PERFORMANCE EVALUATION OF SPARSE SOURCE SEPARATION AND DOA ESTIMATION WITH OBSERVATION VECTOR CLUSTERING IN REVERBERANT ENVIRONMENTS

[1,2]*Shoko Araki,* [1]*Hiroshi Sawada,* [1]*Ryo Mukai, and* [1,2]*Shoji Makino*

shoko@cslab.kecl.ntt.co.jp
[1]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[2]Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo-shi, Hokkaido 060-0814, Japan

ABSTRACT

This paper investigates the effects of real world acoustic environments on sparse source separation and direction of arrival (DOA) estimation performance. The time-frequency mask technique is widely studied as an approach for sparse source separation and DOA estimation. The approach relies on source sparseness, which can easily be affected by, for example, reverberation. In fact, most proposed approaches assume an anechoic condition, which is difficult to maintain in a real acoustic environment. We investigate how the performance of such methods is affected when the problem does not meet the assumed conditions. We show that strong reverberation and a large distance between the sources and sensors degrade the separation performance, however, the DOA estimation performance is not so severely affected.

1. INTRODUCTION

The time-frequency mask approach to blind sparse source separation (BSS) is being widely studied (e.g. [1, 2]). Some authors have also proposed direction of arrival (DOA) estimation methods for sparse sources [3, 4]. The time-frequency mask approach is attractive because it can handle the underdetermined problem where the sources outnumber the sensors. With regard to these problems, we have also already proposed the time-frequency mask approach, which is based on observation vector clustering [5], and a DOA estimation method [6] for sparse sources.

In order to estimate the time-frequency mask for the separation, all the cited methods rely on the assumption of *source sparseness*, and some methods adopt an *anechoic* assumption. Moreover, all the above DOA estimation methods utilize an *anechoic* assumption. In practice, however, these assumptions cannot hold due to such factors as reverberation and noise.

To study the effects of such practical issues, in this paper, we investigate how the performance of a time-frequency mask approach degrades when the problem becomes far from holding the assumptions. We focus particularly on the reverberation variations caused by changing the room reverberation time and the distance between sensors and sources.

2. PROBLEM DESCRIPTION

Suppose that sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{l} \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l),\ j=1,\ldots,M, \quad (1)$$

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. In this paper, we look especially at a situation where the number of sources N can exceed the number of sensors M ($N > M$). We assume that N and M are known, and that the sensor alignment does not cause the spatial aliasing problem. We have two goals. One is to obtain separated signals $\mathrm{y}_k(t)$ that are estimations of s_k solely from M observations. The other is to estimate the DOAs $\mathbf{q}_k$ of signals s_k ($k = 1, \cdots, N$) using the knowledge of the sensor alignment. Here $\mathbf{q}_k$ is a 3-dimensional vector of a unit-norm representing the direction of the source s_k [6].

This paper employs a time-frequency domain approach. Using a short-time Fourier transform (STFT), the convolutive mixtures (1) can be converted to instantaneous mixtures at each frequency f:

$$x_j(f,\tau) \approx \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau), \quad (2)$$

or in vector notation,

$$\mathbf{x}(f,\tau) \approx \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,\tau), \quad (3)$$

where $h_{jk}(f)$ is the frequency response from source k to sensor j, $s_k(f,\tau)$ is the STFT of a source signal s_k, and τ is a time index. We call $\mathbf{x} = [x_1, \ldots, x_M]^T$ *an observation vector* and $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ is a vector of the frequency responses from source s_k to all sensors.

3. METHOD REVIEW

In this section, first we explain the assumptions that have been widely employed for solving the underdetermined

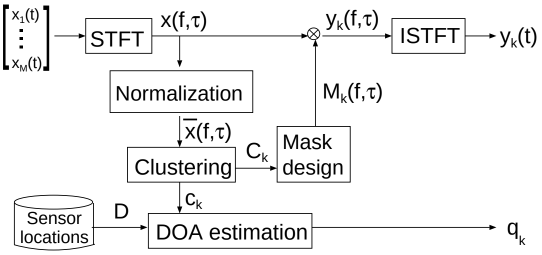


Figure 1: Flow of our method.

problem (e.g. [1, 5]). Then, we briefly describe our separation and DOA estimation methods [5, 6] (Fig. 1).

3.1. Assumptions

ASSUMPTION 1: Source sparseness

We assume the sparseness of sources in the time-frequency domain. When the signals are sufficiently sparse, we can assume that the sources rarely overlap at each time-frequency point, and (3) can be approximated as

$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f)s_k(f,\tau), \quad k \in \{1,\cdots,N\}, \quad (4)$$

where $s_k(f,\tau)$ is a dominant source at the time-frequency point (f,τ). For instance this is true for speech signals in the time-frequency domain [1].

ASSUMPTION 2: Anechoic model

We also assume an anechoic environment. If the source is Dirac's delta function $s_k(t) = \delta(t)$, its observation at the j-th sensor is

$$\begin{aligned} x_j(t) &= h_{jk}(t)\delta(t) \\ &= h_{jk}(t) = \lambda_{jk}\delta(t-\tau_{jk}) \end{aligned}$$

where $\lambda_{jk} \geq 0$ and τ_{jk} are the attenuation and the time delay from source k to sensor j. If we assume an anechoic environment, λ_{jk} and τ_{jk} are determined solely by the geometric condition of the sources and sensors. In the frequency domain, an impulse response $h_{jk}(f)$ is represented by

$$h_{jk}(f) \approx \lambda_{jk} \exp\left[-\jmath 2\pi f \tau_{jk}\right]. \quad (5)$$

3.2. Sparse source separation

Here we employ the method proposed in [5], which employs the above two assumptions as discussed below.

First, we normalize all observation vectors $\mathbf{x}(f,\tau)$ for all time-frequency points (f,τ) such that they form clusters, each of which corresponds to an individual source. The normalization includes phase-normalization with respect to a sensor J and the frequency-normalization,

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{4fc^{-1}d_{\max}}\right] \quad (6)$$

where c is the propagation velocity and $d_{\max}$ is the maximum distance between sensor J and a sensor $j \in \{1,\ldots,M\}$. Then, we apply unit-norm normalization

$$\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau) \,/\, ||\bar{\mathbf{x}}(f,\tau)|| \quad (7)$$

to $\bar{\mathbf{x}}(f,\tau) = [\bar{x}_1(f,\tau),\ldots,\bar{x}_M(f,\tau)]^T$.

Considering the sparseness assumption (4) and the anechoic assumption (5), the normalized vector can be written as

$$\bar{x}_j(f,\tau) \approx \frac{\lambda_{jk}}{A} \exp\left[-\jmath \frac{\pi(\tau_{jk}-\tau_{Jk})}{2c^{-1}d_{\max}}\right], \quad (8)$$

where $A = \sqrt{\sum_{j=1}^{M}\lambda_{jk}^2}$. We can see that the normalized observation vector $\bar{\mathbf{x}}(f,\tau)$ depends only on the source geometry λ_{jk} and τ_{jk} of the source s_k, which is dominant at the time-frequency point (f,τ). This means the normalized observation vectors can be clustered based on the source geometry.

Therefore, the next step is to find clusters $C_1,\ldots,C_N$ formed by all normalized vectors $\bar{\mathbf{x}}(f,\tau)$. After setting appropriate initial centroids $\mathbf{c}_k$ ($k = 1,\cdots,N$), clustering is realized by the following iterative updates:

$$\begin{aligned} C_k &= \{\bar{\mathbf{x}}(f,\tau) \mid k\text{= argmin}_i ||\bar{\mathbf{x}}(f,\tau)-\mathbf{c}_i||^2\} \quad (9) \\ \mathbf{c}_k &\leftarrow E[\bar{\mathbf{x}}(f,\tau)]_{\bar{\mathbf{x}}\in C_k}, \ \ \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||, \quad (10) \end{aligned}$$

where $E[\cdot]_{\bar{\mathbf{x}}\in C_k}$ is a mean operator for the members of a cluster C_k. This minimization can be performed efficiently with the k-means clustering algorithm [7] with a given source number N.

Because each resulting cluster corresponds to an individual source, finally, we obtain separated signals $y_k(f,\tau)$ = $M_k(f,\tau)x_j(f,\tau)$ where j is an arbitrary selected sensor index $j \in \{1,\ldots,M\}$ and

$$M_k(f,\tau) = \begin{cases} 1 & \bar{\mathbf{x}}(f,\tau) \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Then we have time-domain outputs $\mathrm{y}_k(t)$ by using an inverse STFT (ISTFT).

3.3. DOA estimation [6]

When $\mathbf{q}_k$ and $\boldsymbol{d}_j$ are 3-dimensional vectors representing the k-th source DOA and the j-th sensor position, respectively, the delay τ_{jk} in (5) normalized by the delay τ_{Jk} for the sensor J is

$$\tau_{jk} - \tau_{Jk} = c^{-1}(\mathbf{d}_j - \mathbf{d}_J)^T\mathbf{q}_k. \quad (12)$$

On the other hand, from (8) and (10), the delay τ_{jk} can also be derived as

$$\tau_{jk} - \tau_{Jk} = -\frac{2c^{-1}d_{\max}}{\pi}\arg\{\mathbf{c}_k\}_j \quad (13)$$

where $\{\mathbf{c}_k\}_j$ is the j-th component of the centroid $\mathbf{c}_k$. From (12) and (13), we obtain the DOA

$$\mathbf{q}_k = -\frac{2d_{\max}}{\pi}\boldsymbol{D}^{+}\boldsymbol{r}_k. \quad (14)$$

where $\boldsymbol{r}_k = [\arg[\{\mathbf{c}_k\}_1],\cdots,\arg[\{\mathbf{c}_k\}_M]]^T$, $\boldsymbol{D} = [\mathbf{d}_1 - \mathbf{d}_J,\cdots,\mathbf{d}_M - \mathbf{d}_J]^T$, and $\cdot^{+}$ denotes a pseudo-inverse.

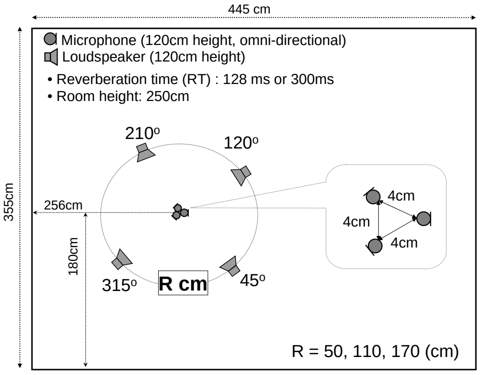

445 cm
355cm
Microphone (120cm height, omni-directional)
Loudspeaker (120cm height)
• Reverberation time (RT) : 128 ms or 300ms
• Room height: 250cm
210°
120°
256cm
180cm
4cm
4cm
4cm
315°
R cm
45°
R = 50, 110, 170 (cm)

Figure 2: Experimental setup.

If rank($\boldsymbol{D}$) ≥ 3, we can estimate the 3-dimensional DOA. Note that the DOA estimation procedure also assumes source sparseness and an anechoic condition.

4. PERFORMANCE INVESTIGATIONS

4.1. General conditions

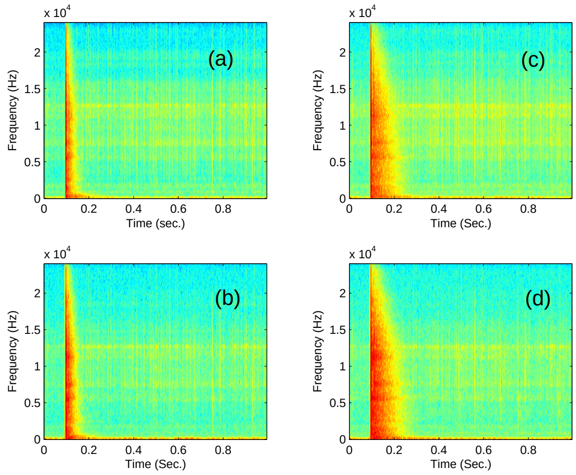
We investigated how the performance of a time-frequency mask approach is affected by reverberation. We performed experiments under an anechoic condition and some reverberant conditions. Here, we only tested the 3-microphone and 4-source case. For the anechoic test, we simulated the mixture by using the anechoic model ((5) and (13)) and the mixture model (1). For the reverberant tests, observations were simulated by following (1) with impulse responses h_{jk} measured in a room (Fig. 2). The room reverberation times RT were 128 and 300 ms. For both RTs, we utilized the same room but changed the wall condition. We also changed the distance R between the sensors and sources. The distances variations were R=50, 110, and 170 cm (see Fig. 2). The sources $\mathrm{s}_k(t)$ were 5-second English speech signals sampled at 8 kHz. We investigated eight speaker combinations and averaged the results for all the outputs. The STFT frame size was 512 and the frame shift was $128(=512/4)$.

4.2. Properties of impulse responses of reverberant conditions

Figure 3 shows example impulse responses under different reverberant conditions; (a) RT=128 ms, R=50 cm, (b) RT=128 ms, R=170 cm, (c) RT=300 ms, R=50 cm, (d) RT= 300 ms, R=170 cm. The impulse response becomes long as RT and R increase. Figure 3 also includes the clarity index [8]:

$$C = 10\log_{10}\frac{\int_0^{80\text{ms}} h^2(t)dt}{\int_{80\text{ms}}^{\infty} h^2(t)dt}$$

which explains the ratio between direct sound and reverberant sound. Small (large) C means the reverberant sound (direct sound) is large. We can see that the clarity C becomes small as the reverberation and distance R increase. That is, when the reverberation is long and R is large, the anechoic assumption becomes corrupted.



Clarity	R=50cm	R=110cm	R=170cm
RT=128ms	45.5	40.8	36.6
RT=300ms	40.5	34.9	32.7

Figure 3: Example impulse responses. (a) RT=128 ms, R=50 cm, (b) RT=128 ms, R=170 cm, (c) RT=300 ms, R=50 cm, (d) RT=300 ms, R=170 cm.

4.3. Source sparseness under reverberant conditions

Here, we investigate how the sparseness changes under some of the reverberant conditions. For the sparseness measure, we employed the approximate W-disjoint orthogonality [9]:

$$r_k(z) = \frac{\sum_{(f,\tau)} ||\Phi_{(k,z)}(f,\tau)s_k(f,\tau)||^2}{\sum_{(f,\tau)} ||s_k(f,\tau)||^2} \times 100 \qquad (15)$$

where $\Phi_{(k,z)}$ is a time-frequency binary mask that has a parameter z

$$\Phi_{(k,z)}(f,\tau) = \begin{cases} 1 & 20\log\left(s_k(f,\tau)/\hat{y}_k(f,\tau)\right) > z \\ 0 & \text{otherwise} \end{cases} \qquad (16)$$

and $\hat{y}_k(f,\tau)$=STFT$\left[\sum_{i=1,i\neq k}^{N} s_i(t)\right]$ (sum of interference components). The approximate W-disjoint orthogonality $r_k(z)$ means the percentage of the energy of source k for time-frequency points where it dominates the other sources by z dB.

Figure 4 shows the $r_k(z)$ values under some reverberant conditions. The sparseness decreases when the contribution of the direct sound is small (see Fig. 3). That is, the sparseness decreases as a result of both the reverberation and distance R.

4.4. Separation performance

The separation performance was evaluated in terms of the signal-to-interference ratio (SIR) improvement. Moreover,

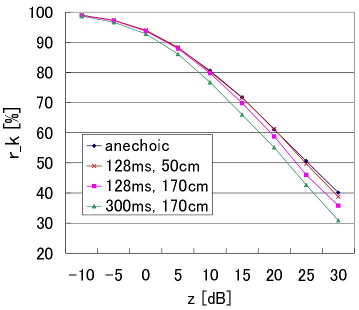


Figure 4: Source sparseness for some reverberant conditions. $r_k(z)$is calculated for $N = 4$.

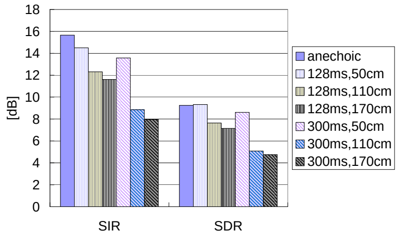


Figure 5: Separation performance for each condition.

we also evaluated the sound quality with the signal to distortion ratio (SDR). The definitions of SIR improvement and SDR can be found in [5].

Figure 5 shows the result for each condition. The performance degrades as the reverberation becomes long. Moreover, performance degradation was observed as the distance R became large. Both phenomena were expected from the clarity values in Fig. 3 and from the sparseness evaluation in the previous subsection. The performance tends to worsen as the direct sound contribution becomes smaller.

4.5. DOA estimation performance

For DOA evaluation, we converted the estimated DOAs $\mathbf{q}_k$ as follows

$$\mathbf{q}_k = [\cos\theta_k \cos\phi_k, \sin\theta_k \cos\phi_k, \sin\phi_k]^T \qquad (17)$$

where θ_k and ϕ_k are the azimuth and the elevation of the k-th source, respectively. Because the elevations ϕ_k were always zero with our setting, we evaluated only the azimuth θ_k.

The azimuth values were set as shown in Fig. 2, however, there were small differences between the drawn DOA and real DOA because they were set manually for every RT and R. Therefore, we evaluated the estimation error

$$\text{Error}_k = |\theta_k - \hat{\theta}_k|$$

where $\hat{\theta}_k$ represents the approximated *true* directions, which were estimated by the MUSIC algorithm [10] when there was only one source signal.

The result is shown in Fig. 6. When the reverberation and distance R are large, the DOA estimation error increases, however, the DOA estimation error is still not very large even under difficult conditions.

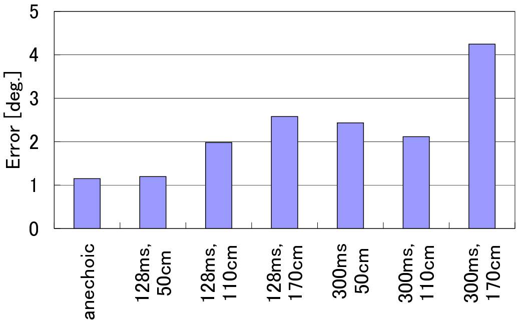


Figure 6: DOA estimation error in degrees. The error bar shows the standard deviations.

5. CONCLUSION

We investigated the time-frequency mask approach for BSS and DOA estimation with respect to reverberation. As our test was only for the 3-microphone and 4-source case, further investigations are required for more complicated cases (more sources, more sensors, longer reverberation, etc.). The performance degradation caused by diffused noise should also be investigated in the future.

6. REFERENCES

[1] Ö. Yılmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] N. Roman, D. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *Journal of Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[3] S. Rickard and F. Dietrich, “DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET,” in *Proc. SSAP2000*, Aug. 2000, pp. 311–314.

[4] M. Matsuo, Y. Hioka, and N. Hamada, “Estimating DOA of multiple speech signals by improved histogram mapping method,” in *Proc. IWAENC2005*, Sept. 2005, pp. 129–132.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *Proc. IWAENC 2005*, Sept. 2005.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, “DOA estimation for multiple sparse sources with normalized observation vector clustering,” in *Proc. ICASSP2006*, May 2006, vol. 5, pp. 33–36.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.

[8] “ISO 3382: Acoustics-Measurement of the reverberation time of rooms with reference to other acoustical parameters,” 1997.

[9] S. Rickard and Ö. Yılmaz, “On the approximate W-disjoint orthogonality of speech,” in *Proc. ICASSP2002*, may 2002, vol. I, pp. 529–532.

[10] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas and Propagation*, vol. 34, pp. 276–280, Mar. 1986.