SPEECH RECOGNITION IN THE BLIND CONDITION BASED ON MULTIPLE DIRECTIVITY PATTERNS USING A MICROPHONE ARRAY

Toshiyuki SEKIYA and Tetsunori KOBAYASHI

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

ABSTRACT

A novel hands free speech recognition method using a microphone array is proposed and is applied to the multi-talk recognition in the blind condition, no prior information about the sound sources and the characteristics of room acoustics. The proposed system is constructed by the cascade of the sound localization system, MUSIC, and the sound segregation system, SMDP (Segregation using Multiple Directivity Patterns) proposed in our previous paper. SMDP is characterized by using redundant directivity patterns. Usually, it is difficult for this sort of cascade system to achieve high performance because the sound localization stage cannot be perfect and errors occurred in this first stage cause serious damages to the segregation stage. Particularly missing the sound source is critical. By arranging the virtual sound sources, we treat the excess sound sources. In the proposed method, contrarily, the errors in the localization stage hardly cause the problems as long as they are insertion. SMDP uses redundant directivity patterns from the first, so it tolerates the insertion errors. The proposed method achieved 70% word accuracy in the double-talk recognition experiment of 20 K vocabulary, which is 18 point better compared to the ICA-based blind source separation with the source-number-given condition.

1. INTRODUCTION

Multi-talk recognition is indispensable to realize various applications of hands free speech recognition, for example, group conversation systems such as humanoid robot, dictation systems of meeting, interfaces of car-navigation systems. Microphone array, which makes active use of spatial information between microphones and sound sources, is very effective to realize these applications[1][2]. We have proposed the novel speech segregation method, called SMDP, based on multiple directivity patterns using a microphone array not only single directivity[3]. High performance source segregation is realized to be robust against the error factor caused by the back-ground noise and the reverberation and so on. Proposed system showed superiority to the conventional array techniques. However the source positions were completely given in the previous experiment.

In this paper, double-talk recognition is realized in the blind condition, where there is no priori information about the sound sources and the characteristics of room acoustics. SMDP is carried out based on the estimated source positions given by MUSIC[4]. Usually segregation performance is influenced by precision of source localization. The source spectra are estimated as the solutions of redundant simultaneous equations in proposed system, therefore insertion of sound sources hardly causes the deterioration of segregation performance. However the missing of sound sources is critical to the segregation performance. The virtual sound sources are arranged at the positions where the sound sources are likely to exist. These positions are decided by analyzing the frequency of estimated positions during the utterance. Multiple directivity patterns are designed from both of virtual and estimated source positions. In this way, disturbance spectrum is estimated to be robust against uncertainty of source localization and is removed by spectral subtraction. Enhancement of target speech is realized.

In the following section, the algorithm of proposed hands free speech recognition method is described in detail. In section 3, conditions and results of continuous speech recognition are described. We give the conclusions in section 4.

2. PROPOSED METHOD

2.1. Formulation of the sound field

Figure.1 shows the diagram of proposed method. We assume the environment where D sound sources exist and the sound field is observed by M microphones. We define the input vector $\boldsymbol{x}(k,t)$ as STFT of the input signal into a microphone array. Using the location vector, $\boldsymbol{x}(k,t)$ is written as follows.

$$\boldsymbol{x}(k,t) = \boldsymbol{A}(k)\,\boldsymbol{s}(k,t) + \boldsymbol{n}(k,t)$$

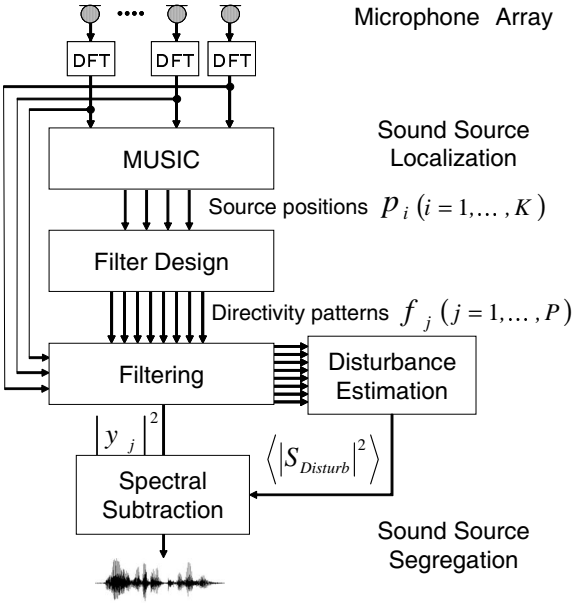

Fig. 1. Diagram of proposed method.

where,

$$\begin{aligned} \boldsymbol{A}(k) &= [\boldsymbol{a}_1(k), \cdots, \boldsymbol{a}_D(k)] \\ \boldsymbol{s}(k,t) &= [s_1(k,t), \cdots, s_D(k,t)]^T \\ \boldsymbol{n}(k,t) &= [n_1(k,t), \cdots, n_M(k,t)]^T \end{aligned}$$

$\boldsymbol{a}_d(k)$ denotes the location vector from d-th source to the microphones. The DFT coefficients of measured impulse responses are usually used as $\boldsymbol{a}_d(k)$. The location vectors are calculated using characteristics of delay between microphones and the source positions to be robust against environmental changes. $s_d(k,t)$ denotes the spectrum of d-th source. $n_m(k,t)$ denotes the spectrum of the back-ground noise and the reverberation at microphone m. $[\cdot]^T$ denotes the transposition. k and t denote the discrete frequency and frame index respectively. From this, to simplify the expression, we omit the symbol k and t.

2.2. Sound source localization

MUSIC is applied to sound source localization. The spatial correlation matrix $\boldsymbol{R} = E\left[\boldsymbol{x} \cdot \boldsymbol{x}^H\right]$ is decomposed into the signal subspace and the noise subspace. MUSIC spectrum is calculated in each frequency band using the eigenvalue decomposition, $\boldsymbol{R} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^{-1}$.

$$\begin{aligned} P_{music}(r,\theta) &= \frac{\boldsymbol{a}^H(r,\theta)\,\boldsymbol{a}(r,\theta)}{\boldsymbol{a}^H(r,\theta)\,\boldsymbol{E}_N\boldsymbol{E}_N^H\boldsymbol{a}(r,\theta)} \\ \boldsymbol{E}_N &= [\boldsymbol{e}_{L+1}, \boldsymbol{e}_{L+2}, \cdots, \boldsymbol{e}_M] \end{aligned}$$

$[\cdot]^H$ denotes the complex conjugate transposition. $\boldsymbol{E}_N$ is the eigenvectors corresponding to the noise subspace, the smallest $M - L$ eigenvalues. The number of sound sources must be known in advance to apply MUSIC. The number of sound sources is estimated by AIC in each frequency band and in each analysis block.

The source positions $p_i\ (i = 1, 2, \ldots, K)$ are estimated by peak picking of MUSIC spectrum in two-dimensional space.

2.3. Sound source segregation

2.3.1. Estimation of source spectrum

Redundant simultaneous equations between amplitudes of source spectra and multiple directivity patterns are set up. Source spectra are estimated as the least squares solutions of these equations. To make the problem simple, let us assume two sound sources exist in the sound field. When a directivity pattern $\boldsymbol{f}_1$ is given to the input vector $\boldsymbol{x}$, average power spectrum of the output y_1 is written as follows[3].

$$\langle |y_1|^2 \rangle = |F_{11}|^2 \langle |s_1|^2 \rangle + |F_{12}|^2 \langle |s_2|^2 \rangle + \varepsilon_1$$

F_{ij} represents the dot product between $\boldsymbol{f}_i$ and $\boldsymbol{a}_j$. ε_i represents the error factor caused by the back-ground noise and the reverberation and the error of the location vector itself. $\langle \cdot \rangle$ denotes the frame-averaging. Equation is only one, nevertheless unknown variables are $\langle |s_1|^2 \rangle$, $\langle |s_2|^2 \rangle$. Source spectra cannot be estimated.

Applying P directivity patterns $f_i\ (i = 1, \ldots, P)$, the redundant simultaneous equations are obtained.

$$\boldsymbol{Y} = \boldsymbol{F} \cdot \bar{\boldsymbol{s}} + \boldsymbol{\varepsilon}$$

where,

$$\begin{aligned} \boldsymbol{Y} &= [\,\langle |y_1|^2 \rangle, \langle |y_2|^2 \rangle, \cdots, \langle |y_P|^2 \rangle\,]^T \\ \bar{\boldsymbol{s}} &= [\,\langle |s_1|^2 \rangle, \langle |s_2|^2 \rangle, \ldots, \langle |s_D|^2 \rangle\,]^T \\ \boldsymbol{\varepsilon} &= [\,\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_P\,]^T \end{aligned}$$

$$\boldsymbol{F} = \begin{pmatrix} |F_{11}|^2 & |F_{12}|^2 & \cdots & |F_{1D}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |F_{P1}|^2 & |F_{P2}|^2 & \cdots & |F_{PD}|^2 \end{pmatrix}$$

$\boldsymbol{Y}$ denotes the average power spectra of the output signals given by P directivity patterns. $\bar{\boldsymbol{s}}$ denotes the average source power spectra. $\boldsymbol{\varepsilon}$ denotes the error factors. $\boldsymbol{F}$ denotes the amplitudes of directivity patterns. As a matter of fact, each equation contains the error factor. The average source power spectra are estimated by minimizing the squared error $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$.

$$\begin{aligned} \min_{\boldsymbol{s}}\ \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \quad &\Rightarrow \quad \nabla_s \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = 0 \\ \bar{\boldsymbol{s}} &= (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{Y} \end{aligned}$$

2.3.2. Speech enhancement

The disturbance spectrum is removed from short-time spectrum by spectral subtraction. Let us assume that s_1 is target source and y_1 represents the short-time spectrum that the target source is emphasized or the signal that the disturbance sources are suppressed. The short-time spectrum of target source $|\hat{s}_1|^2$ is obtained using the estimated disturbance spectrum $\langle |s_2|^2 \rangle$.

$$|\hat{s}_1|^2 = \begin{cases} |y_1|^2 - \alpha \cdot \langle |s_2|^2 \rangle, \\ \qquad\qquad if \quad |y_1|^2 - \alpha \cdot \langle |s_2|^2 \rangle > \beta \\ \beta, \qquad\qquad otherwise \end{cases}$$

α is an amplitude of the subtraction process. β is flooring coefficient.

2.4. Integration

Multiple directivity patterns are designed using source localization results. However source localization cannot be perfect. Insertion and missing of source sources sometimes occurred. Proposed method is robust against the insertion because source spectra are estimated as the least squares solutions of redundant simultaneous equations from the first. Missing of sound sources is quite critical to segregation performance because the number of directivity patterns is limited. Therefore disturbance estimation is unreliable. The virtual sound sources are arranged at the positions where the sound sources are likely to exist. The virtual source positions are decided by analyzing the frequency of estimated positions during the utterance. Multiple directivity patterns are designed using both of the virtual and estimated source positions. This concept sometimes produces the excess sound sources over the actual number. However, insertion of sound sources hardly cause the deterioration of segregation performance as has been mentioned. The reliable disturbance spectrum is estimated and removed by spectral subtraction. In proposed system, the errors of localization stage are tolerated in segregation stage.

3. EXPERIMENT

3.1. Experimental Setup

We recorded the speech data to enable continuous speech recognition. Sampling and quantization is 32 kHz and 16 bits respectively. The microphone array consists of eight omnidirectional microphones. Array form is linear and consistent spacing of 3cm. Figure. 2 shows the recording condition. The reverberation time (RT) is 120ms and 200ms. The loudspeaker arranged in front of the microphone array is the target source. Another loudspeaker is the disturbance source and is moved to vary experimental conditions. Evaluation data is totally recorded in four different conditions.

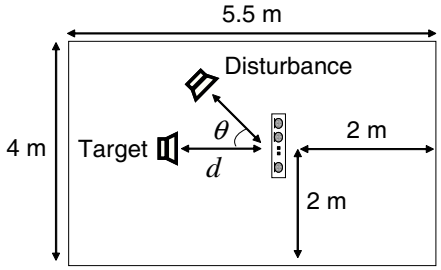

Fig. 2. Recording condition.

As for the target utterances, we select 100 sentences spoken by 23 male speakers from ASJ-JNAS continuous speech corpus. As for the disturbance utterances, we select speech data spoken by other male speakers from ASJ-JNAS. Each utterance is adjusted to almost the same length and the same energy. The SNR is almost 0 dB.

3.2. Speech Processing

3.2.1. Sound source localization

The spatial correlation matrix is calculated every 96ms. Frame length is 32ms and frame shift is 32ms. The period in which the input energy is small is eliminated from calculation of spatial correlation matrix. MUSIC spectrum calculated in each frequency band is added in frequency domain.

$$P_{music}(r, \theta) = \sum_{k=k_1}^{k_2} P_{music}(k, r, \theta)$$

$[k1, k2] = [2000, 4000]$Hz. The location vectors are calculated at intervals of 5 degree in the range of -90 to 90 degree and at intervals of 10 cm in the range of 50 to 150 cm to the microphone array front. Totally the location vectors are calculated at 407 points. Source positions can be estimated with very detailed resolution because we don't use measured impulse responses.

3.2.2. Sound source segregation

In the case that source localization result is within ± 10 degree in front of microphone array, segregation of target source is carried out. The directivity patterns used in this experiment are the DS filter, which emphasizes each sound source, and the DCMP filter, which suppresses each sound source. The number of directivity patterns is twice that of estimated sound sources. Analysis condition is as follows. Frame length is 32ms, frame shift is 8ms and window function is Hamming.

3.2.3. Speech Recognition

The parameters of the acoustic features are as follows. Acoustic features are 12-dimentional MFCC and ΔMFCC and Δpower. Pre-emphasis is done by $1 - 0.97z^{-1}$. Frame length is 25ms and frame shift is 10ms. Window function is Hamming. 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 microphone. The language models are the trigram language models using lexicon of 20 K vocabulary size. In this experiment, the speech data is sampled at 32kHz. On the other hand, the acoustic models are trained with the speech data sampled at 16kHz. Segregated speech is converted to 16kHz sampling rate and converted to acoustic features.

3.3. Evaluation

We apply the proposed method to double-talk recognition in the blind condition. Effectiveness of the proposed method is evaluated in two points.

1. Comparison of the condition that the number of sources is unknown (Blind) and the number of sources is known (Semi-Blind) and the sound source positions are completely known (Completely Known).
2. Comparison with the BSS method based on ICA.

In experiment 2, JADE[5] algorithm in frequency domain is adopted as the BSS method. In this method, source separation is carried out in the semi-blind condition. On the other hand, proposed method is carried out in the blind condition.

3.4. Results

Figure.3 shows the result of experiment 1. Word accuracy in the blind condition is the almost same in the semi-blind condition. From this result, source segregation is carried out to be robust against the uncertainty of source localization. We can confirm that it is very effective to arrange the virtual sound sources at the positions where the sound source are likely to exist but are not estimated practically. Proposed method achieved about 90 % performance compared with the condition that the source positions are completely known. In our preliminary experiment, the result of source localization had the error of about 10 degree in direction. Proposed method showed the robustness against the source localization error. The errors of localization stage are tolerated in segregation stage. Result of the comparison with the BSS method is shown in fig.4. Performance of BSS deteriorated under the long reverberation environment compared with proposed method. The longer the reverberation time is, the more error factors increase. Proposed method is robust against the error factors and has superiority to the BSS method.

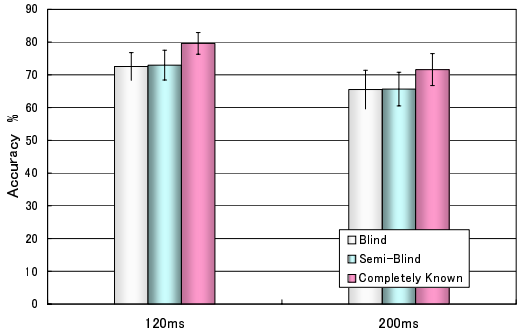

Fig. 3. Evaluation of proposed method. (Each thick bar represents the average performance. Line on the bar represents the maximum and minimum performance.)

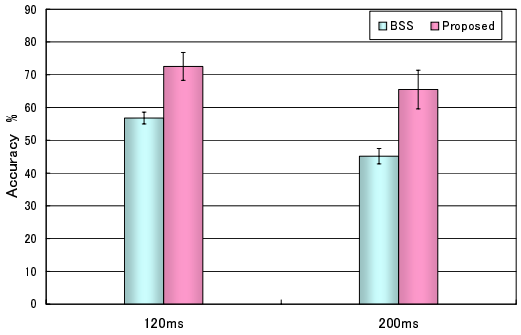

Fig. 4. Comparison of proposed method and BSS method.

4. CONCLUSION

We proposed hands free speech recognition method, which is constructed by the cascade of MUSIC and SMDP. Robust source segregation against the errors in localization stage is realized.Proposed method achieved 70 % word accuracy in double-talk recognition of 20 K vocabulary in the blind condition. From the comparison of the ICA-based BSS method, the great advantage of proposed method was shown, particularly in long reverberation environment.

5. REFERENCES

[1] J. L. Flanagan et al., "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am. 78 (5), pp.1508-1518, 1985.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," IEEE Trans. Antennas Propagatation., vol.AP-30, pp.27-34, Jan. 1982

[3] T. SEKIYA et al., "Speech enhancement based on multiple directivity patterns using a microphone array," Proc. ICASSP2004. vol.1, pp.877-880

[4] R. O. Schmidt,, "Multiple emitter location and signal parameter estimation," IEEE Trans., vol.AP-34, No.3, pp.276-280

[5] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," IEEE on SP, 140, pp.362-370, Dec, 1993