2006 IEEE International Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

A Robust Speech Enhancement System for Vehicular Applications Using H_∞ Adaptive Filtering

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

Abstract—**This work proposes a novel and robust adaptive speech enhancement system, which contains both time-domain and frequency-domain beamformers using H_∞ filtering approach in vehicle environments. A corresponding microphone array data acquisition hardware is also designed and implemented. Traditionally, mutually matched microphones are needed, but this requirement is not practical. To conquer this issue, the proposed system adapts the mismatch dynamics to allow unmatched microphones to be used in an array. Furthermore, to achieve a satisfactory 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 channel effect usually causes the modeling error in a channel recovery process because of the long channel response. The proposed system using the H_∞ filtering approach, which makes no assumptions about noise and disturbance, is robust to the modeling error. Consequently, the proposed frequency-domain beamformer provides a satisfactory performance without the need to retrain the speech recognizer.**

I. Introduction

THE use of mobile phones and electronic systems in vehicles is becoming increasing. Considering driving safety and convince, mobile phones and many in-car electronic systems such as global positioning system (GPS), CD, air conditioner, etc. should not be accessed by hands while driving. Consequently, intelligent hands-free interfaces with speech recognition were proposed in recent years. However, the echo of the far-end speech and environmental noises degrade the recognition performance and result in a low acceptance of hands free to consumers. Therefore, methods such as single-channel [1]-[2] and multi-channel speech enhancement techniques [3]-[8] have been introduced. Although single-channel based methods can reduce the hardware complexity, the performance degrades due to various problems [3].

To overcome the limitation, the microphone array based noise suppression approaches, such as Frost beamformer [4], robust adaptive beamformer [5], and generalized sidelobe canceller (GSC) [6] are proposed. However, these methods still suffer from several non-ideal factors. For example, the microphones are required to be mutually matched and no coherent interference signal exists. Dahl et. al. [7] proposed a finite impulse response (FIR) based normalized least-mean-square (NLMS) filtering approach to perform indirect microphone calibration and to minimize the sound signal distortion due to channel effect by using a pre-recorded speech signal and a desired signal acquired when the environment was quiet. Because the variation between pre-recorded signals and the desired signal contain useful information about the dynamics of channel and microphones' characteristics, this method outperforms other un-calibrated algorithms in real applications [8]. However, the FIR filter using the finite number of taps is unlikely to completely characterize the channel dynamics [9]. Moreover, the NLMS based formulation assumes the disturbance is uncorrelated to the source, zero mean and Gaussian distributed. These will limit the performance of speech enhancement.

On the contrary, the proposed H_∞ filtering approaches are robust to the modeling error caused by finite tap length of FIR filters and have no assumption made regarding the characteristics of environmental noises [10]-[11]. Furthermore, the method of using the pre-recorded signals and the desired signal can suppress the gain from noises to the output and the characteristics of received multi-channel signals can be automatically adjusted to those of the desired signal. Therefore, extra training processes for speech recognizer in vehicles are not needed in this work. For speech communication via hands-free mobile phones, a time-domain beamformer using H_∞ is proposed to produce a more clean and undisturbed speech waveform. Secondly, for speech recognition applications, a frequency-domain beamformer using H_∞ 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 approaches using two microphones outperform dual channel delay-and-sum beamformer with a high-pass filter introduced in [3]. Different choices of the number of the microphone are also compared

The remainder of this paper is organized as follows. The proposed speech enhancement system and the microphone

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)

1-4244-0100-3/06/$20.00 ©2006 IEEE

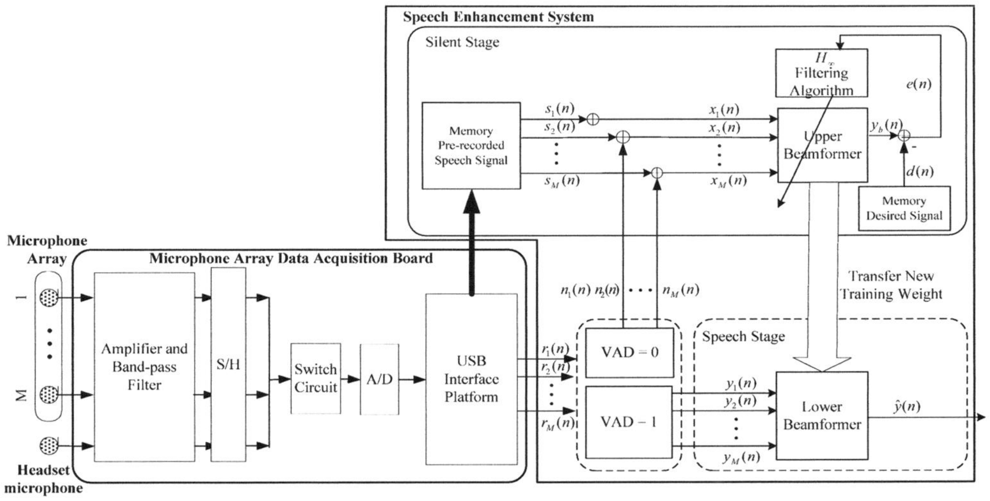


Fig. 1. Overall system architecture

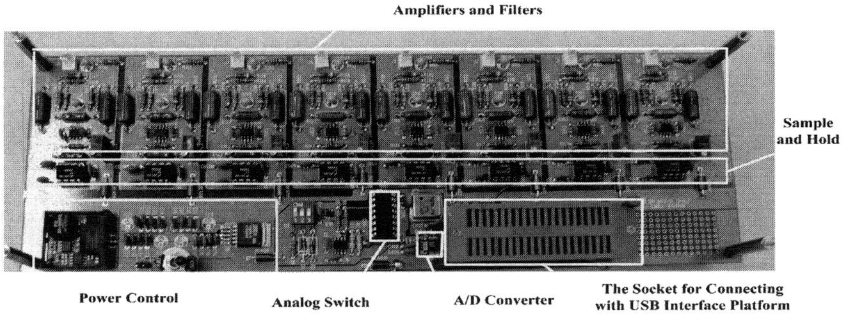


Fig. 2. Microphone array data acquisition board

array data acquisition hardware designed are introduced in section 2. Section 3 presents the two proposed H_∞ filtering based beamformers in both time and frequency-domain. Several representative experiments in a real vehicle are shown and the performances of experimental results are discussed in section 4. Finally, the conclusion is made 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 that the system can digitalize the received sound signals and transmit them in real-time via USB interface. The second sub-system represents 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 is ranged from 50*Hz* to 16*kHz*. The microphone array acquisition board made of a four-layer board can be divided into three parts.

In the first part, the microphone signals are amplified and filtered by six amplifiers and six band-pass filters designed to take the microphone sensitivity and anti-aliasing into consideration. The gain of the M microphones and the headset microphone are set to 60dB and 20dB individually. The second part comprises six sample-and-hold circuits (S/H), one analog switch circuit, and one analog-to-digital (A/D) converter. The third part contains the control and data transmission lines which are controlled by the universal serial bus (USB) interface. Through the control line, the USB interface platform can control the timing of the sample-and-hold circuits, switch, and A/D converter. The switching frequency and the timing of the system can be selected flexibly, and the sampling frequency in this work is set to 8 *kHz*. The converted 16-bit digital data are transmitted in real-time through the USB interface.

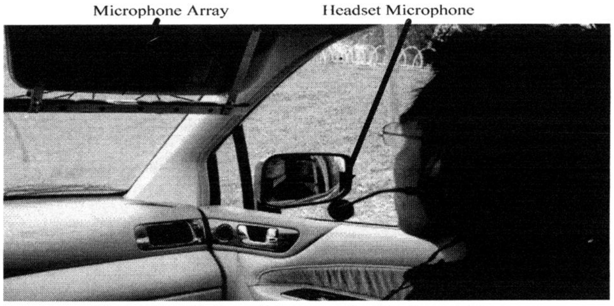


Fig. 3. Installation of the array inside the vehicle

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 for collecting the desired signal, i.e., the user does not need the headset microphone during the online applications.

B. Speech Enhancement System

This system can be separated into two stages, silent stage and speech stage, by a voice activity detector (VAD) which can distinguish whether the received signals contain speech signals or not. The voice activity detection algorithm can be found in reference [12]. If the result of VAD is equal to zero, which means that no speech exists, the system will be run in the silent stage. When the result of VAD is equal to one, the system could be switched to the speech stage.

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 at the same time when the pre-recoded speech signals are collected. Since the headset is close to the mouth, the desired signals contain 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 noise signals without speech signals are recorded online. The environmental noise signals 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 noise signals, $n_1(n),\cdots,n_M(n)$, with the pre-recorded speech signals, $s_1(n),\cdots,s_M(n)$, to construct training signals, $x_1(n),\cdots,x_M(n)$. The training signals are used to adapt the weighting vector using H$_\infty$ 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, $y_1(n),\cdots,y_M(n)$.

III. Proposed Speech Enhancement Approaches

A. Time-Domain Beamformer Using H$_\infty$ 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:

$$d(n) = \boldsymbol{x}^T(n)\boldsymbol{w} + e(n) \tag{1}$$

In this work, italics fonts represent scalars, bold italics fonts represent vectors, and bold upright fonts represent matrices. 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 $\boldsymbol{x}(n) = [\boldsymbol{x}_1(n) \ \cdots \ \boldsymbol{x}_M(n)]^T$ is a $MP\times 1$ training signal vector. $\boldsymbol{x}_i(n) = [x_i(n) \ \cdots \ x_i(n-P+1)]$ is a $P\times 1$ training signal vector. In addition, $\boldsymbol{w}$ 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.

To apply the adaptive H$_\infty$ filtering algorithms, the linear model, as in (1), is transformed into its state space form:

$$\begin{aligned} &\boldsymbol{w}(n+1) = \boldsymbol{w}(n) \\ &d(n) = \boldsymbol{x}^T(n)\boldsymbol{w}(n) + e(n) \quad \text{with } \boldsymbol{w}(n) = \boldsymbol{w} \end{aligned} \tag{2}$$

The criterion in the sense of H$_\infty$ is:

$$\begin{aligned} \min_{\hat{\boldsymbol{w}}(n)} \max_{(e(n),\hat{\boldsymbol{w}}(0))} J = &-\frac{1}{2}\xi^2\mu_0^{-1}\left|\boldsymbol{w} - \hat{\boldsymbol{w}}(0)\right|^2 \\ &+\frac{1}{2}\sum_{n=0}^{N}\left[\left|\boldsymbol{w} - \hat{\boldsymbol{w}}(n)\right|^2 - \xi^2\left|e(n)\right|^2\right] \end{aligned} \tag{3}$$

where μ_0 is a weighting parameter and $\hat{\boldsymbol{w}}(n)$ is the $MP\times 1$ estimated filter coefficient vector. $|.|^2$ denotes the square of the 2-norm. According to [13], the solution of $\hat{\boldsymbol{w}}(n)$ can be approximated by the iteration:

$$\mathbf{M}^{-1}(n+1) = \mathbf{M}^{-1}(n) + \boldsymbol{x}(n)\boldsymbol{x}^T(n) - \xi^{-2}\mathbf{I} \tag{4}$$

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) + \mathbf{M}(n)\boldsymbol{x}(n)\frac{(d(n) - \boldsymbol{x}^T(n)\hat{\boldsymbol{w}}(n))}{(1 + \boldsymbol{x}^T(n)\mathbf{M}(n)\boldsymbol{x}(n))} \tag{5}$$

$$\hat{\boldsymbol{w}}(0) = \boldsymbol{0}, \ \mathbf{M}^{-1}(0) = (\mu_0^{-1} - \xi^{-2})\mathbf{I} \tag{6}$$

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

$$\hat{y}(n) = \boldsymbol{y}^T(n)\hat{\boldsymbol{w}}(n) \tag{7}$$

where $\hat{y}(n)$ is the purified result, and $\boldsymbol{y}(n)$ is the

$MP \times 1$ online recorded polluted speech signal vector acquired by the microphone array.

B. Frequency-Domain Beamforming Using H_∞ Filtering Approach

The unknown estimation disturbance at frame k and frequency ω can be written as:

$$E(\omega,k) = D(\omega,k) - \mathbf{W}^H(\omega,k)\boldsymbol{X}(\omega,k) \tag{8}$$

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 weighting vector at frequency ω. The superscript H denotes Hermitian operation. $\mathbf{X}(\omega,k)$, $\mathbf{N}(\omega,k)$ and $\mathbf{S}(\omega,k)$ represent the frequency-domain training signal vector, the online recorded environmental noise vector, and the pre-recorded speech signal vector, respectively.

In general, 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 [8] proposed an approach called soft penalty frequency domain block beamformer (SPFDBB). However, the NLMS algorithm used in that work [8] limits its performance due to its inherent assumptions on the disturbances and channel dynamics. Consequently, the H_∞ based filtering approach is adopted to improve the performance further. The H_∞ iterative solutions can be shown as:

$$\hat{\boldsymbol{W}}(\omega,k+1) = \hat{\boldsymbol{W}}(\omega,k) + \mathbf{K}(\omega,k)\left\lfloor \boldsymbol{D}_L(\omega,k) - \mathbf{H}(\omega,k)\hat{\boldsymbol{W}}(\omega,k) \right\rfloor \tag{9}$$

$$\mathbf{K}(\omega,k) = \mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\left(\mathbf{I} + \mathbf{H}(\omega,k)\mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\right)^{-1} \tag{10}$$

$$\mathbf{P}^{-1}(\omega,k+1) = \mathbf{P}^{-1}(\omega,k) + \mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k) - \xi^{-2}\mathbf{I}_\mathbf{M} \tag{11}$$

$$\mathbf{H}(\omega,k) = \begin{bmatrix} \boldsymbol{X}(\omega,k) & \cdots & \boldsymbol{X}(\omega,k+L-1) & \mu^{\frac{1}{2}} \sum_{j=k}^{k+L-1} \boldsymbol{S}(\omega,j) \end{bmatrix}^H \tag{12}$$

$$\boldsymbol{D}_L(\omega,k) = \begin{bmatrix} D(\omega,k) & \cdots & D(\omega,k+L-1) & \mu^{\frac{1}{2}} \sum_{j=k}^{k+L-1} D(\omega,j) \end{bmatrix}^T \tag{13}$$

$$\mathbf{P}^{-1}(\omega,1) = \mu_0 \mathbf{I}_\mathbf{M} \text{ and } \hat{\boldsymbol{W}}(\omega,0) = [0 \quad 0 \quad \cdots \quad 0]^T \tag{14}$$

where $\boldsymbol{W}(\omega)$ denotes an unknown weighting vectors, the superscript * denotes the complex conjugate, and $\mathbf{H}(\omega,k)$ is a $(L+1) \times M$ dimensional matrix at kth block. $\mathbf{I}_\mathbf{M}$ is an identity matrix with dimension $M \times M$. The value of ξ during the iteration is chosen as $\delta eig\left(\mathbf{P}^{-1}(\omega,k) + \mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k)\right)$ to keep ξ greater than the minimum value.

Consequently, the purified output signal at kth block can be obtained by the following equation:

$$\hat{y}(\omega,k) = \hat{\boldsymbol{W}}^H(\omega,k)\hat{\mathbf{X}}(\omega,k) \tag{17}$$

where $\hat{y}(\omega,k)$ and $\hat{\mathbf{X}}(\omega,k)$ is the $M \times L$ online recorded polluted speech signal matrix. The step k is chosen as $0, L, 2L, 3L, \cdots$ to perform the adaptation process every L frames.

IV. Experimental Results

A. Experimental Conditions and Parameters

The experiment was 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. In addition, the distance between the microphone array and the mouth of the speaker who sits in passenger seat is about 0.62 *m*. To show the performance of the proposed approaches, 341 pairs of the vehicle identification numbers and ten conditions (C1-C10 of Table I) were used. The average SNR's in the ten conditions are shown in Table I and a music piece containing vocal sound is played repeatedly by 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 which contains lowest channel distortion. The first and second microphones are utilized for dual microphone case ($M = 2$) and the first, second, and third microphones are used when $M = 3$ and so on. For comparison purpose, the delay-and-sum beamformer with a high-pass filter (DS+HP) introduced in [3] is implemented.

B. Time-Domain Performance Evaluation

Instead of using signal to noise ratio (SNR), two performance indices, signal recover ratio (SRR) and noise

TABLE I

TEN EXPERIMENTAL CONDITIONS AND ISOLATED AVERAGE SNR

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/hr*	Off	4.20	C6	20 *Km/hr*	On	-0.08
C2	40 *Km/hr*	Off	2.84	C7	40 *Km/hr*	On	-2.19
C3	60 *Km/hr*	Off	2.72	C8	60 *Km/hr*	On	-2.28
C4	80 *Km/hr*	Off	-1.90	C9	80 *Km/hr*	On	-4.75
C5	100 *Km/hr*	Off	-3.04	C10	100 *Km/hr*	On	-5.40

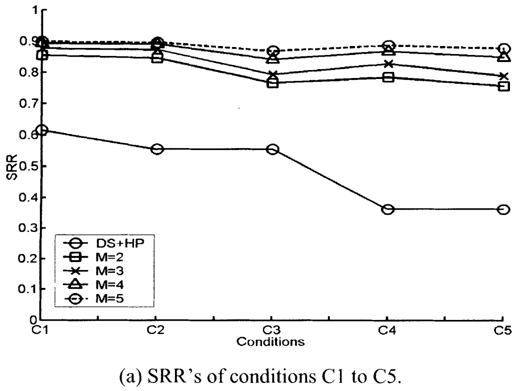
power ratio (NPR), are defined to evaluate the degree of signal distortion and noise suppression. This is because a higher SNR does not imply that the signal distortion is low. SRR is defined as:

$$SRR(n) = \frac{\operatorname{cov}\left(\left(d(n),\left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right)\right)\right)}{\sqrt{\operatorname{cov}\left(d(n),d(n)\right) \times \operatorname{cov}\left(\left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right),\left(\boldsymbol{w}(n)^T \boldsymbol{s}(n)\right)\right)}} \quad (18)$$

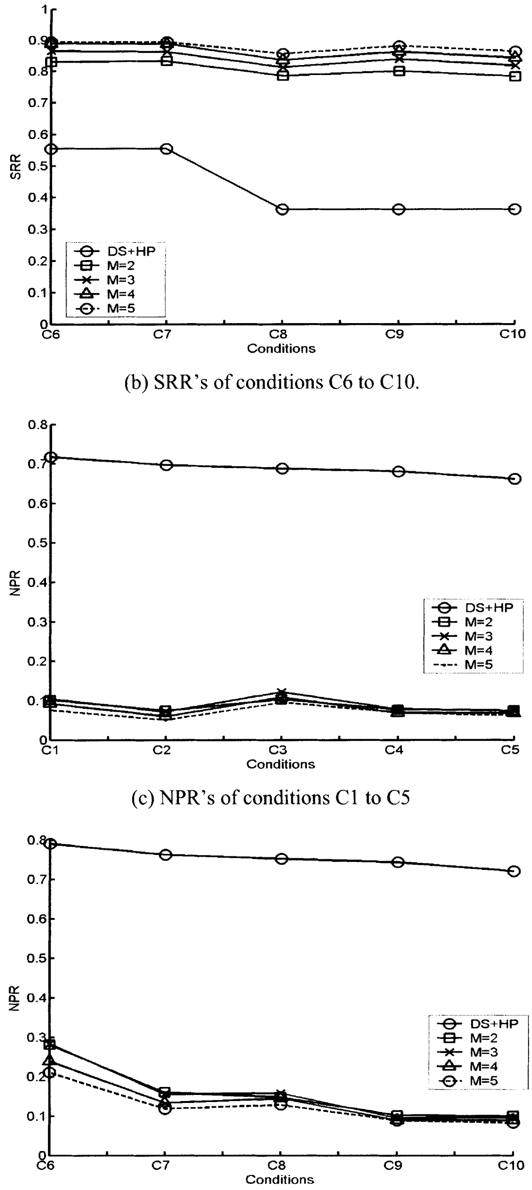
where $\operatorname{cov}(\cdot)$ is the covariance operation. Further, NPR is defined as:

$$NPR(n) = \sum_{n=1}^{V} \left[\boldsymbol{w}(n)^T \boldsymbol{n}(n)\right]^2 \Big/ \sum_{n=1}^{V} n_1^2(n) \quad (19)$$

where V in (19) denotes the length of the desired signal. SRR is defined as the correlation coefficient between the desired signal, and the recovered signal ($\boldsymbol{w}(n)^T \boldsymbol{s}(n)$). Consequently, a higher value of SRR means a 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. The smaller value of NPR represents a more clean speech signal. The order of the time-domain beamformer using H_∞ filtering approach was set to 128, and μ_0 and ξ were set to 0.9 and 0.95 individually. The values of SRR and NPR after the DS+HP and time-domain beamformer using H_∞ filtering technique for the ten testing conditions are illustrated in Fig. 5. As shown in Fig. 5 (a) and Fig. 5 (b), the SRR's of the proposed approach is higher than those of the DS+HP when two microphones are utilized 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 error. Moreover, the high-pass filter in 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 also outperform the traditional DS+HP in conditions C1 to C10. The values of NPR in C6 to C10 are larger than those in C1 to C5 because turning on the in-car audio system raises the complexity of the noise. The improvement of SRR and NPR are consistent with the number of microphone used. It means a larger number of microphones could provide a better sound quality.



(a) SRR's of conditions C1 to C5.



(b) SRR's of conditions C6 to C10.

(c) NPR's of conditions C1 to C5

(d) NPR's of conditions C6 to C10

Fig. 5. SRR's and NPR's 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, HTK [14]. In the experiments, μ_0 and ξ were set to 0.9 and 0.95 individually 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 32ms speech signal which gives a total of 512 samples. The best possible recognition rate using the desired signal is 97.15%. A baseline of the recognition rate using the first microphone only is established. As shown in Fig. 6 and 7, the baseline performance is poor as expected. When only the car noises exist (conditions C1-C5), the DS+HP can improve the recognition rate in 15.52% to 25.25% range compared with the baseline. Because the DS+HP only attempts to suppress the noise signals instead of dealing with the channel distortion, the performance cannot be satisfactory unless the recognizer is re-trained. As indicated by Fig. 6, the improvement using the proposed method over DS+HP becomes more significant when the environmental noise is higher. The improvements are more significant when the music is turned on (Fig. 7). The recognition rate for DS+HP drops because it can only suppress a small amount of the wideband music signal. Comparing Fig. 6 and Fig. 7, the proposed method maintains a similar recognition performance at a vehicle speed without and with music in the background.

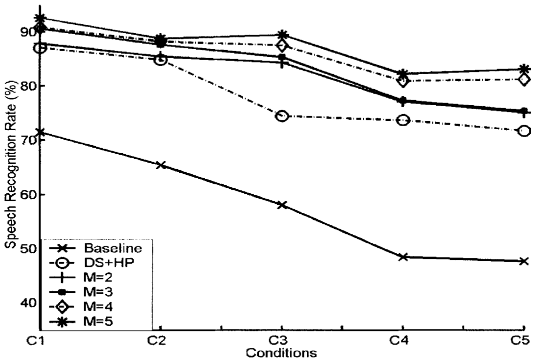


Fig. 6 Speech recognition rate of Conditions 1 to 5

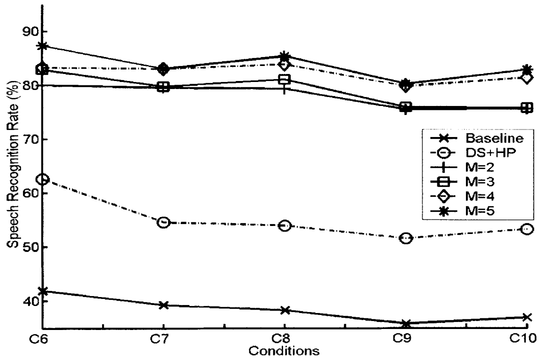


Fig. 7 Speech recognition rate of Conditions 6 to 10

V. Conclusion

A time-domain and a frequency-domain adaptive beamformer using H_∞ filtering approaches are proposed. The methods can be applied as a hands-free speech acquisition interface for communication or speech recognition in a vehicle. The performance indexes (SRR, NPR, and speech recognition rate) of different numbers of microphone are introduced and compared to provide design tradeoff among the number of microphone used, performance and circuit complexity. The experimental results show that the proposed system could improve the communication quality and the speech recognition rate significantly without the time consuming re-training process for the speech recognizer.

References

[1] 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.

[2] 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. 855-861April 2005.

[3] 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.

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

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

[6] 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.

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

[8] J. S. Hu and Chieh-Cheng 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.

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

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

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

[12] J. Ramírez, J.C. Segura, C. Benítez, d.l. Torre, Ángel; et. al. “Efficient voice activity detection algorithms using long-term speech information,” *Speech Communication*, vol. 42, pp. 271-287, April 2004.

[13] X. Shen and L. Deng, “A dynamic system approach to speech enhancement using the H_∞ filtering algorithm,” *IEEE Trans. Speech and Audio Process.*, vol. 7, pp. 391-399, July 1999.

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