The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Speech Signal Enhancement under Multiple Interferences Using Transfer Function Ratio Beamformer

Jwu-Sheng Hu, *Member, IEEE*, and Chia-Hsing Yang, *Student Member, IEEE*

***Abstract*—In many practical environments, the desired speech signal is usually contaminated not only by stationary noise but also nonstationay interferences, such as competing speech. This paper proposes a speech enhancement method which can extract desired speech in a multiple interferences and reverberant environment. The proposed method uses transfer function ratio beamformer and multi-channel adaptive filter algorithm. The virtual sound source concept is proposed to simplify the theoretical treatment for multiple competing speeches. In addition, a transfer function ratio estimation method in a more practical scenario is also proposed. The experiments are performed in a real room acoustic environment.**

I. Introduction

IT is important for robot to understand spoken language and respond to auditory events. However, the speech signal of interest is usually contaminated by competing speech, reverberation or background noise. Microphone array based techniques have been proposed to solve the speech enhancement problems for more than three decades [1-3].

In recent years, microphone array have been widely used for the robot audition system [4], [5]. Takeda *et al.* [6] proposed a step-size parameter adaptation technique of multi-channel semi-blind independent component analysis (ICA) for a barge-in-able robot audition system. For the barge-in-able robot, the user can interrupt and begin speaking while the robot is speaking. For robot audition, the recognition of the front talker is critical for smooth interactions. Hence, Kim *et al.* [7] presented an enhanced speech detection method which can separate and recognize speech signals originating from the front even in the noisy environment. The robot audition system consists of a voice activity detection based on the complex spectrum circle centroid and a maximum signel-to-noise beamformer. The environment may contain various types of noise, such as diffused noise, directiondal noise, and noise from the robot. Hosoya *et al.* [8] proposed a noise reduction method which consists of four-stage signal processing using a square microphone array with four microphones.

Despite the effort to enhance target speech for robot audition, a robust interface is still considered a difficult problem due to the variety of environments. In particular, some aggressive enhancement techniques could result in distorted speech signals and degrade the recognition performance, especially under competing speech situation. This paper considers speech enhancement problem under multiple speech sources in a reverberant and noisy environment condition and we focus on reconstructing the desired speech while suppressing competing speech sources and stationary noise using beamformer based technique rather than ICA based method. In a reverberant environment, the transfer function (TF) from source to microphone should be explicitly modeled [9] to replace the simple delay assumption. However, estimating the TF in a real environment is a complicated work. Rather than estimating the TF, this paper uses the transfer function ratio (TFR) based beamformer [10] for noise reduction.

For mobile robots, it is cumbersome and impractical to analyze the TFR of each interference signal. Therefore, this paper proposes the virtual sound source perspective explained by singular value decomposition (SVD) method to simplify the complexity of multiple interference signals. This paper proposes a two-stage speech enhancement algorithm using the TFR beamformer and the multi-channel adaptive filter algorithm. The TFR beamformer can be considered a prefilter to filter out the major component of the virtual sound source first and the residual noise from TFR beamformer output can be suppressed by multi-channel adaptive filter for dual-objectives optimization. In addition, this paper considers the TFR estimation in a more practical scenario and the proposed TFR estimation can be referred to [11]. The proposed algorithm is implemented in the frequency domain and the performance is evaluated in the real environment. The proposed algorithm is also tested by an automatic speech recognition system (ASR) for the application consideration.

II. Problem Formulation

A. Problem Description

Consider P speech sources and M microphones in the reverberant and noisy environment ($M > P$). The received signal of the m-th microphone can be written as:

$$x_m(t) = \sum_{p=1}^{P} a_{mp}(t) \otimes s_p(t) + n_m(t) \tag{1}$$

where each symbol in (1) represents:

$\otimes$ convolution operation;

$a_{mp}(t)$ the transfer function from the p-th sound source to the m-th microphone;

$s_1(t)$ the desired speech signal;

Manuscript received March 10, 2010. This work was supported in part by the National Science Council of Taiwan, ROC under grant NSC 96-2628-E-009-163-MY3

J.S. Hu and C.H.Yang are with Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC. (e-mail: {jshu@cn.; chyang.ece92g@}nctu.edu. tw)

$s_2(t) \sim s_P(t)$ the nonstationary interfering speech signals (competing speech signals);

$n_m(t)$ the (directional or omni-directional) stationarynoise of the *m*-th microphone.

Typically, the transfer function $a_{mp}(t)$ is assumed to be time-invariant over the observation period. In this paper, the competing speech signals, $s_2(t) \sim s_P(t)$, are regarded as interference signals. Applying the short time Fourier transform (STFT) operation to (1) yields:

$$X_m(k,\omega) = \sum_{p=1}^{P} A_{mp}(\omega) S_p(k,\omega) + N_m(k,\omega) \tag{2}$$

where *k* is the frame number and ω is the frequency band. $X_m(k,\omega)$, $S_p(k,\omega)$ and $N_m(k,\omega)$ are the STFT of the respective signals. $A_{mp}(\omega)$ is the time-invariant transfer function from the *p*-th source to the *m*-th microphone. The objective of this work is to reconstruct the desired speech from the microphone received signal.

B. Virtual Sound Source Perspective

When the desired speech signal and the stationary noise are absent, the microphone received signal can be expressed in the matrix form as:

$$\boldsymbol{X}_1(k,\omega) = \mathbf{A}_1(\omega)\boldsymbol{S}_1(k,\omega) \tag{3}$$

where

$$\boldsymbol{X}_1(k,\omega) = \begin{bmatrix} X_1(k,\omega) \\ X_2(k,\omega) \\ \vdots \\ X_M(k,\omega) \end{bmatrix} \in C^{M\times 1} \quad , \quad \boldsymbol{S}_1(k,\omega) = \begin{bmatrix} S_2(k,\omega) \\ S_3(k,\omega) \\ \vdots \\ S_P(k,\omega) \end{bmatrix} \in C^{(P-1)\times 1}$$

$$\mathbf{A}_1(\omega) = \begin{bmatrix} A_{12}(\omega) & A_{13}(\omega) & \cdots & A_{1P}(\omega) \\ A_{22}(\omega) & A_{23}(\omega) & & A_{2P}(\omega) \\ \vdots & \vdots & & \vdots \\ A_{M2}(\omega) & A_{M3}(\omega) & \cdots & A_{MP}(\omega) \end{bmatrix} \in C^{M\times(P-1)}$$

Assume the rank of the transfer function matrix $\mathbf{A}_1(\omega)$ is *R* and $\mathbf{A}_1(\omega)$ can be decomposed by SVD:

$$\mathbf{A}_1(\omega) = \mathbf{U}(\omega)\mathbf{D}(\omega)\mathbf{V}^{\mathrm{H}}(\omega) \tag{4}$$

where

$$\mathbf{U}(\omega) = \begin{bmatrix} \boldsymbol{u}_1(\omega) & \boldsymbol{u}_2(\omega) & \cdots & \boldsymbol{u}_R(\omega) \end{bmatrix} \in C^{M\times R}$$

$$\mathbf{V}(\omega) = \begin{bmatrix} \boldsymbol{v}_1(\omega) & \boldsymbol{v}_2(\omega) & \cdots & \boldsymbol{v}_R(\omega) \end{bmatrix} \in C^{(P-1)\times R}$$

$$\mathbf{D}(\omega) = \begin{bmatrix} \sigma_1(\omega) & 0 & \cdots & 0 \\ 0 & \sigma_2(\omega) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_R(\omega) \end{bmatrix} \in C^{R\times R}$$

$\sigma_r(\omega)$ are the nonzero singular values of $\mathbf{A}_1(\omega)$ with $\sigma_1(\omega) \ge \sigma_2(\omega) \ge \cdots \ge \sigma_R(\omega) > 0$. $\boldsymbol{v}_r(\omega)$ and $\boldsymbol{u}_r(\omega)$ are the input and output singular vectors of $\mathbf{A}_1(\omega)$ respectively which construct the interference subspace. The idea of virtual sound source is characterized as the following (from (3) and (4)) and (3) can be rewritten as:

$$\begin{aligned} \boldsymbol{X}_1(k,\omega) &= \sum_{i=1}^{R} \sigma_i(\omega)\boldsymbol{u}_i(\omega)\boldsymbol{v}_i^{\mathrm{H}}(\omega)\boldsymbol{S}_1(k,\omega) \\ &= \left(\boldsymbol{A}_V(\omega) + \boldsymbol{\Delta}_V(k,\omega)\right) S_V(k,\omega) \end{aligned} \tag{5}$$

where

$$S_V(k,\omega) = \sigma_1(\omega)\boldsymbol{v}_1^H(\omega)\boldsymbol{S}_1(k,\omega)$$

$$\boldsymbol{A}_V(\omega) = \begin{bmatrix} A_{1V}(\omega) \\ A_{2V}(\omega) \\ \vdots \\ A_{MV}(\omega) \end{bmatrix} = \boldsymbol{u}_1(\omega) \, , \, \boldsymbol{\Delta}_V(k,\omega) = \begin{bmatrix} \Delta_{1V}(k,\omega) \\ \Delta_{2V}(k,\omega) \\ \vdots \\ \Delta_{MV}(k,\omega) \end{bmatrix} = \sum_{i=2}^{R} \alpha_i(k,\omega)\boldsymbol{u}_i(\omega)$$

$$\alpha_i(k,\omega) = \frac{\sigma_i(\omega)\boldsymbol{v}_i^{\mathrm{H}}(\omega)\boldsymbol{S}_1(k,\omega)}{\sigma_1(\omega)\boldsymbol{v}_1^{H}(\omega)\boldsymbol{S}_1(k,\omega)}$$

From (5), the MIMO acoustic system of (3) can be treated as the single-input multiple-output (SIMO) acoustic system. The single input is the virtual sound source $S_V(k,\omega)$ with the TF $\boldsymbol{A}_V(\omega) + \boldsymbol{\Delta}_V(k,\omega)$. The virtual sound source is formed by mapping the interference signals $\boldsymbol{S}_1(k,\omega)$ along the most sensitive input direction $\boldsymbol{v}_1(\omega)$ which in turn is scaled by the maximum singular value $\sigma_1(\omega)$. The TF of the virtual sound source consists of two parts, time-invariant part $\boldsymbol{A}_V(\omega)$ and time-varying part $\boldsymbol{\Delta}_V(k,\omega)$. This paper considers that $\boldsymbol{A}_V(\omega)$ is constructed by the highest gain output direction $\boldsymbol{u}_1(\omega)$ and $\boldsymbol{\Delta}_V(k,\omega)$ is the linear combination of $\boldsymbol{u}_2(\omega) \sim \boldsymbol{u}_R(\omega)$ with time-varying coefficients $\alpha_i(k,\omega)$.

III. Suppress Interference and Stationary Noise Signals

This section presents the proposed TFR based beamformer and multi-channel adaptive filter for suppressing the noise signals ($S_2(k,\omega) \sim S_P(k,\omega)$ and $N_m(k,\omega)$). According to section II-B, equation (2) can be written as

$$\begin{aligned} X_m(k,\omega) &= \sum_{p=1}^{P} A_{mp}(\omega) S_p(k,\omega) + N_m(k,\omega) \\ &= A_{m1}(\omega) S_1(k,\omega) + \sum_{p=2}^{P} A_{mp}(\omega) S_p(k,\omega) + N_m(k,\omega) \\ &= A_{m1}(\omega) S_1(k,\omega) + \left(A_{mV}(\omega) + \Delta_{mV}(k,\omega)\right) S_V(k,\omega) + N_m(k,\omega) \end{aligned} \tag{6}$$

For the virtual sound source components, we consider $A_{mV}(\omega)S_V(k,\omega)$ and $\Delta_{mV}(k,\omega)S_V(k,\omega)$ to be the principal part and residual part respectively, since $\boldsymbol{A}_V(\omega)$ is the highest gain output direction of the transfer function matrix $\mathbf{A}_1(\omega)$ and $A_{mV}(\omega)S_V(k,\omega)$ is constructed by the principal interference subspace. If the sound source number is two, i.e., *P*=2, then the residual part is zero.

A. Transfer Function Ratio Based Beamformer

The system architecture for suppressing the interference and stationary noise signals is shown in Fig. 1. This paper uses the TFR and multi-channel adaptive filter techniques for noise reduction problems. Hence, *M* microphones received signals are separated into *M*-1 microphone pairs for the subsequent signal processing. It is supposed that the TFRs defined in (7)

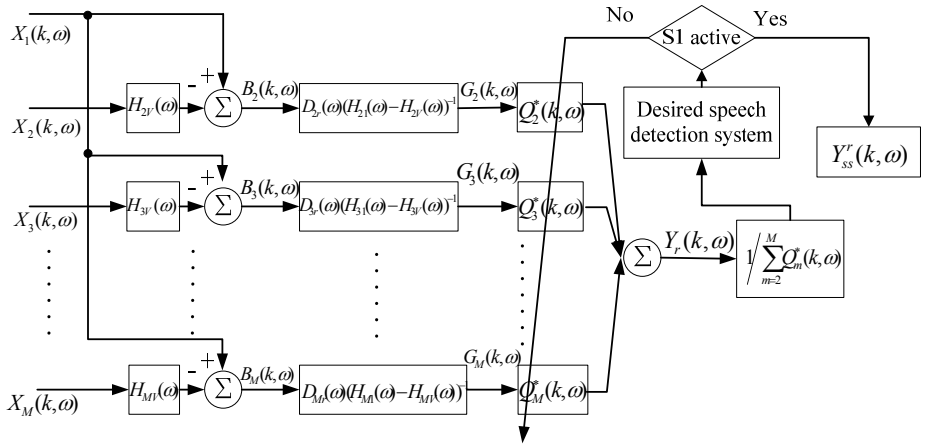

Fig. 1. The system architecture for interference signal and stationary noise suppression

have been identified using the method introduced in [11]. The TFRs for the desired speech and virtual sound source are defined as

$$H_{m1}(\omega)=\frac{A_{11}(\omega)}{A_{m1}(\omega)}\ ,\quad H_{mV}(\omega)=\frac{A_{1V}(\omega)}{A_{mV}(\omega)},\qquad m=2,3,\cdots,M \tag{7}$$

First, this paper employs the TFR of the virtual sound source to remove the principal part of the virtual sound source for each microphone pair:

$$\begin{aligned}
&B_m(k,\omega)=X_1(k,\omega)-\left(\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\right)X_m(k,\omega)\\
&=\left(A_{11}(\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}A_{m1}(\omega)\right)S_1(k,\omega)+N_1(k,\omega)\\
&\quad-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}N_m(k,\omega)+\left(\Delta_{1V}(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\Delta_{mV}(k,\omega)\right)S_V(k,\omega)\\
&=A_{m1}(\omega)\left(\frac{A_{11}(\omega)}{A_{m1}(\omega)}-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\right)S_1(k,\omega)+N_1(k,\omega)\\
&\quad-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}N_m(k,\omega)+\left(\Delta_{1V}(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\Delta_{mV}(k,\omega)\right)S_V(k,\omega)\\
&\text{for}\ \ m=2,3,\cdots,M
\end{aligned} \tag{8}$$

Equation (8) represents that the spatial null is placed toward the principal part direction of the virtual sound source by using two microphones. If the sound source number is two ($\Delta_{mV}(k,\omega)=0$), equation (8) means that the spatial null is placed toward the only competing speech directly. The TFR beamformer output $B_m(k,\omega)$ consists of 3 terms: distorted desired speech signal, residual virtual sound source and stationary noise. Since the TFR $H_{m1}(\omega)$ and $H_{mV}(\omega)$ are known and we assume $(H_{m1}(\omega)-H_{mV}(\omega))$ is non-zero.

To mitigate the distortion on the desired speech signal, (8) is multiplied by $D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}$ as:

$$\begin{aligned}
&G_m(k,\omega)=B_m(k,\omega)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}\\
&=A_{r1}(\omega)S_1(k,\omega)\\
&+\left(N_1(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}N_m(k,\omega)\right)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}\\
&+\left(\Delta_{1V}(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\Delta_{mV}(k,\omega)\right)S_V(k,\omega)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}
\end{aligned} \tag{9}$$

where

$$D_{mr}(\omega)=\frac{A_{r1}(\omega)}{A_{m1}(\omega)}=H_{m1}(\omega)\,(H_{r1}(\omega))^{-1}\qquad r=1,2,\cdots,M$$

$D_{mr}(\omega)$ is used to adjust the desired speech signal distortion to the same reference and r is the reference microphone number which we can select.

The noise part of output signal $G_m(k,\omega)$ still contains the residual part of the virtual sound source and stationary noise, and hence the multi-channel adaptive filter stage is employed here to minimize the noise in $G_m(k,\omega)$. Let us sum all the output signals $G_m(k,\omega)$ with the weighting function $Q_m(k,\omega)$:

$$\begin{aligned}
&Y_r(k,\omega)=Q_2^*(k,\omega)G_2(k,\omega)+\cdots+Q_M^*(k,\omega)G_M(k,\omega)\\
&=A_{r1}(\omega)S_1(k,\omega)\sum_{m=2}^{M}Q_m^*(k,\omega)\\
&+\sum_{m=2}^{M}Q_m^*(k,\omega)\left(N_1(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}N_m(k,\omega)\right)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}\\
&+\sum_{m=2}^{M}Q_m^*(k,\omega)\left(\Delta_{1V}(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\Delta_{mV}(k,\omega)\right)S_V(k,\omega)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}
\end{aligned} \tag{10}$$

where * represents the complex conjugation. The noise components can be cancelled if

$$\boldsymbol{Q}^{\mathrm{H}}(k,\omega)\boldsymbol{Z}(k,\omega)=-Q_2(k,\omega)Z_2(k,\omega) \tag{11}$$

where H represents conjugation transpose;

$$\begin{aligned}
&Q_2(k,\omega)=1\\
&\boldsymbol{Q}(k,\omega)=\left[Q_3(k,\omega)\ \ \cdots\ \ Q_M(k,\omega)\right]^{\mathrm{T}}\\
&\boldsymbol{Z}(k,\omega)=\left[Z_3(k,\omega)\ \ \cdots\ \ Z_M(k,\omega)\right]^{\mathrm{T}}
\end{aligned}$$

$$Z_m(k,\omega)=\left(N_1(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}N_m(k,\omega)\right)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}$$

$$+\left(\Delta_{1V}(k,\omega)-\frac{A_{1V}(\omega)}{A_{mV}(\omega)}\Delta_{mV}(k,\omega)\right)S_V(k,\omega)D_{mr}(\omega)(H_{m1}(\omega)-H_{mV}(\omega))^{-1}$$

The solution of $\boldsymbol{Q}(k,\omega)$ can be found by using adaptive algorithm suggested in section III-B when $S_1(k,\omega)$ is silent (desired speech inactive periods). Once the weight vector $\boldsymbol{Q}(k,\omega)$ is obtained, the beamformer output can be given as:

$$Y_{ss}^{r}(k,\omega)=\frac{Y_r(k,\omega)}{\sum_{m=2}^{M}Q_m^*(k,\omega)} \tag{12}$$

B. Multi-channel Adaptive Algorithm

For the real environment, it is unlikely to remove the noise completely and hence the output signal $Y_r(k,\omega)$ can be expressed as:

$$Y_r(k,\omega)=A_{r1}(\omega)S_1(k,\omega)\sum_{m=2}^{M}Q_m^*(k,\omega)+e_n(k,\omega) \tag{13}$$

where $e_n(k,\omega)$ is the residual noise and it is anticipated that the desired speech signal components are dominant compared to the residual noise. Therefore, equation (12) can be written as:

$$Y_{ss}^{r}(k,\omega)=A_{r1}(\omega)S_1(k,\omega)+\frac{e_n(k,\omega)}{\sum_{m=2}^{M}Q_m^*(k,\omega)} \tag{14}$$

According to (11), the error signal at frequency ω and frame *k* is written as:

$$\varepsilon_Z(k,\omega)=-Q_2(k,\omega)Z_2(k,\omega)-\boldsymbol{Q}^{\mathrm{H}}(k,\omega)\boldsymbol{Z}(k,\omega) \tag{15}$$

The optimal set of filter coefficients vectors $\boldsymbol{Q}(k,\omega)$ can be found using the formula:

$$\min_{Q}\varepsilon_Z(k,\omega)\varepsilon_Z^*(k,\omega) \tag{16}$$

To avoid amplifying the term $e_n(k,\omega)$ in (14) and arriving at a trivial solution of (16), a penalty function is added into (16) as:

$$\min_{Q}\varepsilon_Z(k,\omega)\varepsilon_Z^*(k,\omega)+\mu\varepsilon_N(k,\omega)\varepsilon_N^*(k,\omega) \tag{17}$$

where μ is the penalty parameter;

$\varepsilon_N(k,\omega)=\beta-\boldsymbol{Q}^{\mathrm{H}}(k,\omega)\boldsymbol{O}$

$\boldsymbol{O}=\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^{\mathrm{T}}\in R^{(M-2)\times 1}$

β is a constant larger than one to ensure $\sum_{m=2}^{M}Q_m^*(k,\omega)$ not amplifying the noise term in (14). Thus, the normalized least-mean-square (NLMS) solution of (17) is given by:

$$\hat{\boldsymbol{Q}}(k+1,\omega)=\hat{\boldsymbol{Q}}(k,\omega)+\frac{\lambda\left(\varepsilon_Z^*(k,\omega)\boldsymbol{Z}(k,\omega)+\mu\varepsilon_N^*(k,\omega)\boldsymbol{O}\right)}{\boldsymbol{Z}^{\mathrm{H}}(k,\omega)\boldsymbol{Z}(k,\omega)+\mu\boldsymbol{O}^{\mathrm{H}}\boldsymbol{O}} \tag{18}$$

where λ is the small positive step size. Notably, if the solution $\hat{\boldsymbol{Q}}(k,\omega)$ can meet the constraint of (17), the noise components in (10) can not only be minimized but also be attenuated again using (12).

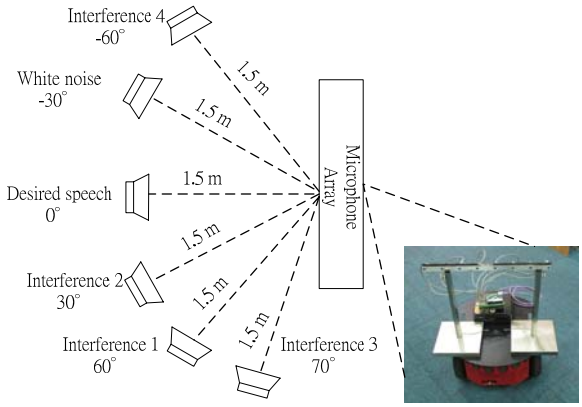

Fig. 2. Arrangement of microphone array and sources

TABLE I
FIVE KINDS OF EXPERIMENTAL CONDITIONS

C1	Desired speech at 0° and stationary noise at -30°
C2	Desired speech at 0°, stationary noise at -30° and interference signal at one of (30°, 60°, 70°, -60°)
C3	Desired speech at 0°, stationary noise at -30° and interference signals at two of (30°, 60°, 70°, -60°)
C4	Desired speech at 0°, stationary noise at -30° and interference signals at three of (30°, 60°, 70°, -60°)
C5	Desired speech at 0°, stationary noise at -30° and interference signals at 30°, 60°, 70° and -60°

IV. EXPERIMENTAL RESULTS

The proposed algorithm was tested in a real environment with dimensions 10 m × 6 m × 3.6 m and the reverberation time at 1000 Hz is 0.52 second. A uniform linear microphone array of eight un-calibrated microphones separated by 0.05 m was constructed for this experiment. The amplified microphone signals were sampled at 8 kHz and 16 bits. The microphone array was placed on a mobile robot at a distance of 2 m from the wall. The arrangement of microphone array and sound sources is shown in Fig. 2. The desired speech signal at 0° consists of sentences from TCC-300 database [12] spoken by 150 males and 150 females. The interference signals 1, 2 and 4 are speech signals spoken by 3 females and interference signal 3 is the speech signal spoken by a male. The position of each sound source is fixed in this experiment. Two speech enhancement algorithms, delay and sum beamformer (DSB) [1] and reference signal based adaptive beamformer (RAB) implemented in frequency domain [13] are adopted to compare with the proposed algorithm. For RAB and the proposed algorithm, we assume the perfect desired speech detection system exists, allowing the adaptive noise cancellation system to adapt filter weight during inactive periods of desired speech. The STFT size is 1024 with 320 shift samples and 64 zero padding samples. The parameters of λ, μ and β are set to 0.2, 1 and $2+2i$ and the step size of all adaptive algorithm is set to 0.2. Five

conditions denoted from C1 to C5 for the experiments are listed in Table I. The average SINR is defined as

$$\mathrm{SINR}=\frac{E\left\{\left[a_{11}(t)\otimes s_1(t)\right]^2\right\}}{\sum_{p=2}^{P}E\left\{\left[a_{1p}(t)\otimes s_p(t)+n_1(t)\right]^2\right\}} \quad (19)$$

where $E(\cdot)$ is the expectation operation.

A. *Interference signals and stationary noise suppression evaluation*

This section evaluates the interference signals and stationary noise suppression ability of the proposed algorithm and hence the output $Y_{ss}^{r}(k,\omega)$ is sent for waveform assessment. For the proposed algorithm, the reference microphone number *r* is set to one. For the RAB algorithm, the pre-recorded speech signals of the first microphone are chosen as the desired signal and hence the minimum criterion can be written as

$$\min_{Q}\left[D(k,\omega)-\boldsymbol{Q}^{\mathrm{H}}(k,\omega)\hat{\boldsymbol{X}}(k,\omega)\right]\left[D(k,\omega)-\boldsymbol{Q}^{\mathrm{H}}(k,\omega)\hat{\boldsymbol{X}}(k,\omega)\right]^{*} \quad (20)$$

where $\hat{\boldsymbol{X}}(k,\omega)$ is the vector containing the linear combination of present microphone received signal and $A_{m1}(\omega)\tilde{S}_1(k,\omega)$. $\tilde{S}_1(k,\omega)$ is the representative speech signal at 0° in Fig. 3 and $A_{m1}(\omega)\tilde{S}_1(k,\omega)$ are the pre-recorded speech signals which can be recorded when the environment is quiet. $D(k,\omega)$ is the desired signal set to $A_{11}(\omega)\tilde{S}_1(k,\omega)$ in this section. The filter weight $\boldsymbol{Q}(k,\omega)$ can be trained when the desired speech signal is inactive.

Two objective performance indices are used to measure the waveform property directly. The first is frequency weighted segmental SINR (FWSINR) defined as

$$\mathrm{FWSINR(dB)}=\frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{\sum_{\omega}C(k,\omega)10\log_{10}(\sigma_{s,\omega}^2/\sigma_{in,\omega}^2)}{\sum_{\omega}C(k,\omega)}\right) \quad (21)$$

where *k* is the frame when the desired speech signal is active. $C(k,\omega)$ is the frequency weighting at frame *k* for the ear's critical bands ω [14]. Within the band ω, $\sigma_{s,\omega}^2$ is the signal component power of the reference signal $x_{1,s}(t)$ and $\sigma_{in,\omega}^2$ is the power of noise signal $x_{1,s}(t)-g_y y(t)$ for the same segment. Note that $x_{1,s}(t)$ is the signal component recorded by the first microphone, g_y is the gain factor and $y(t)$ is the output of the algorithm.The second quality measure is segmental noise level (segNL)

$$\mathrm{segNL(dB)}=\frac{1}{K}\sum_{k=1}^{K}\left(10\cdot\log_{10}(\sum_{i=1}^{I}y^2(i+kI))\right) \quad (22)$$

where $y(t)$ is the algorithm output when $s_1(t)$ is silent and $s_2(t)\sim s_P(t)$ and $n_m(t)$ are all active. *I* is the length of the frame.

The third quality measure is log spectral distortion (LSD) defined as

$$\mathrm{LSD}=\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{W}\sum_{\omega=1}^{W}\left(20\cdot\log_{10}|A_{11}(\omega)S_1(k,\omega)|-20\cdot\log_{10}|Y(k,\omega)|\right)^2} \quad (23)$$

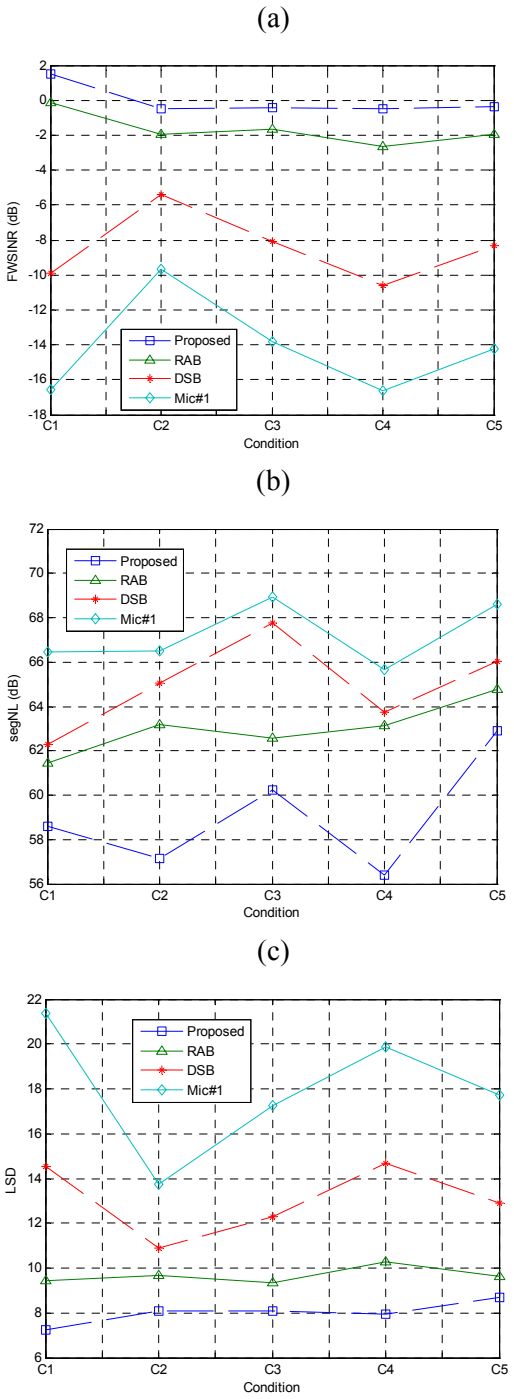

Fig. 3. (a) FWSINR results
(b) segNL results (c) LSD results

where $Y(k,\omega)$ is the STFT of the algorithm output. Note that a lower LSD level corresponds to a better performance. The experimental results are shown in Fig. 3 and Mic#1 represents the contaminated speech recorded by the first microphone. The ranges of average input SINR are -3 dB to 3 dB. The test sentences for each figure are 200. As can be seen, the best performance is obtained by the proposed algorithm and the DSB performs worst. Since the DSB aligns only the direct path signal, it does not take reflections into account and no nulls are placed directly in interference signal directions.

For the RAB algorithm, the finite impulse response coefficients $\boldsymbol{Q}(k,\omega)$ are trained to achieve two objectives simultaneously during the desired speech inactive periods: to suppress the interference and stationary noise signals, and to adjust the distorted desired speech of each microphone $A_{m1}(\omega)\widetilde{S}_1(k,\omega)$ to the same channel effect $A_{11}(\omega)\widetilde{S}_1(k,\omega)$. However, the finite number of taps and NLMS adaptive algorithm are unlikely to achieve these two objectives fully at the same time especially for complex channel dynamics. (e.g., competing speeches are present). It is unlike the proposed algorithm which separates these two objectives. The proposed algorithm suppresses competing speech and adjusts desired speech channel effect first using TFR techniques and then minimizes the residual noise with multi-channel adaptive filter. This is the reason why RAB performs better than DSB but worse than the proposed algorithm.

B. *Automatic Speech Recognition Tests*

ASR systems are sensitive to additive noise and speech distortion, especially for the competing speech. Therefore, this section utilizes the ASR rates to measure the performance of the proposed algorithm. The ASR system [15] that we use is the Hidden Markov Model (HMM) based Mandarin keyword spotting recognition system. The feature vector is 26-dimensional Mel Frequency Cepstral Coefficients (MFCC) and the TCC-300 database with some white noise is used for training. The testing database is speaker independent 3332 words spoken by 11 female and 18 male and each word consists of one Chinese name. The vocabulary size is 121 Chinese names. The testing data are played at the desired speech position in Fig. 2 and the time domain speech enhancement output is sent directly to the ASR system for further processing. The recognition result is considered correct when the output Chinese name of the ASR system is completely the same with the input. The correct rates, when tested on the clean 3332 words ($s_1(t)$) and on the received signal of the first microphone ($a_{11}(t) \otimes s_1(t)$), are 100 % and 90.49 % respectively. The recognition results for different conditions are summarized in Table II and the correct rate of each condition is obtained by using random 500 Chinese names from the testing database. As can be seen, the proposed algorithm has the best correct rate and Table II also indicates that ASR system can be considered an application of the proposed algorithm.

TABLE II
ASR CORRECT RATES (%)

Input SINR	Condition	Proposed	RSAB	DSB	Mic#1
-3 ~ 3 (dB)	C1	77.6	47.8	39.2	5.4
	C2	66.8	54.8	50	13.6
	C3	63	48.2	42.8	11
	C4	41.4	28.6	22	2.4
	C5	67.2	60.8	57.6	16

V. CONCLUSION

This paper proposes speech enhancement method to perform desired speech extraction and multiple competing speeches and stationary noise signals suppression. Unlike the generalized sidelobe canceler (GSC) structure comprised of three building blocks [3] or the RAB structure [13] which minimizes the noise signals and equalizes the channel effect using only adaptive filters, this paper proposes a two-stage speech enhancement algorithm using the TFR beamformer and the multi-channel adaptive filter algorithm. The virtual sound source concept which transforms the multiple competing speeches from MIMO to SIMO acoustic system is presented to simplify the complicated acoustic system and a novel TFR estimation method for more practical scenario is also derived [11]. The performance of the proposed algorithm was tested in a real, noisy and reverberant environment and we also showed the improvement on correct rate using Mandarin ASR system. An efficient desired speech signal detection system is a further research topic.

ACKNOWLEDGMENT

The authors would like to thank Professor Sin-Horng Chen of National Chiao-Tung University, Taiwan for providing the speech recognition software.

REFERENCES

[1] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] O.L. Frost III, "An algorithm for linear constrained adaptive array processing," *Proc. IEEE*, vol.60, no.8, pp. 926–935, Aug. 1972.

[3] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag*., vol.AP-30, no.1, pp. 27–34, Jan. 1982.

[4] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, H. G. Okuno, "Robust Recognition of Simultaneous Speech By a Mobile Robot," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 742-752, 2007.

[5] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound Source Separation of Moving Speakers for Robot Audition," *in Proc. of the IEEE Int. Conf. Acoust., Speech, Signal Process*., pp. 3685–3688, Apr. 2009.

[6] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata and H. G. Okuno "Step-size Parameter Adaptation of Multi-channel Semi-blind ICA with Piecewise Linear Model for Barge-in-able Robot Audition," *in Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems.,* pp.2277-2282, Oct. 2009.

[7] H.D. Kim, J. Kim, K. Komatani, T. Ogata and H. G. Okuno "Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments," *in Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems.,* pp.1705-1711, Sept. 2008.

[8] K. Hosoya, T. Ogawa and T. Kobayashi "Robot auditory system using head-mounted square microphone array," *in Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems.,* pp.2736-2741, Oct. 2009.

[9] E. E. Jan and J. Flanagan, "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors," *in Proc. of the IEEE Int. Conf. Acoust., Speech, Signal Process*., pp.917–920, May 1996.

[10] G. Reuven, S. Gannot and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech and Language Process*., vol.16, no.4, pp.711-727, May 2008.

[11] Website: http://140.113.150.64/Transfer Function Ratio Estimation.pdf

[12] The Association for Computational Linguistics and Chinese Language Processing, Website: http://www.aclclp.org.tw/corp.php

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

[14] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[15] S.H. Chen, Website: http://www.speech.cm.nctu.edu.tw/