Processing of speech signals using a microphone array for intelligent robots

J Hu*, **C C Cheng**, and **W H Liu**
Department of Electrical and Control Engineering, National Chiao Tung University, Taiwan, Republic of China

The manuscript was received on 22 January 2004 and was accepted after revision for publication on 11 November 2004.

DOI: 10.1243/095965105X9461

Abstract: For intelligent robots to interact with people, an efficient human–robot communication interface is very important (e.g. voice command). However, recognizing voice command or speech represents only part of speech communication. The physics of speech signals includes other information, such as speaker direction. Secondly, a basic element of processing the speech signal is recognition at the acoustic level. However, the performance of recognition depends greatly on the reception. In a noisy environment, the success rate can be very poor. As a result, prior to speech recognition, it is important to process the speech signals to extract the needed content while rejecting others (such as background noise). This paper presents a speech purification system for robots to improve the signal-to-noise ratio of reception and an algorithm with a multidirection calibration beamformer.

Keywords: beamforming, beamformer, DOA, microphone array, robot hearing, speech enhancement

1 INTRODUCTION

With the advent of computing power of microprocessors and digital signal processors, the possibility of constructing an intelligent robot to perform complex tasks is not such a far-reaching goal. Among various features offered by an intelligent robot, the communication interface is still an on-going research topic. It is generally believed that the interface should not be restricted to keyboard, mouse, or remote controller, but also to the nature language instead. For these reasons, robot hearing research has received much attention over the years. Chun and Caudell [**1**] tried to use the inferior colliculus structure and the head related transfer function (HRTF) information combined with the image processing technique to find general rules of human hearing. Schauer and Gross [**2**] use interaural time difference (ITD) and interaural intensity difference (IID) signals to perform a 360° direction of arrival (DOA) estimation. Speech recognition will inevitably be incorporated into an intelligent robot to make it understand what people say or which command is given. Although speech recognition can have high accuracy in a quiet environment, undesirable signal components due to the ambient noise and channel distortion render the recognizer unusable for real-world applications. An adaptive microphone array system is thus designed to purify the polluted signal and to improve the recognition rate.

Using adaptive microphone array algorithms for enhancing speech reception in a noisy environment has been developed for many years. Earlier approaches, such as the Frost beamformer [**3**], GSC [**4**], and the robust adaptive beamformer [**5**], are only good in the ideal case. The ideal case here means that the microphones are mutually matched and the environment is a free space. To cope with these limitations, Hoshuyama *et al.* [**6**] proposed two robust constraints on the blocking matrix design. Weinstein [**7**] proposed a new channel estimation method for standard GSC architecture in the frequency domain. However, its estimation accuracy would be decreased by a louder noise and circuit noise. Dahl and Claesson [**8**] proposed an adaptive algorithm which calibrates both the microphone mismatch and channel effect using *a priori* information. This *a priori* information is a set of speech data recorded by the same microphone array in a

* *Corresponding author: Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China. email: jshu@cn.nctu.edu.tw*

quiet environment. It then serves as a reference signal to update the coefficients of the filters when the speaker is silent (or non-speech segments) and the environment is noisy. With this *a priori* information, the calibration problem would be solved implicitly. Dahl's algorithm is suitable in the car environment where the speaker's position is fixed (e.g. the driver). To apply the algorithm for mobile robots, it is necessary to record reference signals from all directions since the speaker's position might not be fixed. In this paper, a beamforming architecture modified from the method proposed by Dahl and Claesson [**8**] is constructed by using a beam-steer filter with only one set of pre-recorded speech source. As a result, the memory requirement and the effort of pre-recording are reduced tremendously. This modified architecture could be more suitable for a robot hearing application.

The direction of the speaker must be known before the beam is formed in the speaker direction. In a noisy environment, the conventional delay estimation method in the time domain [**9**] or in the frequency domain [**10–13**] is not able to obtain satisfactory results. In order to make a sound source direction available, a customized wide-band eigenstructure-based DOA estimation algorithm is proposed in this system. This method is based on a blind DOA estimation algorithm called MUSIC (multiple signals classification) [**14**], with modifications to decrease the computing time and increase the accuracy of the DOA estimation.

The overall system is shown in Fig. 1. The first part consists of a speech activity detection to decide when the adaptive beamformer should be switched on or off. The second part is a DOA estimation and adaptation of the upper beamformer. By incorporating DOA knowledge the beam-steer filter is used to steer the direction of the beam for acquiring clean speech of a speaker. Because the target is a speech signal, a broadband beam-steer filter is needed. The third part is to apply the beamformer computation to increase the signal-to-noise ratio (SNR).

The paper is organized as follows. The customized wide-band eigenstructure-based DOA estimation algorithm will be described in section 2. Section 3 discusses the modified beamformer, speech activity detection, and the beam-steer filter. Section 4 provides experimental results of the DOA and beamformer obtained with the speaker in several different directions. Finally, a conclusion will be given in section 5.

2 DIRECTION OF ARRIVAL (DOA) ESTIMATION

The idea of a blind DOA estimation algorithm called MUSIC [**14**] is adopted in this platform to detect the speaker's direction. The received signal contains d sources and can be presented as

$$x_m(t) = \sum_{k=1}^{d} a_{mk} s_k(t - \tau_{mk}) + n_m(t) \tag{1}$$

Generally, sources here may include speech source and interference signals from the acoustic environment. Noise $n_m(t)$ is referred to non-directional interference signals such as electronic noise (called non-directional noise in the following context). In order to express the delay relations into the phase shift, the received signal is transformed into the

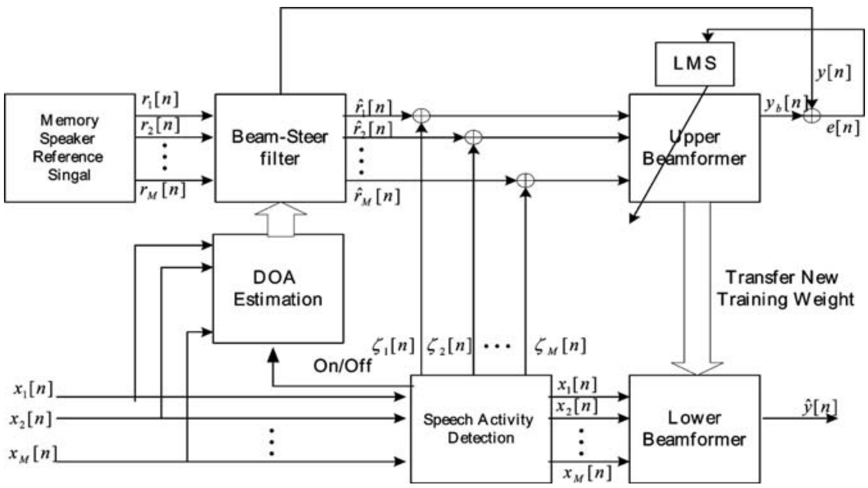


Fig. 1 Overall system structure

frequency domain over a finite observation interval T

$$X_m(\omega_l) = \frac{1}{T}\int_{-T/2}^{T/2} x_m(t)\,\mathrm{e}^{-\mathrm{j}\omega_l t}\,\mathrm{d}t$$
$$\omega_l = \frac{2\pi}{T} l, \qquad \text{for } l = 1, \ldots, L \tag{2}$$

where ω_1 and ω_L are the lowest and highest frequencies included in bandwidth B.

The original model can be described as

$$X_m(\omega_l) = \sum_{k=1}^{d} a_{mk} S_k(\omega_l)\,\mathrm{e}^{-\mathrm{j}\omega_l \tau_{mk}} + N_m(\omega_l) \tag{3}$$

Rewrite equation (3) in matrix form as

$$\boldsymbol{X}(\omega_l) = \mathbf{A}(\omega_l)\boldsymbol{S}(\omega_l) + \boldsymbol{N}(\omega_l) \tag{4}$$

where

$$\boldsymbol{X}^T(\omega_l) = [X_1(\omega_l), \ldots, X_M(\omega_l)]$$
$$\boldsymbol{N}^T(\omega_l) = [N_1(\omega_l), \ldots, N_M(\omega_l)]$$
$$\boldsymbol{S}^T(\omega_l) = [S_1(\omega_l), \ldots, S_d(\omega_l)]$$
$$\mathbf{A}(\omega_l) = \begin{bmatrix} a_{11}\,\mathrm{e}^{-\mathrm{j}\omega_l\tau_{11}} & \ldots & a_{1d}\,\mathrm{e}^{-\mathrm{j}\omega_l\tau_{1d}} \\ \vdots & & \vdots \\ a_{M1}\,\mathrm{e}^{-\mathrm{j}\omega_l\tau_{M1}} & \ldots & a_{Md}\,\mathrm{e}^{-\mathrm{j}\omega_l\tau_{Md}} \end{bmatrix}$$

Note that each column presents the delay relations caused by different sources between microphones, the ith column vector of $\mathbf{A}(\omega_l)$ being denoted by $\boldsymbol{A}_i(\omega_l)$ and referred to as the direction vector.

Suppose noises are mutually independent. If the noise correlation matrix is the diagonal matrix $\sigma^2(\omega_l)\mathbf{I}$, the received signal correlation matrix can be described as

$$\mathbf{R}_{xx}(\omega_l) = \mathbf{A}(\omega_l)\mathbf{R}_{ss}(\omega_l)\mathbf{A}^H(\omega_l) + \sigma^2(\omega_l)\mathbf{I} \tag{5}$$

where

$$\mathbf{R}_{ss}(\omega_l) = E[\boldsymbol{S}(\omega_l)\boldsymbol{S}^H(\omega_l)]$$

and the eigenvalue decomposition

$$\mathbf{R}_{xx}(\omega_l) = \sum_{i=1}^{M} [\lambda_i(\omega_l) - \sigma_n^2(\omega_l)]\boldsymbol{E}_i(\omega_l)\boldsymbol{E}_i^H(\omega_l) \tag{6}$$

with eigenvalues $\lambda_1(\omega_l) \geqslant \lambda_2(\omega_l) \geqslant \cdots \geqslant \lambda_M(\omega_l)$. From equations (4) and (5), the source part correlation matrix is

$$\boldsymbol{C}_{xx}(\omega_l) = \boldsymbol{A}(\omega_l)\boldsymbol{R}_{ss}(\omega_l)\boldsymbol{A}^H(\omega_l) = \sum_{i=1}^{d} [\lambda_i(\omega_l) - \sigma_n^2(\omega_l)]\boldsymbol{E}_i(\omega_l)\boldsymbol{E}_i^H(\omega_l) \tag{7}$$

and the rank of $\mathbf{C}_{xx}(\omega_l)$ is d. Then the following equations can be derived

$$RangeSpace\,(\mathbf{C}_{xx}(\omega_l)) = span\,\{\boldsymbol{A}_1(\omega_l), \ldots, \boldsymbol{A}_d(\omega_l)\} = span\,\{\boldsymbol{E}_1(\omega_l), \ldots, \boldsymbol{E}_d(\omega_l)\}$$
$$RangeSpace\,(\mathbf{A}(\omega_l))^{\perp} = span\,\{\boldsymbol{E}_{d+1}(\omega_l), \ldots, \boldsymbol{E}_M(\omega_l)\}$$

Combining the equations above, the signal subspace can be defined as

$span\,\{\boldsymbol{E}_1(\omega_l), \ldots, \boldsymbol{E}_d(\omega_l)\}$ is the source subspace

$span\,\{\boldsymbol{E}_{d+1}(\omega_l), \ldots, \boldsymbol{E}_M(\omega_l)\}$
is the non-directional noise subspace

Because the source subspace is orthogonal to the non-directional noise subspace

$$\boldsymbol{E}_j^H(\omega_l)\boldsymbol{A}_i(\omega_l) = 0, \qquad i = 1, \ldots, d;\ j = d+1, \ldots, M \tag{8}$$

By equation (8), a non-directional noise projection matrix $\mathbf{P}_N(\omega_l)$ can be established as

$$\mathbf{P}_N(\omega_l) = \sum_{i=d+1}^{M} \boldsymbol{E}_i(\omega_l)\boldsymbol{E}_i^H(\omega_l) \tag{9}$$

The number of sources d can be determined by the distribution of eigenvalues. The DOA can be detected by projecting the direction vector on to the non-directional noise projection matrix when

$$\mathbf{P}_N(\omega_l)\boldsymbol{A}_i(\omega_l) = 0 \tag{10}$$

Usually, the maximum d values are regarded as the d source directions

$$\frac{1}{(1/L)\sum_{l=1}^{L} \|\boldsymbol{E}_j^H(\omega_l)\boldsymbol{A}_i(\omega_l)\|_2^2} = \frac{1}{(1/L)\sum_{l=1}^{L} \boldsymbol{A}_i^H(\omega_l)\mathbf{P}_N(\omega_l)\boldsymbol{A}_i(\omega_l)} \tag{11}$$

The computing requirement of equation (11) can be reduced by considering only significant frequencies of concern. The selection criterion is based on the assumption that non-directional noises are mutually independent. Therefore, the non-diagonal components of correlation matrix exclude non-directional noise terms. It means the following terms in the correlation matrix (5) should be small

$$R_{x_i x_j}(\omega_l) = \sum_{p=1}^{d}\sum_{o=1}^{d} a_{ip} a_{jo} R_{s_p s_o}(\omega_l), \qquad \forall i \neq j \tag{12}$$

Then the Q significant frequencies $\hat{\omega}_1, \ldots, \hat{\omega}_Q$ can be selected as

$$\hat{\omega}_q = \left\langle \sum_{i=1}^{M}\sum_{j=i+1}^{M} |R_{x_i x_j}(\omega_l)| \right\rangle_q \tag{13}$$

As a result, the d source directions can be estimated by searching maximum d values of

$$J(\theta_i) = \frac{1}{(1/Q)\sum_{q=1}^{Q} \boldsymbol{A}_i^H(\hat{\omega}_q)\mathbf{P}_N(\hat{\omega}_q)\boldsymbol{A}_i(\hat{\omega}_q)} \tag{14}$$

Searching the spectrum for d peaks to determine the direction of arrival still requires plenty of process time when the accuracy requirement is high. This is the drawback of this method, which requires further improvements. Although there is the root-finding MUSIC [**15**] algorithm to calculate the DOA without searching the spectrum, a uniform-shaped array is needed. Because the shape of the microphone array on the robot may change with different applications, the root-finding method is not implemented in the proposed platform.

3 SPEECH ENHANCEMENT

3.1 The modified beamformer approach

The approach could be arranged in the following steps:

Step 1 is to pre-record the speech source.

Step 2 is speech activity detection described in section 3.2.

Step 3 is to adjust the pre-recorded speech source by the beam-steer filter in order to produce the correct reference signals. The DOA information is obtained by the MUSIC algorithm mentioned above. Generally, the MUSIC spectrum contains both directional information of the speaker and an interference signal during the speech segment. In order to determine the speaker's direction, the MUSIC spectrum is computed contiguously and then the speaker's direction can be obtained by comparing the spectrums before and after the speech activity is detected. The design of the beam-steer filter will be mentioned in section 3.3 and the modified reference signals are denoted as $\hat{r}_1[n], \ldots, \hat{r}_M[n]$.

In step 4, the weighting matrix of the upper beamformer is modified in the non-speech segments, and the newly updated weighting matrix is passed to the lower beamformer in the speech segments. The LMS method is used here to perform the adaptation in the non-speech segments. If the speech segments are detected, the data would flow through the lower beamformer and then the output data sequence $\hat{y}[n]$ could be produced. Assume that the order of the weighting vector in each microphone is F. The adaptation of LMS algorithm is

$$\begin{aligned}
&\boldsymbol{w}[k+1] = \boldsymbol{w}[k] + \mu(y[k] - y_b[k])(\hat{\boldsymbol{r}}[k] + \boldsymbol{\zeta}[k])\\
&\boldsymbol{w}^{\mathrm{T}}[k] = [w_{11}[k], \ldots, w_{1F}[k-F-1]\\
&\qquad\qquad w_{21}[k], \ldots, w_{MF}[k-F-1]]\\
&\boldsymbol{\zeta}^{\mathrm{T}}[k] = [\zeta_1[k], \ldots, \zeta_M[k]]\\
&\hat{\boldsymbol{r}}^{\mathrm{T}}[k] = [\hat{r}_1[k], \ldots, \hat{r}_M[k]]
\end{aligned} \tag{15}$$

3.2 Speech activity detection

Two possible speech detection methods, energy-based and entropy-based [**16**], can be used. They are based on the assumption that the noise is static stationary or slowly varying in time. The entropy-based method is chosen in this paper because it is able to detect voice activity in a low SNR environment.

Observation of the spectrogram of very noisy speech signals shows that the speech segments are more organized than noise segments. Because of this fact, Shannon's entropy [**17**] can be used to measure the organization of the speech signals and was defined as

$$H(G) = -\sum_{u=1}^{U} f(g(u)) \log_2[f(g(u))] \tag{16}$$

where $f(g(u))$ is the probability density function of a speech signal of symbol u. The concept of entropy applied to speech activity detection is based on the assumption that the signal is more organized in speech segments than in non-speech segments. The measure of entropy is redefined in the spectral domain as

$$\begin{aligned}
&H(|G(\omega, z)|^2)\\
&\quad = -\sum_{l=1}^{L} \frac{|G(\omega_l, z)|^2}{\sum_{l=1}^{L}|G(\omega_l, z)|^2} \log\left[\frac{|G(\omega_l, z)|^2}{\sum_{l=1}^{L}|G(\omega_l, z)|^2}\right]
\end{aligned} \tag{17}$$

where z means the zth frame and

$$\begin{aligned}
|G(z)|^2 = [&|G(\omega_1, z)|^2, \ldots, |G(\omega_2, z)|^2, \ldots,\\
&|G(\omega_L, z)|^2]^{\mathrm{T}}
\end{aligned}$$

is the magnitude spectrum for frame z. When the input is a white noise, $H(|G(\omega, z)|^2)$ is maximized and the maximum value is $\log(\omega)$. On the other hand, $H(|G(\omega, z)|^2)$ is minimized when the input is a pure tone and the minimum value is zero. The dynamic of $H(|G(\omega, z)|^2)$ is thus bounded between 0 and $\log(\omega)$ and the entropy of the non-speech segments should be larger than that of the speech segments.

Figure 2 shows the waveform for the utterance '*nine three eight*' (in Mandarin) contaminated by

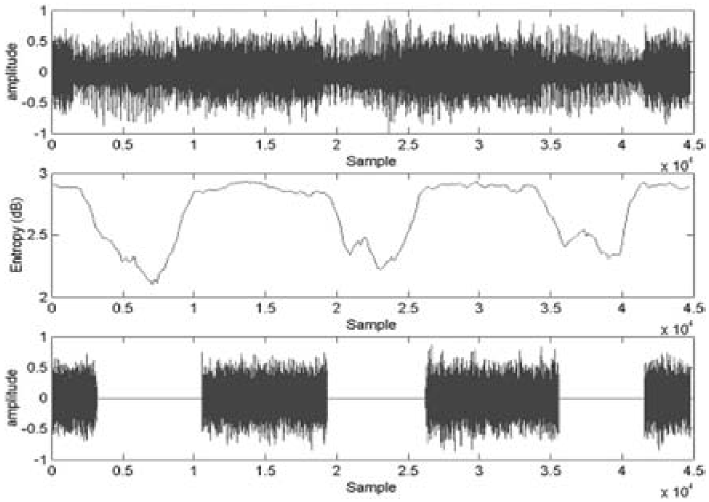


Fig. 2 Noisy signal at an SNR of −5 dB in white Gaussian noise for '*nine three eight*', measured entropy distribution, and the detection of non-speech segments with a fixed threshold of 2.85

white Gaussian noise with a global SNR of −5 dB, measured entropy distribution, and the detection of non-speech segments with a fixed threshold of 2.85. The entropy detection shows an acceptable detection of non-speech segments in highly noisy conditions.

3.3 Beam-steer filter

A simple delay-and-sum algorithm is used for the beam-steering filter. To cope with the fractional delay problem, an optimal fraction delay FIR filter design technique [**18**] is implemented. Without loss of generality, the signals are assumed to have no frequency components above $\alpha\pi$ rad/s $(0<\alpha<1)$ and the optimal estimation $\hat{c}(i)$ through linear combination of the sample values is

$$\hat{c}(i) = \sum_{v=0}^{V} h_v c(v) \tag{18}$$

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \\ h_V \end{bmatrix} = \begin{bmatrix} K(0,0) & K(0,1) & \dots & K(0,V) \\ K(1,0) & K(1,1) & \dots & K(1,V) \\ K(2,0) & K(2,1) & \dots & K(2,V) \\ \vdots & \vdots & & \vdots \\ K(V,0) & K(V-1,1) & \dots & K(V,V) \end{bmatrix}^{-1} \times \begin{bmatrix} K(0,i) \\ K(1,i) \\ K(2,i) \\ \vdots \\ K(V,i) \end{bmatrix} \tag{19}$$

where $K(t, s) = \alpha \sin c[\alpha(t-s)]$.

4 EXPERIMENTAL RESULTS

A uniform, linear array using six microphones is constructed for the experiment. The larger spacing between the microphones could achieve a better beamforming result, but the MUSIC algorithm needs a smaller spacing to prevent the spatial aliasing effect in the lower frequency range. Because the frequency range, 0–2400 Hz, contains the major information of the speech source, the spacing between the microphones is chosen as 7 cm. The amplified microphone signals are sampled by a 16 kHz, 16 bits A/D (analogue-to-digital) card and the computing platform is a Pentium III 550 MHz PC. The array is mounted on an easel with a height of 1 m and 3 m to the nearest wall. The environment is a 20 m × 15 m room full of office furniture to simulate a real environment. The interference signals in the experiment are mutually uncorrelated white noise. The first scenario (Fig. 3) tests the performance under a

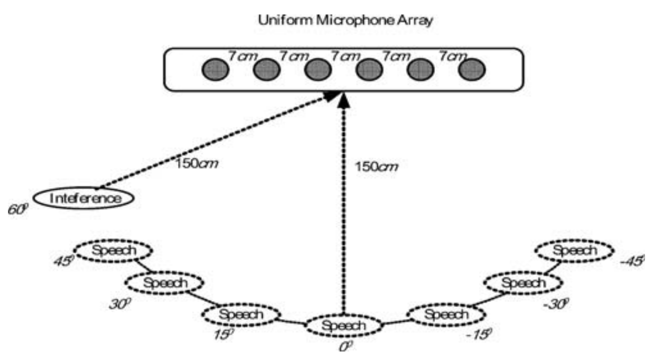


Fig. 3 Testing scenario 1: array of six microphones in a noisy environment

fixed interference signal and different speech source directions. Loudspeakers are used to produce these signals. The interference signal comes from 60° with a distance of 150 cm. The second scenario (Fig. 4) tests the performance under a fixed speech source and a different number of interference signals. Other than the performance of the proposed algorithm (Fig. 1), the original adaptive beamformer proposed by Dahl and Claesson [**8**] is also tested for comparison. The results are shown in the following sections.

4.1 Scenario 1

4.1.1 DOA result

Table 1 shows the statistics of the estimation result of the proposed DOA algorithm where the SNR in different angles can be seen in Table 2. This result is compared with the DOA algorithm that processes all frequencies in a signal bandwidth. Although the proposed algorithm chooses only ten significant frequencies to estimate the power spectrum (as listed in left half of the table), the statistical result shows that it has a better accuracy than the algorithm that processes all frequencies in the signal bandwidth. In Fig. 5, the dotted line and the solid line represent the estimated MUSIC spectrum in the non-speech segment and in the speech segment. By comparing these two spectrums the speaker source direction can be determined.

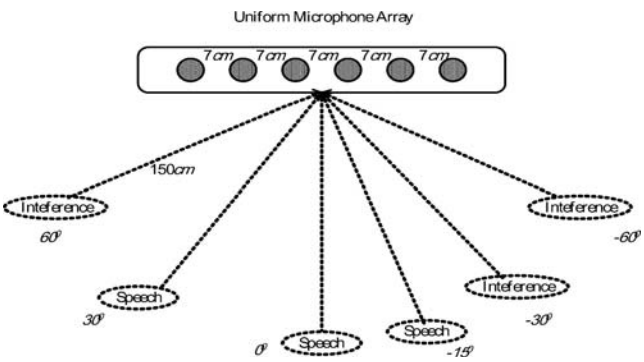


Fig. 4 Testing scenario 2: array of six microphones in a noisy environment

Table 2 Beamforming result with order 30

Correct angle (deg)	Input SNR (dB)	Original beamformer (dB)	Modified beamformer (dB)
45	5.7539	22.3684	21.4832
30	5.6336	21.2468	20.2601
15	4.0356	19.4224	19.1934
0	4.3570	20.3941	20.3941
−15	3.5473	21.3124	21.0396
−30	4.5161	23.9333	22.3824
−45	4.0351	21.7139	20.9475

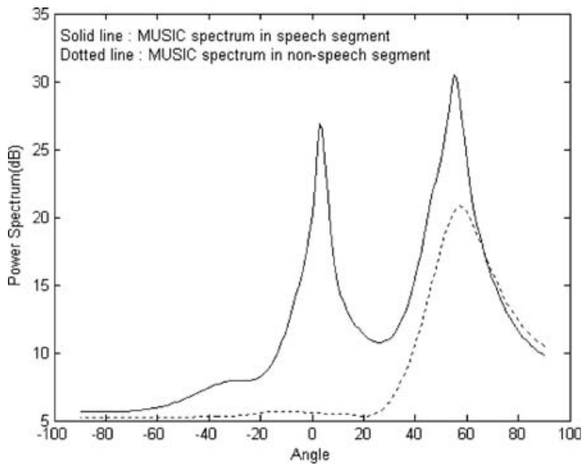


Fig. 5 Customized DOA spectrum

4.1.2 Beamforming result

Tables 2 to 4 show the SNR improvements in the experiments when the filter tap length in the beamformer is 30, 60, and 90. For the modified algorithm, the beam-steer filter's tap length is 4 (section 3.3). The results show a little degradation of the modified algorithm compared with the original one by Dahl and Claesson. However, the modified algorithm only

Table 1 Customized DOA estimation result

Correct angle (deg)	Number of frequencies selected			
	Ten significant frequencies are selected		All frequencies are selected	
	Mean	Standard deviation	Mean	Standard deviation
−45	−43.7619	1.3381	−43.8571	2.1974
−30	−30.2381	2.644	−30.4762	3.0922
−15	−15	2.4698	−14.4762	3.4441
0	2.9524	3.7878	2.6667	5.0133
15	14.8095	2.2939	14.3333	3.3066
30	29.5238	2.9431	29.4286	3.0589
45	43.4762	1.4703	43.0476	2.4388

Table 3 Beamforming result with order 60

Correct angle (deg)	Input SNR (dB)	Original beamformer (dB)	Modified beamformer (dB)
45	5.7539	22.3891	22.0821
30	5.6336	22.3814	21.3591
15	4.0356	20.9760	19.2551
0	4.3570	20.5921	20.5921
−15	3.5473	22.4586	21.5892
−30	4.5161	24.5836	22.4966
−45	4.0351	22.9700	22.0310

Table 4 Beamforming result with order 90

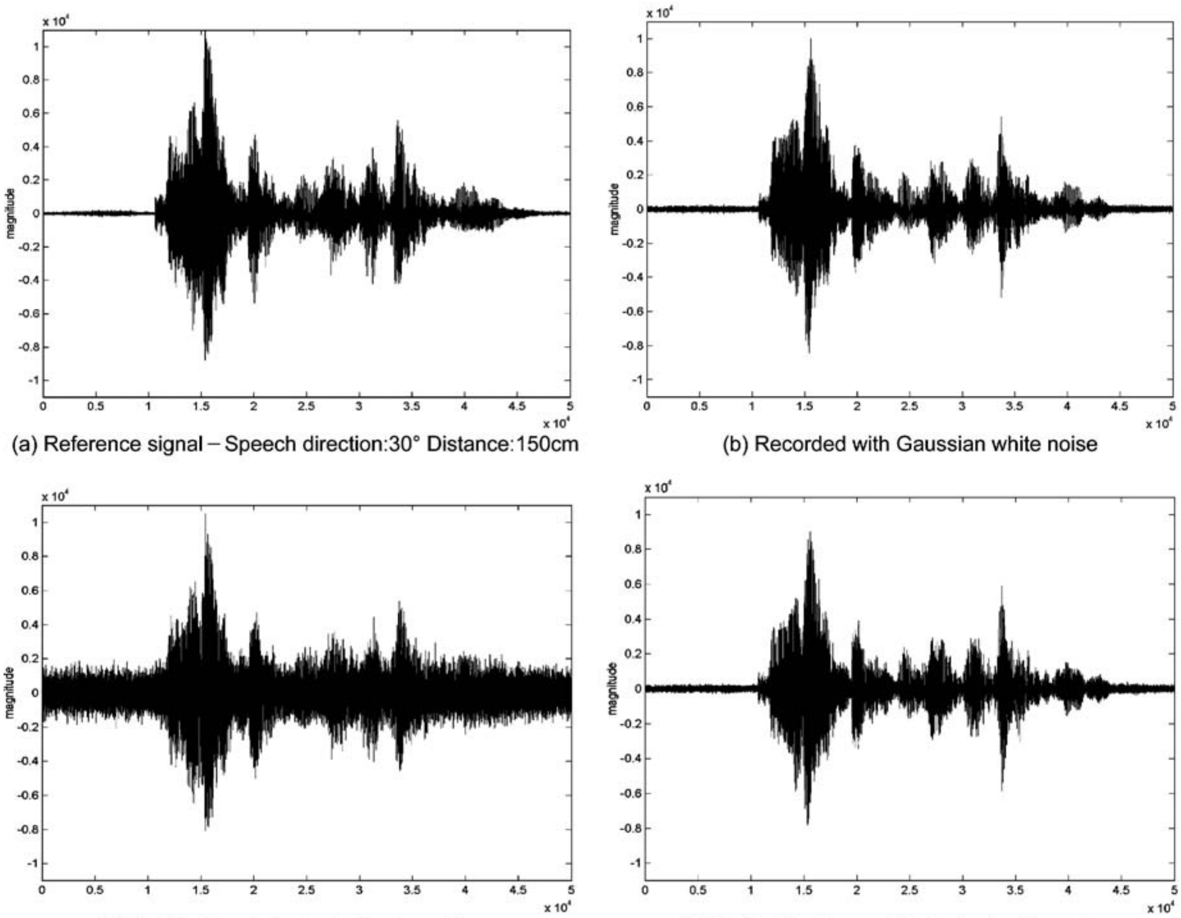
Correct angle (deg)	Input SNR (dB)	Original beamformer (dB)	Modified beamformer (dB)
45	5.7539	21.3245	21.0223
30	5.6336	22.8585	21.4578
15	4.0356	21.9316	19.3706
0	4.3570	21.7993	21.7993
−15	3.5473	23.0127	21.4250
−30	4.5161	25.3235	22.3848
−45	4.0351	22.9967	22.2750

records one set of the source signal at 0°. This shows that with correct DOA information, a simple delay-and-sum beam-steering, can simulate the source signal well in different directions for the adaptive algorithm to be effective. However, this does not mean that the delay-and-sum beam-steering captures the spatial characteristics accurately. In other words, performance may be degraded due to other uncertainties such as misplacement of sensors or mismatch in the delay time. Figure 6 shows the time-domain waveforms of the source signal, the interference, and the enhanced results. In general, the SNR can be enhanced to about 19.2–25 dB from about 3.5–5.7 dB. With the increase of the filter tap length, the SNR is improved, as shown in Fig. 7.

4.2 Scenario 2

4.2.1 DOA result

In this scenario, a speaker source is fixed in one direction with different interference signals from other directions. As shown in Table 5, the standard deviation of the DOA estimation increases with the number of interference signals. This is because



(a) Reference signal – Speech direction:30° Distance:150cm

(b) Recorded with Gaussian white noise

(c) Purified by original adaptive beamformer

(d) Purified by the modified adaptive beamformer

Fig. 6 Waveform of the beamforming result with order 60

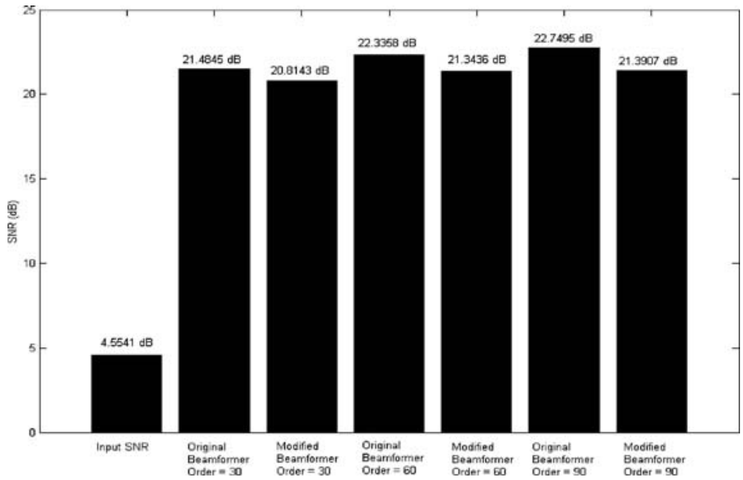


Fig. 7 Average SNR

Table 5 DOA result in scenario 2

Correct angle (deg)	Without interference signals		Interference signals at 60° and −30°		Interference signals at 60°, −30°, and −60°	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
0	1.45	1.3168	2.8	4.1624	2.9	6.5042
30	31.85	1.6944	29.25	5.8658	28	6.8133
−15	−14.7	1.7501	−17.7	4.2932	−18.6	5.0928

increasing the number of interference signals leads to a lower SNR and less degrees of freedom in the noise subspace. Although the estimation accuracy decreases in the complex environment, it still remains in an acceptable range.

4.2.2 Beamforming result

Tables 6 and 7 are the beamforming results with a 60th-order weighting vector applied for each microphone. Compared with Table 3, the modified beamformer still works well by increasing the number of interference signals.

4.3 Improvement of the MFCC error distance

Besides the noise power reduction, another important point that should be considered is whether the cepstrum feature of the reference signal is changed

Table 6 Beamforming result with noise angles of 60° and −30°

Correct angle (deg)	Input SNR (dB)	Original beamformer (dB)	Modified beamformer (dB)
30	2.8234	20.0548	18.9461
0	1.2637	17.3820	17.3820
−15	0.4372	17.9555	16.7834

Table 7 Beamforming result with noise angles of 60°, −30°, and −60°

Correct angle (deg)	Input SNR (dB)	Original beamformer (dB)	Modified beamformer (dB)
30	−0.2980	16.8331	15.7307
0	−1.8639	14.4653	14.4653
−15	−2.6842	14.8471	13.2040

after processing. The purified signal may be used to perform speech recognition in order to understand voice commands for robots. If the feature of recorded speech is changed after processing, the proposed beamformer would not be suitable when speech recognition is required. Because the Mel-frequency cepstral coefficient (MFCC) is the most popular feature for speech recognition, minimizing the cepstral error distance would increase the speech recognition rate. The cepstral error distance is defined as

$$E_c = \sum_{p=1}^{P} \|\boldsymbol{MFCC}_{\mathrm{pure}}(p) - \boldsymbol{MFCC}_{\mathrm{comparison}}(p)\|_2^2 \tag{20}$$

Figure 8 shows the MFCC of one frame. The solid line denotes the MFCC of the pre-recorded speech source in the ideal situation for speech recognition. When the reference signal is recorded in a noisy

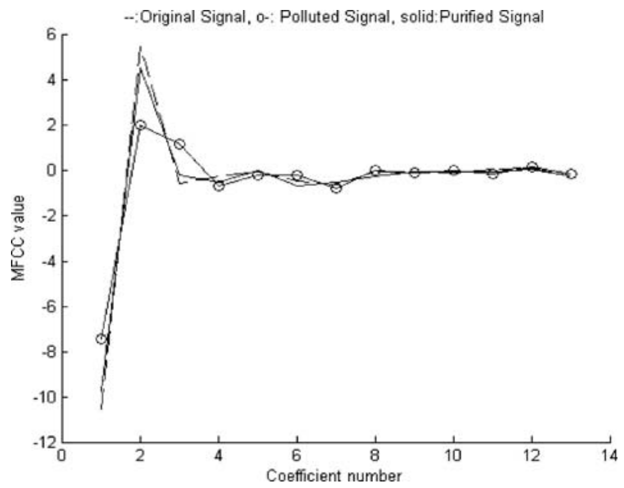


Fig. 8 MFCC distance

environment as scenario 1, the average cepstral error distance increased to 10.699 (–○– line), which means the cepstrum feature of the reference signal is changed by environmental noise and channel distortion. After the contaminated signal is processed by the proposed beamformer, the average cepstral error distance drops to 0.8941 (solid line), which greatly reduces the influence of the interference.

5 CONCLUSION

A microphone array with a customized wide-band eigenstructure-based DOA estimation algorithm and a modified beamformer is proposed in this paper. The experimental result shows that this customized DOA can detect the speaker direction with an acceptable error range. Further, the modified beamformer can also reduce the cepstral distance, overcome the calibration problem caused by the mismatch between microphones, and enhance the SNR. With a beam-steer filter, the request of extra memory needed to form a beam in an arbitrary direction is greatly decreased, and the beam direction is infinite. The modified beamformer is easy to implement and the hardware cost is low compared with other robust beamformers.

REFERENCES

1 **Chun, G. D.** and **Caudell, T. P.** A model for auditory localization in robotic systems based on the neurobiology of the inferior colliculus and analysis of HRTF data. In Proceedings of the International Joint Conference on *Neural Networks* (*IJCNN '01*), 2001, Vol. 2, 15–19 July 2001, pp. 1107–1111.

2 **Schauer, C.** and **Gross, H.-M.** Model and application of a binaural 360 degree sound localization system. In International Joint INNS–IEEE Conference on *Neural Networks*, Washington DC, 14–19 July, 2001.

3 **Frost, O. L.** An algorithm for linear constrained adaptive array processing. *Proc. IEEE*, August 1972, **60**(8), 926–935.

4 **Griffiths, L. J.** and **Jim, C. W.** An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propagation*, January 1982, **AP-30**, 27–34.

5 **Henry, C.** Robust adaptive beamforming. *IEEE Trans. Acoust. Speech, Signal Processing*, October 1987, **ASSP-35**, 1365–1376.

6 **Hoshuyama, O., Sugiyama, A.,** and **Hirano, A.** A robust adaptive beamformer for microphone arrays with blocking matrix using constrained adaptive filters. *IEEE Trans. Signal Processing*, October 1999, **47**(10).

7 **Gannot, S., Burshtein, D.,** and **Weinstein, E.** Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Processing*, August 2001, **49**, 1614–1626.

8 **Dahl, M.** and **Claesson, I.** Acoustic noise and echo cancelling with microphone array. *IEEE Trans. Vehicular Technol.*, September 1999, **48**(5), 1518–1526.

9 **Abdallah, S., Montrésor,** and **Baudry, M.** Speech signal detection in noisy environment using a localentropic criterion. In *Eurospeech*, Rhodes, Greece, September 1997.

10 **Knapp, C. H.** and **Carter, G. C.** The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech, Signal Processing*, August 1976, **ASSP-24(4)**, 320–327.

11 **Brandstein, M. S.** and **Silverman, H. F.** A robust method for speech signal time-delay estimation in reverberant rooms. In *ICASSP-97*, Vol. 1, April 1997.

12 **Hu, J., Su, T. M., Cheng, C. C., Liu, W. H.,** and **Wu, T. I.** A self-calibrated speaker tracking system using both audio and video data. In IEEE Conference on *Control Applications*, September 2002.

13 **Hu, J., Cheng, C. C., Liu, W. H.,** and **Su, T. M.** A speaker tracking system with distance estimation using microphone array. In IEEE/ASME International Conference on *Advanced Manufacturing Technologies and Education*, August 2002.

14 **Schmidt, R. O.** Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation*, **AP-34**, 276–280.

15 **Rao, B. D.** and **Hari, K. V. S.** Performance analysis of root-MUSIC. *Acoust. Speech, Signal Processing*, 1989, **ASSP-37**, 1939–1949.

16 **Junqua, J.-C., Mak, B.,** and **Reaves, B.** A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. Speech and Audio Processing*, July 1994, **2**(3), 406–412.

17 **Gokhale, D. V.** Maximum entropy characterization of some distributions. In *Statistical Distributions in Scientific Work* (Eds Patil, Kotz, and Ord). 1975, Vol. 3, pp. 299–304 (M.A. Reidel, Boston, Massachusetts).

18 **Yu, S. H.** and **Hu. J.** Optimal synthesis of a fractional delay FIR filter in a reproducing kernel Hilbert space. *IEEE Signal Processing Lett.*, June 2001, **8**(6).

APPENDIX

Notation

a_{mk}	amplitude from the kth speech source to the mth microphone
$\mathbf{A}(\omega_l)$	direction matrix in frequency ω_l
$\boldsymbol{A}_i(\omega_l)$	direction vector in frequency ω_l
$c(v)$	undelayed original signal
$\hat{c}(i)$	estimated delay signal
$\mathbf{C}_{xx}(\omega_l)$	source part correlation matrix in frequency ω_l
d	number of sources
D	number of significant frequencies
DOA	direction of arrival
$e[n]$	error signal
E_c	MFCC error distance
$\boldsymbol{E}_1(\omega_l), \ldots, \boldsymbol{E}_M(\omega_l)$	eigenvectors of $\mathbf{R}_{xx}(\omega_l)$
$f(\cdot)$	probability density function
$G = [g(1), \ldots, g(U)]$	speech signal of U symbols
h_v	vth component of the beam-steer filter
$H(.)$	entropy
HRTF	head related transfer function
IID	interaural intensity difference
ITD	interaural time difference
$J(\theta_i)$	cost function for a DOA estimation at θ_i
$K(.)$	sinc function
L	number of frequency components
LMS	least mean square
M	number of microphones
$\boldsymbol{MFCC}_{\text{comparison}}(p)$	MFCC of the polluted signal or the processed signal in the pth frame
$\boldsymbol{MFCC}_{\text{pure}}(p)$	MFCC of the original signal in the pth frame
MUSIC	multiple signals classification
$n_1(t), \ldots, n_M(t)$	non-directional noises from microphone 1 to M in the continuous time domain
$N_1(\omega_l), \ldots, N_M(\omega_l)$	non-directional noises from microphone 1 to M in frequency ω_l
$\boldsymbol{N}(\omega_l)$	non-directional noise vector in the frequency domain
P	frame number of calculated data
$\mathbf{P}_N(\omega_l)$	non-directional noise projection matrix in frequency ω_l
$r_1[n], \ldots, r_M[n]$	pre-recorded speech sources from microphone 1 to M in the discrete time domain
$\hat{r}_1[n], \ldots, \hat{r}_M[n]$	modified reference signals from microphone 1 to M in the discrete time domain
$\hat{\boldsymbol{r}}[k]$	modified reference signal vector at the kth iteration
$\mathbf{R}_{ss}(\omega_l)$	source correlation matrix in frequency ω_l
$R_{s_p s_o}(\omega_l)$	correlation between source p and source o in frequency ω_l
$\mathbf{R}_{xx}(\omega_l)$	received signal correlation matrix in frequency ω_l
$R_{x_i x_j}(\omega_l)$	correlation between received signal i and received signal j in frequency ω_l
$s_1(t), \ldots, s_d(t)$	sources in the continuous time domain
$S_1(\omega_l), \ldots, S_d(\omega_l)$	sources in frequency ω_l
$\boldsymbol{S}(\omega_l)$	source vector in frequency ω_l
SNR	signal-to-noise ratio
T	finite observation interval
U	number of symbols
V	order of the beam-steer filter
$\boldsymbol{w}[k]$	weighting vector at the kth iteration
$x_1[n], \ldots, x_M[n]$	practical received signals from microphone 1 to M in the discrete time domain
$x_1(t), \ldots, x_M(t)$	practical received signals from microphone 1 to M in continuous time domain
$X_1(\omega_l), \ldots, X_M(\omega_l)$	practical received signals from microphone 1 to M in frequency ω_l
$\boldsymbol{X}(\omega_l)$	practical received signal vector in frequency ω_l
$y[n]$	desired signal
$y_b[n]$	output data signal of the upper beamformer
$\hat{y}[n]$	output data signal of the lower beamformer

$\zeta_1[n], \ldots, \zeta_M[n]$	environmental noises from microphone 1 to M in the discrete time domain	τ_{mk}	time delay from the kth speech source to the mth microphone
		ω	frequency value
$\boldsymbol{\zeta}[k]$	environmental noise vector at the kth iteration	ω_c	central frequency
		ω_l	lth frequency component
$\lambda_1(\omega_l), \ldots, \lambda_M(\omega_l)$	eigenvalues of $\mathbf{R}_{xx}(\omega_l)$	$\hat{\omega}_q$	$\boldsymbol{q}$th significant frequency component
μ	step size for the LMS algorithm	$\langle . \rangle_q$	$\boldsymbol{q}$th biggest values