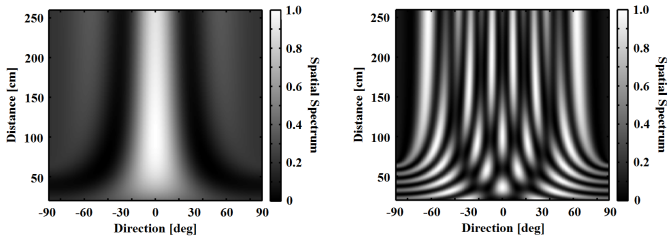
Suitable spatial resolution at frequency bands based on variances of phase differences for real-time talker localization

Kohei Hayashida*, Masato Nakayama†, Takanobu Nishiura† and Yoichi Yamashita†
*Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan.
E-mail: cm012063@ed.ritsumei.ac.jp Tel: +8177-561-5075
†Colleage of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan
Tel: +8177-561-5075

***Abstract*—Conventional near-field talker localization methods with microphone-array calculate spatial spectrum in each scanning position of discretized space and each frequency. Hence, elapsed time is increased and real-time processing is difficult. Real-time processing is important for achieving the natural interaction with the speech interfaces. To overcome this problem, we newly propose a talker localization method based on Multi-resolution Scanning in Frequency Domain (MSFD). MSFD utilizes lower spatial resolution in the lower frequency band and higher spatial resolution in the higher frequency band to reduce elapsed time. We also propose a calculation method for suitable spatial resolution at each frequency on the basis of the variances of phase differences among microphones. The results of evaluation experiment indicated that the proposed MSFD could reduce the elapsed time without degrading the localization accuracy.**

I. Introduction

A microphone-array is effective item at capturing distant-talking speech with high quality in noisy environments. It captures the target speech by steering the directivity on the basis of the direction of the target talker. Techniques have been developed not only to estimate the talker directions in the far-field but also to localize the talker positions in the near-field [1], [2], [3], [4]. The talker localization in the near-field is especially important for developing useful speech interfaces such as speech-controlled machines, humanoid robots, acoustic surveillance systems, and so on. Sound source localization methods based on Time Delay Of Arrival (TDOA) estimation [5], [6] have already been developed [1], [2]. These methods could localize sound source with fewer elapsed time, and real-time processing is realized easily. However, the localization accuracy depends on the estimation accuracy for the TDOA at each paired microphone. For the accurately sound source localization in near field, Two-Dimensional MUltiple Signal Classification (2D-MUSIC) [3] has already been developed. However, 2D-MUSIC needs more elapsed time for calculating spatial spectrum in each scanning position and each frequency, and the localization accuracy of this method is degraded when the number of microphones is few. In our former research, we proposed two-dimensional cross-power spectrum phase analysis with multiple microphones (multiple channel 2D-CSP) [4] for realizing accurately sound source localization with fewer elapsed time and fewer number of microphones. Multiple channel 2D-CSP localizes sound source more accurately than TDOA based method and 2D-MUSIC, and the elapsed time of this method is less than that of 2D-MUSIC [4]. However, it is still difficult to realize real-time processing with multiple channel 2D-CSP because it calculates spatial spectrum in each scanning position of discretized space and each frequency. Real-time processing is important for achieving the natural interaction with the speech interfaces. To overcome this problem, we newly propose a talker localization method based on Multi-resolution Scanning in Frequency Domain (MSFD). We also propose a calculation method for suitable spatial resolution at each frequency on the basis of the variances of phase differences among microphones. MSFD utilizes lower spatial resolution in the lower frequency band and higher spatial resolution in the higher frequency band. Variances of phase differences among microphones based on changing sound source location are small in the lower frequency band and large in higher frequency band. Therefore, the proposed method can reduce the elapsed time without degrading the localization accuracy by reducing the number of scanning positions in the lower frequency band. We aim to reduce the elapsed time without degrading the localization accuracy with the proposed method.



(a) Lower frequency (b) Higher frequency

Fig. 1. Spatial spectra in each frequency (source location $(\theta, r) = (0, 100)$)

II. Multi-resolution scanning in frequency domain for proposed method

Variances of phase differences between microphones based on changing sound source location are small in the lower

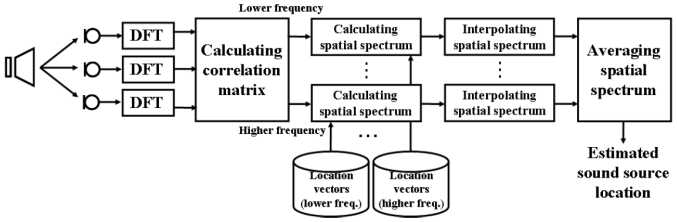
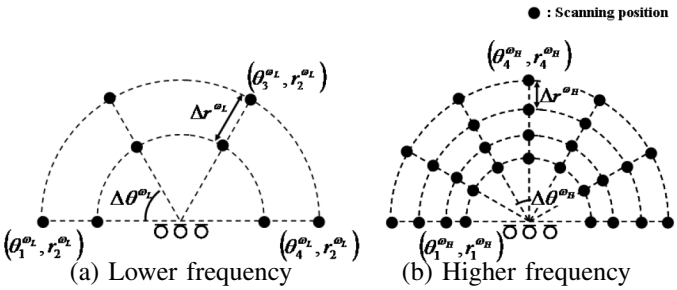


Fig. 2. Processing flow of the proposed multi-resolution scanning



(a) Lower frequency (b) Higher frequency

Fig. 3. Spatial discretization of the proposed method

frequency band and large in the higher frequency band. These differences appear in a peak width of spatial spectra. Figure 1 shows spatial spectra in each frequency. Spatial spectrum in the lower frequency band (Fig. 1 (a)) has a wider peak. On the other hand, spatial spectrum in the higher frequency band (Fig. 1 (b)) has a narrower peak. Therefore, we utilize a peak width of spatial spectra at each frequency for deciding spatial resolutions at each frequency. Figures 2 and 3 show the processing flow in the sound source localization for the proposed method and the spatial discretization of the proposed method. The proposed method estimates a talker location by using lower spatial resolution in the lower frequency band and higher spatial resolution in the higher frequency band to reduce elapsed time without degrading localization accuracy.

1) Calculating spatial resolutions in each frequency

For the direction, variances of phase difference among microphones are maximized in the direction of the center of a microphone-array at faraway sound source location. Hence, the proposed method calculates the spatial resolution for the direction on the basis of far-field assumption. At first, the spatial spectrum $P_\omega(\theta)$ at discrete frequency ω and direction θ is calculated with the algorithm for using the sound source localization. Then, peak width $\bar{\theta}_\omega$ is calculated from $P_\omega(\theta)$ and spatial resolution for the direction $\Delta\theta^\omega$ is determined by Eq. (1).

$$\Delta\theta^\omega \quad = \quad \alpha\bar{\theta}_\omega. \tag{1}$$

The symbol α denotes coefficient for calculating spatial resolution for the direction.

Also, for the distance, variances of phase difference among microphones are maximized in the center of a microphone-array at nearby sound source location. Hence, the proposed method calculates the spatial resolution for the distance on the basis of near-field assumption. At first, the spatial spectrum $P_\omega(r)$ at discrete frequency ω and distance r is calculated with the algorithm for using the sound source localization. Then, the proposed method calculates a peak width $\bar{r}_\omega$ on the basis of $P_\omega(r)$, and the scanning resolution for distance Δr^ω is derived from Eq. (2).

$$\Delta r^\omega \quad = \quad \beta\bar{r}_\omega. \tag{2}$$

The symbol β denotes coefficient for calculating spatial resolution for the distance.

The proposed method performs these calculations before the sound source localization. The proposed method localizes sound source with spatial resolutions for the direction and the distance that were calculated in Step 1).

2) Calculating spatial spectra

The spatial spectrum $P_\omega(\theta^\omega, r^\omega)$ is calculated at each frequency ω. The symbols θ^ω and r^ω denote scanning direction and scanning distance at frequency ω. These are derived from Eqs. (3) and (4).

$$\theta^\omega \in \theta_1^\omega, \cdots, \theta_{N_\theta^\omega}^\omega \ (\theta_i^\omega = \theta_1^\omega + (i-1)\Delta\theta^\omega), \tag{3}$$

$$r^\omega \in r_1^\omega, \cdots, r_{N_r^\omega}^\omega \ (r_i^\omega = r_1^\omega + (i-1)\Delta r^\omega). \tag{4}$$

The symbol N_θ^ω denotes the number of scanning directions at frequency ω, N_r^ω the number of scanning distances at frequency ω.

3) Interpolating spatial spectra

Since spatial resolutions of the spatial spectra differ between each frequency, the proposed method cannot calculate the correct averaged spatial spectrum. Therefore, we utilize the cubic spline interpolation [7] to generalize the spatial resolution of spatial spectrum.

4) Calculating averaged spatial spectrum

The averaged spatial spectrum $\overline{P(\theta, r)}$ is calculated from Eq. (5) on the basis of interpolated spatial spectra $P_\omega^{'}(\theta, r)$.

$$\overline{P(\theta, r)} \quad = \quad \sum_{\omega=\omega_L}^{\omega_H} P_\omega^{'}(\theta, r)/(\omega_H - \omega_L + 1). \tag{5}$$

The symbols $[\omega_L, \omega_H]$ are indices for the lower and upper bounds of the frequency range, and (θ, r) denote the scanning direction and the scanning distance after the interpolation.

5) Estimating sound source location

The estimated sound source location $(\hat{\theta}, \hat{r})$ is determined from Eq. (6).

$$(\hat{\theta}, \hat{r}) = \underset{(\theta,r)}{\operatorname{argmax}}(\overline{P(\theta, r)}). \tag{6}$$

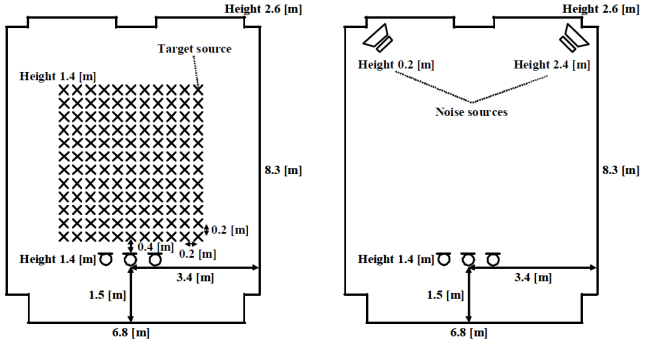
III. Evaluation experiments

A. Experimental conditions

We carried out the evaluation experiments in a real noisy environment to evaluate the localization accuracy and the elapsed time of the conventional and proposed methods. Table I shows the experimental conditions. Figure 4 (a) shows the placement of the sound sources and three microphones in a conference room. There were 132 different sound source positions, each located 0.2 [m] apart from another. We used a mouth simulator as loudspeaker to simulate the radiation characteristics of someone speaking. Figure 4 (b) shows the placement of noise sources and the three microphones. We used two loudspeakers

TABLE I
EXPERIMENTAL CONDITIONS.

Environments	Conference room ($T_{[60]} = 0.4$ [s])
Number of microphones	3 mics.
Distance between microphones	0.3 [m]
Sound sources	Speech signals (2 [speaker] * 10 [word])
Noise source	White noise
Signal to Noise Ratio	10 and 20 [dB]
Frequency range for localization	0.3 $\sim$ 3.3 [kHz]
FFT length	1024 [sample]
Frame length	512 [sample]
Frame shift	128 [sample]
Number of frames	16 [frame]
Signal length	2432 [sample] (0.152 [sec])
Scanning direction	-90 $\sim$ 90 [deg]
Scanning distance	0.2 $\sim$ 2.6 [m]
Sampling frequency	16 [kHz]
Quantization	16 [bit]



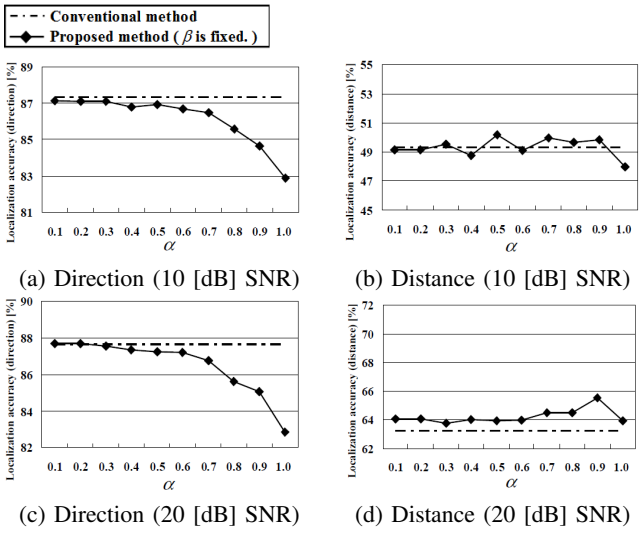
(a) Speaker and mics. location. (b) Diffused noise and mics. location.

Fig. 4. Configurations in experimental environment.

to create the diffused noise environment. One loudspeaker, placed at 0.2 [m] off the ground, created a diffused noise environment by emitting toward the floor, and the other, placed at 2.4 [m] off the ground, also did the same by emitting toward the ceiling. We designed evaluation signals by adding speech and noise in 10 and 20 [dB] SNR conditions.

We evaluated the conventional and proposed methods by the localization accuracy and the elapsed time. The localization accuracy is evaluated by the estimation accuracy for the direction and the distance. The estimation accuracy for the direction represents the number of sound sources localized within tolerance, which was 3 [degree]. The estimation accuracy for the distance represents the number of sound sources localized within tolerance, which was 0.2 [m].

In this experiment, the spatial spectra were calculated by cross-power spectrum phase analysis with multiple microphones [4]. The frequency range for sound source localization was 0.3 $\sim$ 3.3 [kHz], which are the dominant frequencies of the human voice. We used a laptop PC for the evaluation with Core-i5 2.67 GHz CPU and 4 Gbytes of memory. The conventional and the proposed methods were implemented with Matlab R2010b. For the conventional method, the spatial resolution for the direction $\Delta\theta$ is 1.0 [deg], and that for the distance Δr is 0.1 [m]. The conventional method utilized the same spatial resolution in all frequency bands. On the other hand, the proposed method utilized different spatial resolutions in the frequency domain. We evaluated the proposed method in two steps to determine suitable α and β in Eqs. (1) and (2). We first determined a suitable value of α by fixed β and changing α. Next, we determined a suitable value of β by fixed α and changing β.



(a) Direction (10 [dB] SNR) (b) Distance (10 [dB] SNR)

(c) Direction (20 [dB] SNR) (d) Distance (20 [dB] SNR)

Fig. 5. Localization accuracy (fixed β).

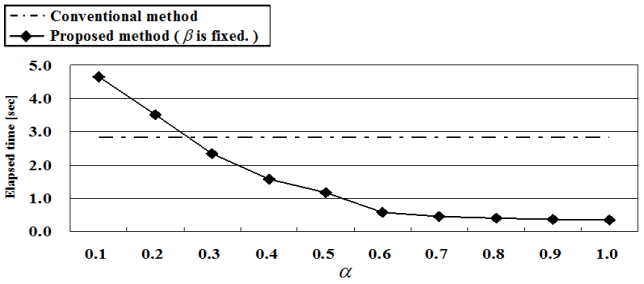


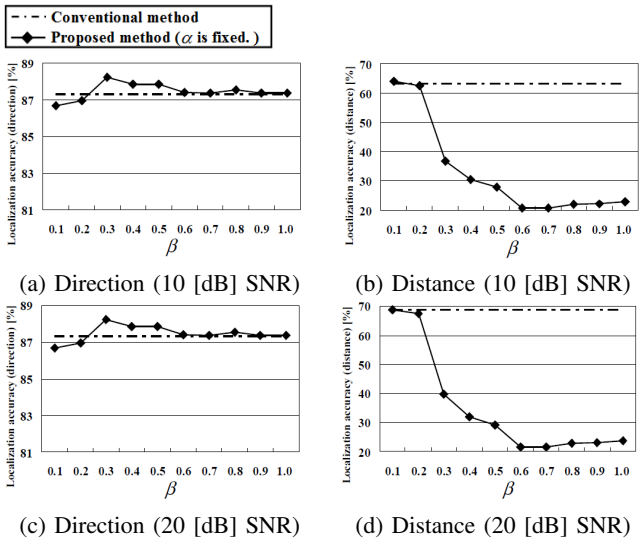
Fig. 6. Elapsed time (fixed β).

B. Experimental results for localization with fixed β

For the proposed method in this section, α changes from 0.1 through 1.0 in steps of 0.1, and β is always fixed in 0.1. The proposed method utilized different spatial resolutions at each frequency band. Figure 5 shows the localization accuracy in 10 and 20 [dB] SNR conditions. In Fig. 5 (a) and (c), the estimation accuracies for the direction of the proposed method with fixed β are nearly the same as those of the conventional method when α is less than 0.5. Figure 6 shows elapsed times of the conventional and the proposed methods. The proposed method ($\alpha = 0.5$) could reduce the elapsed time by 59 % more than the conventional method. we determined the suitable α is 0.5 and utilized it for the next evaluation.

C. Experimental results for localization with fixed α

For the proposed method in this section, β changes from 0.1 through 1.0 in steps of 0.1, and α is always fixed in 0.5. Figure 7 shows the localization accuracy in 10 and 20 [dB] SNR conditions. In Fig. 7 (b) and (d), the estimation accuracies for the distance of the proposed method are nearly



(a) Direction (10 [dB] SNR) (b) Distance (10 [dB] SNR)

(c) Direction (20 [dB] SNR) (d) Distance (20 [dB] SNR)

Fig. 7. Localization accuracy (fixed α).

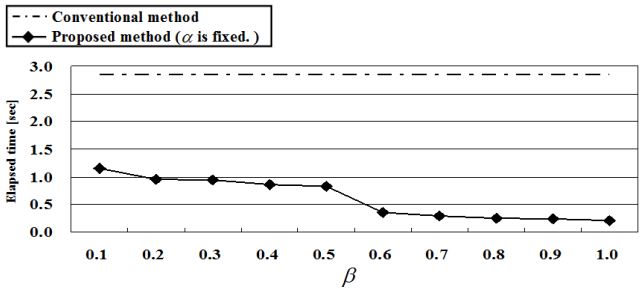


Fig. 8. Elapsed time (fixed α).

the same as those of the conventional method when β is less than 0.2. Figure 8 shows elapsed time of the conventional and the proposed methods. The proposed method ($\beta = 0.2$) could reduce the elapsed time by 66 % more than the conventional method. The results of evaluation experiments indicated that the proposed method could reduce the elapsed time without degrading the localization accuracy.

D. Discussions

From the result of Figs. 5 and 7, we determined that the suitable α and β are 0.5 and 0.2, respectively. This means the suitable spatial resolution for the distance is higher than that for the direction. This is caused by variances of phase differences between microphones based on changing sound source location. Variances of phase differences between microphones based on changing sound source location are small in the distance and large in the direction. Therefore, the error by the interpolation in lower spatial resolution for distance is larger than that in lower spatial resolution for direction. From these facts, the lower spatial resolution for direction more effectively reduces the elapsed time without degrading the localization accuracy than the lower spatial resolution for distance.

Figure 5 (a) and (c) indicate similar tendencies for the decrease in estimation accuracy with the increase in α. Figure 7 (b) and (d) also indicate similar tendencies for the decrease in estimation accuracy with the increase in β. From these facts, we can confirm that suitable values for α and β do not depend on SNR. Therefore, the suitable values for α and β can be calculated by the simulation that assumes a clean environment without noise. Moreover, they can be calculated when the number and the placement of microphone-arrays are decided. In conclusion, these results indicated that the proposed method could completely determine suitable values for α and β before the talker localization in a real environment. Hence, the proposed method effectively reduces the elapsed time without degrading the localization performance.

IV. Conclusions

For achieving useful speech interface, sound source localization should be finished in real-time. In this study, we propose a localization method based on multi-resolution scanning in frequency domain. The proposed method localizes the sound source by using lower spatial resolution in the lower frequency band and higher spatial resolution in the higher frequency band to reduce elapsed time. The results of evaluation experiments in diffused noisy environments indicated that the proposed method effectively reduces elapsed time without degrading localization performance. In future work, we intend to evaluate the proposed method under various kinds of noise environment with multiple sound sources. Also, we will try to improve the localization performance.

V. Acknowledgements

This work was partly supported by Grants-in-Aid for Scientific Research funded by The Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

[1] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan and M.H. Bianchi, "A DSP implementation of source location using microphone arrays", *Proc. SPIE*, vol. 2846, pp. 88–99, 1996.

[2] J.M. Valin, F. Michaud and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filterling", *Robotics and Autonomous Systems*, vol. 55, pp. 216–228, 2007.

[3] F. Asano, H. Asoh and T. Matsui, "Sound source localization and separation in near field," *IEICE Trans. Fundamentals*, vol. E83-A, no. 11, pp. 2286–2294, 2000.

[4] K. Hayashida, M. Morise and T. Nishiura, "Near field sound source localization based on cross-power spectrum phase analysis with multiple channel microphones", *Proc. INTERSPEECH 2010*, pp. 2758–2761, 2010.

[5] C. H. Kanapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[6] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. ICASSP94*, pp. 273–276, 1994.

[7] R.W. Schager and L.R. Rabiner, "A digital signal processing approach to interpolation", *Proc. IEEE*, vol. 61, pp. 692–702, 1973.