HYBRID APPROACH FOR MULTICHANNEL SOURCE SEPARATION
COMBINING TIME-FREQUENCY MASK WITH MULTI-CHANNEL WIENER FILTER

Shoko Araki  Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper discusses a hybrid approach for the multi-channel source separation, where both a time-frequency (t-f) mask and a multi-channel Wiener filter (WF) are utilized. T-f mask based approaches have been widely studied, because they can separate signals with a low calculation cost. However, the separated signals with a t-f mask usually contain a non-linear distortion. On the other hand, a new multi-channel WF framework employing a spatial covariance matrix model has recently been proposed. With the WF method, we can obtain separated signals of better quality than with a t-f mask, however, the method is computationally expensive because it requires many iterations for the optimization. In this paper, in order to take advantages of both approaches, we explain the hybrid algorithm by introducing the t-f mask concept to the WF approach. Then we show that the hybrid approach achieves high performance without the iterative calculation for the WF. We also present the way for applying the hybrid method to the case where the number of sources is unavailable.

Index Terms— Source separation, source number estimation, time-frequency mask, multi-channel Wiener filter, EM algorithm

1. INTRODUCTION

Blind source separation (BSS) is an approach for estimating source signals that uses only the mixed signal information observed at each microphone. The BSS technique for speech discussed in this paper has many applications, including hands-free teleconference systems and preprocessing for an automatic speech recognizer.

For the BSS, t-f mask based approaches that cluster the t-f bins according to the sparseness of sources in the time-frequency (t-f) domain (e.g., [1–3]) have been widely studied. This may be because the t-f mask approach can be applied to the underdetermined case where we have more sources than microphones, and because it is usually computationally inexpensive. We can also solve the permutation problem, which is usually problematic for a t-f domain approach, by synchronizing the obtained t-f masks in the all frequency bins [3]. Moreover, the t-f mask approach can estimate the number of sources in the microphone observation [4, 5]. The t-f mask approach seems versatile, however, separated signals with a t-f mask usually contain a non-linear distortion [6] that is called the musical noise.

On the other hand, a new multi-channel WF framework has recently been proposed [7]. The method classifies the t-f components by relying on the spatial covariance matrix, that includes the spatial characteristics of each source. It has been shown that the method achieves high performance in reverberant conditions by assuming the full-rank spatial covariance matrix. Moreover, because it separates the signals by using multi-channel WFs, the separated signals has less musical noise than those with a t-f mask approach.

The effectiveness of the method has been confirmed, however, the method is computationally expensive because it requires many iterations each of which requires a relatively high computational cost, and the method can be applied only when the number of sources $N_s$ is available.

In order to take advantages of both approaches, we propose a hybrid approach (Fig. 1) that combines the t-f clustering and the WF [7] using t-f mask estimation. For this purpose, we introduce the t-f mask into the WF approach. By considering the dominant source index at each t-f slot as a hidden variable, we can bring the t-f mask concept into the WF approach. We describe the maximum likelihood (ML) based optimization method by defining the likelihood function and the auxiliary function, which have not been explicitly shown in the original work [7].

We then explain two interesting advantages of the hybrid method. The first issue is that we can reduce the computational cost of the WF approach, because with the hybrid method we can easily include the results obtained with existing t-f clustering based approaches, that are computationally inexpensive, into the WF approach. We will see that we can obtain high performance with no iterative calculation for the WF estimation. The second interesting point is that we can apply the hybrid method to the case where the number of sources is unknown. This became possible due to the introduction of the dominant source index at each t-f slot.

Experimental results will show that the hybrid approach achieves high performance with the just one time WF calculation. We will also examine the applicability of the hybrid method to the case where the number of sources is unknown.

2. ASSUMPTIONS ON SOURCE AND OBSERVATION

Let us formulate the task. Suppose that $n(=1,...,N_n)$ and $f(=1,...,N_f)$ are time and frequency indices of a t-f slot, and that $N_s \geq 2$ speech signals $s_{n,k}^{(1)}, \ldots, s_{n,k}^{(N_s)}$ are mixed and the observed at $N_m$ microphones,

\[ x_{n,f} = \sum_{l=1}^{N_n} s_{n,k}^{(l)} c_{n,f}^{(l)} = \sum_{l=1}^{N_n} h_{f}^{(l)} s_{n,f}^{(l)}, \quad (1) \]

where $x_{n,f} = [x_{n,f}^{(1)}, \ldots, x_{n,f}^{(N_m)}]^T$ is an observation vector, $x_{n,f}^{(l)}$ is the observation at $n$-th microphone, $c_{n,f}^{(l)} = [c_{n,f}^{(l)}(1), \ldots, c_{n,f}^{(l)}(N_m)]^T$ is the spatial image of the $l$-th source, $c_{n,f}^{(l)}$ represents the contribution

![Fig. 1. Framework of the hybrid algorithm.](image)
of source $s_{n,f}$ at microphone $m$, $h_f^{(l)} = [h_1^{(l)}, \ldots, h_{N_m}^{(l)}]^T$, and $h_{m}^{(l)}$ represents the impulse response from source $l$ to microphone $m$. In this paper, without loss of generality, we consider a stereo case $N_m = 2$. Our goal is to obtain estimates of the source images $c_{n,f}$ from the observation $x_{n,f}$ without information on the speech sources $s_{n,f}$, or the mixing process $\mathbf{H}_f^{(l)}$. The number of sources $N_s$ could be unavailable.

We assume that the source image $c_{n,f}^{(l)}$ is a complex-Gaussian random vector with zero mean and covariance matrix $v_{n,f}^{(l)}B_f^{(l)}$ \cite{7},

$$p(c_{n,f}^{(l)};B_f^{(l)};v_{n,f}^{(l)}) = N(c_{n,f}^{(l)};0,v_{n,f}^{(l)}B_f^{(l)})$$

where the non-negative scalar $v_{n,f}^{(l)}$ denotes the spectral power of the $l$th source, and the $(N_m \times N_m)$ matrix $B_f^{(l)}$ is the spatial covariance matrix that represents the spatial characteristics of the mixing process. That is, the source model is parameterized by $\theta^{(l)} = (v_{n,f}^{(l)},B_f^{(l)})$, where $a_{n,f}^{(l)}$ denotes a set of variables $a_{n,f}$ for all $n$ and $f$.

We also assume the sparseness of the sources \cite{1}, that is, we assume that at most one source is dominant at each t-f slot $(n,f)$. Now the observed signal (1) becomes

$$x_{n,f} = c_{n,f}^{(l)} + e_{n,f}^{(l)} \approx c_{n,f}^{(l)}$$

where $z_{n,f}$ is an index of the dominant source at the t-f slot $(n,f)$, and $e_{n,f}^{(l)}$ is noise for the dominant source. This equation means that the observed signal $x_{n,f}$ depends only on the dominant source $c_{n,f}^{(l)}$, and the non-dominant source images $c_{n,f}^{(l)}(l \neq z_{n,f})$ are observed as 0. Letting $\theta = (\theta^{(l)})$, where $(\theta^{(l)})$ means a set of $\alpha$ for all $l$, this assumption can be formulated as:

$$p(\theta^{(l)};\{z_{n,f}\};z_{n,f};\theta) = N(c_{n,f}^{(l)};0,v_{n,f}^{(l)}B_f^{(l)}) \prod_{l' \neq z_{n,f}} \delta(c_{n,f}^{(l')})$$

where $\delta$ is the Dirac delta function, and

$$U_{n,f}^{(z_{n,f})} = e_{n,f}^{(z_{n,f})}e_{n,f}^{(z_{n,f})}H = \sum_{l \neq z_{n,f}} v_{n,f}^{(l)}B_f^{(l)}$$

represents the noise covariance for the dominant source $c_{n,f}^{(z_{n,f})}$. That is, we regard the non-dominant sources as the random Gaussian noise for the dominant source.

In order to estimate the source images $c_{n,f}^{(l)}$, we have to determine the parameters in (2): $\theta = (\theta^{(l)}) = (\{v_{n,f}^{(l)},B_f^{(l)}\})$. The next section explains a maximum likelihood (ML) approach for estimating these parameters.

### 3. HYBRID ALGORITHM

Figure 1 shows a block diagram of the hybrid algorithm. The hybrid approach attempts to realize an efficient computation with the TF clustering, and to obtain the high quality separated signals with WF separation. Both approaches are combined by the t-f mask estimation concept. In this section, the WF separation and the mask estimation steps are discussed in Section 3.1, and the TF clustering step will be described in Section 3.2.

#### 3.1. WF separation and mask estimation

In this subsection, we incorporate the t-f mask concept into the WF approach. First, we define the log-likelihood function for the complete data $\mathcal{D} = \{x_{n,f};z_{n,f};\{c_{n,f}^{(l)}\}\}$ as follows:

$$\mathcal{L}(\theta;\mathcal{D}) = \sum_{n} \sum_{f} \log \int_{c} L_{n,f} dc$$

where

$$L_{n,f} = \sum_{z_{n,f}} p(\{x_{n,f},z_{n,f};\{c_{n,f}^{(l)}\}\};\theta)$$

and $\int_{c} dc$ denotes the marginalization with respect to the hidden variable $c_{n,f}^{(l)}$. In (8), we consider the mixture weight $p(z_{n,f})$ in our model, unlike the method in \cite{7}. The term $p(\{c_{n,f}^{(l)};z_{n,f}\};\theta)$ in (8) is given by the source image model (2), and the term $p(\{x_{n,f};\{c_{n,f}^{(l)}\}\};z_{n,f};\theta)$ is by (4).

#### 3.1.1. EM algorithm

The likelihood function (6) can be maximized by using the EM algorithm. Here, the dominant source index $\{z_{n,f}\}$ and the source images $\{c_{n,f}^{(l)}\}$ are dealt with as hidden variables.

**E-step (Mask estimation):** The auxiliary function $Q$ is given as

$$Q(\theta;\theta') = \sum_{n} \sum_{f} \sum_{z_{n,f}} \int_{c} p(\theta^{(l)};\{z_{n,f}\};z_{n,f};\theta) \log p(D;\theta) dc$$

At the E-step, we calculate this $Q$ function by calculating

$$p(\{c_{n,f}^{(l)}\};z_{n,f};x_{n,f}) = \prod_{l} p(\{c_{n,f}^{(l)}\};z_{n,f};x_{n,f})$$

The second term of the right-hand side of (10), $M_{n,f}^{(l)} = p(z_{n,f} = l|x_{n,f})$, can be obtained as

$$M_{n,f}^{(l)} = \frac{p(x_{n,f},z_{n,f} = l; \{B_f^{(l)}\}, \{v_{n,f}^{(l)}\})}{\sum_{l'} p(x_{n,f},z_{n,f} = l'; \{B_f^{(l')}\}, \{v_{n,f}^{(l')}\})}$$

where

$$g_{n,f}^{(l)} = \exp \left(-\frac{1}{2} \frac{\|x_{n,f}\|^2}{U_{n,f}^{(l')}|^{1/2}|v_{n,f}^{(l')}B_f^{(l')}|^{-1/2}} - \frac{1}{2} \frac{\|x_{n,f}\|^2}{U_{n,f}^{(l)}|^{1/2}|v_{n,f}^{(l)}B_f^{(l)}|^{-1/2}} \frac{1}{2} \log |v_{n,f}^{(l)}B_f^{(l)}| \right)$$

**M-step (WF separation):** At the M-step, we calculate $\theta = (B_f^{(l)}), (v_{n,f}^{(l)})$ which maximize the $Q$ function. The terms related to $B_f^{(l)}$ and $v_{n,f}^{(l)}$ in the $Q$ function can be given as

$$Q' = \sum_{n} \sum_{f} \left(-\frac{1}{2} \text{Tr}(M_{n,f}^{(l)}(v_{n,f}^{(l)}B_f^{(l)})^{-1}R_{n,f}) - \frac{1}{2} \log |v_{n,f}^{(l)}B_f^{(l)}| \right)$$

\footnote{In the E-step, there is no need to calculate explicitly the first term of the right-hand side of (10).}
where
\[ \hat{R}^{(l)}_{n,f} = r^{(l)}_{n,f} + \hat{c}^{(l)}_{n,f} (c^{(l)}_{n,f})^H, \]  
(13)
\[ \hat{c}^{(l)}_{n,f} = W^{(l)}_{n,f} x_{n,f}, \]  
(14)
\[ r^{(l)}_{n,f} = (I - W^{(l)}_{n,f}) v^{(l)}_{n,f} B^{(l)}_f, \]  
(15)
\[ W^{(l)}_{n,f} = v^{(l)}_{n,f} B^{(l)}_f X^{-1}_{n,f}, \]  
(16)
\[ X_{n,f} = \sum_{l} v^{(l)}_{n,f} B^{(l)}_f = U^{(-1)}_{n,f} + v^{(l)}_{n,f} B^{(l)}_f. \]  
(17)
In the above equations, \( \hat{c}^{(l)}_{n,f} \) in (14) gives us the expectation value of the source image \( c^{(l)}_{n,f} \) (i.e., separated signal), and it is calculated by using a multi-channel WF (16).

From the function \( Q' \), the update rules for \( B^{(l)}_f \) and \( v^{(l)}_{n,f} \) can be derived as:
\[ a^{(l)}_{n,f} = \frac{1}{N_m} \text{Tr} \left( (B^{(l)}_f)^{-1} \hat{R}^{(l)}_{n,f} \right), \]  
(18)
\[ v^{(l)}_{n,f} = M^{(l)}_{n,f} a^{(l)}_{n,f}, \]  
(19)
\[ B^{(l)}_f = \hat{R}^{(l)}_{n,f} P_{1} = \frac{1}{N_n} \sum_{n=1}^{N_n} a^{(l)}_{n,f} \hat{R}^{(l)}_{n,f}. \]  
(20)
The update rule for the mixture weight \( p(z_{n,f} = l) \) is
\[ p(z_{n,f} = l) = \sum_{n} \sum_{f} M^{(l)}_{n,f} / N_n N_f, \]  
(21)
where \( N_n \) and \( N_f \) are the number of time frames and frequency bins.

After the iterative parameter estimation, the final estimates of the source images \( c^{(l)}_{n,f} \) are obtained by using (14).

In the algorithm, the posterior \( M^{(l)}_{n,f} \) in (11) can be interpreted as a t-f mask. As shown in (19), the source variance \( v^{(l)}_{n,f} \) depends on the posterior \( M^{(l)}_{n,f} \), that is, the WF \( W^{(l)}_{n,f} \) in (16) is influenced by a t-f mask \( M^{(l)}_{n,f} \). It is worth noting that, when \( M^{(l)}_{n,f} = 1 \) for all t-f slot \( (n,f) \), the algorithm is equivalent to the multi-channel WF approach in [7].

3.2. TF clustering

Because the hybrid approach includes the mask estimation, we can include the existing t-f clustering based approach (e.g., [1–3]) into the hybrid framework as the TF clustering step. This paper employs the mixing vector \( h_f \) clustering based algorithm [3] for the TF clustering step, because it achieves relatively high separation performance [6]. In Fig. 1, \( \phi \) denotes a parameter set for the TF clustering.

As the TF clustering can provide the variance estimations of separated signals in a computationally efficient way, we utilize this advantage to the hybrid framework. In concrete terms, as the \( \psi \) in Fig. 1, the TF clustering step outputs the variance estimation \( v^{(l)}_{n,f} = M^{(l)}_{n,f} |x^{(l)}_{n,f}|^2 \) where \( x^{(l)}_{n,f} \) is the first channel observation.

As in [4, 5], because the TF clustering also estimates the t-f mask \( M^{(l)}_{n,f} \), it is possible to estimate the number of signals \( N_s \) at the TF clustering step in a straightforward way. However, such an estimation of \( N_s \) at the TF clustering step in the hybrid approach is beyond the scope of this paper.

3.3. Algorithm summary

The following is the overall procedure for the hybrid approach.

1. Run the TF clustering to estimate \( v^{(l)}_{n,f} \) (see Section 3.2).
2. Initialize \( B^{(l)}_f = \frac{1}{N_n} \sum_{n} x_{n,f} x_{n,f}^H |x_{n,f}|^2 \).
3. Iterate the following steps until convergence is achieved.
   (a) Update \( M^{(l)}_{n,f} \) with (11).
   (b) Calculate \( \hat{R}^{(l)}_{n,f} \) with (13)-(17), and update \( v^{(l)}_{n,f} \) and \( B^{(l)}_f \) with (19) and (20), and \( p(z_{n,f}) \) by (21).
4. Calculate the separated signal by using (14).

In the experiments, we also evaluate the separation performance without the TF clustering. For this setting, we just skip step 1 and we initialize the variance at step 2 by \( v^{(l)}_{n,f} = M^{(l)}_{n,f} |x_{n,f}|^2 \), where \( M^{(l)}_{n,f} \) is a random t-f mask. We refer to this parameter setting as the random initialization (RI).

The way for estimating the number of sources with the hybrid approach will be shown in Section 4.2.

4. EXPERIMENTS

We performed experiments with measured impulse responses in a room, whose reverberation times \( T_R \) were 250 or 400 ms. The microphone spacing was 4 cm and the number of sources was \( N_s = 2 \) or 3. \( N_s \) was given in Section 4.1, and unknown in Section 4.2. Mixtures were made by convolving the measured room impulse responses and 8-second English speech signals. For each reverberation condition, we tried six speaker combinations. The sampling rate was 8 kHz, and the frame size for STFT and the frame shift were 256 (128 ms) and 256 (32 ms), respectively. The permutation problem was solved by synchronizing the mask \( M^{(l)}_{n,f} \) of all frequency bin with the method in [3].

The performance was evaluated in terms of the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR), which are defined in [8].

4.1. When the number of sources \( N_s \) is given

This section shows that, when the number of sources \( N_s \) is given, the hybrid approach can obtain high performance with the small iteration number for the WF separation.

Figure 2 indicates an example performance for each iteration. As shown in the figure, the SDR is almost saturated at the first iteration. Even when we iterate the algorithm 50 times, the performance gain is small. From this result, we can conclude that we can obtain sufficient performance with the non-iterative WF calculation. This greatly reduces the calculation time for the WF separation step.

Figure 3 summarizes the performance with the TF clustering itself, with the hybrid approach (HB) with one or 50 iteration(s) ("HB1" or "HB50"), and performance of the HB with the RI where iteration number was 50 ("HB50 (RI)"). It should be noted that, with the RI we require many iterations to achieve source separation. It can be seen that the HB1 can improve the performance from the TF, especially the SAR values. It can also be seen from the SDR values that the HB1 outperforms the HB50 and the HB50 (RI), which requires 50 times as expensive computational cost as HB1. From this result, we can say that we can reduce the computational cost for the WF separation step by using the TF clustering result.

2The image-to-spatial distortion ratio (ISR) was also evaluated in the experiments, but not shown in this paper.
4.2. When number of sources $N_s$ is unknown

In this section, we show that we can estimate the number of sources $N_s$ and obtain separated signals even when $N_s$ is unknown.

Even when $N_s$ is unavailable, we can still run the hybrid algorithm by preparing a sufficient number of classes. That is, we prepare $L > N_s$ classes and replace $\sum_{z_{n,f}=1}^{N_s} p_{n,f}$ with $\sum_{z_{n,f}=1}^{L} p_{n,f}$ in the algorithm. In this section we set $L = 8$. Here, as we set initial $\theta$ values with the RI, we need a large number of iteration. In this section, the number of iteration was 100.

Figure 4 shows example mixture weight parameters $p(z_{n,f})$ which are calculated from the t-f masks $M_{n,f}^{(s)}$ by (21) after the convergence. We can see that, for example when $N_s = 2$, two ($= N_s$) have large values and other values become small. So, we can estimate the number of sources by counting the number of dominant mixture weights

$$p(z_{n,f}) > 1/L.$$  (22)

These results show that, with the HB, we can estimate both the number of sources $N_s$ and the source images $e_{n,f}$, simultaneously.

5. CONCLUSION

In this paper, we proposed the hybrid method of the t-f mask and the multi-channel WF approaches. By bringing the t-f mask concept into the WF approach, we can reduce the calculation load of the WF approach, and can estimate the number of sources. We confirmed that, when the number of sources is given, the hybrid approach achieves reasonably high performance without iterating the WF calculation.

We also obtained encouraging results for the case where the number of sources is unavailable.

6. REFERENCES


