SOUND SOURCE SEPARATION OF OVERCOMPLETE CONVOLUTIVE MIXTURES USING GENERALIZED SPARSENESS

Masahito Togami, Takashi Sumiyoshi, and Akio Amano

{mtogami,t-sumiyo,amano}@crl.hitachi.co.jp
Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

We propose a sound source separation method that works well even if there are more sources than mixtures and signals are recorded in a reverberant room. The proposed method is based on generalized sparseness, where the number of active sources is assumed to vary from 1 to the number of mixtures M at each time-frequency point, and the proposed $sparseness\ estimator$ estimates the most suitable number of active sources. Approaches using binary masks assume that only one source is active at each time-frequency point. However, when more than two sources are active, separated signals are greatly distorted with musical noise. The separated signals by the shortest-path algorithm are less distorted than those obtained by binary masks. However, when there are fewer than M active sources and noise, the shortest-path algorithm overestimates the source signal's value. To overcome the overestimation and distortion problems, the proposed method does not fix the number of sources as one or M. Instead, those are estimated at each time-frequency point. Experimental results in a room (reverberation time = 100 ms) indicate that NRR (Noise Reduction Ratio) of signals separated by our proposed method outperform those of binary masks and the shortest-path algorithm by about 3-5db.

1. INTRODUCTION

Sound source separation is an essential function for communication tools such as hands-free phones that are used in a noisy environment.

Independent Component Analysis(ICA)[1] is a popular separation method. However, when there are more sources than mixtures, ICA cannot separate all of the sources. In this paper, we propose a separation method of overcomplete convolutive mixtures, where there are more sources than mixtures, and those sources reverberate. The proposed method is based on generalized sparseness, where the number of active sources is assumed to vary from 1 to the number of mixtures M at each time-frequency point.

Most of conventional methods use the assumption that there is only one active source[2], or there are as many active sources as M[3][4][5] at each time-frequency point. However, the assumption that the number of active sources is a fixed number is a problem, because the number of active sources varies at each time-frequency point.

Approaches that combine sparseness and a mixing matrix (or ICA) have been proposed[6][7]. These approaches estimate whether the number of active sources is one or two. Our proposed sparseness estimator estimates whether the number of active sources is one, two, or...M, even if $M \geq 3$.

Furthermore, in the proposed method, the time-averaged cost function of the sparseness estimator is proposed. This function is defined on the hypothesis that data points of a speech signal are correlated.

In this paper a mixing matrix is supposed to be known. This matrix can be estimated by clustering approach[6].

2. PROBLEM STATEMENTS AND NOTATION

2.1. Mixing process

Let $\boldsymbol{X}(f,\tau)$ be the observed M-dimensional vector, $\boldsymbol{A}(f)$ be the mixing ($M \times N$) matrix, $\boldsymbol{S}(f,\tau)$ be the source N-dimensional vector, and $\boldsymbol{N}(f,\tau)$ be a white Gaussian noise, where f is the frequency, and τ is the frame index. The mixing process is

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

2.2. A posteriori probability of the separated signals

We can obtain maximum likelihood values of original sources as follows:

$$\hat{\boldsymbol{S}}(f,\tau) = \underset{\boldsymbol{S}(f,\tau)}{\operatorname{argmax}}\, P(\boldsymbol{S}(f,\tau)|\boldsymbol{X}(f,\tau), \boldsymbol{A}(f)). \tag{2}$$

Letting $\boldsymbol{N}(f,\tau)$ be a white Gaussian noise and the probability of $\boldsymbol{S}(f,\tau)$ be the uniform Laplacian, the log-posteriori probability of $\boldsymbol{S}(f,\tau)$ is

$$\begin{aligned}&\log P(\boldsymbol{S}(f,\tau)|\boldsymbol{X}(f,\tau), \boldsymbol{A}(f)) = \\ &-\alpha\|\boldsymbol{A}(f)\boldsymbol{S}(f,\tau) - \boldsymbol{X}(f,\tau)\|^2 - \|\boldsymbol{S}(f,\tau)\|_1,\end{aligned} \tag{3}$$

where $\alpha \propto \frac{1}{\sigma^2}$, σ^2 is the variance of $\boldsymbol{N}(f,\tau)$, the second term is the $l1$-norm in this paper. $\boldsymbol{S}(f,\tau)$ which maximizes equation 3 cannot be obtained straightforwardly.

2.3. Conventional approach: binary masks

When only one source is active at each time-frequency point, the active source's index $\hat{i}$ and estimated value $\hat{s}(f,\tau)$ are obtained as follows:

$$\{\hat{i}, \hat{s}(f,\tau)\} = \underset{i,s(f,\tau)}{\operatorname{argmin}} \|\boldsymbol{a}_i(f)s(f,\tau) - \boldsymbol{X}(f,\tau)\|^2 \quad (4)$$

, where $l1$-norm is regarded as constant at a time-frequency point, and $\boldsymbol{a}_i(f)$ is the i-th column of $\boldsymbol{A}(f)$.
When more than two sources are active, the separated signals by binary masks are greatly distorted with musical noise.

2.4. Conventional approach: $l1$-norm minimization

$l1$-norm minimization[3] assumes that noise is absent, and $\hat{\boldsymbol{S}}(f,\tau)$ can be otained as follows:

$$\hat{\boldsymbol{S}}(f,\tau) = \underset{\boldsymbol{S}(f,\tau), where \boldsymbol{X}(f,\tau)=\boldsymbol{A}(f)\boldsymbol{S}(f,\tau)}{\operatorname{argmin}} \|\boldsymbol{S}(f,\tau)\|_1. \quad (5)$$

$\boldsymbol{S}(f,\tau)$ is a complex-valued vector and cannot be obtained by linear programming. The shortest-path algorithm[3][5] assumes that there are as many active sources as M and chooses $\boldsymbol{S}(f,\tau)$ that minimizes $\|\boldsymbol{S}(f,\tau)\|_1$. However, when there are less active sources than M and noise, the shortest-path algorithm overestimates the source signal's value.

3. APPROACH

3.1. Generalized sparseness

Conventional approaches assume that the number of active sources is fixed at one or M. However, the number of active sources varies at each time-frequency point.
If there are $L(< M)$ sources, separated signals can be obtained as follows:

$$\hat{\boldsymbol{S}}_{L,j}(f,\tau) = \underset{\boldsymbol{S}(f,\tau)\in\Omega_{L,j}}{\operatorname{argmin}} \|\boldsymbol{A}(f)\boldsymbol{S}(f,\tau) - \boldsymbol{X}(f,\tau)\|^2 \quad (6)$$

$$\hat{\boldsymbol{S}}_{L}(f,\tau) = \underset{\hat{\boldsymbol{S}}_{L,j}(f,\tau)}{\operatorname{argmin}} \|\boldsymbol{A}(f)\hat{\boldsymbol{S}}_{L,j}(f,\tau) - \boldsymbol{X}(f,\tau)\|^2 \quad (7)$$

, where Ω_L is a set of N-dimensional complex-valued vectors that have $N-L$ zero-valued elements, and $\Omega_{L,j}$ is the jth Ω_L's subset in which the same index elements are zero valued. $\hat{\boldsymbol{S}}_{L,j}(f,\tau)$ can be obtained straightforwardly as follows:

$$\hat{\boldsymbol{S}}'_{L,j}(f,\tau) = (\boldsymbol{A}'^{*}(f,\tau)\boldsymbol{A}'(f,\tau))^{-1}\boldsymbol{A}'^{*}(f,\tau)\boldsymbol{X}(f,\tau), \quad (8)$$

where $\boldsymbol{A}'(f,\tau)$ is the $M \times L$ matrix in which zero-valued elements' mixing vectors of $\boldsymbol{A}(f,\tau)$ are eliminated, and $\hat{\boldsymbol{S}}'_{L,j}(f,\tau)$ is the L-dimensional vector in which zero-valued elements of $\hat{\boldsymbol{S}}_{L,j}(f,\tau)$ are eliminated.

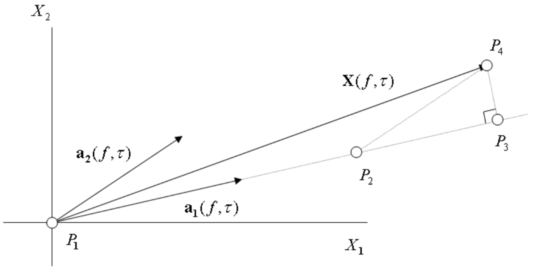

Figure 1: *The shortest-path algorithm's overestimation problem: $M = 2$. $\boldsymbol{a}_1(f)$ and $\boldsymbol{a}_2(f)$ are mixing vectors. $\boldsymbol{X}(f,\tau)$ is the observed vector.*

If there are $L = M$ sources, separated signals can be obtained by the shortest-path algorithm.
Ω_M contains $\Omega_1, \ldots, \Omega_{M-1}$. However, when there are $L < M$ active sources and noise, the shortest-path algorithm overestimates a source signal's value. This problem is shown in Fig. 1. Let the first source vector be P_1P_3, the second source vector be a zero-valued vector, and the noise vector be P_3P_4. Estimations of the source vectors by the shortest-path algorithm are P_1P_2 and P_2P_4. In this situation, P_2P_3 of the first source vector is lost and the second source vector is overestimated. Estimations of the source vectors by equation 7 are P_1P_3 and a zero-valued vector, so the estimation obtained by using equation 7 is more suitable than the estimation obtained by using the shortest-path algorithm in this situation.
However, this does not mean that equation 7 is always more suitable than the shortest-path algorithm. In fig. 1, when there are two active sources and noise is absent, the shortest-path algorithm is more suitable than equation 7. The most suitable model, $L_{suitable}$, at each time-frequency point is defined as

$$L_{suitable} \overset{\text{def}}{=} \underset{L}{\operatorname{argmin}} \|\boldsymbol{S}_{correct}(f,\tau) - \hat{\boldsymbol{S}}_L(f,\tau)\|^2, \quad (9)$$

where $\boldsymbol{S}_{correct}(f,\tau)$ is the correct-valued vector. We also define the suitable model factor, $P(L)$, as

$$P(L) \overset{\text{def}}{=} \frac{\sum_\tau \sum_f \delta(L_{suitable} = L)\|\boldsymbol{X}(f,\tau)\|^2}{\sum_\tau \sum_f \|\boldsymbol{X}(f,\tau)\|^2}, \quad (10)$$

where $\delta(true) = 1, \delta(false) = 0$. If only one model $L(1 \leq L \leq M)$ is always most suitable, only $P(L)$ is non-zero valued. Fig. 2 shows that the most suitable L varies at each time-frequency point. The proposed method is based on generalized sparseness, where the number of active sources is assumed to vary from 1 to the number of mixtures M at each time-frequency point.

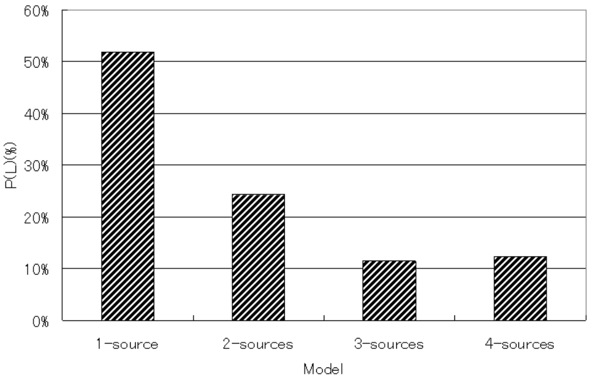

Figure 2: *Suitable model factor $P(L)$: M=4, N=5 case. "1-source" - "4-sources" are cases L = 1-4 of equation 10.*

3.2. Sparseness estimator

We propose a sparseness estimator that estimates the most suitable estimation of $\hat{\boldsymbol{S}}_L(f,\tau)$ at each time-frequency point. The sparseness estimator is based on the hypothesis of generalized sparseness. The sparseness estimator assumes that the probability of noise is white Gaussian. In fig. 1, when $\|P_1P_2\| + \|P_2P_4\| > \|P_1P_3\| + \alpha\|P_3P_4\|^2$, the sparseness estimator estimates there is one signal source and noise, and when $\|P_1P_2\| + \|P_2P_4\| < \|P_1P_3\| + \alpha\|P_3P_4\|^2$, the sparseness estimator estimates that there are two sources and noise is absent. The α is inversely proportional to σ^2. The σ^2 is the variance of noise. This concept is equivalent to the original MAP concept (equation 3). However, MAP searches for the most suitable estimation over all the N-dimensional complex-valued vectors. Contrary to MAP, the proposed sparseness estimator searches the most suitable estimation over estimations of equation 7 and the estimation of the shortest-path algorithm. Therefore, the proposed sparseness estimator can obtain the solution easily, but MAP cannot obtain solutions without using a recursive algorithm. The computational cost of MAP is significantly higher than the proposed sparseness estimator. The number of active sources is estimated as follows:

$$L_{min} = \underset{1\le L\le M}{\operatorname{argmin}} \alpha\|\boldsymbol{A}(f)\hat{\boldsymbol{S}}_L(f,\tau) - \boldsymbol{X}(f,\tau)\|^2 + \|\hat{\boldsymbol{S}}_L(f,\tau)\|_1. \qquad (11)$$

Separated signals $\hat{\boldsymbol{S}}(f,\tau)$ are

$$\hat{\boldsymbol{S}}(f,\tau) = \hat{\boldsymbol{S}}_{L_{min}}(f,\tau). \qquad (12)$$

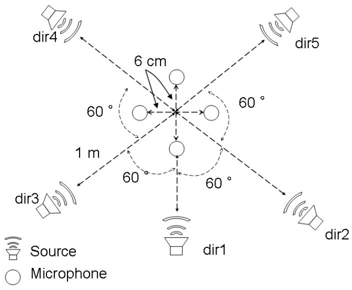

Figure 3: *Recording environment (reverberation time 100 ms).*

3.3. Time-averaged cost function

The original definition of MAP has been independently formulated for each data point. However data points of a speech source are correlated. We assume that values of the cost functions of the sparseness estimator also change slowly. Therefore, we propose a time-averaged cost function. Equation 7 is replaced as follows.

$$j_{min} = \underset{j}{\operatorname{argmin}} \sum_{m=-k}^{k} \gamma(m)\mathrm{err}_{L,j}(f,\tau+m), \qquad (13)$$

where $\mathrm{err}_{L,j}$ is $\|\boldsymbol{A}(f)\hat{\boldsymbol{S}}_{L,j}(f,\tau) - \boldsymbol{X}(f,\tau)\|^2$ and $\gamma(m)$ is the weight vector for time averaging. $\hat{\boldsymbol{S}}_L(f,\tau)$ is

$$\hat{\boldsymbol{S}}_L(f,\tau) = \hat{\boldsymbol{S}}_{L,j_{min}}(f,\tau). \qquad (14)$$

Time-averaging sparseness estimator estimates the suitable number of active sources as follows:

$$L_{min} = \underset{L}{\operatorname{argmin}} \sum_{m=-k}^{k} \gamma(m)\Big(\alpha\mathrm{err}_{L,j_{min}}(f,\tau) + l1_{L,j_{min}}(f,\tau)\Big), \qquad (15)$$

where $l1_{L,j_{min}}(f,\tau)$ is $\|\hat{\boldsymbol{S}}_{L,j_{min}}(f,\tau)\|_1$.

4. EXPERIMENT

4.1. Conditions

The performance of the proposed method was evaluated by a five sources separation problem in a reverberant room whose reverberation time is 100 ms. There are 4 mixtures. The recording environment is shown in Fig. 3.

The source signals used for the experiment were Japanese speech signals that were sampled at 11.025 Hz, and 200 sentences were sent from each direction. The mixing matrix is given in this experiment. The measure of evaluation is NRR $= -10\log_{10}\frac{\sum_t(\hat{s}(t)-s(t))^2}{\sum_t(n(t))^2}$, where $s(t)$ is the source signal, $n(t)$ is the noise signal, and $\hat{s}(t)$ is the separated signal. High-NRR signals are low-noise signals.

4.2. Results

The experimental results are shown in Table 1.

Table 1: Performance of source separation(the average of NRRs[db] of 200 sentences): "bin" is the separation algorithm based on equation 4, "two" and "three" are the algorithms based on equation(7)(L=2,3), "short" is the shortest-path algorithm, "p1" is the separation algorithm using the proposed sparseness estimator, and "p2" is the separation algorithm using the proposed time-averaging sparseness estimator with optimized parameters, $k = 5$, $\gamma(0) = 1, \gamma(\pm1) = 0.7, \gamma(\pm2) = 0.5, \gamma(\pm3) = 0.3, \gamma(\pm4) = 0.2, \gamma(\pm5) = 0.1$, and $\alpha = 1.43$.

	bin	two	three	short	p1	p2
dir1	12.7	15.2	14.4	12.2	17.4	18.0
dir2	12.8	14.7	11.8	12.0	16.2	16.5
dir3	12.1	14.5	11.6	11.3	15.9	16.3
dir4	13.2	16.2	12.7	13.6	17.0	17.1
dir5	14.0	16.8	13.1	15.0	17.7	17.9

As can be seen, NRRs of the proposed sound source separation method using the sparseness estimator outperformed those of binary masks and the conventional shortest-path algorithm by about 3-5db. Furthermore, the proposed method outperformed the methods reported in columns "two" or "three" . This indicates that the estimation of the most suitable model at each time-frequency point is effective. In comparison to "p1" , "p2" gives us higher-NRR signals. This indicates that time-averaging is also effective. Examples of separated signals otained by the proposed time-averaging sparseness estimator are illustrated in Fig. 4.

5. CONCLUSION

We proposed a sound source separation method based on the generalized sparseness assumption. We have shown that the proposed method is more effective than the algorithm using binary masks and the conventional shortest-path algorithm in a reverberant room(reverberation time = 100 ms).

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," John Wiley & Sons, 2001.

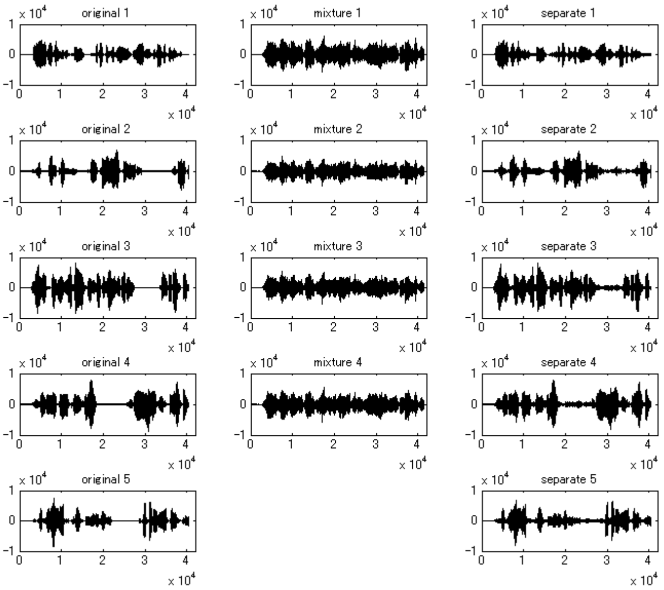

Figure 4: *Examples of separated signals obtained by the proposed method. First column: five speech signals. Second column: four mixtures. Third column: five separated signals obtained by the proposed method. The horizontal axis is time. The vertical axis is amplitude.*

[2] 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.

[3] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time fourier transform," $Proc.ICA2000$, pp. 87-92, 2000.

[4] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," $Proc.ICASSP2003$, pp. 581-584, 2003.

[5] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," $Proc.ICA2004$, pp. 652-660, 2004.

[6] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," IEICE Trans. Fundamentals, vol. E88-A, no.7, pp.1693-1700, 2005.

[7] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," $Proc.IWAENC2003$, pp. 271-274, 2003.