ASR for human-symbiotic robot “EMIEW2” with Mechanical Noise and Floor-Level Noise Reduction

Takashi Sumiyoshi, Masahito Togami, Yasunari Obuchi

Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan
{takashi.sumiyoshi.bf; masahito.togami.fe; yasunari.obuchi.jx}@hitachi.com

Abstract

A human-symbiotic robot called “EMIEW2” and its auditory function which includes two noise reduction methods against self-generated mechanical noise and external floor-level noise is introduced. The former type of noise is produced by the robot itself, and this is a difficult problem because it can be loud, non-stationary, and have a wide frequency band. We adopt a maximized SNR technique, in which noise correlation matrix is selected from noise clusters that are learned from the pre-recorded noise signals. The latter type of noise, which can occur when robots are used in office environments, is also a problem, and we addressed it by expanding the beamforming area from one dimension (azimuth angle) to the two dimensions (azimuth and elevation angles). We evaluated these methods in a 100-word speech recognition task and we show that both methods are effective for improving the speech recognition rate.

Index Terms: speech recognition, speech localization, speech separation, robot auditory

1. Introduction

Recently, robots have been considered to be not only important industrial products, but also assistants and companions for human beings. Such robots are expected to communicate with humans with voices and gestures, so they must recognize human’s words accurately under the real world (noisy) conditions.

We previously developed a human-symbiotic robot called “EMIEW” (“Excellent Mobility and Interactive Existence as a Workmate”) [1][2] that can support people in various aspects of their lives. EMIEW has a self-balancing two-wheeled mobility system with an inverted pendulum control algorithm, which realizes its maximum speed of 6 km/h. It also has a speaker, a head and arms actuated by motors to communicate with voices and gestures. We have developed a distant-speech recognition technology for EMIEW that has a microphone-array system consisting of eight microphones positioned on its head and body [3]. In a demonstration at EXPO 2005 in AICHI, JAPAN, EMIEW acted as a waiter and successfully recognized the audience’s requests in an environment with approximately 70dB of ambient noise.

In 2007, we announced a new version of EMIEW, “EMIEW2” [4]. EMIEW2 has a smaller and lighter body than the legacy EMIEW, so it is expected to be able to work safely in an office without obstructing human activity. EMIEW2 has 14 microphones on its head, as shown in Fig. 1, as well as improved speech processing engines so that it can communicate with humans more effectively.

In this paper, we introduce the auditory function of EMIEW2 and explain the technology applied for distant-speech recognition. Then, we explain the methods used to solve two problems of self-generated mechanical noise and floor-level noise, that is, external noise occurring at low-lying positions, and finally we evaluate the effectiveness of these technologies for robots like EMIEW2 with a speech recognition task.

2. Mechanical Noise Reduction

2.1. Mechanical Noise of EMIEW2

In robot speech recognition, not only external noise but also mechanical noise from the robot itself reduces the recognition precision.

We recorded noises (sampled at 16 kHz) occurred in three typical situations described below.

**Stop** An ambient noise only at 16 kHz occurred in three typical situations described below.

**Stay** Kept standing without moving arms and a head

**Motion** Kept moving arms in standing position

These three noises and a human voice spoken from one meter in front of EMIEW2 are illustrated in Fig. 2. The SNR and power levels of these sounds are listed in Table 1. Here, SNR is the ratio of the **Voice** power to the noise power of each type.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR(dB)</th>
<th>Avg(dB)</th>
<th>Max(dB)</th>
<th>Min(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>16.6</td>
<td>-47.2</td>
<td>-41.3</td>
<td>-54.2</td>
</tr>
<tr>
<td>Stay</td>
<td>2.7</td>
<td>-33.3</td>
<td>-28.3</td>
<td>-38.5</td>
</tr>
<tr>
<td>Stay-NF</td>
<td>2.9</td>
<td>-33.5</td>
<td>-28.4</td>
<td>-39.5</td>
</tr>
<tr>
<td>Motion</td>
<td>3.0</td>
<td>-33.6</td>
<td>-27.9</td>
<td>-39.8</td>
</tr>
<tr>
<td>Voice</td>
<td>N/A</td>
<td>-30.6</td>
<td>-21.5</td>
<td>-61.3</td>
</tr>
</tbody>
</table>

These data show that the SNR is poor and the noise frequency band is wide for both **Stay** and **Stop**. In the **Stay** case, the high power stationary noise was observed at around 830 Hz.
and it is considered that the noise was produced by the actuator. This stationary noise will not become the critical problem for ASR, but after removing it by a notch filter, the SNR still remains poor (shown as Stay-NF in Table 1). We can say that almost all of the noise is non-stationary and the mechanical noise reduction technique is strongly required.

2.2. Mechanical Noise Reduction for EMIEW2

As seen in Fig. 2, we find that self-generated mechanical noise is loud, non-stationary, and have a wide frequency band. Such type of noise is not easily removed like 1ch noise reduction technique.

One of common methods to reduce mechanical noise is to use reference signals recorded from microphones arranged near the noise sources. Robots like EMIEW2, however, have many noise sources, for example, the motors on its feet and in its arm joints. Thus, using many reference signals corresponding to these noise sources would require very high production and computational cost.

Another way is to use microphone array techniques like independent component analysis (ICA) based methods [5] [6], but in case of changing the impulse responses of the noise sources rapidly, they can not work effectively.

We have proposed a noise reduction method in which a filter is selected from pre-recorded noise filter models with criteria for maximizing the SNR [12]. The target signal is extracted with the maximizing SNR filter as follows:

\[ y(f, \tau) = \mathbf{w}_{SNR}(f, \tau) \mathbf{x}(f, \tau), \]

where \( \mathbf{x}(f, \tau) \) and \( y(f, \tau) \) denote the input 14-ch signals and the output target signal respectively. \( \mathbf{w}_{SNR}(f, \tau) \) is calculated as follows:

\[ \mathbf{w}_{SNR}(f, \tau) = \arg \max_{\mathbf{w}(f, \tau)} \mathbf{w}(f, \tau)^* \mathbf{R}_s(f, \tau) \mathbf{w}(f, \tau) \]

\[ = \lambda \max e ig(\mathbf{R}_s(f, \tau)^{-1} \mathbf{R}_n(f, \tau)), \quad (2) \]

\( \mathbf{R}_s(f, \tau) \) and \( \mathbf{R}_n(f, \tau) \) denote the correlation matrix of the target signal and noise signal respectively. In this paper, we assume that only the target sound and mechanical noise existed. So, \( \mathbf{R}_s(f, \tau) \) is calculated as \( \mathbf{R}_s(f, \tau) = E[\mathbf{x}(f, \tau) \mathbf{x}(f, \tau)^*] \).

Furthermore, \( \mathbf{R}_s(f, \tau) \) is obtained by selecting the cluster of a pre-learned noise correlation matrix. In the pre-learning phase, pre-recorded noise is clustered into \( C \) clusters by the k-means algorithm. First, noise vectors \( \mathbf{n}(f, \tau) = \)

\[ (n_0(f, \tau), ..., n_{13}(f, \tau)) \] are normalized as follows:

\[ \tilde{\mathbf{n}}(f, \tau) = \frac{\mathbf{n}(f, \tau)\mathbf{n}_0(f, \tau)}{\mathbf{n}(f, \tau)^*\mathbf{n}_0(f, \tau)} \]

(3)

Then, they are clustered into \( C \) clusters by k-means with the following Euclidean distance function, \( D(x, y) = |x - y| = \sqrt{\sum (x_i - y_i)^2} \). Finally, the noise correlation matrix of each cluster is estimated from the elements of the cluster as follows:

\[ \mathbf{R}_{n,i}(f, \tau) = \sum_{\text{index}(f, \tau) = i} (\mathbf{n}(f, \tau)\mathbf{x}(f, \tau)^*), \quad (4) \]

where \( \text{index}(f, \tau) \) denotes the cluster index obtained by the k-means clustering. Then, the maximizing SNR filter can be obtained as follows:

\[ \mathbf{w}_{SNR}(f, \tau) = \lambda \max e ig(\mathbf{R}_{n,i}(f, \tau)^{-1} \mathbf{R}_s(f, \tau)), \quad (5) \]

Finally, gain adjustment between the frequency bins is done with the method described in [7].

3. Floor-level Noise Reduction

3.1. Floor-level Noise

When EMIEW2 recognizes human’s voice at rooms and corridors in an office, variety of noises from various locations will suffer it. The sound source localization and separation algorithm of the legacy EMIEW is a one-dimensional; that is, it uses only an azimuth angle, varying the steering vectors, to estimate the direction of sounds. It can reduce floor-level noise to some extent, but it is possible to improve the noise reduction process using two-dimensional (azimuth and elevation) localization and separation.

We focused the human footsteps and things moving on casters or wheels which comes from a lower position than EMIEW2, and is therefore referred to as “floor-level noise”. We tried to reduce these noises by two-dimensional speech localization and separation to improve ASR accuracy.

We recorded the noise of footsteps and the noise from the casters on chairs as the chairs were moved laterally about 1.2 m from the front of EMIEW2, as indicated by the waveform in Figs. 3. These sounds were recorded in room that was not noise-controlled, with a standard non-carpeted office floor.

3.2. Floor-level Noise Reduction for EMIEW2

To realize two-dimensional speech localization and separation, following processes are done for the input waveform sequentially: 1) Preprocessing, 2) Localization, 3) Sound Separation. The 2) Localization process is enhanced for two-dimensional processing.

Figure 2: Waveforms and spectrum 1)Stop, 2)Stay, 3)Motion, 4)Voice

Figure 3: Waveforms and spectrum (footsteps, casters on chair)
Table 2: SNRs and powers of each floor-level noise types

<table>
<thead>
<tr>
<th></th>
<th>Steps</th>
<th>Casters</th>
<th>Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR(dB)</td>
<td>4.3</td>
<td>1.5</td>
<td>N/A</td>
</tr>
<tr>
<td>Avg(dB)</td>
<td>-37.9</td>
<td>-35.1</td>
<td>-33.6</td>
</tr>
<tr>
<td>Max(dB)</td>
<td>-34.8</td>
<td>-30.5</td>
<td>-27.9</td>
</tr>
<tr>
<td>Min(dB)</td>
<td>-41.9</td>
<td>-41.5</td>
<td>-39.8</td>
</tr>
</tbody>
</table>

3.2.1. Preprocessing

The 14-ch waveforms sent by EMIEW2 are divided into time frame sequences. Each frame length is 10 ms. Then, each frame waveform is converted into spectra \( x(f, \tau) = (x_0(f, \tau), ..., x_{13}(f, \tau)) \) by short time Fourier transform. Here, \( f \) denotes the frequency bin, and \( \tau \) denotes the frame index.

3.2.2. Localization

Next, the power of the direction of each frequency bin and each frame \( P_i(f, \tau) \) is calculated from 14-ch spectra using the modified delay sum beam forming (DSBF) method [3], which is an improved method of DSBF based on the sparseness of the human voice.

\[
P_i(f, \tau) = \max_{l \in \Lambda} |a_i(f)^* x(f, \tau)|
\]  

(6)

Here, \( a_i(f) \) denotes the steering vector of direction \( l \) and frequency bin \( f \), and \( \Lambda \) denotes the search area. In the case of EMIEW2, sounds are assumed to come from the directions in which the azimuth angle is from -180 to 180 degrees, and elevation angle is from -50 to 50 degrees, and the search resolution is 10 degrees. Therefore, in one-dimensional sound source localization, \( \Lambda = \{\theta \in [-180, -170, ..., 160, 170]\} \), and in two-dimensional sound source localization, \( \Lambda = \{\theta, \phi \in [-180, -170, ..., 160, 170, \phi = -50, -40, ..., 40, 50]\} \). We adopted the fast two-dimensional sound source localization algorithm called “SPIRE-MDSBF” [4] for EMIEW2’s localization.

Next, sound segments are extracted from \( P_i(f, \tau) \) by finding the sequence \( Seq_i = (v_i, \{d_{i,t} | t = 0, ..., L_i\}) \) where \( L_i > T_L, \sum_j d_{i,j} > T_P(t = 0, ..., L_i) \) and \( |d_{i,t} - d_{i,t+1}| < T_D(t = 0, ..., L_i) \). These conditions are used to distinguish continuous sounds from instant noise, and thresholds \( T_L, T_P \) and \( T_D \) are experimentally defined. \( v_i \) denotes the frame index of the beginning position, \( L_i \) denotes the length, and \( d_{i,t} \) denotes the directions of each frame \( t \) of the segment. The segment index \( i \) is numbered accumulatively, and the set of the indices of the active segments at \( \tau \) is denoted as \( N(\tau) \), and its center direction is denoted as \( N\tilde{d}(\tau) \).

3.2.3. Sound Separation

Next, waveforms are extracted for each extracted segment. Each segment is treated as a speech phrase or continuous directional noise from one sound source. First, the directions of sound sources \( Dir(\tau, i) \) that has the ith largest power at frame \( \tau \) are estimated based on the directional power \( DP_i(\tau) = \sum_j a_i(f)^* x(f, \tau) \).

\[
Dir(\tau, i) = \arg\max_{l \in \Lambda, f \in Dir(\tau, j), j < i} DP_i(\tau)
\]  

(7)

Then, each frequency bin is assigned to one candidate direction, whose index is denoted as \( FD(f, \tau) \).

\[
FD(f, \tau) = \arg\max_{l \in Dir(\tau, i), 0 < c < k} a_i(f)^* x(f, \tau)
\]  

(8)

Then, the segment indices are assigned for each frequency bin based on the direction of the sound source candidate \( FD(f, \tau) \), the direction of each segment \( \omega \in N\tilde{d}(\tau) \), and the distance between directions function \( Dist(\alpha, \beta) = \sqrt{(\alpha\theta - \beta\theta)^2 + (\alpha\phi - \beta\phi)^2} \) as follows:

\[
MD(f, \tau) = \min_{\omega \in N\tilde{d}(\tau)} Dist(FD(f, \tau), \omega),
\]  

(9)

\[
Seg(f, \tau) = \begin{cases} 
\arg\min_{\omega \in N\tilde{d}(\tau)} Dist(FD(f, \tau), \omega) & \text{if } MD(f, \tau) > T_S \\
-1 & \text{if } MD(f, \tau) \leq T_S
\end{cases}
\]  

(10)

Here, \( \alpha\theta \) and \( \alpha\phi \) denotes the azimuth and elevation angle of \( \alpha \) respectively. \( Seg(f, \tau) \) denotes the segment index of frequency bin \( f \) at frame \( \tau \). If the distance to the nearest segment is larger than given threshold \( T_S \), the frequency bin is regarded to not to belong to any active segments, and \( Seg(f, \tau) \) is set to -1.

To separate the sound in one segment from sounds in other segments, the minimum variance beamforming (MVBF) technique[11] is used. The correlation matrices used by MVBF are calculated for each segment and “outer-segment” frame by frame. The correlation matrices and the filters \( w_i(f, \tau) \) of each segment are calculated as follows:

\[
R_i(f) = E[|x(f, \tau)|^2] \delta(Seg(f, \tau), i)], \\
R_{out}(f) = E[|x(f, \tau)|^2] \delta(Seg(f, \tau), -1),
\]  

(11)

\[
R_i(f) = \sum_{j \in N(\tau), j \neq i} R_j(f) + R_{out}(f),
\]  

(12)

\[
w_i(f, \tau) = \frac{R_i^{-1}(f) a_i(f)}{a_i(f)^* R_i^{-1}(f) a_i(f)}
\]  

(13)

where \( \delta(a, b) \) is the Kronecker delta function.

Finally, the sound waveform in the segment \( y_i(f, \tau) \) is obtained as \( y_i(f, \tau) = w_i(f, \tau)^* x(f, \tau) \).

4. Evaluation

4.1. Speech Recognition with Mechanical Noise Reduction

First, we evaluate the proposed mechanical noise reduction method based on maximizing SNR. The test set is synthesized from a speech database and a noise database. The speech database consists of utterances of 100 popular Japanese family names spoken by 25 females and 15 males. Each family name was about 3 to 4 mora, for example, “Suzuki,” “Takahashi,” and so on. Each utterance was convoluted by the impulse response recorded one meter in front of EMIEW2 to make simulated waveforms of the 14 channels. The noise database was the same as that described in section 2. We simply mixed the speech waveforms and the noise waveforms to make a test set. In the first test case, where the noise reduction was inactive (denoted as “Inactive” in Table 3), the 1st channel of each test set waveform was input to the speech recognition system. In the second test case, where the noise reduction was active (denoted as “Active” in Table 3), the 14-channel waveform was processed using our method, and the output waveform was input to the speech recognition system.
Table 3: Recognition rates in mechanical noise environment(%)  

<table>
<thead>
<tr>
<th>Noise condition</th>
<th>Noise reduction</th>
<th>Inactive</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without noise</td>
<td>96</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Stop</td>
<td>69</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Stay</td>
<td>23</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Motion</td>
<td>14</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

We trained the mechanical noise models using the same types of noise waveforms (recorded separately from the noise data used to create the test set). The length of each waveform was about 10 seconds, and number of filters was fixed at eight.

We use the tied-state left-to-right triphone hidden Markov model (HMM) for speech recognition. Each HMM consists of three states, which are tied to 3000 states, each of which has a 16 mixture Gaussian distribution of 25-dimensional Mel frequency cepstral coefficient (MFCC) features (12 base MFCC features, 12 velocity features, and 1 velocity log power). The HMMs are trained by the HMM Tool Kit (HTK) [9] using the Japanese Newspaper Article Sentences (JNAS) speech database [8]. For this evaluation, we used a Julius decoder [10]. The vocabulary size was 100 and was identical to the content of the test set.

The result of the recognition examination is shown in Table 3. It is clear that out method succeeded in improving the recognition rate in an environment with various types of mechanical noise.

4.2. Speech Recognition with Floor-level Noise Reduction

The test set was synthesized in the same manner as in the previous experiment, using this floor-level noise. There were three test cases: firstly, no sound source separation was carried out, and the 1st channel of each test set waveform was input to the speech recognition system (denoted as “Baseline” in Table 4.) In second and third cases, one-dimensional and two-dimensional sound source separation was done (denoted as “1-dim” and “2-dim” in Table 4). We used 1000 utterances for this examination, and the same acoustic model, vocabulary set, and decoder for the speech recognition as previously mentioned.

The recognition results are shown in Table 4. For both Steps and Casters, the two-dimensional sound source separation exceeded the one-dimensional method by absolute 11.8% and 6.2% respectively, but for Casters, the recognition rate was far from good. This is because the noise was loud, had a wider frequency bandwidth, and continued to sound. The assumption of sparseness is not true in this case, and the target voice was corrupted.

Table 4: Recognition rates of floor-level noise environment(%)  

<table>
<thead>
<tr>
<th>Noise condition</th>
<th>Speech separation method</th>
<th>Baseline</th>
<th>1-dim.</th>
<th>2-dim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without noise</td>
<td></td>
<td>97.3</td>
<td>96.0</td>
<td>95.4</td>
</tr>
<tr>
<td>Steps</td>
<td></td>
<td>22.3</td>
<td>62.5</td>
<td>74.3</td>
</tr>
<tr>
<td>Casters</td>
<td></td>
<td>2.8</td>
<td>17.9</td>
<td>24.1</td>
</tr>
</tbody>
</table>

5. Conclusions

5.1. Conclusions

We have introduced a human-symbiotic robot called “EMIEW2” and its auditory function, and explained the two kinds of noise that can occur and cause problems: self-generated mechanical noise and external floor-level noise. To address the first problem, we adopt a mechanical noise reduction method based on a maximized SNR technique, in which a noise correlation matrix is selected from noise clusters that are pre-learned from the same type of recorded noise. To the second problem, we expanded the beamforming area from one dimension to two dimensions to reduce floor-level noise. We evaluated each method and proved that these methods can effectively improve the speech recognition rate for simulated input waveforms.

5.2. Future Works

We have already combined these two methods into one, which is intended to extract the target voice from multi-channel signals contaminated by other directional noise and mechanical noise [12]. We are planning to evaluate the combined method focusing on the speech recognition rate and conduct a field examination to further evaluate EMIEW2 functions, focusing on problems that occur in the real environment.

6. References