STEPWISE PHASE DIFFERENCE RESTORATION METHOD FOR SOUND SOURCE LOCALIZATION USING MULTIPLE MICROPHONE PAIRS

Masahito Togami, Takashi Sumiyoshi, Akio Amano

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

We propose a new methodology of sound source localization named **SPIRE** (Stepwise Phase dIfference REstoration) that is able to localize sources even if they are neighboring in a reverberant environment. Localizing sound sources in reverberant environments is difficult, because the variance of the direction of an estimated sound source increases in reverberant environments. The major feature of our proposed method is restoration of a microphone pair's phase difference(M1) by using the phase difference of another microphone pair(M2) under the condition that the distance between M1's microphones is longer than the distance between M2's microphones. This restoration process makes it possible to reduce the variance of an estimated sound source direction and to solve the spatial aliasing problem that occurs with the M1's phase difference. The experimental results in a reverberant environment (reverberation time=about 300ms) indicate that our proposed method can localize sources even if they are neighboring (even if the difference in the sources' directions equals 10 degree).

Index Terms—

DOA estimation, Phase estimation, Acoustic applications, Acoustic arrays

1. INTRODUCTION

Sound source localization is an essential function for interactive robots, TV conference systems, or voice activity detectors on mobile phones. For these applications, more than two sources occasionally exist at the same time. Localization methods that are able to localize multiple sources are desirable for these applications. Conventional methods based on signal subspace, e.g., MUSIC, have been proposed[4]. These methods can localize sources even when there are multiple sources. However, the computational cost of these methods is proportional to the spatial resolution of localization, so localization at high-resolution is difficult. Sound source localization methods based on the assumption of sources' sparseness, e.g., DUET, have been proposed[1], and extensions of DUET have also been proposed, e.g., 360-degree localization [2], 3-dimensional localization [3]. The computational cost of these methods is independent of the spatial resolution of localization. The longer the distance is between a microphone pair, the higher the performance of these methods is in a reverberant environment. However, when the distance between a microphone pair exceeds a certain distance defined by the sources' maximum frequency (aliasing distance), these methods are not able to localize sources because of the spatial aliasing[5]. The distance between a microphone pair needs to be as long as possible, but the distance cannot be longer than the aliasing distance, so the upper bound of localization performance of these methods is restricted. In this paper, we propose a new methodology of sound source localization named SPIRE (Stepwise Phase dIfference REstoration) that uses multiple microphone pairs which have different distance each other. The upper bound of localization performance of our proposed method is restricted by the maximum distance. However, the maximum distance can be longer than the aliasing distance if at least one microphone-pair distance is shorter than the aliasing distance. The estimation of the azimuth is the issue addressed in this paper. Results for both a linear array and a nonlinear array are discussed.

2. PROBLEM STATEMENTS AND NOTATION

2.1. Mixing process

Let $\boldsymbol{X}(f,\tau) = [x_1(f,\tau), \ldots, x_M(f,\tau)]$ be the observed M-dimensional vector, $\boldsymbol{A}(f) = [\boldsymbol{a}_1(f), \ldots, \boldsymbol{a}_N(f)]$ the mixing ($M \times N$) matrix, $\boldsymbol{S}(f,\tau) = [s_1(f,\tau), \ldots s_1(f,\tau)]$ the source N-dimensional vector, where f is the frequency, and τ is the frame index. The mixing process is

$$\boldsymbol{X}(f,\tau) = \boldsymbol{A}(f)\boldsymbol{S}(f,\tau). \quad (1)$$

2.2. Assumption based on sparseness

When only one source is active at each time-frequency point, the observed vector $\boldsymbol{X}(f,\tau)$ can be approximated as follows:

$$\boldsymbol{X}(f,\tau) = \boldsymbol{a}_i(f)s_i(f,\tau), \quad (2)$$

where i is the index of the active source at frame τ and frequency f. This sparseness assumption is appropriate for human speech.

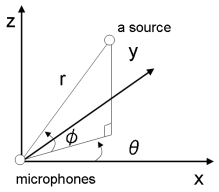

Fig. 1. Coordinate system

2.3. Phase difference

The phase difference of the j-th microphone pair is

$$\sigma_j = arg(\frac{x_{j,1}}{x_{j,2}}), \tag{3}$$

where $arg(x)$ is a function that outputs the phase of x, and this function's domain of definition is from -π to π; $j,1$ is one microphone's index of the j-th microphone pair, and $j,2$ is the other microphone's index of the j-th microphone pair. Fig.1 shows the coordinate system. In this paper, ϕ is set to be 0. Assuming that sources are the plane waves, when the j-th microphone pair is parallel to the y axis, the azimuth estimated by the j-th microphone pair $\hat{\theta}$ is

$$\hat{\theta} = arcsin\frac{\sigma_j}{2\pi f d_j c^{-1}}, \tag{4}$$

where d_j is the distance between the j-th microphone pair, and c is the sound velocity. When only one source i is active, and the elevation ϕ of the source is set to be 0, $\hat{\theta}$ is equivalent to the azimuth of source i.

2.4. Azimuth histogram

Estimation of azimuth $\hat{\theta}_{f,\tau}$ is assumed to be obtained at each time frequency point. Estimation of the sources' azimuths are obtained by peak-searching of the histogram of $\hat{\theta_{f,\tau}}$ for all time frequency points. The histogram is obtained as follows:

$$P(k) = \sum_{f,\tau} \sigma(\sin\hat{\theta_{f,\tau}} \geq -1+k\Delta)\sigma(\sin\hat{\theta_{f,\tau}} < -1+(k+1)\Delta), \tag{5}$$

where $\Delta = \frac{2}{L}$, $\sigma(\mathbf{x})$ is 1 when $\mathbf{x}$ is true, and $\sigma(\mathbf{x})$ is 0 when $\mathbf{x}$ is false.

3. AN ANALYSIS OF LOCALIZATION BASED ON SPARSENESS

3.1. Variance of the estimated azimuth

Assuming that there are a directional source s and a diffused noise n, the i-th input signal is

$$x_i(f,\tau) = s_i(f,\tau) + n_i(f,\tau). \tag{6}$$

The phase difference of the i-th input signal and the j-th input signal is

$$arg\frac{x_i}{x_j} = arg\frac{s_i+n_i}{s_j+n_j} \tag{7}$$

$$= arg\frac{s_i}{s_j} + arg(1+\frac{n_i}{s_i}) - arg(1+\frac{n_j}{s_j}) \tag{8}$$

The first term of (8) is

$$arg\frac{s_i}{s_j} = 2\pi f d \sin\theta c^{-1}, \tag{9}$$

where θ is the source's azimuth. In (8), $arg(1+\frac{n_i}{s_i})$ is a function of SNR and is not correlated with $arg(1+\frac{n_j}{s_j})$. Accordingly, the variance of phase difference $arg\frac{x_i}{x_j}$ is a function of SNR and is independent of the distance between the i-th and j-th microphones. The variance of sin of the estimated azimuth $\hat{\sin\theta} = \frac{c}{2\pi f d} arg\frac{x_i}{x_j}$ is proportional to $\frac{1}{d^2}$. Consequently, the longer the distance between a microphone pair is, the smaller the variance of $\hat{\sin\theta}$ is, and the higher the localization performance of sound source localization is.

3.2. Spatial aliasing problem

Let f_{max} be the maximum frequency of signals. When the distance between a microphone pair d is longer than $\frac{c}{2f_{max}}$ (aliasing distance), the spatial aliasing problem occurs[5]. In this case, the phase range of $\frac{x_i}{x_j}$ is larger than 2π, but the range of arg is 2π, so after arg operation, information about the source's azimuth is lost, and the source's azimuth cannot be estimated.

4. APPROACH

The longer the distance is between a microphone pair, the higher the localization performance becomes. However, when the distance is longer than the aliasing distance, the spatial aliasing problem occurs, so the distance cannot be set to be longer than the aliasing distance. Consequently, the upper bound of localization performance is restricted. Our proposed method **SPIRE** clears up this problem. Multiple microphone pairs that are sorted in ascending sequence of the distance between microphones are used in our proposed method. The method consists of two stepwise processes. In the *estimation process*, the phase difference of the i-th microphone pair is estimated. Then, in the *restoration process*, the phase difference of the i-th microphone pair is restored using the restored phase difference of the $i-1$-th microphone pair. In the restoration process, the spatial aliasing problem is settled, and the variance of the estimation azimuth can be reduced.

We explain SPIRE for linear microphone arrays and nonlinear microphone arrays. The explanation is based on the assumption that sources are placed on the $\phi = 0$ plane.

4.1. Algorithm for linear microphone arrays

The number of microphone pairs is defined as L. The L microphone pairs are sorted in ascending sequence of the distance between each microphone pair. Let $\hat{\sigma_{-1}}$ be 0, and d_{-1} be 1.

Estimation process:

The phase difference of the i-th microphone pair is obtained as follows:

$$\sigma_i = arg(\frac{x_{i,1}}{x_{i,2}}). \tag{10}$$

Restoration process:

When the microphone distance of the i-th microphone pair is longer than the aliasing distance, the direction of the source cannot be obtained by phase difference σ_i. In this case, $\sigma_i + 2n\pi$ is the true phase difference. Then the i-th true phase difference is restored by using the phase argument σ_i and the $i-1$-th restored phase difference $\hat{\sigma_{i-1}}$ as follows.

$$\frac{\hat{\sigma_{i-1}} d_i}{d_{i-1}} - \pi \leq \sigma_i + 2n_i\pi \leq \frac{\hat{\sigma_{i-1}} d_i}{d_{i-1}} + \pi. \tag{11}$$

$$\hat{\sigma_i} = \sigma_i + 2n_i\pi. \tag{12}$$

Let $V(x)$ be the variance of x. Assuming that noise is additive white Gaussian noise, when n_i is correctly estimated, $V(\hat{\sigma_i})$ is $\frac{d_{i-1}^2}{d_i^2}$ times smaller than $V(\hat{\sigma_{i-1}})$. The estimation and restoration processes are executed from $i = 0$ to $i = L-1$. Consequently, $\hat{\sigma_{L-1}}$ is obtained. The estimated location of a sound source at time=τ and frequency=f is

$$\hat{\theta} = arcsin \frac{\sigma_{L-1}}{2\pi f d_{L-1} c^{-1}}. \tag{13}$$

The estimated locations of multiple sound sources are obtained by peak-searching of the azimuth histogram made by $\sin\hat{\theta}$.

4.2. Proposed algorithm for nonlinear microphone arrays

We expand our proposed algorithm for sound source localization for 360-degree localization. Instead of multiple microphone pairs, we use sub-microphone arrays. The number of sub-microphone arrays is U. The sub-microphone arrays are sorted in ascending sequence of the maximum distance of each sub-microphone array. L_l is defined as the number of microphone pairs of the l-th sub-microphone array.

Estimation process:

Let $\boldsymbol{p}_i = \begin{bmatrix} x \\ y \end{bmatrix}$ be the position vector of the i-th microphone. Here, j_1 and j_2 are microphone-indexes of the l-th sub-microphone array's j-th microphone pair; $\boldsymbol{d}_j = \boldsymbol{p}_{j_1} - \boldsymbol{p}_{j_2}$. $\boldsymbol{D} = [\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{l-1}]$ is a distance-matrix; and $\boldsymbol{D}^+$ is the Moore and Penrose generalized inverse matrix of $\boldsymbol{D}$. Let $\boldsymbol{q} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$ be the position vector of a source. Also, $\boldsymbol{r} = [arg_0, \ldots, arg_{l-1}]^T$ is an argument vector, where arg_j is a phase difference of the j-th microphone pair of the l-th sub-microphone array. The vector $\boldsymbol{q}$ is estimated by $\hat{\boldsymbol{q}}$ as follows [3]:

$$\hat{\boldsymbol{q}} = \boldsymbol{D}^+ \boldsymbol{r} (2\pi f c^{-1})^{-1} \tag{14}$$

Restoration process:

When more than one distance of the microphone pairs of the l-th sub-microphone array are longer than $\frac{c}{2f_{max}}$, the spatial aliasing problem occurs. In this case, $\hat{\boldsymbol{q}}$ should be $\boldsymbol{D}^+(\boldsymbol{r} + 2\pi\boldsymbol{n})(2\pi f c^{-1})^{-1}$. Also, $\boldsymbol{n} = [n_0, \ldots, n_{l-1}]$, where n_i is integer-valued. The proposed method estimates $\boldsymbol{n}$ using the $l-1$-th sub-microphone array. Let $\boldsymbol{n}_{-1}$ be 0, and let $\hat{\boldsymbol{r}}_{l-1}$ be $\boldsymbol{0}$. Here, $\boldsymbol{n}_l$, which fulfills the following equation, is obtained.

$$\boldsymbol{D}_l \boldsymbol{D}_{l-1}^+ \hat{\boldsymbol{r}}_{l-1} - \pi\boldsymbol{1} \leq_{each} \boldsymbol{r}_l + 2\pi\boldsymbol{n}_l \leq_{each} \boldsymbol{D}_l \boldsymbol{D}_{l-1}^+ \hat{\boldsymbol{r}}_{l-1} + \pi\boldsymbol{1}, \tag{15}$$

where $\boldsymbol{x} \leq_{each} \boldsymbol{y}$ means that each element of $\boldsymbol{y}$ is larger than or equivalent to each element of $\boldsymbol{x}$, and the vector $\boldsymbol{1}$ is a vector whose elements have 1 value. The l-th sub-microphone array's argument vector $\boldsymbol{r}_l$ is restored by the following equation.

$$\hat{\boldsymbol{r}}_l = \boldsymbol{r}_l + 2\pi\boldsymbol{n}_l. \tag{16}$$

The estimation and restoration processes are executed from $l = 0$ to $l = L-1$. Consequently, the final phase difference vector $\hat{\boldsymbol{r}}_{L-1}$ is obtained. The estimated location of a sound source at time=τ and frequency=f is

$$\hat{\boldsymbol{q}}_{L-1} = \boldsymbol{D}_{L-1}^+ \hat{\boldsymbol{r}}_{L-1} (2\pi f c^{-1})^{-1}. \tag{17}$$

The estimated locations of multiple sound sources are obtained by peak-searching for azimuth histogram made by $\hat{\boldsymbol{q}}_{L-1}$.

5. EXPERIMENT

5.1. Conditions

The proposed algorithm was tested using a linear array (LA) and a nonlinear array (NLA). Fig. 2 shows the LA microphone array positions. An NLA consisting of three concentric equilateral triangle arrays (distances of one side are set to be 1 cm, 3 cm, and 9 cm) was also used. Experiments were done in a reverberant room (reverberation time = about 300 ms). The distance between sources and microphones was set at 1 m. The sources were placed on the $\phi = 0$ plane. Signals were sampled at 32 kHz. The frame size of FFT (Fast Fourier Transform) was 2048 pt. Frame shift was 512 pt. Original sound sources were three Japanese voices. The length of the source sounds was about 2 seconds. The histogram had 200 partitions. The normalized azimuth histogram, $P'(k) = \frac{P(k)}{\sum_i P(i)}$, was used in this experiments.

5.2. Results

The proposed method for the LA was tested using a 180-degree localizing problem of two neighboring sources (placed

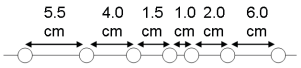

Fig. 2. Microphone array position of a linear array: the distances of microphone pairs used were 1.0 cm, 1.5 cm, 2.0 cm, 2.5 cm, 3.0 cm, 4.5 cm, 5.5 cm, 6.5 cm, 8.5 cm, 10.5 cm, 14.5 cm and 20.0 cm.

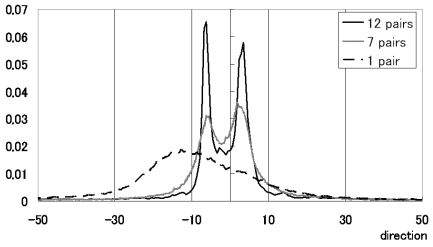

Fig. 3. Azimuth histogram of a linear microphone array: two sources are placed at $-5°$,$5°$, and "1pair" means 1 microphone pair (1.0 cm) was used (equivalent to DUET); "7 pairs" means 7 microphone pairs (1.0 cm-5.5 cm) were used by the proposed method; "12 pairs" means 12 microphone pairs (1.0 cm-20.0 cm) were used by the proposed method.

at $-5°$,$5°$). The azimuth histogram of a linear microphone array is shown at Fig.3. Two neighboring sound sources were located using our proposed method (12 pairs), but were not located by conventional DUET (1 pair). A comparison of two cases using 12 pairs and 7 pairs indicates that the more microphone pairs is, the sharper the histogram becomes. Experimental results for NLA (360-degree localization) are shown at Fig.4 (three sources at $75°$, $85°$, $95°$), Fig.5 (three sources at $-125°$, $-65°$, $95°$) is shown. A comparison of the results using three arrays with those of the selection case (selection of the best sub-microphone array at each frequency point) indicates that the stepwise restoration of the argument vectors is effective.

6. CONCLUSION

In this paper, a new sound source localization method **SPIRE** (Stepwise Phase dIfference REstoration) was proposed. Algorithms for both linear microphone arrays and nonlinear microphone arrays were shown. Experimental results showed that SPIRE succeeded in localizing three sound sources whose locations were near one another (source locations differed by about 10 degrees) in a reverberant room (reverberation time=about 300ms).

7. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE* Trans.SP, vol.52, no.7, pp. 1830-1847, 2004.

[2] M. Matsuo, Y. Hioka, N. Hamada, "Estimating DOA of multiple speech signals by improved histogram mapping method," in *Proc. IWAENC2005*, pp.129-132, 2005.

[3] S. Araki, H. Sawada, R. Mukai, S. Makino, "DOA Estimation for multiple sparse sources with normalized observation vector clustering," *Proc. ICASSP2006*, vol.V, pp.33-36, 2006.

[4] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE* Trans. Antennas and Propagation, vol.34, no.3, pp.276-280, 1986.

[5] D. H. Johnson and D. E. Dudgeon, "Array Signal Processing- Concepts and Techniques," PTR Prentice Hall, New Jersey, USA, 1993.

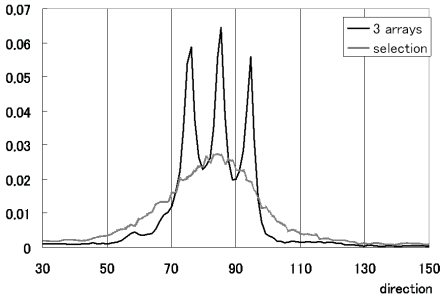

Fig. 4. Azimuth histogram of a nonlinear microphone array: three sources are placed at $75°$,$85°$, and $95°$, and "3 arrays" means the result using the proposed method with 3 sub-microphone arrays (1cm, 3cm, 9cm). "selection" means the result by the selection of 1 sub-microphone array at each frequency with the maximum distance ($\leq \frac{c}{2f}$).

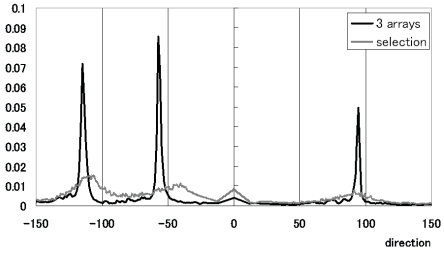

Fig. 5. Azimuth histogram of a nonlinear microphone array: three sources are placed at $-125°$,$-65°$, and $95°$, and "3 arrays" means the result using the proposed method with 3 sub-microphone arrays (1cm, 3cm, 9cm). "selection" means the result by the selection of 1 sub-microphone array at each frequency with the maximum distance ($\leq \frac{c}{2f}$).