Recognition of three simultaneous utterance of speech by four-line directivity microphone mounted on head of robot

Naoya Mochiki, Toshiyuki Sekiya, Tetsuji Ogawa and Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
{mochiki,sekiya,ogawa,koba}@pcl.cs.waseda.ac.jp

Abstract

A sound source separation method using four-line directivity microphones mounted on a head of a robot is proposed and applied to speech recognition under existence of two disturbances of speech. Sound source separation methods using microphones mounted on robot heads generally used strict head-related transfer functions (HRTF). We propose a robust sound source separation that does not require an estimate of a strict HRTF. Our method takes advantage of a sound pressure difference with the robot head acting as a sound barrier. The enhancement of the difference in the target speech is performed by signal processing of three layers: two-line SAFIA, two-line Spectral Subtraction and their integration. The experimental results of three simultaneous utterance recognition with vocabulary of 20K show that the proposed method is effective in achieving 71% error reduction.

1. Introduction

We propose hands-free speech recognition with microphones mounted on a robot head.

In the future, we expect robots to operate in a real environment and in a situation where several people will talk simultaneously. In such a situation, we demand high performance sound source separation methods to extract only the target sound.

Nakadai proposed a mixed speech recognition method based on a HRTF using microphones mounted on a robot head[1]. In most cases, the method uses two microphones to separate a sound source with high performance, yet problems may occur due to sound source positioning and sound source frequencies. In addition, there is the limitation of the shape of a robot head, because an estimate of a strict HRTF is necessary.

In this paper, we propose a robust sound source separation which does not require an estimate of a strict HRTF and limitations on the head shape. Our method mounted four-line directivity microphones on the robot head and took advantage of a sound pressure difference which occurs by using the robot head acting as a sound barrier. We use this sound pressure difference and extract only a target sound source by signal processing three layers. The three layers consists of SAFIA[2], Spectral Subtraction(SS)[3] and integration. In addition, with our method, we extract only a target speech from three simultaneous utterances of speech and evaluate it with continuous speech recognition of vocabulary size 20K.

In the following section, the mounting conditions of the four-line microphone onto the robot head is described. In section 3, the algorithm of the proposed method is described. In section 4, the conditions and results of the continuous speech recognition for three simultaneous utterance of speech in a real environment is described. We give the conclusions in section 5.

2. Mounted condition of microphones

As for a directivity microphone, Auditechnica ATM15a was used. In the experiment, the exterior shell of the robot head was used instead of the actual frame of the robot head. As depicted in Fig.1(a), 1(b), two microphones are mounted on each side: One is directed to the frontal side of the face and the other is directed perpendicularly sideward. As follows, we refer to the two frontal-directed microphones RF-Mic(Right-Front-Microphone), LF-Mic(Left-Front-Microphone). We refer to the two side-directed microphones RR-Mic(Right-Right-Microphone), LL-Mic(Left-Left-Microphone).

Figure 2 shows a purpose of four-line microphone arrangement. We assume an environment where three sound sources exist: the sound coming from the frontal...
The robot head works as a sound barrier

Figure 2: Purpose of four-line microphone arrangement

We use a sound pressure difference which occurs between these channels and realize the sound source separation that does not require an estimate of a strict HRTF.

3. Proposed method

By signal processing in three layers, we remove a sound source coming from the right and left direction and extract only a target sound source coming from the front. Figure 3 shows the diagram of the proposed method.

3.1. First layer

Figure 4 shows signal processing of the first layer. The first layer used SAFIA with the input of RF-Mic and LL-Mic or LF-Mic and RR-Mic. SAFIA used the input of the two channels. First, we decided which channel is dominated by each frequency band. Next, we select only the dominant frequency band and remove the recessive frequency band.

With the arrangement of the four-line microphone and the use of the robot head as a sound barrier, the spectrum received by LF-Mic are written as \((S_F, S_R, S_L)\) and the Spectrum received by RR-Mic are written as \((S_F^S, S_R, S_L^S)\).

When we compare LF-Mic with RR-Mic, the spectrum of \(S_R\) received by LF-Mic is more inferior to the spectrum of \(S_R\) received by RR-Mic. On the contrary, the spectrum of \(S_F, S_L\) received by RR-Mic is more inferior to the spectrum of \(S_F, S_L\) received by LF-Mic. Therefore, we use SAFIA with LF-Mic and RR-Mic. when we compare the power spectrum of each frequency with the input of the two channels, we can separate the spectrum of \((S_F, S_L)\) and \(S_R\), so that a spectrum band of the
channel which is inferior can be removed. In addition, by using SAFIA with RF-Mic and LL-Mic, we can separate the spectrum of \((S_F, S_R)\) and \(S_L\).

### 3.2. Second layer

In the second layer, we use the spectrum of \(S_R\) or \(S_L\) which we extract from the first layer and extract the spectrum of \(S_F\) by SS.

For example, in using the spectrum of \(S_L\) which we extracted from the first layer, we remove the spectrum of \(S_L\) included in a mixture spectrum of \((S_F, S_L)\) by SS and extract only a target spectrum of \(S_F\). The Mixture spectrum of \((S_F, S_L)\) denoted \(S_F + L\) and \(\hat{S}_F\) is written as follows.

\[
\hat{S}_F = \begin{cases} 
|S_{F+L}|^2 - \alpha \cdot |S_L|^2 \cdot e^{j\phi}, & \text{if } |S_{F+L}|^2 - \alpha \cdot |S_L|^2 > 0 \\
0, & \text{otherwise}
\end{cases}
\]

\(\alpha\) is an amplitude of the subtraction process. \((\alpha = 1.0)\) \(\phi\) is an appropriate phase function. For example, we can use the phase of the spectrum of \(S_{F+L}\). Likewise, in using the spectrum of \(S_R\) which we extracted from the first layer, we remove the spectrum of \(S_R\) included in a mixture spectrum of \((S_F, S_R)\) by SS and extract only a target spectrum of \(S_F\).

### 3.3. Third layer

After the second layer, we obtain the two estimates of \(S_F\): one is RF origin and the other is LF origin. At the third layer, we integrate these two estimates of \(S_F\), and refine \(\hat{S}_F\).

Here we try two methods for the integration. One is to add two estimates. The spectrum of \(S_F\) is averaged by this. We call this method "Addition".

The other is to select inferiors of the two \(S_F\) estimates by the spectral band.

Two spectra of \(S_F\) in the second layer tend to be larger than the real one because they are still contaminated with disturbance speech. Therefore, inferior is expected to be closer to the real one. We call this method "Minimization".

### 4. Experiment

#### 4.1. Conditions

We recorded three simultaneous speech data to enable continuous speech recognition. Input signal is sampled by 32kHz with 16 bit data. We placed three loudspeakers in an arrangement as depicted in Fig.5, acting as the three sound sources in place of a person’s utterance. First, a loudspeaker for a target sound source was placed in front of the robot. Next, we define the location of the robot as the center, and the two loudspeakers representing the disturbance sounds were placed according to the angle of \(\theta\) (\(\theta = 60, -60\) deg). The distance between each loudspeaker and the robot is \(d\). \((d=100\, \text{cm})\)

As evaluation data, we selected one hundred sentences spoken by twenty male speakers from the ASJ-JNAS continuous speech corpus [4]. In the experiment, two different speech signals were played simultaneously. The utterance length and volume of the two different speeches were almost equivalent. The SNR was almost 0 dB.

#### 4.2. Speech recognition

We evaluate recognition performance for speech data which is extracted by six different methods of (F) from
We carried out the continuous speech recognition with vocabulary size of 20K on the processed speech data. The acoustic features is shown in Table1. The acoustic models are trained with 20 K sentences spoken by about 100 male speakers from ASJ-JNAS corpus. The training data is recorded with close-talk microphones. As for the language models, we use trigram language models using a lexicon size of 20 K. Decoder is SKOOD, developed by our group. In addition, we attempt to enable speech recognition to be robust to spectral distortion by MLLR. In the proposed method, the disturbance can be removed. But the recovered target speech contains some particular noise. The discontinuity of the spectrum causes the musical noise in the recovered target speech. This causes a degation of performance. To improve the performance, we adopt the MLLR-based acoustic model adaptation with the recovered speech which contains the characteristic of the proposed method. 80 sentences are used for the adaptation. Test set consists of 100 sentences.

4.3. Results

Figure. 7 shows the results of continuous speech recognition on three simultaneous speech utterances. In the first layer, by using SAFIA(B) the recognition rate was only -0.2% word accuracy.

When we use SS process in the second layer, the recognition rate was 55.4% word accuracy. This achieved 55% error reduction compared to (B).

When we use a process of integration in the third layer, we could improve performance even further. As for the processing of the third layer, we can see that Minimization(E) is more effective than Addition(D). Minimization is effective in the removal of a disturbance sound source. By signal processing the three layers, SAFIA, SS and Minimization, the recognition rate of (E) reached up to 68.7% in word accuracy. This achieved 30% and 71% error reduction compared to (C) and (A).

Furthermore, the MLLR adaptation of the recovered speech which was processed by the three layers improved performance. This achieved 78% error reduction compared to Non-processing(A). The performance of 76.5% in word accuracy was achieved in three simultaneous speech recognition.

5. Conclusion

We mounted four-line directivity microphones on the robot head and proposed a robust sound source separation which does not require an estimate of an strict HRTF and limitations on the shape of the head. In a real environment, experimental results showed the effectiveness of the proposed method. The error reduction was 71% compared to Non-Processing. Furthermore, the recognition performance was improved by MLLR adaptation of the recovered speech.

6. References