Noisy Environment-Aware Speech Enhancement for Speech Recognition in Human-Robot Interaction Application

Sheng-Chieh Lee
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
leesc@icwang.ee.ncku.edu.tw

Bo-Wei Chen
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
chenbw@icwang.ee.ncku.edu.tw

Jhing-Fa Wang
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
wangjf@mail.ncku.edu.tw

***Abstract*—In this study, we introduce a noisy environment-aware speech enhancement system, which can be used in human-robot interaction (HRI) application for command recognition. In order to effectively filter different noises and improve speech recognition rates, the proposed system adopts automatic noise cancellation that is combined with independent component analysis (ICA) and subspace speech enhancement (SSE). Furthermore, it can automatically decide when to use noise reduction according to SNRs of the detected noisy speeches at any time (using proposed noisy environment-aware determination). The experimental results show that our proposed system is suitable for various types of noisy environments, and it is capable of improving the speech quality for recognition. Our proposed system can enhance SNRs by about 20dB, which is higher than those of original noisy speeches.**

***Keywords*—Aware computing, noise reduction, human-robot interaction, blind source separation, subspace speech enhancement, microphone array**

I. INTRODUCTION

To date, a great deal of research has been proposed to deal with noises in speech signal processing fields, especially in the preprocessing stage of speech recognition. Although noise reduction can improve recognition performance, however, in some cases, noise signals are not so obvious in environments. Therefore, noise cancellation is not applicable in certain applications, and it could cause distortion to speech signals. On the other hand, noise signals still may occur in one place from time to time. In this circumstance, a recognition system must have a certain mechanism to handle with various noises.

In this study, we present a speech enhancement system, which is installed in a human-robotic interaction (HRI). It is capable of deciding whether the received speech signal should be enhanced or not when we perform speech recognition module. In our HRI system, a new scheme "noisy environment-aware determination" that can automatic choose noise thresholds is introduced. Furthermore, we also integrate independent component analysis (ICA) with subspace speech enhancement (SSE), improving speech quality.

The remainder of this study is organized as follows: Section II reviews the previous work of noise reduction. In Section III, we present the details of the proposed system. The experimental results are described in Section IV. Section V summarizes the conclusions of this study.

II. PREVIOUS WORK

In the following descriptions, we give a brief introduction to the previous work that is used in noisy environments.

A. Independent Component Analysis

ICA is often used in blind signal separation (BSS) [1-2]. Under the assumptions of ICA, each of source signals should be independent of each other. By using ICA, each original source signal can be separated from mixed signal. The equation of BSS is shown as follows, where x_1 and x_2 are the received mixed signals; s_1 and s_2 are original source signals; **A** is an unknown mixing matrix.

$$\begin{aligned}\mathbf{X}&=\begin{bmatrix}x_1\\x_2\end{bmatrix}\\&=\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}\begin{bmatrix}s_1\\s_2\end{bmatrix}\\&=\mathbf{AS}\end{aligned} \tag{1}$$

In order to separate the source signal from the mixed one, it is required to calculate a de-mixing matrix to obtain the primary separated signal. The equation of de-mixing matrix is shown below, where y_1 and y_2 are the separated signals, and matrix **W** is the de-mixing matrix. According to (1) and (2), the separated signals should similar to the original source ones.

$$\begin{aligned}\mathbf{Y}&=\begin{bmatrix}y_1\\y_2\end{bmatrix}\\&=\begin{bmatrix}w_{11}&w_{12}\\w_{21}&w_{22}\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}\\&=\mathbf{WX}\end{aligned} \tag{2}$$

According to the central limit theorem, the non-Gaussian characteristic of an individual non-Gaussian distribution signal is larger than that of a mixing signal, which is composed of several non-Gaussian distribution signals. Hence, the classical

978-1-4244-6588-0/10/$25.00 ©2010 IEEE

ICA exploited high order statistics and information theory to measure non-Gaussianity, and then used optimal algorithms to calculate a de-mixing matrix. In order to accelerate computation of ICA, a novel algorithm named Fast ICA is proposed in [3]. Fast ICA adopted the negentropy to estimate the non-Gaussian characteristic of signal. The negentropy formula is defined as follows, where y is the estimated signal and y_{Gauss} is the Gaussian distribution signal with the same covariance matrix of signal y; $H(\cdot)$ is the entropy calculation.

$$J(y) = H(y_{\text{Gauss}}) - H(y) \tag{3}$$

When the estimated signal is Gaussian distribution, the value of negentropy becomes zero. After we estimate the non-Gaussian characteristic of a signal, the de-mixing matrix can be obtained by iteration.

Although ICA can separate noises from received signals, it is still not enough when we deal with noisy environments. Therefore, we combine another technique, which is called subspace speech enhancement (SSE) [4] to improve noise reduction performance.

B. Subspace Speech Enhancement

According to the theory of SSE, the observed signal vectors can be divided into two vectors, of which the first one consists of clean speech, and the other one is composed of noise signal. Nevertheless, there are still few residual noises in the subspace vector of clean speeches, such as the white noise, which exists in each frequency band.

In order to remove the noise element from the subspace of clean speech, it is necessary to use a filter to obtain the signal, which has no noise. The estimation can be estimated by the subtraction of the original signal and the filtered signal, which is shown in (4), where F means the filter, y is the original speech signal, n is the noise signal, and $\mathbf{I}$ is an identity matrix. Equation (4) can be also expressed by two parameters δ_y and δ_n. δ_y is the speech distortion, which is caused by the filter processing, and δ_n is the residual noise in the filtered signal.

$$\delta = F\cdot(y+n) - y = (F-\mathbf{I})\cdot y + F\cdot n = \delta_y + \delta_n \tag{4}$$

We calculate the variance of these two parameters and take these variance ones as the criteria of error estimation. Finally, we optimize filter processing according to the measurement defined in (4).

There are two major concerns when we optimize the filter in (4). The first one is that the degree of speech distortion should be as minimum as possible, because it may decrease speech recognition rates. The second one is that the residual noise should be suppressed so that the recognition result will not be influenced by noise. These two optimal conditions of the filter are shown in (5), where F means objective filter, $\bar{\delta}_y$ and $\bar{\delta}_n$ are the variance of the speech distortion and residual noise, σ^2 is the variance of noise, and γ is the controlled parameter. According to (5), we can obtain an optimal filter by Lagrange multiplier method.

$$\min_{\mathrm{F}} \ \bar{\delta}_y \quad \text{subject to: } \bar{\delta}_n \le \gamma\sigma^2, 0 \le \gamma \le 1 \tag{5}$$

III. PROPOSED METHOD

The proposed system is shown in Fig. 1, which includes three stages: The first one is noisy environment-aware determination processing, the second one is noise reduction processing, and the third one is speech recognition processing. The proposed system is set on an HRI humanoid robot, which is described in Section IV later.

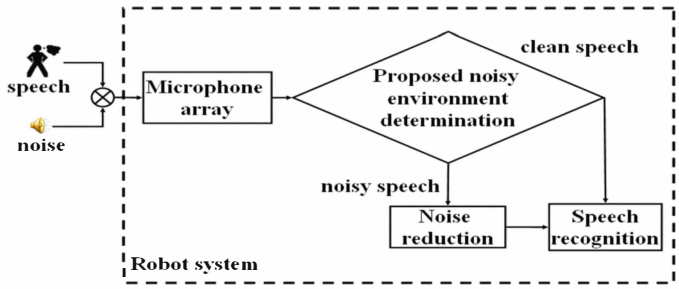

Figure 1. Proposed noisy environment-aware speech recognition system

In the noisy environment-aware determination processing, we use a microphone array, which is replaced on the humanoid robot, to receive the noisy speech signal and calculate the SNR value. If the SNR value is equal or smaller than a threshold, then the received signal is passed to the noise reduction procedure. Otherwise, it is immediately passed to speech recognition processing.

A. Noisy Environment-Aware Determination

The proposed noisy environment-aware procedure is shown in Fig. 2. In the noisy environment-aware determination, we use SNRs to measure the signal power ratio between speeches and noises. Environmental noise is assumed to be stationary and continually occurred. While the system operates, we use the microphone array to record environmental noise signals and calculate the signal power value in advance. Then, the system continues to record speakers' voices.

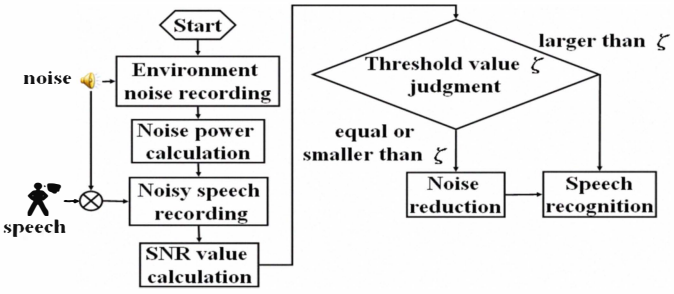

Figure 2. Proposed noisy environment-aware precedure

Because we suppose that environmental noise is continuous and stationary, the noisy speech can be taken as the summation of the clean speech and noise signal (see (6)), where $x[t]$ means the noisy signal, $s[t]$ and $n[t]$ are the clean speech and noise, and L is the length of the noisy speech.

$$x[t] = \sum_{t=0}^{L-1} (s[t] + n[t]) \tag{6}$$

SNRs can be calculated by (7), where σ_s^2 and σ_n^2 are the spectral power of the speech and noise, and parameter ζ is a threshold of noise reduction. If the SNR is equal or smaller than ζ, it means the energy of noise is so obvious that this noisy speech should be enhanced. On the other hand, when the SNR is larger than ζ, it implies that noise is not quite apparent and can be passed to the speech recognition stage.

$$SNR = 10\log_{10}\frac{\sigma_s^2}{\sigma_n^2} = 10\log_{10}\frac{\sigma_x^2 - \sigma_n^2}{\sigma_n^2} \tag{7}$$

B. *Noise Reduction Processing*

In Section II, we have described a noise cancellation method that integrates ICA and SSE. At this stage, we introduce noise reduction preprocessing, which is used before ICA and SSE (see details in Fig. 3).

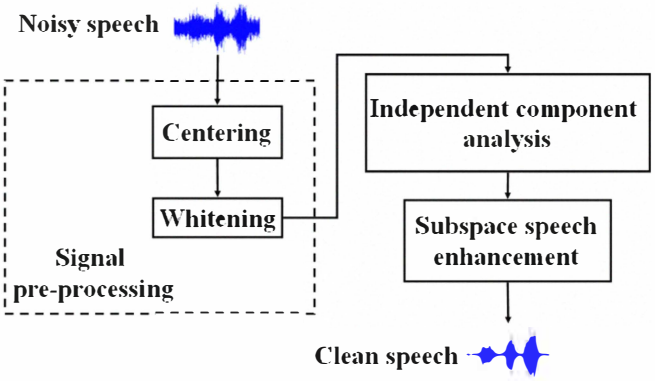

Figure 3. Noise reduction procedure

When running noise cancellation with ICA, we have to make sure that source signals are independent of each other. Accordingly, we employ signal centering and whitening to solve this problem. The purpose of preprocessing is to transfer dependent source signals into uncorrelated ones.

The centering process is shown in (8), where x is a mixed signal, and $E\{x\}$ is a mean value of the mixed signal.

$$\bar{x} = x - E\{x\} \tag{8}$$

As for the whitening process, its purpose is to estimate a whitening matrix so that the mixed signal multiplied by whitening matrix can become uncorrelated, and the covariance matrix of the signal can be an identity matrix. The whitening calculation is defined as (9), where **W** is a whitening matrix, $E\{\hat{x}\hat{x}^T\}$ is a covariance matrix of signal $\hat{x}$, and **I** is an identity matrix.

$$\begin{aligned} \hat{x} &= \mathbf{W} \cdot \bar{x} \\ E\{\hat{x}\hat{x}^T\} &= \mathbf{I} \end{aligned} \tag{9}$$

C. *Speech Recognition Processing*

At the speech recognition processing stage, we simply use the hidden Markov model toolkit (HTK) as a speech recognizer [5].

We have trained several acoustic models for recognizing speech commands. When a speech signal is transferred into the recognizer, it estimates the speech features and compares with speech content according to a lexicon, grammar rules, and the acoustic models. After the comparison, it takes the closest speech content as the recognition result.

IV. Experiments Results

A. *Experimental Setup*

In order to setup the experimental environment, firstly, we embed our proposed system in an interactive robot (a 16-DOF humanoid robot, which is made by Innovati company (www.innovati.com.tw)). The illustration of this humanoid robot is shown in Fig. 4.

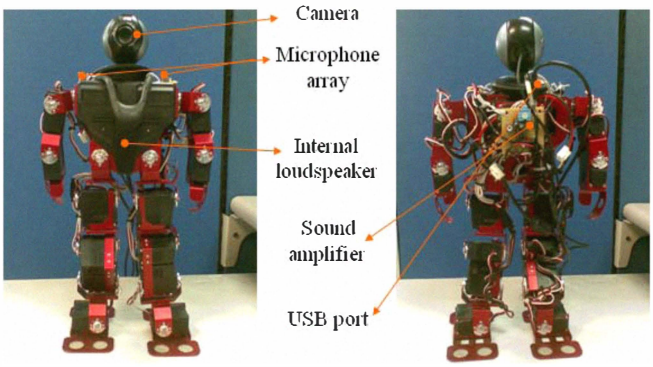

Figure 4. Illustration of the humanoid robot

We utilize a microphone array of two omni-directional microphones (placed on the shoulder of the robot), and its spacing is 10 cm. In addition, a sound amplifier circuit is also attached to the robot in order to increase sound volume.

While evaluating noise reduction performance, we utilize the noisex-92 database [6] as the testing noise signals. There are five types of noises: White, pink, babble, car, and factory noises. Ten speeches of three different speakers (two males and one female) are recorded as the source signals. Totally, there are 10 × 3 = 30 speech segments, and the length of each is about three to five seconds.

The distance between the speakers and the robot is 1.5 m, and the distance between noises and the robot is 2 m. The noisy environment determination parameter is set to 10.

B. *Evaluation Results*

In the experimental results, in order to compare SNR values of original noisy speech and enhanced speech conveniently, we record noisy speeches, of which the SNRs are 0dB, 5dB, and 10dB. After noise reduction, we calculate SNRs and segment SNRs of enhanced speech. Segment SNR is described in (10), where N means the number of frames, and m is the frame index.

$$SegSNR = \frac{1}{N}\sum_{m=1}^{N} SNR_m \quad (10)$$

Table I is the comparison between average SNR values and segment SNR values of enhanced speeches. Compared with the original signals, the average SNR of the enhanced signals is increased by 20dB to 30dB. Among all the results, the most evident improvement is car noise; we observe that it reaches 37dB on average, which is 32dB higher than the original ones. As for the segment SNRs, each enhanced signal is 30dB higher than original ones. Both results imply the efficacy of our proposed method.

Table II lists the comparison of the recognition rates between the original speeches and enhanced speeches. It shows that our proposed system can improve the recognition rates by about 15 % to 35 % after enhancement.

Judging from Tables I and II, we find that our noisy environment-aware speech recognition system can suppress several types of noise in noisy environments and improve speech recognition rates effectively.

TABLE I. SNRS OF THE ENHANCED SPEECH SIGNALS

Noise Type	Average SNR and Segment SNR Values		
	Average SNR values of original signals	*Average SNR values of enhanced signals*	*Segment SNR values of enhanced signals*
White	5dB	27.01dB	30.26dB
Pink	5dB	26.61dB	29.79dB
Babble	5dB	25.69dB	29.66dB
Car	5dB	30.77dB	33.78dB
Factory	5dB	28.35dB	31.11dB
Average	5dB	27.69dB	30.92dB

TABLE II. COMPARISON OF RECOGNITION RATES

Noise Type	Noisy Speech Recognition Rate	
	Original	*Proposed*
White	51.11%	68.89%
Pink	37.78%	62.22%
Babble	12.22%	45.56%
Car	42.22%	65.56%
Factory	16.67%	53.33%
Average	32.00%	59.11%

V. CONCLUSIONS

In this study, we proposed a noisy environment-aware speech enhancement system. We utilized a simple and efficient method to detect the noisy environment. This detection can avoid overfiltering noises in quiet environments. The experimental results show that our system can remove background noise and enhance speech recognition rates effectively. Besides, our system is capable of improving SNRs by 30dB on average. The experiments clearly demonstrate the performance of the proposed system and its feasibility.

REFERENCES

[1] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," IEEE Trans. Signal Processing, vol. 44, pp. 2768-2779, 1996.

[2] J. F. Cardoso and Comon, "Independent component analysis, a survey of some algebraic methods," in IEEE Symp. Circuits and Systems, 1996, pp. 93-96.

[3] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. Neural Networks, vol. 10, pp. 626-634, 1999.

[4] E. Y. and V. Trees, "A signal subspace approach for speech enhancement," IEEE Trans. Speech and Audio Processing, vol. 3, pp. 251-266, 1995.

[5] Young S., Everman G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V., and Woodland P., "HTKbook (V3.4)," ed. Cambridge city: Cambridge University Engineering Dept., 2006.

[6] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems " in Speech Communication, 1993, pp. 247-251.

[7] J. L. Drury, J. Scholtz, and H. A. Tanco, "Awareness in human-robot interactions," in IEEE Conf. Systems, Man and Cybernetics, 2003, pp. 912-918.

[8] C. T. Ishi, S. Matsuda, T. Kanda, T. Jisuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "A robust speech recognition system for communication robots in noisy environments," IEEE Trans. Robotics, vol. 24, pp. 759-763, 2008.

[9] D. A. K., "Understanding and using context," in Personal and Ubiquitous Computing, 2001, pp. 4-7.

[10] 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 environment," in IEEE/RSJ Conf. Intelligent Robots and Systems, 2008, pp. 1705-1711.

[11] K. Kosuge and Y. Hirata, "Human-robot interaction," in IEEE Conf. Robotics and Biomimetics, 2004, pp. 8-11.

[12] K. Park, S. J. Lee, H.-Young Jung, and Y. Lee, "Human-robot interface using robust speech recognition and user localization based on noise separation device," in IEEE Symp. Robot and Human Interactive Communication, 2009, pp. 328-333.

[13] B. N. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in IEEE Work. Mobile Computing Systems and Applications, 1994, pp. 85-90.

[14] P. Zhang, K. K. Lee, and Y. Xu, "Context-aware robot service coordination system," in IEEE Conf. Robotics and Biomimetics, 2005, pp. 410-415.