respectively and the soft penalty λ is set to 2. In addition, the frame number L is set to 40. The window contains 256 zero-padded samples and a 32*ms* speech signal, giving a total of 512 samples. Fig. 6 illustrates the processed frame and the overlapping condition.

TABLE III
PARAMETERS OF HTK

Recognition kernel	HTK ver.3.0
Model	HMM
Feature Vector	12 th order MFCC + 12th order ΔMFCC
Training data Set	1001 clean pairs of vehicle identification numbers
Recognition Task	341 pairs of the vehicle identification numbers

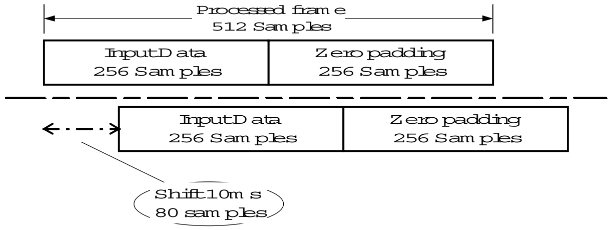


Fig. 6. Processed frame and overlapping condition

The best possible recognition rate using the desired signal is 97.15%. A baseline of the recognition rate is established using only the first microphone. As shown from Figs. 7 and 8, the baseline performance is poor, as would be expected. When only car noises are present (conditions C1-C5), the DS+HP improved the recognition rate by 15.52-25.25% over the baseline. Because the DS+HP only attempts to suppress the noises instead of dealing with the channel distortion, the performance cannot be satisfactory unless the recognizer is retrained. As indicated in Fig. 7, the improvement over the DS+HP from the proposed method becomes more significant when the environmental noise is louder. The improvements are most significant when the music is turned on (Fig. 8). The DS+HP has a poor recognition rate for music, because it could only suppress a small part of the wideband music signal. A comparison of Figs. 7 and 8 indicates that the proposed method maintains a similar recognition performance at a given vehicle speed both with and without music playing in the background.

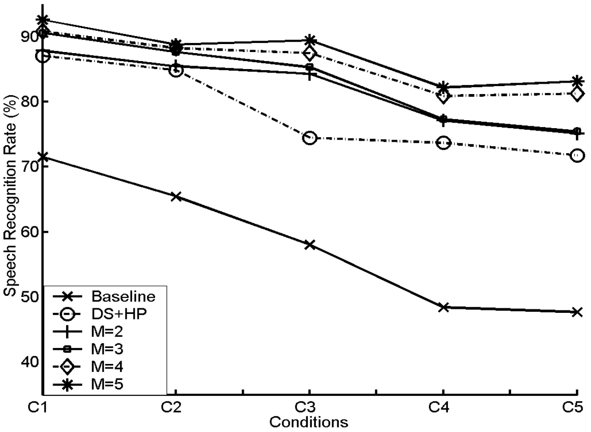


Fig. 7. Speech recognition rate of Conditions 1 to 5

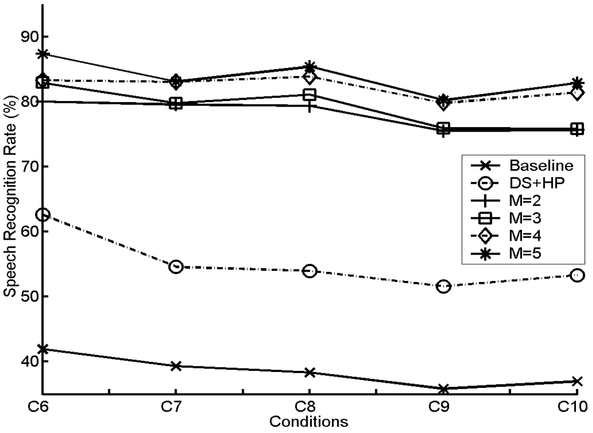


Fig. 8. Speech recognition rate of Conditions 6 to 10

V. CONCLUSION

A speech enhancement system is proposed and implemented in this work. Because the system includes a self-calibration mechanism, unmatched microphones and electronic circuits can be utilized to reduce the hardware cost. The performance indexes (SRR, NPR, and speech recognition rate) of different numbers of microphones are introduced and compared to provide design tradeoff among the number of microphones, performance, and circuit complexity. The real in-vehicle experimental results demonstrate that the proposed system can significantly improve the speech quality and the speech recognition rate without requiring time-consuming retraining processes for the speech recognizer.

ACKNOWLEDGMENT

This work is supported in part of the National Science Council of Taiwan under the grant number NSC 93-2218-E-009-031 and the Ministry of Education, Taiwan under the grant number 91-1-FA06-4-4.

REFERENCES

[1] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition," *Proc. European Conference on Speech Communication and Technology, EUROSPEECH99*, pp. 2255-2258, Sep. 1999.

[2] M. Matassoni, M. Omologo, P. Svaizer, "Use of real and contaminated speech for training of a hands-free in-car speech recognizer", *Proc. European Conference on Speech Communication and Technology, EUROSPEECH 2001*, pp. 61-66, Sep. 2001.

[3] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Noise-robust hands-free speech recognition on PDAs using microphone array technology," *Autumn Meeting of the Acoustical Society of Japan*, Sendai, Japan, pp. 51-54, Sep. 2005

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

[5] A. Kawamura, Y. Iiguni, and Y. Itoh, "A noise reduction method based on linear prediction with variable step-size," *IEICE Tran. Fundamentals*, vol. E88-A, no. 4, pp. 855-861, April 2005.

[6] O. L Frost, "An Algorithm for Linear Constrained Adaptive Array Processing," *Proc. IEEE*, vol. 60, no. 8, pp.926-935, Aug. 1972.

[7] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol. AP-30, pp. 27-34, Jan. 1982.

[8] H. Cox, R. M. Zeskind, and M. M. Owen., "Robust Adaptive Beamforming," *IEEE Trans. Acoust. Speech and signal Processing*, vol. ASSP-35, pp. 1365-1376, Oct. 1987.

[9] S. Gannot, D. Burshtein, and E. Weinstein "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, pp. 1614-1626, Aug. 2001.

[10] M. Dahl, and I. Claesson "Acoustic noise and echo canceling with microphone array," *IEEE Trans. Vehicular Technology,* vol. 48, pp.1518 -1526, Sept. 1999.

[11] J. S. Hu and C. C Cheng, "Frequency domain microphone array calibration and beamforming for automatic speech recognition,", *IEICE Trans. Fundamentals*, vol. E88-A, no. 9, pp. 2401-2411, Sep. 2005.

[12] S. Ahn and H. Ko, "Background noise reduction via dial-channel scheme for speech recognition in vehicular environment," *IEEE Trans. Consumer Electronics*, vol. 51, no. 1, pp. 22-27, Feb. 2005.

[13] T. Giannakopoulos, N. A. Tatlas, T. Ganchev, and I. Potamitis, "A practical, real-time speech-driven home automation front-end," *IEEE Trans. Consumer Electronics*, vol. 51, no. 2, pp514 - 523, May 2005.

[14] H. Kuttruf, *Room acoustics*, London: Elsevier, 1991, chapter 3, pp. 56.

[15] W. Zhuang, "Adaptive H infinity channel equation for wireless personal communications," *IEEE Trans. Vehicular Technology*, vol. 48, no. 1, pp. 126-136, January 1999.

[16] B. Hassibi, and T. Kailath, "H_∞ adaptive filtering," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 2, pp. 949-952, May 1995.

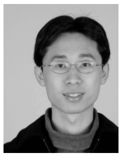
[17] J. Ramírez, J.C. Segura, C. Benítez, d.l. Torre, Ángel, and R. Antonio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271-287, April 2004.

[18] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in presence of noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 406–412, July 1994.

[19] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H∞ filtering algorithm," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 391-399, July 1999.

[20] Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk/)

Jwu-Sheng Hu received the B.S. degree from the Department of Mechanical Engineering, National Taiwan University, Taiwan, in 1984, and the M.S. and Ph.D. degrees from the Department of Mechanical Engineering, University of California at Berkeley, in 1988 and 1990, respectively. He is currently a Professor in the Department of Electrical and Control Engineering, National Chiao-Tung University, Taiwan, R.O.C. His current research interests include microphone array signal processing, active noise control, intelligent mobile robots, embedded systems and applications.



Chieh-Cheng Cheng was born in 1978. He received the B.S. and Ph.D. degrees in Electrical and Control Engineering from National Chiao Tung University, Taiwan, ROC, in 2000 and 2006. He is the championship of TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office in 2001. His research interests include sound source localization, microphone array signal processing, adaptive signal processing, pattern recognition, speech signal processing, and echo and noise cancellation.

Wei-Han Liu was born in Kaohsiung, Taiwan in 1977. He received the B.S. and M.S. degree in Electrical and Control Engineering from National Chiao Tung University, Taiwan, ROC in 2000 and 2002. He is currently a Ph.D. candidate in Department of Electrical and Control Engineering at National Chiao Tung University, Taiwan, ROC. He is the championship of TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office in 2001. His research interests include sound source localization, microphone array signal processing, adaptive signal processing, speech signal processing, and robot localization.

Chia-Hsing Yang was born in 1981. He received the B.S. degree and the M.S. degree in Electrical and Control Engineering from National Chiao Tung University, Taiwan, ROC in 2003 and 2005,respectively. He is currently a Ph.D. student in Department of Electrical and Control Engineering at National Chiao Tung University, Taiwan, ROC. His research interests include sound source localization, microphone array signal processing, and adaptive signal processing.