SOURCE EXTRACTION FROM SPEECH MIXTURES WITH NULL-DIRECTIVITY PATTERN BASED MASK

Shoko Araki, Shoji Makino, Hiroshi Sawada and *Ryo Mukai*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: {shoko, maki, sawada, ryo}@cslab.kecl.ntt.co.jp

ABSTRACT

We propose a method for extracting source signals from mixtures when the N sources outnumber the M sensors ($N > M$). Previously, Yılmaz and Rickard [1] employed the sparseness property of source signals and extracted each signal using a time-frequency binary mask (BM). However, the use of binary masks means that the extracted signals have too much discontinuous zero-padding, and they contain loud musical noise.

In order to reduce musical noise, we propose a new source extraction method that employs a spatially continuous (SC) time-frequency mask. First, using the source sparseness, we estimate the direction of arrival (DOA) of each source. Next, we make a null beamformer (NBF) which has a small gain for the DOAs of signals to be removed, and then we design an SC mask using the directivity pattern of the NBF. Even if we have $M < N$ sensors for N sources, we can make an SC mask by *assuming* $V = N$ sensors. Since this SC mask is not binary, the discontinuous zero-padding to the extracted signals decreases.

Experimental results show that our method can extract signals with little distortion without serious deterioration in the separation performance measured by using SIR even in a real reverberant environment of T_R=130 ms.

PROPOSED METHOD

[Step1] DOA estimation: First, the observed mixed signals $x_j(n)$ ($j = 1, \cdots, M$) are converted into time-frequency signals $X_j(f, m)$ with a short time Fourier transform (STFT), where f is the frequency and m is the time-dependence of the STFT. By assuming that the signals are sufficiently sparse, we can classify the observation sample points $X_j(f, m)$. To classify the observations, we use the estimated DOA $\theta(f, m) = \arccos \frac{\varphi(f,m)c}{2\pi f d}$ where $\varphi(f, m) = \angle \frac{X_i(f,m)}{X_j(f,m)}$ ($i \neq j$) is the phase difference between two sensors, d is the sensor space, and c is the speed of sound. The DOA has N clusters (Fig. 1) and each cluster corresponds to one source. We estimate the DOAs of sources using the centroid of each cluster, $\tilde{\theta}_1, \tilde{\theta}_2, \cdots, \tilde{\theta}_N$.

[Step2] Design of SC mask: Then using the estimated DOAs $\tilde{\theta}_i$, we make an NBF that makes nulls towards given $N - 1$ directions formed by $V = (N - 1) + 1$ (virtual) sensors. Here, V is not necessarily equal to M; V can be greater than M. Here, we assume that the number of sources N is known or estimated beforehand, e.g., from a histogram such as that in Fig. 1. We also assume that the sensors are arranged linearly. The NBF $\boldsymbol{W}(f)$ can be made as follows.

First we form a ($V \times V$) matrix $\boldsymbol{H}_{\text{NBF}}(f)$ whose ji-th element $H_{\text{NBF}ji}(f) = \exp(\jmath 2\pi f \tau_{ji})$ and calculate $\boldsymbol{W}(f) = \boldsymbol{H}_{\text{NBF}}^{-1}(f)$. Here, $\tau_{ji} = d_j c^{-1} \cos\hat{\theta}_i$, d_j is the position coordinate of the j-th virtual sensor, $\{\hat{\theta}_i\ (i = 2, \cdots, V)\}$ are the DOAs of signals to be removed and $\hat{\theta}_1 = \tilde{\theta}_k$ where $\tilde{\theta}_k$ is the DOA of a signal to be extracted. The directivity pattern of the NBF that extracts the signal from $\hat{\theta}_1 = \tilde{\theta}_k$ can be obtained by

$$F^k(f, \theta) = \sum_{j=1}^{V} W_{1j}(f) \exp(\jmath 2\pi f d_j c^{-1} \cos\theta). \tag{1}$$

Then we make an SC time-frequency mask by using this $F(f, \theta)$,

$$\text{[SC 1]} \qquad M^k_{\text{SC1}}(f, m) = F^k(f, \theta(f, m)) \tag{2}$$

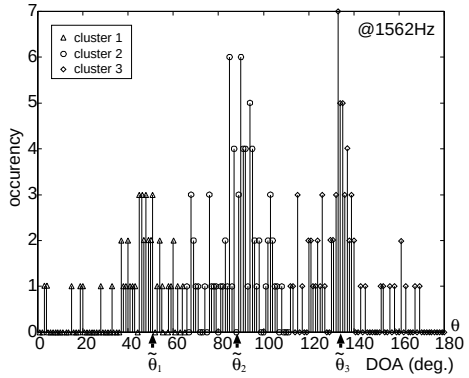

Figure 1: Example histogram. A male-male-female combination with STFT frame size T = 512. T_R = 0 ms.

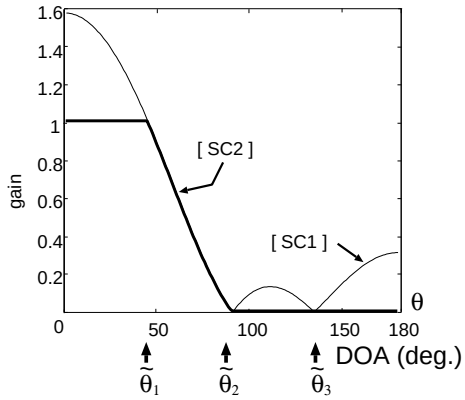

Figure 2: Example mask patterns.

and its modified version [2],

$$\text{[SC 2]} \qquad M^k_{\text{SC2}}(f,m) = \begin{cases} c_e & \theta(f,m) \in \text{extraction area} \\ F^k(f,\theta(f,m)) & \theta(f,m) \in \text{transition area} \\ c_r & \theta(f,m) \in \text{removal area} \end{cases} \tag{3}$$

where c_e is a constant (e.g., $F^k(f,\tilde{\theta}_k)$) and c_r is a small constant (e.g., the minimum value of the directivity pattern). Figure 2 shows example SC mask patterns.

[Step3] Source extraction: Using the SC mask, we obtain the extracted signal by $Y_k(f,m) = M^k_{\text{SC}}(f,m)X_J(f,m)$ where $X_J(f,m)$ is one of the observations and $J \in \{1,\cdots,M\}$.

RESULTS

We examined our method and compared its performance with that of a binary mask (BM) approach. Here we define a time-frequency binary mask

$$\text{[BM]} \qquad M^k_{\text{BM}}(f,m) = \begin{cases} 1 & \tilde{\theta}_k - \Delta \le \theta(f,m) \le \tilde{\theta}_k + \Delta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where Δ is an extraction range parameter and here we used the standard deviation σ_k of cluster k for this parameter. We conducted experiments using three sources and two sensors ($N = 3$, $M = 2$). Table 1 shows the separation results for a male-male-female speaker combination. We tried three speaker positions and the averaged results are shown. With BM, the signal to distortion ratio (SDR) values were unsatisfactory, and a large musical noise was heard. In contrast, with SC masks, we were able to obtain high SDR values without serious degradation of the separation performance, i.e., the signal to interference ratio (SIR). Some sound examples can be found at [3]. It should be noted that it remains difficult to separate signals at the center position with any method.

Table 1: Separation performance in SIR [dB] and SDR [dB]. (a) T_R=0 ms, (b) T_R=130 ms. A male-male-female combination,

(a)	SIR1	SDR1	SIR2	SDR2	SIR3	SDR3
BM	16.6	10.1	9.0	12.6	13.5	8.9
SC1	15.2	11.2	5.8	17.1	14.2	11.6
SC2	16.1	11.6	7.6	16.0	15.3	11.9

(b)	SIR1	SDR1	SIR2	SDR2	SIR3	SDR3
BM	11.3	4.5	5.5	11.9	4.7	5.5
SC1	10.2	4.8	2.0	16.2	8.2	7.4
SC2	12.0	5.0	3.4	15.1	8.6	8.3

References

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

[2] S. Araki, S. Makino, H. Sawada and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," *ICA2004, (Lecture Notes in Computer Science 3195)*, pp. 898–905, Springer-Verlag, Sept. 2004.

[3] http://www.kecl.ntt.co.jp/icl/signal/araki/scm.html