ESTIMATING POSITIONS OF MULTIPLE ADJACENT SPEAKERS BASED ON MUSIC SPECTRA CORRELATION USING A MICROPHONE ARRAY

Hidetomo Tanaka and Tetsunori Kobayashi

Dept. EECE, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, JAPAN
{tanaka,koba}@tk.elec.waseda.ac.jp

ABSTRACT

In this paper, we propose an improved method of estimating the positions of two speakers using a microphone array. A well-known method, MUSIC, can be used to estimate speaker positions with high precision. However, in the special case that the speakers are closely located, the conventional MUSIC-based method sometimes fails to identify the existence of certain speakers, because close peaks in the MUSIC spectrum cannot be resolved. To overcome this difficulty, we propose a new method utilizing a cross-correlation between space-spectra calculated by MUSIC. Experimental results in a real environment have shown that the proposed method is effective in resolving the approximate positions of adjacent speakers.

1. INTRODUCTION

In this paper, we describe a sound localization method specifically for resolving adjacent speakers.

Hands-free speech recognition, in which the microphone is mounted on a terminal side rather than the user's body, has a wide range of applications, including situations in which many users share the system and may speak simultaneously. A personal robot serving a family in a house hold is a good example. To realize hands-free speech recognition requires speech enhancement of speech recorded with a microphone at distance. One method that promises to solve this problem is a combination of sound direction estimation and beam-forming using a microphone array, by which the sound direction is estimated, and then a beam is formed against the estimated direction. In this type of solution, the accuracy of sound direction estimation is essential for realizing good overall system performance.

To date, many sound localization methods have been proposed using microphone arrays, including the delayed-sum array,[1] maximum entropy, linear prediction, MUSIC.[2] In particular, it is well known that the MUSIC method, which uses the eigenvalues of a correlation matrix, realizes superior performance compared to the other methods. However, the conventional MUSIC-based method is effective only for a single source or multiple separated sources. In the case that the speakers are closely located, MUSIC sometimes fails to identify the existence of certain speakers because multiple peaks in the MUSIC spectrum cannot be distinguished when the occur close together.

In this paper, we introduce a new method, based on the cross-correlation of the space-spectrum calculated by MUSIC, as a solution to the problem of resolving adjacent speaker positions.

First, the problem of estimating sound source positions using the MUSIC method is described. We then describe the details of the proposed method. Finally, we present experiments on the estimation of the position of multiple adjacent speakers in the real environment and discuss the results.

2. CONVENTIONAL ESTIMATION OF SOUND SOURCE POSITION BY MUSIC

2.1. Calculation of MUSIC spectrum

This section explains how to apply MUSIC to estimate sound source positions by the conventional method.

A correlation matrix of the elements in the microphone array is represented by $R(n,k) = E[\mathbf{x}(n,k) \cdot \mathbf{x}^T(n,k)]$, $\mathbf{x}(n,k) = [X_1(n,k) \cdots X_M(n,k)]$, where $X_m(n,k)$ denotes short time Fourier transformation coefficients at microphone m, frame n, and discrete frequency k. M is the number of elements. Decomposing the correlation matrix R to eigenvectors, $R\mathbf{e}_i = \lambda_i \mathbf{e}_i (i = 1, \ldots, M)$, there is a relationship that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L > \lambda_{L+1} = \cdots = \lambda_M = \sigma^2$($\sigma^2$:power of thermal noise). Finally, the number of sound sources L is estimated from the eigenvalues using AIC.[3]

Since $\mathcal{S} = span\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_L\}$ is a signal subspace and $\mathcal{N} = span\{\mathbf{e}_{L+1}, \mathbf{e}_{L+2}, \cdots, \mathbf{e}_M\}$ is a noise subspace, if the correlation matrix is connected with a sound from position (r, θ), there is an orthogonal relationship between $\mathcal{N}$ and mode vector $\mathbf{a}_k(r, \theta)$.

Mode vector $\mathbf{a}_k(r, \theta)$ is defined as $\mathbf{a}_k(r, \theta) = [a_{1,k}(r, \theta), a_{2,k}(r, \theta), \cdots, a_{M,k}(r, \theta)]^T$, where $a_{m,k}(r, \theta)$ is the transform function from position (r, θ) to a microphone element m at discrete frequency k. From these definitions, the MUSIC spectrum for a given sound source position is expressed as

$$P_{music}(k, r, \theta) = \frac{\mathbf{a}_k^H(r, \theta)\mathbf{a}_k(r, \theta)}{\mathbf{a}_k^H(r, \theta) E_N E_N^H \mathbf{a}_k(r, \theta)} \quad (1)$$

$$E_N \stackrel{\text{def}}{=} [\mathbf{e}_{L+1}, \mathbf{e}_{L+2}, \cdots, \mathbf{e}_M]$$

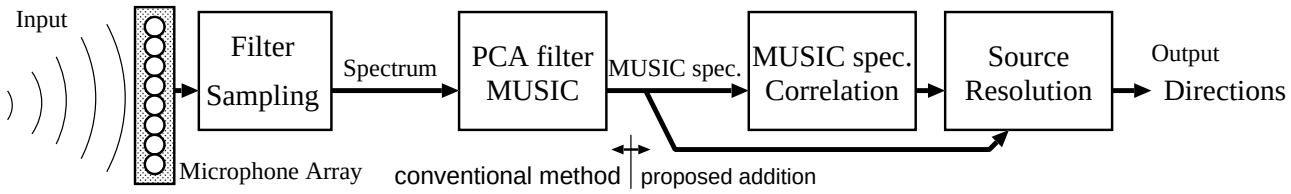

Figure 2: the proposed method

2.2. Search Sound Source

This section describes a method of estimating the position of the sound source using the MUSIC spectrum calculated in section 2.1. The MUSIC spectra of each discrete frequency are summed using the following equation:

$$P_{normal}(\theta) = \sum_{\theta} \sum_{k} P_{music}(r, \theta) \tag{2}$$

After taking the sum, a spectrum such as that in Fig. 1 can be obtained. The sound source position is estimated by extracting the peak of the MUSIC spectrum.

It is easy to find the peak of each speaker using the conventional method when the speakers are significantly separated. Two peaks can easily be resolved in the example '0 and 60 [deg]' (two speakers, at 0 and 60 degrees, respectively). However, when the two speakers are close together, it is difficult to find clear peaks for the corresponding speakers in the MUSIC spectrum. Only one peak can be resolved in the example '0 and 10 [deg]' in Fig. 1. This particular example shows that estimating the position of multiple adjacent speaker is very difficult by the conventional method.

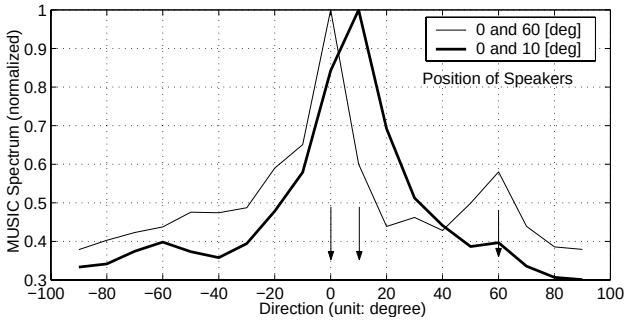

Figure 1: Music Spectra $P_{normal}(\theta)$ for two men's voices

3. PROPOSED METHOD

In this section, we explain the proposed method for estimating the positions of tightly grouped speakers. The method is based on the fact that each sound source has a different pattern of harmonics. An overview of the method is shown in Fig. 2.

3.1. Distinction by Harmonics

It is reasonable to assume that the spectral structures of the voice of two different persons differ. Therefore, if two MUSIC spectra taken in different directions exhibit strong similarities, we can determine that only one speaker exists. Differences in the MUSIC spectra taken in two directions will be an indication of the existence of more than one sound source.

3.2. Harmonics Structure of MUSIC Spectrum

Figure 3 shows an example of MUSIC spectra taken in the directions of two adjacent positions, $A = (r, \theta)$ and $B = (r, \theta')$ $\theta \simeq \theta'$, for one speaker's voice. Position A is the real location of the speaker, and there is no sound source at position B. Each line in the figure is the MUSIC spectrum for the corresponding position. This example shows that the MUSIC spectra for adjacent positions are very similar in the case of a single sound source.

Figure 4 shows an example of MUSIC spectra for two adjacent locations for two speakers. The speakers stood at position A and B. This example shows that the MUSIC spectra are significantly dissimilar, even when the positions of multiple sound sources are very close.

3.3. Correlation of Music Spectrum

In this experiment, we focus on sound source direction detection. The MUSIC spectrum in a given direction is given by

$$P(k, \theta) = \sum_{r} P_{music}(k, r, \theta) \tag{3}$$

To decide whether two adjacent spectra are attributable the same source, our method uses a correlation between the MUSIC spectra of two positions. The correlation is defined as Eq. (4); α_m and α_n are coefficients for the normalization of $C(\theta_m, \theta_n)$. As microphone arrays perform better at higher frequencies, the correlation is calculated using $\pi/2$ to π of the discrete normalized frequency k. The determination of this range depends on the shape of the array. If the spectra for different directions θ_m and θ_n are strongly affected by the one speaker's voice, $C(\theta_m, \theta_n)$ is closer to 1. If not, $C(\theta_m, \theta_n)$ is closer to 0.

$$C(\theta_m, \theta_n) = \frac{1}{\alpha_m \cdot \alpha_n} \sum_{k=\frac{1}{2}\pi}^{\pi} P(k, \theta_m) P(k, \theta_n) \tag{4}$$

$$C(\theta_m, \theta_m) = 1 \quad \Rightarrow \quad \alpha_m = \sqrt{\sum_{k=\frac{1}{2}\pi}^{\pi} P(k, \theta_m)^2} \tag{5}$$

3.4. Score of Sound Source

The correlations discussed in the previous section are utilized for the detection of the differences caused by different sound sources. Thus, the correlations are not a criteria for whether a sound exists or not. The MUSIC spectrum peaks at positions of sound sources. The low value of the correlation implies that the spectra are from different sound sources. Inverting the correlation to $1 - C(\theta_m, \theta_n)$, the score

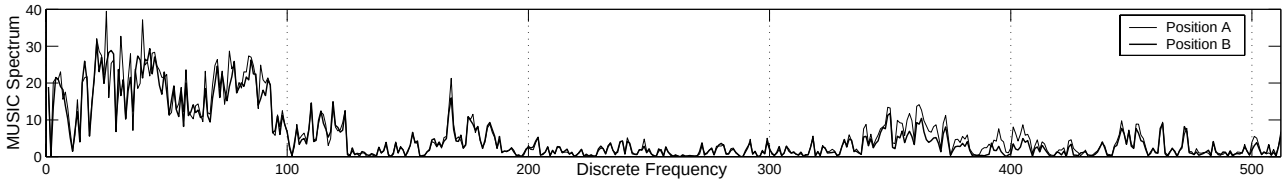

Figure 3: MUSIC Harmonics (One Voice)

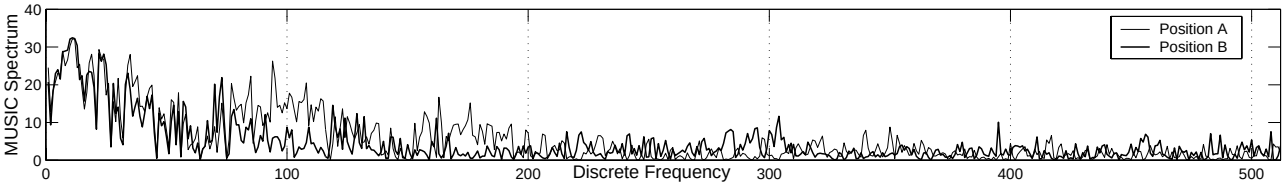

Figure 4: MUSIC Harmonics (Two Voices)

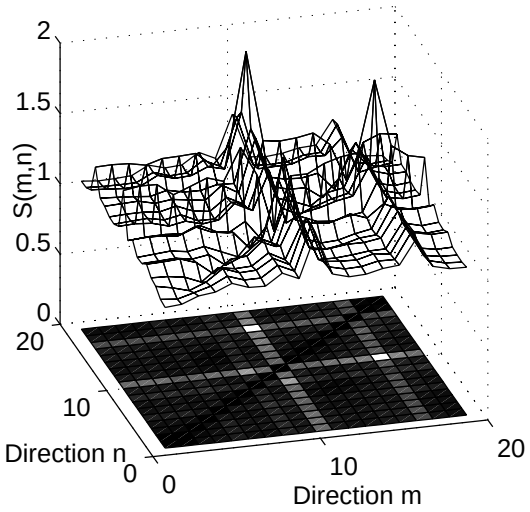

Figure 5: S(m,n) (two separated voices)

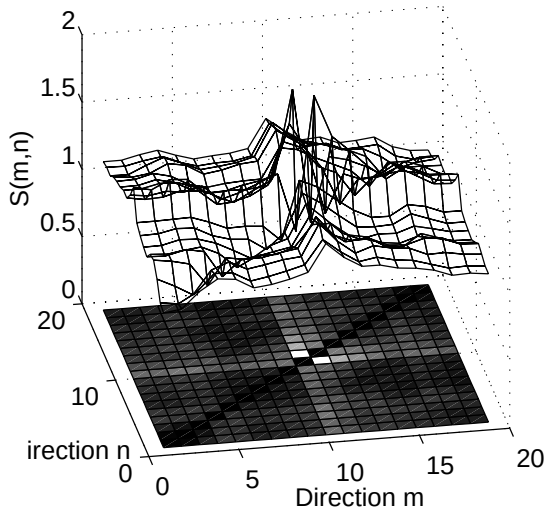

Figure 6: S(m,n) (two closed voices)

of the sound source is a product of the inverted correlation and the MUSIC spectrum, i.e.

$$S(\theta_m, \theta_n) = (1 - C(\theta_m, \theta_n)) \cdot P_{normal}(\theta_m) P_{normal}(\theta_n) \quad (6)$$

Scores calculated using Eq. (6) are shown in Figs. 5 and 6. $S(\theta_m, \theta_n)$ is a symmetric matrix because $S(\theta_m, \theta_n)$ is calculated from the correlation. If the directions θ_m and θ_n are the directions of sound sources, $S(\theta_m, \theta_n)$ becomes large. If the rows and columns are the directions of a sound source, the row and column of a sound source have higher values compared with the other elements of S. In the case of Fig. 6, $S(\theta_m, \theta_n)$ has maximum values at $m = 10$ ($\theta_m : 0$ [deg]), $n = 11$ ($\theta_n : 10$ [deg]), showing that sound sources exist at 0 [deg] and 10 [deg].

3.5. Estimation of Direction

The algorithm for estimating the direction of a sound source is as follows: Θ_s and Θ_c are the selected directions and the directions of candidates, respectively.

1. Set $\Theta_s = \{\}$, $\Theta_c = \Theta = \{\theta_1, \theta_2, \cdots, \theta_N\}$, where N is the number of directions.
2. Select θ_{first} from Θ_c and move it from Θ_c to Θ_s. $\theta_{first} = \arg\max_{\theta_m} \sum_{\theta_n \in \Theta} S(\theta_m, \theta_n)$, $\theta_m \in \Theta_c$
3. $\theta_p \leftarrow \theta_{first}$
4. Select a direction θ_n that satisfies the following from Θ_c and move it from Θ_c to Θ_s.
   - $\theta_n = \arg\max_{\theta_m} S(\theta_m, \theta_p)$, $\theta_m \in \Theta_c$
   - $S(\theta_n, \theta_p) \geq mean(S(\theta_m, \theta_p))$, $\theta_m \in \Theta$
   - $P_{normal}(\theta_n) \geq mean(P_{normal}(\theta))$, $\theta \in \Theta$
   - $\forall \theta \in \Theta_s$, $S(\theta_n, \theta_p) \geq mean(S(\theta_m, \theta))$, $\theta_m \in \Theta$
5. $\theta_p \leftarrow \theta_n$
6. Repeat from step 4 until the direction θ_n is found.
7. Elements of Θ_s are the estimated directions of the frame. Repeat $1 \sim 6$ for each frame.
8. Smooth the result. Cancel directions that are not selected for more than 0.05 seconds.

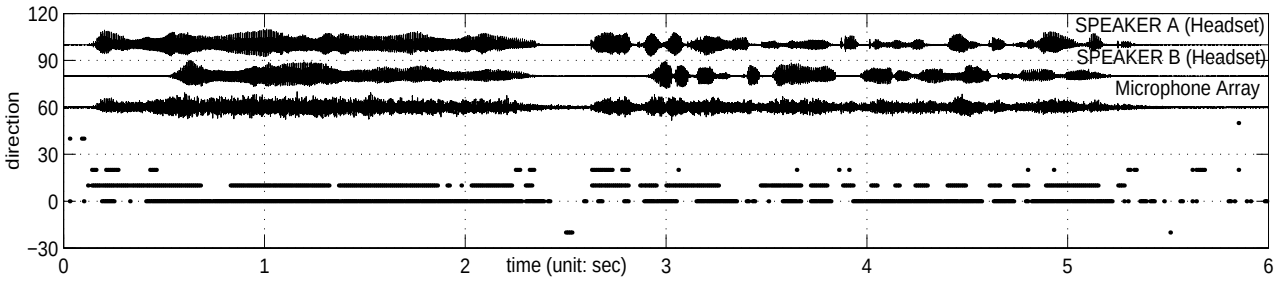

Figure 8: Estimated position of speakers (positions of speakers are 10 [deg] and 0 [deg])

Table 1: Conditions of microphone array and experiment

array form	linear and consistent same spacing
	8 elements spacing 3cm
element	non-directional condenser microphone
sampling	32kHz,16bit
frame length	1024 samples(32ms)
	hamming window
frame shift	320 samples: 100 frames per second
voice	2 men's voice
position	shown at Fig. 7
distance	60,30,20 and 10 degree
mode vector	65536 point measured with TSP signal[4]
	impulse length 1024 samples

4. EXPERIMENT AND RESULTS

4.1. Conditions

The arrangement of the microphone array and the conditions of this experiment are shown in Table 1. The mode vectors were measured at 209 points, in as are from -90∼90 [deg] at 10 [deg] intervals, and with speaker separations of 50∼150 [cm] in 10 [cm] intervals. The samples were measured using TSP (Time Stretched Pulse). After measuring, the effect of reverberation was removed. The utterances of the speakers in the experiment were Japanese sentences.

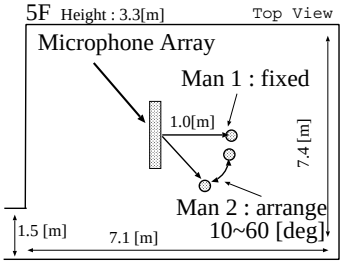

Figure 7: Microphone array and source positions

4.2. Results

The results of estimating the sound source directions using the proposed method are shown in Fig. 8. The figure shows the sound sources as three waves. The upper and middle waves are the voice of two speakers recorded with headsets, and the lower wave is the wave recorded by an element of the microphone array.

When there is no speaking, ambient and extraneous sounds are detected and included in the estimation process, causing the result to be unstable. However, when the men are speaking, the directions were estimated accurately. The number of estimations for each sound source direction in the recording are shown in Table 2.

Table 2: Precision of estimation of speaker direction

Distance [deg]	Total [frame]	Correct [frame]	Wrong [frame]	Precision [%]
10	827	766	61	92.62
20	654	634	20	96.94
30	700	628	72	89.71
60	724	681	43	94.06

5. CONCLUSION

We have proposed a practical method for estimating speaker direction using a correlation between MUSIC spectra. Experimental results show that the proposed method is effective enough to estimate the position of multiple speakers even if they are located closely together.

The proposed method can be applied not only to straight arrays, but also to other types of arrays. Here, the mode vectors are reusable, in contrast to root-MUSIC. Our future work will include applying the proposed method to speech enhancement and speech recognition in hands-free environments.

REFERENCES

[1] J.L. Flanagan, J.D. Johnston, R. Zahn, G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", J. Acoust. Soc. Am. 78 (5), pp. 1508-1518, 1985.

[2] R.O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation", IEEE Trans., vol. AP-34, No. 3, pp. 276-280, 1986.

[3] H. Wang et al., "Coherent Signal-Subspace Processing for the Detection and Estimation of Angles of Arrival of Multiple Wide-Band Sources", IEEE, ASSP-33 No. 4, pp. 823-831, 1985.

[4] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", J. Acoust. Soc. Am. Vol. 97 (2), pp. 1119-1123, 1995