22

Automatic Speech Recognition of Human-Symbiotic Robot EMIEW

Masahito Togami, Yasunari Obuchi, and Akio Amano
Central Research Laboratory, Hitachi Ltd.
Japan

1. Introduction

Automatic Speech Recognition (ASR) is an essential function of robots which live in the human world. Many works for ASR have been done for a long time. As a result, computers can recognize human speech well under silent environments. However, accuracy of ASR is greatly degraded under noisy environments. Therefore, noise reduction techniques for ASR are strongly desired.

Many approaches based on spectral subtraction or Wiener filter have been studied. These approaches can reduce stationary noises such as fan-noise, but cannot reduce non-stationary noise such as human-speech.

In this chapter, we propose a novel noise reduction technique using a microphone-array. A microphone-array consists of more than one microphone. By using a microphone-array, robots can obtain information about sound sources' direction. When directions of noise sources and the desired source are different from each other, even if noises are non-stationary, noises can be reduced by spatial filtering with a microphone array. In this chapter, a new estimation method of direction of sources, MDSBF (modified delay and sum beam-former), is proposed. Then spatial filtering method using MDSBF named SB-MVBF (Sparseness Based Minimum Variance Beam-Former) is proposed.

The proposed noise reduction technique is implemented in a human-symbiotic prototype robot named EMIEW (Excellent Mobility and Interactive Existence as Workmate). It is shown that ASR technique with SB-MVBF is more accurate than ASR technique with the conventional method (MVBF) under noisy environments.

2. Human Symbiotic Robot EMIEW

Human symbiotic robot EMIEW (Excellent Mobility and Interactive Existence as Workmate) has been developed since 2004 by Hitachi Ltd (Hosoda et al., 2006).

EMIEW was designed as an assistant and a co-worker of human. Appearance of EMIEW is shown in Fig.1. When conventional robots live with human, one of major problems is lack of mobility. People can walk at about a few km/h, but robots before EMIEW cannot walk so rapidly. Maximum speed of him is about 6 km/h : the speed of a rapidly walking person. Therefore EMIEW can walk with human. Furthermore, it can avoid obstacles, so can move safely.

Source: Human-Robot Interaction, Book edited by Nilanjan Sarkar,
ISBN 978-3-902613-13-4, pp.522, September 2007, Itech Education and Publishing, Vienna, Austria

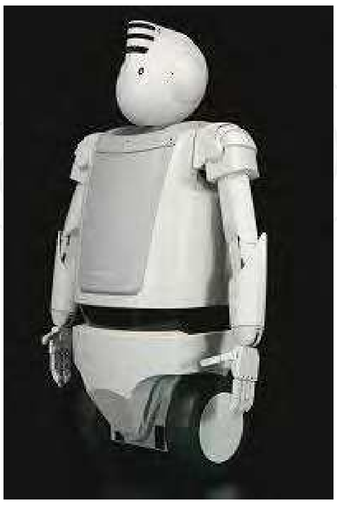

Figure 1. Appearance of EMIEW: body height is about 130cm. maximum speed is about 6 km/h. EMIEW has 8 microphones around his ears and neck. EMIEW is able to communicate with people even under noisy environment

For human symbiotic robots, in addition to mobility, communication capability is also important. It is desirable that robots can communicate with human in natural languages. EMIEW has speech synthesis function (Kitahara, 2006) and it can recognize human speech. In the future, EMIEW will work under noisy environments such as train-stations, airports, streets, and so on. Therefore, it is necessary that EMIEW can talk with humans under such environments.

We developed automatic speech recognition technology under noisy environments. We demonstrated this technology at EXPO 2005 AICHI JAPAN. Noise level of demonstration areas was from 70 db(A) to 80 db(A). It was verified that EMIEW can talk with guests at such environments.

3. Noise Reduction Technique for ASR

Automatic speech recognition (ASR) is a computational technology, which recognizes human speech which is recorded by microphones using pre-learned acoustic model.

Recognition performance of ASR is high for speeches which are recorded under noise-less and anechoic rooms. However, it is known that recognition performance of ASR is greatly degraded when human speech is convolved with noise or reverberation. Therefore, conventionally, noise reduction techniques have been studied (Boll, 1979) (Frost, 1972) (Aoki et al., 2001) (Hoshuyama et al., 1999). Microphone input signal is expressed as follows:

$$x(t) = s(t) + n(t) \tag{1}$$

, where t is the time-index, x(t) is the microphone input signal, s(t) is desired source signal, n(t) is the noise signal. Spectrum of speech signal is known to be stationary for a few dozen milliseconds. Therefore, many noise reduction approaches convert time domain expression to time-frequency domain expression by using short time Fourier transform as follows:

$$x(f,\tau) = s(f,\tau) + n(f,\tau) \tag{2}$$

,where f is the frequency index, τ is the frame index. When speech and noise is uncorrelated, power spectral of input signal is represented as follows:

$$E[|x(f,\tau)|^2] = E[|s(f,\tau)|^2] + E[|n(f,\tau)|^2]. \tag{3}$$

Spectral Subtraction (SS) (Boll, 1979) is the major noise reduction technique. SS subtracts time-averaged noise power spectral as follows:

$$\hat{s}(f,\tau) = \sqrt{|x(f,\tau)|^2 - \frac{1}{L}\sum_{\tau}|n(f,\tau)|^2}\,\frac{x(f,\tau)}{|x(f,\tau)|} \tag{4}$$

,where L is the number of averaged noise power spectral, $\hat{s}(f,\tau)$ is the output signal of SS. SS can reduce spectral-stationary noise such as fan-noise, but when noise is non-stationary (such as speech like noise), noise component cannot be reduced. To make things worse, in this case, the output signal of SS is greatly degraded by musical noise compared to the original speech. For this problem, noise reduction approaches using multi microphone elements (microphone array) have been widely studied. Direction of arrival (DOA) of sources can be estimated with a microphone array. One-channel noise reduction approaches such as SS cannot use DOA information. If DOA of noise and desired source are different from each other and DOA of desired source is given, even when noise is non-stationary, noise component can be reduced by spatially ``NULL'' beam-former such as MVBF (Minimum Variance Beam-Former) (Frost, 1972). However, when the given DOA of desired source is not accurate, the desired source is reduced or degraded. This problem is called signal cancellation problem.

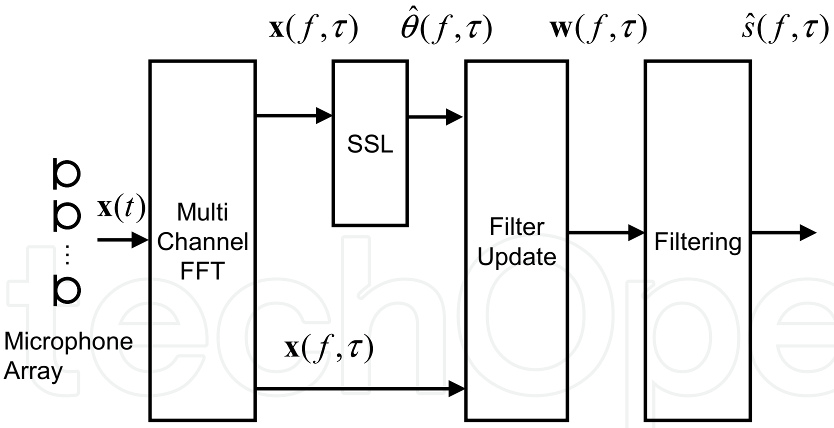

Figure 2. The noise-reduction block diagram of the proposed method at each frame

We will propose a novel noise reduction approach based on source's sparseness named SB-MVBF (Sparseness Based Minimum Variance Beam-Former). To solve signal cancellation problem, the spatial ``NULL'' beam-former is updated only when DOA of multi channel input signal is far from the given DOA of the desired source. The noise-reduction block diagram of the proposed method is shown in Fig. 2. Multi channel input signals of a microphone array are transformed to frequency domain signals by FFT. DOA of input signal at each time-frequency point is estimated by Sound Source Localization (SSL). Filters for noise reduction are updated in ``Filter Update” block. Only when DOA of input signal is far from DOA of desired source, filters are updated in ``Filter Update” Block. Finally, noise is filtered by updated filters in ``Filtering” block.

In the following sections, we explain the proposed sound source localization method of each frequency component: MDSBF (Modified Delay and Sum Beam-Former) and adaptation method of noise reduction filter based on MDSBF: SB-MVBF (Sparseness Based Minimum Variance Beam-Former) and automatic speech recognition (ASR) based on SB-MVBF are shown.

3.1 Modified Delay and Sum Beam-Former (MDSBF)

Let M be the number of microphones, and $x_i(f,\tau)$ be the input signal of the i-th microphone at frame τ and frequency f. Sound source localization localizes direction of arrival of sources by the multi-channel input vector $\mathbf{X}(f,\tau) = [x_1(f,\tau), x_2(f,\tau), \ldots, x_M(f,\tau)]$. From now on, the suffix (f,τ) is omitted.

For simplicity, we assume that there is only one source at each time frequency point. In this case, the multi-channel input vector $\mathbf{X}$ is expressed as the following equation.

$$\mathbf{X} = \mathbf{a}(r,\theta)s \tag{5}$$

, where s is the source signal, r is distance between the source and the microphones, and θ is source’s direction. The variable s is independent from the microphone index. The vector $\mathbf{a}(r,\theta) = [a_1, \ldots, a_M]$ is called steering vector. Each element is calculated as follows:

$$a_i = A_i e^{-2\pi f\rho} \tag{6}$$

, where A_i is the attenuation coefficient from the source position to the i-th microphone position, and ρ is time delay from the source position to the i-th microphone position. When microphones are sufficiently distant from the source position, A_i is independent from the microphone index, and it only depends on distance between the source and the microphones. Time-delay ρ is calculated as follows:

$$\rho = \frac{r}{c} + \lambda_i(\theta) \tag{7}$$

,where r is distance between the source and microphones, the term $\lambda_i(\theta)$ depends on source’s direction and the microphone index, but it is independent from distance between the source and microphones.

Based on equation (5), we obtain the following inequality.

$$\frac{|\mathbf{a}(r,\theta)^* \mathbf{X}|}{|\mathbf{a}(r,\theta)||\mathbf{a}(r,\theta)|} = |\mathbf{X}| \geq \frac{|\mathbf{a}(\tilde{r},\tilde{\theta})^* \mathbf{X}|}{|\mathbf{a}(\tilde{r},\tilde{\theta})||\mathbf{a}(r,\theta)|} = |\mathbf{X}| \frac{|\mathbf{a}(\tilde{r},\tilde{\theta})^* \mathbf{a}(r,\theta)|}{|\mathbf{a}(\tilde{r},\tilde{\theta})||\mathbf{a}(r,\theta)|} \tag{8}$$

Therefore, source’s distance and direction are estimated as follows:

$$\hat{r}, \hat{\theta} = \underset{\tilde{r}, \tilde{\theta}}{\arg\max} \left| \mathbf{a}(\tilde{r}, \tilde{\theta})^{*} \mathbf{X} \right| \tag{9}$$

,where l2-norm of $\mathbf{a}(\tilde{r}, \tilde{\theta})$ is normalized to 1. By using equation (7), the vector $\mathbf{a}(\tilde{r}, \tilde{\theta})$ is expressed as follows:

$$\mathbf{a}(\tilde{r}, \tilde{\theta}) = e^{-2\pi f \frac{r}{c}} \hat{\mathbf{a}}(\tilde{\theta}) \tag{10}$$

,where $\hat{\mathbf{a}}(\theta) = [e^{-2\pi f \lambda_0(\theta)}, \ldots, e^{-2\pi f \lambda_i(\theta)}, \ldots, e^{-2\pi f \lambda_M(\theta)}]$. Therefore, equation (9) can be transformed as follows:

$$\hat{\theta} = \underset{\tilde{\theta}}{\arg\max} \left| \hat{\mathbf{a}}(\tilde{\theta})^{*} \mathbf{X} \right|. \tag{11}$$

Therefore, in this case, we can estimate only source’s direction.

When there are more than one source at the same time frequency point, we cannot obtain sources’ direction of arrival by Equation 11. However, it is known that speech is sparse signal in the time-frequency domain and few frequency components of multiple sources have big power at the same time point (Aoki et al., 2001). Therefore, it is considered to be the rare case that multiple sources are mixed in the same time-frequency point. Based on this sparseness assumption, we can estimate source’s direction of arrival (DOA) at each time-frequency point by equation 11.

When sources are sparse, DOA of the input vector at each time-frequency point corresponds to the true source’s angle, but this angle is variable with respect to each time-frequency point. Therefore, multiple sources’ DOA is obtained by peak-searching in the histogram of the estimated DOA at all time-frequency points.

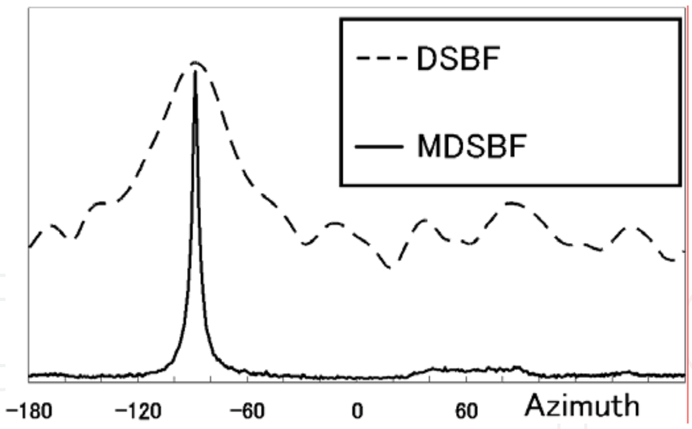

Figure 3. The DOA histogram at one source case: azimuth of the source is about -90 degree. Both DSBF and MDSBF succeeded to localize the source’s DOA. However, the peak of DOA histogram by MDSBF is sharper than that by DSBF

Experimental results of DOA histogram at one source case (in Fig. 3) and two sources case (in Fig. 4) are shown. Reverberation time is about 300 milliseconds. Comparison to conventional delay and sum beam-former (DSBF) is shown. DOA histogram by DSBF has only one peak in the two source case. However, DOA histogram by MDSBF has two peaks in the two source case.

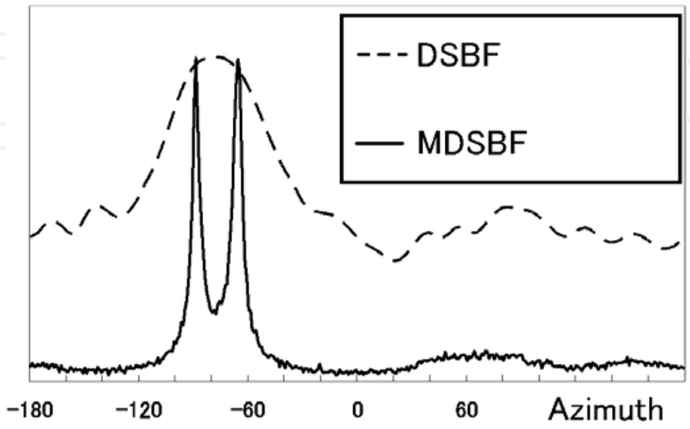

Figure 4. The DOA histogram at two source case: azimuth of the sources are about -90 degree and -70degree. MDSBF succeeded localization of sources' DOA. However, DSBF failed localize sources' DOA.

Success probability of DOA estimation by MDSBF was also checked. When there are only one source (ratio of one source (S1) 's power to the other source(S2) 's power is more than 30db), successful DOA estimation of MDSBF was 79%.

3.2 A Novel Adaptation Method : SB-MVBF

When steering vectors of desired source and noise are given, filtering process is simply expressed in Fig. 5.

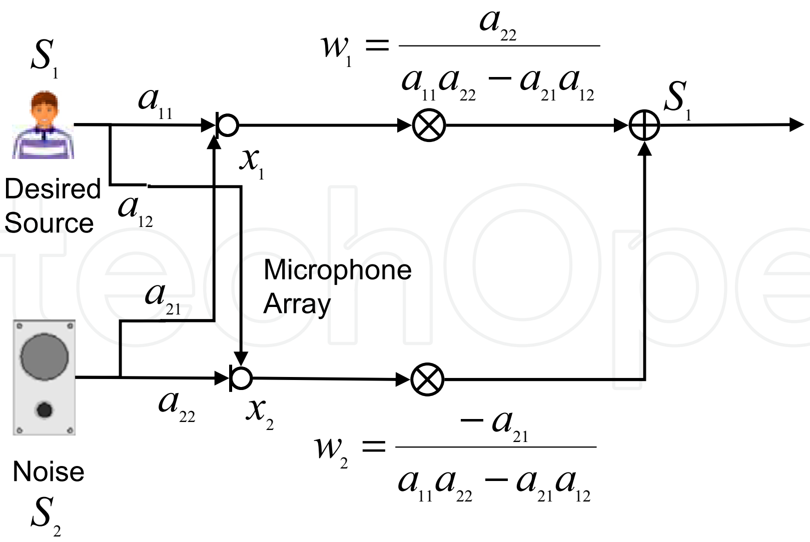
S_1
$w_1 = \frac{a_{22}}{a_{11}a_{22} - a_{21}a_{12}}$
S_1
a_{11}
x_1
Desired Source
a_{12}
Microphone Array
a_{21}
a_{22}
x_2
$w_2 = \frac{-a_{21}}{a_{11}a_{22} - a_{21}a_{12}}$
Noise
S_2

Figure 5. Filtering process under the condition that steering vectors are known

However, at least the steering vector of noise is not given and it is time-variable. Therefore, the steering vector of noise needs to be estimated at the beginning and updated when DOA of noise is changed. By Minimum Variance Beam-Former (MVBF) (Frost, 1972) , even when the steering vector of noise is unknown, noise reduction filter can be obtained as follows:

$$w = \frac{aR^{-1}}{aR^{-1}a^{*}} \tag{12}$$

,where R is the correlation matrix of the input vector x and is defined as follows:

$$R = E\left[xx^{*}\right]. \tag{13}$$

MVBF needs only desired steering vector a and correlation matrix of input vectors. The desired steering vector a can be calculated when DOA of desired source is given.

The filter w passes sources whose steering vector completely matches with given desired steering vector a, and reduce sources whose steering vector are different from a.

However, even when DOA of desired source is given based on prior knowledge such as ``desired speaker is in front of the robot'', actually the location of the speaker is different from given DOA. Additionally, given steering vector a is different from the actual steering vector because of reverberation.

Therefore, in this ``steering vector mismatch case”, the filter made by MVBF cancels desired source (signal cancellation problem). This signal cancellation problem frequently occurs when the biggest component in correlation matrix R corresponds to desired signal. When the biggest component in correlation matrix R corresponds to noise signal, this signal cancellation problem does not occur. Therefore, correlation matrix R needs to be updated when desired source is absent. However, DOA of noise is time-variable. Therefore, correlation matrix R needs to be always updated.

To fill these requirements, time-variable coefficient to update correlation matrix R is proposed. Conventional MVBF updates correlation matrix R as follows:

$$\mathbf{R}_{\tau+1} = \beta\mathbf{R}_{\tau} + (1-\beta)\mathbf{x}_{\tau}\mathbf{x}_{\tau}^{*}. \tag{14}$$

Correlation matrix R is updated with the time-invariable coefficient. However, when the biggest source in the input vector is desired source, updating the correlation matrix R is unfavorable. Therefore, proposed SB-MVBF (Sparseness Based Minimum Variance Beam-Former) uses the time-variable coefficient to update correlation matrix R as follows:

$$\mathbf{R}_{\tau+1} = \alpha(\tau)\mathbf{R}_{\tau} + (1-\alpha(\tau))\mathbf{x}_{\tau}\mathbf{x}_{\tau}^{*} \tag{15}$$

,where correlation matrix R is updated with the time-variable coefficient $\alpha(\tau)$. This coefficient set to be 1 when desired source is likely to be inactive , and set to be 0 when desired source is likely to be active.

Estimation of desired source’s status (active/inactive) is done by results of sound source localization. Proposed sound source localization MDSBF can accurately estimate DOA of sources at each time- frequency point. Therefore, when estimated DOA of one time-frequency point is far from DOA of desired source, it is likely that desired source is inactive in this time-frequency point.

SB-MVBF sets the time-variable coefficient to be 1 when estimated DOA by MDSBF is far from DOA of desired source and set it to be 0 when estimated DOA by MDSBF is close to DOA of desired source. An example of separated signal at an room (reverberation time is 300 ms) is shown in Fig. 6.

3.3 Evaluation of ASR Under Noisy Environment

SB-MVBF reduces noise. However, residual noise exists in noise reduced signal, performance of automatic speech recognition is degraded when the acoustic model of ASR is made by clean speeches. Therefore the acoustic model is adapted by speeches convolved with remained noise. In this experiment, the acoustic model is adapted to noise reduced signals by the proposed method.

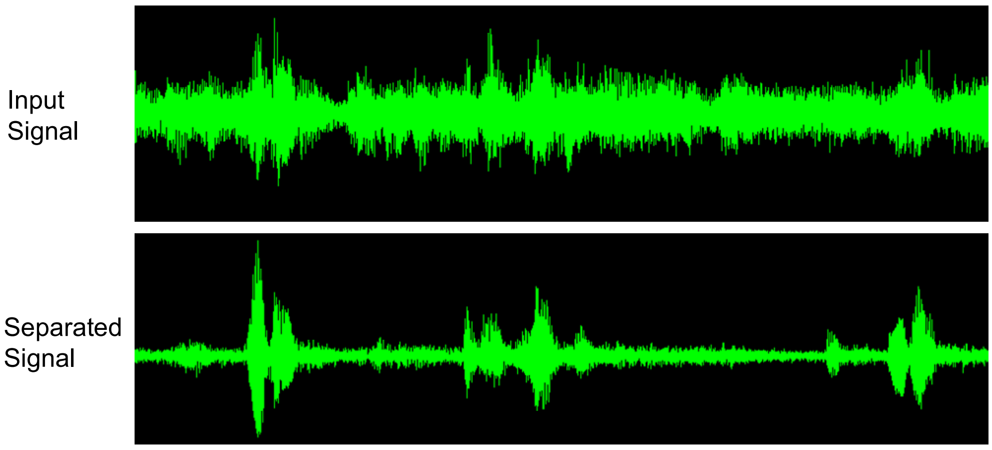

Figure 6. Input signal and separated signal: separated signal has less noisy than input signal

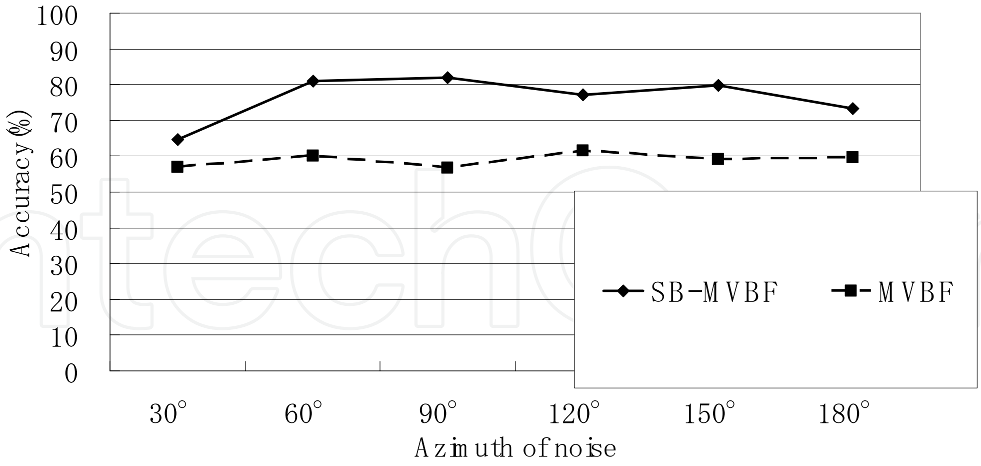

Figure 7. experimental results of ASR: Accuracy of ASR with Proposed method (SB-MVBF) is about from 10% to 20% higher than that with conventional MVBF

Acoustic features are 14-order LPC cepstrum ,14-order delta-LPC cepstrum and 1-oder delta power. Total dimension of features is 29. This experiment of ASR was done under 5db SNR (Signal to Noise Ratio) condition. The recognition vocabulary consists of 10-digits. The number of speakers is 80. The experimental result is shown in Fig. 7. In this experiment, desired source is in front of the microphone array (azimuth=0 degree). Direction of noise is variable from 30 degree to 180 degree.

3.4 Demonstration at EXPO 2005 AICHI JAPAN

Appearance of demonstration at EXPO 2005 AICHI JAPAN is shown in Fig. 8. EMIEW recognized guests' order under noisy environment (noise level= from 70 db(A) to 80 db(A)).

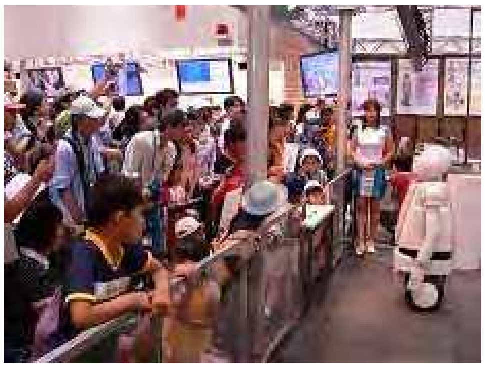

Figure 8. demonstration at EXPO 2005 AICHI JAPAN: EMIEW recognized guests' speech under noisy environment

5. Conclusion

We explained noise reduction technique and automatic speech recognition (ASR) under noisy environments. Human symbiotic robot EMIEW succeeded recognition under noisy environment at EXPO 2005 AICHI JAPAN.

For high accuracy of ASR under noisy environment, noise reduction technique is necessary. In this chapter, robust noise reduction technique with a microphone array was proposed. Proposed Modified Delay and Sum Beam-Former (MDSBF) can localize sources more accurately than conventional Delay and Sum Beam-Former (DSBF) . A novel adaptation method of Minimum Variance Beam-Former (MVBF) with time-variant coefficient (SB-MVBF) is proposed. Performance of ASR with proposed method was shown to be higher than conventional MVBF.

6. Acknowledgment

This work was done through entrustment by the New Energy and Industrial Technology Development Organization (NEDO).

7. References

Hosoda, Y.; Egawa, S. Tamamoto, J. Yamamoto, K. Nakamura, R. & Togami, M. (2006). Basic design of human-symbiotic robot EMIEW, *Proceedings of IROS 2006*, pp. 5079-5084

Kitahara, Y. (2006). Development of High Quality and Intelligent Speech Synthesis Technology. *Hitachi Review* , Vol.88, No. 06, pp. 60-65 (in Japanese)

Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. ASSP*, Vol.27, No.2, pp. 113-120

Togami, M.; Sumiyoshi, T. & Amano, A. (2006). Sound source separation of overcomplete convolutive mixtures using generalized sparseness, *CD-ROM Proceedings of IWAENC2006*

Frost, III, O.L. (1972). An algorithm for linearly constrained adaptive array processing, *Proceedings IEEE* Vol.60, No.8, pp. 926-935.

Griffith, L.J. & Jim, C.W. (1982). An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Anntenas Propagation*, Vol.30, i.1, pp. 27-34

Aoki, M. ; Okamoto, M., Aoki, S., Matsui, H., Sakurai, T. & Kaneda, Y. (2001). Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, *Acoust.Sci & Tech.* Vol.22, No.2, pp. 149-157

Hoshuyama, O. ; Sugiyama, A. & Hirano, A. (1999). A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, *IEEE Trans. Signal Processing* , Vol.47, No.10, pp.2677-2684

Human Robot Interaction
Edited by Nilanjan Sarkar

ISBN 978-3-902613-13-4
Hard cover, 522 pages
Publisher I-Tech Education and Publishing
Published online 01, September, 2007
Published in print edition September, 2007

Human-robot interaction research is diverse and covers a wide range of topics. All aspects of human factors and robotics are within the purview of HRI research so far as they provide insight into how to improve our understanding in developing effective tools, protocols, and systems to enhance HRI. For example, a significant research effort is being devoted to designing human-robot interface that makes it easier for the people to interact with robots. HRI is an extremely active research field where new and important work is being published at a fast pace. It is neither possible nor is it our intention to cover every important work in this important research field in one volume. However, we believe that HRI as a research field has matured enough to merit a compilation of the outstanding work in the field in the form of a book. This book, which presents outstanding work from the leading HRI researchers covering a wide spectrum of topics, is an effort to capture and present some of the important contributions in HRI in one volume. We hope that this book will benefit both experts and novice and provide a thorough understanding of the exciting field of HRI.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821