ESTIMATING SOUND SOURCE DEPTH USING A SMALL-SIZE ARRAY

Satoshi Esaki†, Kenta Niwa†‡, Takanori Nishino§ and Kazuya Takeda†

† Graduate School of Information Science, Nagoya University, Nagoya, Aichi, Japan 464-8603
‡ NTT Cyber Space Laboratories, NTT Corporation, Tokyo, Japan 180-8585
§ Graduate School of Engineering, Mie University, Tsu, Mie, Japan 514-8507

ABSTRACT

A method for estimating the sound source depth, i.e., the distance between a source and receiver, using a small-size array is proposed. The proposed method uses the spatial distribution pattern of quasi-independent signal components obtained by the frequency-domain independent component analysis (FDICA) as the cue for depth estimation. The quasi-independent components are calculated by applying FDICA to array signals with very high redundancy, for example, 60 microphone signals for a pair of sources; therefore, signal components associated with reflection signals are obtained even though they are correlated with the direct signal. Experimental evaluation using a small-size microphone array with a large number of elements confirms that the average (RMS) estimation error of the proposed method is 0.33 m, which is sufficiently accurate for our applications.

Index Terms— Sound source depth, DOA, Virtual sources, Frequency-domain independent component analysis, Selective listening point audio

1. INTRODUCTION

With the rapid spread of 3D video content in the market, related technologies have become important fields in acoustic signal processing. The authors have been developing a selective listening point (SLP) audio system, which reproduces a sound field at a position specified by the user such as a freeview point in TV [1]. In previous works, we proposed a framework for SLP audio and showed that the simultaneous estimation of source location and its signal (SESLS) is the key technology of the SLP audio system [2]. In our previous system, we applied frequency-domain independent component analysis (FDICA) to a large number of array signals so that an acoustic field is decomposed into a set of narrowband quasi-independent signals, which we called virtual sources. Then SESLS was carried out by clustering virtual sources over frequencies and direction of arrival (DOA).

In our previous system, we used seven distributed linear arrays enclosing all acoustic sources, so that source locations can be estimated by combining the DOA at each linear array. However, the use of such a large array system introduces several practical limitations. In this paper, we propose a method for estimating the distance between a sound source and receiver, hereafter referred to as the depth of the source, using a small-size array. The novelty of the proposed approach is the use of the spatial distribution pattern of virtual sound sources that are calculated as by-products of our SESLS algorithm.

Since acoustic source localization is a fundamental problem in various acoustic systems, many algorithms have been developed targeting several different conditions. The most fundamental issue in source localization is estimating the DOA using signals captured at a sensor array. MUSIC is one of the most general methods of DOA estimation [3]. Theoretically, MUSIC formulates DOA estimation as the problem of finding the optimal signal subspace of the correlation matrix calculated using the sensor signals.

On the other hand, little research on source depth estimation has been reported. A simple approach to source depth estimation is combining the DOAs at multiple array positions using trigonometry [4, 5]. A time difference of arrival (TDOA) approach for distributed sensor signals was proposed as a more general approach [6, 7, 8, 9]. When sensors can be placed at given positions so as to surround sound sources, these methods give reasonable results. When a near sound field is assumed, MUSIC can be extended to 2D-MUSIC, which operate with a single array [10]; however, to estimate the distance when the distance to the sound source is in 1-3 m, the size of the array must be more than 1 m.

As suggested in [11], the human auditory system can perceive the distance to a speaker using the power ratio between the direct and echo sounds as a cue. Under realistic room environments, echo signals should have relevant information on the source depth. However, very little research on utilizing echo signals for estimating the source depth has been carried out. The key idea in the proposed depth estimation method is to use the spatial distribution pattern of virtual sources that we used in our SESLS system as a cue for estimating the power ratio between direct and echo signals. The most important merit of the proposed method is that the source depth can be estimated using a small-size device with dimensions of less than 10 cm.

The rest of this paper is as follows. The basic idea of the estimating source depth in a realistic sound field is outlined in Section 2. The proposed algorithm and a few issues regarding its implementation are discussed in Section 3. Section 4 we report an experimental evaluation and its results under our typical target conditions: the size of the room is that of a standard classroom, each sound source is 0.5-3 m distant from the receiver, and there are a pair of sound sources (at a given time) that do not move or move slowly, for example, at less than 4 km/h. In Section 5, we summarize the results and clarify remaining issues.

2. ESTIMATING SOURCE DEPTH IN A REFLECTIVE FIELD

Figure 1 shows the basic idea of estimating the source depth using a small-size array (b), in contrast to using a conventional distributed array system (a). Under the assumption of an anechoic environment, the direct sound is the only cue for estimating the source depth and the location of the source is estimated by trigonometry involving the source and sensors. Therefore, distributed arrays or a single but large-size array must be used. On the other hand, under realistic conditions, contamination of the direct with echo signals is unavoidable.

Each echo signal has a different DOA at the array, and the DOA is determined by the positions of the sound source, the receiver, and reflectors.

Since the reflection sounds are highly correlated with each other as well as with the direct sound, estimating the DOA of echo signals is not a trivial problem. In this study, we use the distribution of quasi-independent sources obtained by frequency domain ICA (FDICA). We cannot expect that each reflection signal will be separated by FDICA, even using a large number of input signals, because of their correlations. However, in previous experiments on SESLS, we confirmed that most signal components in FDICA results represent synthetic sound sources of the direct sound and echo signals with the same time delay (Fig. 1). Since the DOA of a synthetic source is determined by the power ratio between the direct and reflection signals, the distribution of DOAs is more concentrated around the direct sound direction when the power of the direct sound becomes more dominant. Therefore, the distribution pattern of quasi-independent sources in the FDICA result can be an effective cue for depth estimation. Fig. 2 shows examples of the DOA distributions of quasi-independent sources at depths of 0.5 m and 3 m. From the figures, it can be seen that the DOA distribution is more concentrated in the source direction when the source depth is closer and that the distribution is an effective cue for depth estimation.

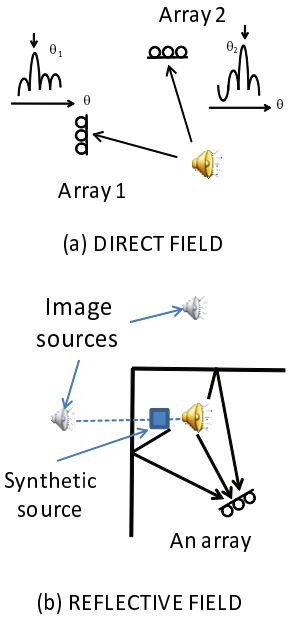

Fig. 1. Basic idea of utilizing reflection signal as the cue for depth estimation. Although the DOAs of image sources are determined by the source position, their estimation is not trivial. FDICA results do not correspond directly to the image sources but to *synthetic sources* of the direct and echo signals.

3. ALGORITHM

3.1. Estimating synthetic sources

In previous works on our SLP audio system, we developed SESLS algorithms that encode large array signals into a pair of quasi-independent source clusters and their DOA centroids. We estimated the synthetic source distribution using the SESLS algorithm. The basic algorithm can be summarized as follows:

- Acoustic signals are captured with a large (typically more than 40 elements), array system so that the correlations among microphones are calculated with sufficient redundancy.
- The linear dependence between different microphone signals is reduced by applying PCA to the spatial correlation matrix $R = E[X(k,i)X^{\mathrm{H}}(k,i)]$. At this stage, Q (typically 20) principal components are extracted.
- By applying FDICA to the Q principal signals, Q quasi-independent signals are estimated together with their corresponding ICA weight vectors in every frequency channel k ($k \leq K$), where K is the number of frequency channels. In other words, $Q \times K$ signal components and their associated ICA weight vectors, from which the DOA of each source can be calculated, are determined.

In previous works, we then clustered these $Q \times K$ synthetic sources into M clusters that are relevant in order to render and reproduce the acoustic field. Subjective evaluation results showed the effectiveness of applying this SESLS algorithm to SLP audio [2]. In this study, we calculate the directional distribution of the estimated synthetic sources, $f(\vec{x})$, as a cue for depth estimation. Here, the 2D variable $\vec{x} = (x_1, x_2)$ is the combination of the azimuth and elevation angles, denoted by x_1 and x_2, respectively.

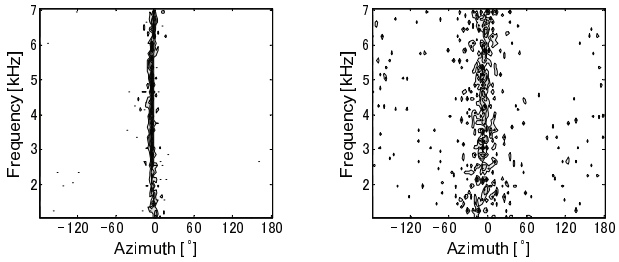

Fig. 2. Examples of the directional (azimuth) distributions of quasi-independent components of frequency channels measured for sources at depths of 0.5 m (left) and 3 m (right).

3.2. Modeling distribution of the synthetic sources

In general, the distribution of synthetic sources, $f(\vec{x})$, depends on the relative positions among the real source, the receiver, and reflectors; however, we assume that it depends only on the direction of the real source $\vec{x}_0$ and its depth r, i.e., $f(\vec{x}) = f(\vec{x}; \vec{x_0}, r)$. We further assume that the distribution is centered in the real source direction and that the source direction can be estimated by averaging the synthetic source directions. Thus, the distribution is represented by the centered variable $\vec{x} - \vec{\eta}_f$ and the function $f_c(\vec{x}; \vec{x}_0, r) = f(\vec{x} - \vec{\eta}_f; \vec{x}_0, r)$. Although determining the analytic form of the function is difficult, we can train distributions at some grid points of $\vec{x}_0$ and r using training data. Finally, the source depth is estimated by finding the closest distribution $f_c(\vec{x}; \vec{x}_0, r)$ to the distribution, $g_c(\vec{x}) = g(\vec{x} - \vec{\eta}_g)$ that is calculated from the input signals,

$$\hat{r} = \arg\min_r D(g_c(\vec{x}) \| f_c(\vec{x}; \vec{x}_0, r)).$$

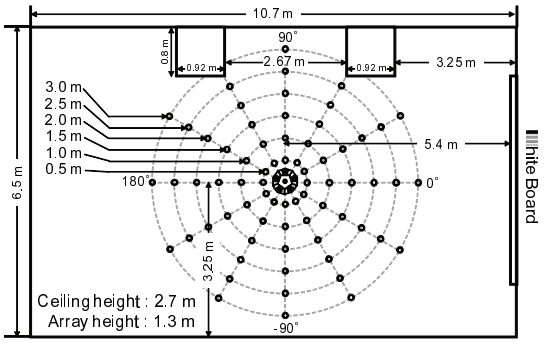

Fig. 3. Experimental setup of source and receiver positions. The receiver is located at the center of the room. Seventy source positions, shown by small circles, are tested. The room is a typical classroom for an audience of 40-70 with tile carpets on the floor. Desks and chairs were removed from the room.

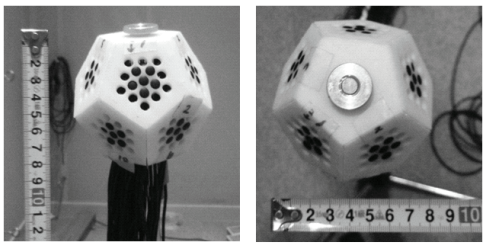

Fig. 4. Dodecahedral microphone array used for the experiment. Microphones can be installed on 10 faces, i.e., all faces except the top and bottom faces, and the maximum number of microphones is 160. Here, six microphones are installed around the center of each face.

4. EXPERIMENTAL EVALUATION

4.1. Experimental setup

Figure 3 shows the experimental setup of the source and receiver positions. A typical classroom whose T_{60} is 300 ms is used for the experiment. The dodecahedral microphone array system shown in Fig. 4 is used as a receiver. The microphone array system is small, 8 cm in diameter, but has a large number (60) of elements and is effective for real-world ICA applications [12]. As shown in Fig. 3, we measured a set of impulse responses from 72 source positions to all 60 elements of the receiver. Source positions are sampled within a circular area of 6 m diameter at intervals of 30° and 0.5 m. Since the receiver has no directivity, the variation across the source direction simulates differences in primary reflection conditions, for example, location relative to the whiteboard and columns. By convolving the impulse responses with the test utterances of 8 speakers, we obtained 33600 signals in total for use in the experiment. The signal sampling rate was 40 kHz and the STFT frame length and frame shift were 25.6 ms and 6.4 ms, respectively.

The distribution of the synthetic sources was calculated for 70 source positions using the test utterances. In this experiment, a 2D histogram f_{ij} is used to represent the distributions, where i and j are indices for the azimuth and elevation directions, respectively. For both directions, each bin is regularly separated 3°.

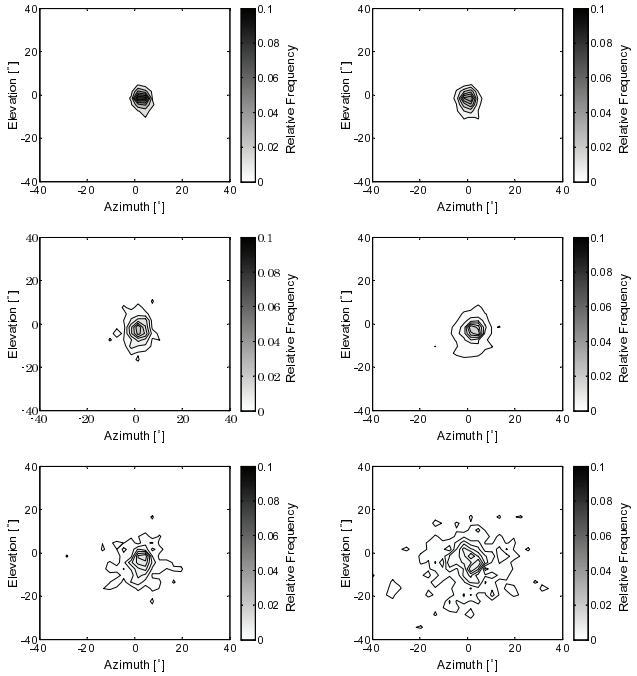

Fig. 5. Measured distributions of synthetic source locations at depths of 0.5 m, 1.0 m, 1.5 m, 2.0 m, 2.5 m and 3.0 m (from top left to bottom right). The source direction is 0°.

4.2. Distributions of the synthetic sources

Figure 5 shows measured distributions of the synthetic source locations for the same source direction of 0° but for different source depths. In the figure, the dispersion of the distribution is clearly observed when the source depth increases. When the source depth is 0.5 m, 25% of the synthetic sources are located in the same direction as the real source. However, when the source depth is 3 m, less than 0.3% of the sources are located in the same direction. From these results, it is predicted that the distribution pattern of synthetic sources can be used a relevant cue for depth estimation.

4.3. Depth estimation

Finally, depth estimation is implemented as a pattern classification problem. Seventy histograms, $f_{ij}(\theta, r)$, calculated using training data are used as templates of the distribution at the source position. Then, the histogram calculated for test data, g_{ij}, is compared against the templates using KL divergence as the similarity between histograms, i.e.,

$$D_{\mathrm{KL}}(f\|g) = \sum_{ij} f_{ij}(\theta, r) \log \frac{f_{ij}(\theta, r)}{g_{ij}} + \sum_{ij} g_{ij} \log \frac{g_{ij}}{f_{ij}(\theta, r)}.$$

The evaluation is performed on the basis of the leave-one-out paradigm using 8 speakers' utterances; therefore, 560 trial results are averaged. The experimental result is shown in Figure 6. Since there was no error in estimating the source direction θ using the average

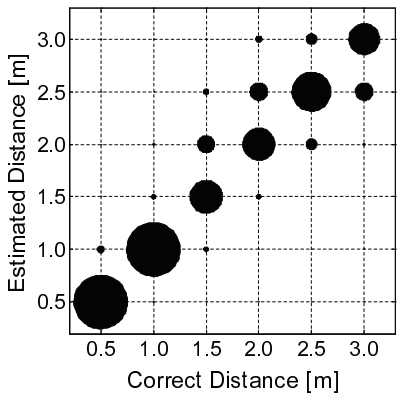

Fig. 6. Depth estimation accuracy plotted as a confusion matrix. The size of each circle shows the number of decisions. The average error in depth estimation (RMS) is 0.33 m.

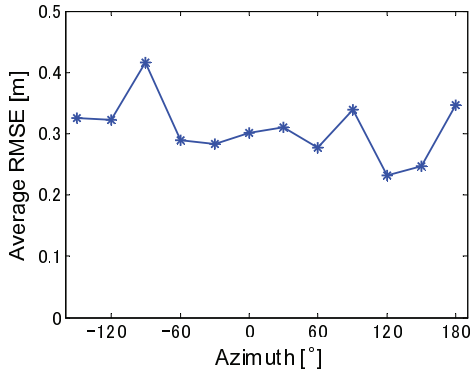

Fig. 7. Depth estimation error for different source angles. The performance is degraded when the distance to the nearest wall decreases. However, there is not a large deviation in the performance among directions.

direction of the synthetic sources, only depth accuracy is depicted. The average accuracy in discriminating the six depth classes, {0.5 m, 1 m, 1.5 m, 2 m, 2.5 m, 3 m}, was 66%. The average (RMS) error of the depth estimation is 0.33 m. This means that the location of a sound source located within a distance of 3 m can be estimated using a small-size array with less than 50 cm error, which is sufficient accuracy for our applications.

5. DISCUSSION

In this paper, we proposed a source depth estimation method for a small-size microphone array, which uses the distribution pattern of quasi-independent signal components that corresponds to the synthetic sources of direct and echo signals. Although the feasible accuracy, i.e., 0.33m RMS error, of the proposed depth estimation method was experimentally confirmed, there are several remaining issues regarding the generalization of the method. First, a theoretical analysis should be performed on the relationship between the synthetic source location and the correlation matrix of array signals. Second, we did not discuss the applications of the proposed method to multiple source conditions in the paper. We have obtained preliminary results under a two-source condition by fitting Gaussian mixture model (GMM) to the distribution. However, extending the proposed method to deal with an unknown number of sources is an important future work. The third issue is the robustness of the method to changes in acoustic conditions. Although we expect that the method works under different source/receiver arrangements, at least, as shown in Fig. 7, the performance is not sensitive to the source direction. However, it should be confirmed through careful experiments. Particularly, when the location of the receiver/source is very close to the wall, the behavior of the synthetic source location may be different from that of the reported experiments.

Acknowledgements: This work has been partially supported by the JST/CREST program.

6. REFERENCES

[1] M.P. Tehrani, K. Niwa, N. Fukushima, et al., "3DAV integrated system featuring arbitrary listening-point and viewpoint generation," *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pp. 855–860, 2008.

[2] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into a 3d sound field representation for selective listening point audio based on blind source separation," *ICASSP 2008*, pp. 181–184, 2008.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] M.S. Brandstein and H.F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

[5] P. Bergamo, S. Asgari, H. Wang, et al., "Collaborative sensor networking towards real-time acoustical beamforming in free-space and limited reverberance," *IEEE Trans. Mobile Computing*, vol. 3, no. 3, pp. 211–224, 2004.

[6] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J.Applied Sig. Proc.*, vol. 2006, pp. 170–170, 2006.

[7] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J.Applied Sig. Proc.*, vol. 2003, pp. 338–347, 2003.

[8] H. Do and H.F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," *ICASSP 2010*, pp. 125–128, 2010.

[9] P. Teng, A. Lombard, and W. Kellermann, "Disambiguation in multidimensional tracking of multiple acoustic sources using a gaussian likelihood criterion," *ICASSP 2010*, pp. 145–148, 2010.

[10] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and separation in near field," *IEICE Trans. Fundamentals*, vol. 83, no. 11, pp. 2286–2294, 2000.

[11] A.W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[12] M. Ogasawara, T. Nishino, and K. Takeda, "A small dodecahedral microphone array for blind source separation," *ICASSP 2010*, pp. 229–232, 2010.