Chapter 1

Source Separation and DOA Estimation for Underdetermined Auditory Scene

Nozomu Hamada and Ning Ding

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/56013

1. Introduction

In human-machine communication the separation of a target speech signal and localization of it in noisy environments are very important tasks. [1] For carrying out these tasks recent advanced sensor array signal processing is promising technology. [2] It utilizes the collection of multi-channel acoustic data by an array of microphones for detecting and producing output signals which is much more intelligible and suitable for communication and automatic speech recognition. [3]

BSS problem

Blind source separation (BSS) aims to estimate source signals by only using their mixed signals without any a priori information about mixing process and acoustic circumstances. The cocktail-party problem is one of the typical BSS problems. [1] Basically, the BSS problem can be solved by exploiting intrinsic properties of speech signals. Depending on the inherent properties there have been proposed lots of methods for BSS problems on speech signals. Among them the most widely applied approaches are the following two.

1. Independent component analysis (ICA)[4]-[8], and
2. Time-Frequency sparseness of source signals [9]-[14].

The ICA-based separation relies on *statistical independence* of speech signals in time-domain [5] [7] as well as in frequency-domain [6]. In addition, [8] proposed a dynamic recurrent separation system by exploiting the spatial independence of located sources as well as temporal dependence. On the other hand the second approach exploits the *sparseness* of speech signals in time-frequency (T-F) domain where only small number of T-F components are dominant in representing a speech signal. The T-F sparseness leads the disjoint property of T-F domain

components, called W-disjoint orthogonality (WDO) property [11] [12], between speech signals. It means that at most one source dominates at every T-F points, in another word; different speech signals rarely generate the same frequency at the same time.

Though ICA approach performs well even in a reverberant condition, it is difficult to solve the underdetermined case in which the number of sources is greater than the number of sensors. Additionally, the frequency-domain ICA [6] the permutation ambiguity of its solution is a serious problem. It needs to align the separated frequency components that originate from the same source.

The T-F masking method which is the most popular sparseness-based approach is the topic concerned in this chapter. The representative method is known as DUET (Degenerate Unmixing Estimation Technique) [11]. A flow of conventional sparseness-based separation can be summarized as follows.

Sparseness-based T-F masking

Observed signals in T-F domain:

Transform time domain acoustic observations during few seconds to the T-F domain signals by applying short time Fourier transform (STFT) where a sparse representation of speech signal is obtained. [15] Thus the T-F components of a speech signal distribute in T-F domain without overlapping with T-F components of other speech signals.

Features of T-F cells:

As known in auditory scene analysis interaural time differences and level differences are significant spatial features of sources. [1] These localization cues are estimated from the differences in the direction and the distance of speakers. Actually, in microphone array the geometric parameters of sources can be obtained from phase differences and attenuation ratios at the mixture T-F cells.

Clustering T-F cells:

Under the WDO assumption the distribution of feature vectors obtained at all T-F cells makes as many clusters as the number of sources. The essential task of separation therefore turns out to cluster the feature vectors. The preliminary clustering method adopted in [9] - [12] is to make the histogram of features and to find the peaks corresponding the sources. Each T-F cell in the mixed signal is thereby associated with one peak depending on the distance in the cell's feature space.

Masking T-F cells:

Utilizing the clustering results individual binary masks are applied to the T-F domain spectrogram to detect the components that originate from individual sources.

Inverse transform:

A set of masked T-F components are inversely transformed by STFT and then it provides restored speech signal.

Remarks:

1. T-F domain sparseness in speech signals is also employed as a separation principle in the context of single channel or monaural signal source separation problem where harmonic structure in spectrogram is crucial for segregation.[16] [17]

2. Associated with the features of T-F cells conventionally used features are summarized in [13] and the features are evaluated from the separation performance point of view.

3. Clustering scheme in T-F masking would be crucial for high separation ability. Subsequent studies after DUET-like approaches [11][12], maximum-likelihood (ML) based method for real-time operation [18], k-means algorithm or hierarchical clustering, and EM algorithm [19] have been proposed. The method called MENUET [13] applies k-means algorithm to a vector space consisting of the signal level ratio and the frequency-normalized phase difference with appropriately weighting terms for effective clustering. They solve the optimization problem by adopting an efficient iterative update algorithm. In [14] k-means algorithm is applied clustering spatial features for arbitrary sensor array configuration even with wider sensor distance where spatial aliasing may occur. Their clustering procedure is divided into two steps, the first one of which is applicable to the non-aliasing or lower frequency band and the second one treats the remaining aliasing occurred frequency band.

DOA estimation

Localization of acoustic sources using microphone array system is a significant issue in many practical applications such as hands-free phone, camera control in video conference system, robot audition, and so on. The latter half of this chapter focuses on the Direction-Of-Arrival (DOA) estimation of sources. Since this monograph interests in speech signals, we make no mention of the methods addressed for narrow-band signals, for instance in radar/sonar processing. There have been proposed a large number of DOA estimation methods for broadband signals [20], [21]. Typical array processing approaches are;

1. Generalized Cross-Correlation (GCC) methods [22]

2. Subspace approaches using spatial covariance matrix of observed signals [23]

3. T-F domain sparseness-based approaches [11],[24]-[27]

4. ICA separation based approaches [28]

The first category of GCC method is to estimate the delay time that maximizes a generalized cross-correlation function between the filtered outputs of the acquired signals at microphones. The phase transform (PHAT) method [22] exploits the fact that the Time-Delay-Of-Arrival (TDOA) information is conveyed in the phase. Although GCC methods are usually performed well and are also computationally efficient for single source case, it does not cope with multiple sources case in which this chapter interests.

The second category is the subspace analysis applying a narrowband signal model. The analysis uses the properties in the spatial covariance matrix of multichannel array observa-

tions. The MUSIC-like algorithms are well-known methods for narrowband target signals. For broadband signals such as speech, several frequency-domain approaches have been proposed. The subspace-based approaches for small number of sensors have to overcome two drawbacks, one of which is the limited precision for DOA estimation, and the other is that it is unable to deal with the underdetermined case.

Sparseness-based approaches

The third category of the DOA estimation algorithms is based on sparseness of speech signals and is closely related to the BSS. Source sparseness assumption implies WDO or its weaker condition TIFROM [24]. These conditions are the crucial properties to solve DOA problems for underdetermined multiple sources. The BSS approach associated with these assumptions is a group of T- F masking framework. In DUET-like methods [9]-[14], the delay time or the frequency-normalized ratio of the frequency-domain observations at each T-F point is used to compute the TDOA. An alternative DOA estimation method proposed by Araki et al. [27], in the context of k-means algorithm, estimates DOA as the individual centroid of each cluster of normalized observation vectors corresponding to an individual source. The DEMIX [25] algorithm introduces a statistical model in order to exploit a local confidence measure to detect the regions where robust mixing information is available. The computational cost of DEMIX would be high due to performing the principal component analysis for every local scatter plot of observation vectors at individual T-F points.

For addressing robust cocktail-party speech recognitions the localization cue such as TDOA or spatial direction evaluated at each T-F cell has a central role. As in [29][30], integrating approaches the segregation/localization of sound sources and speech recognition against background interferences are significant CASA (Computational Auditory Scene Analysis) front-ends.

DOA Tracking

Not only estimating but also tracking sound sources draws lots of attentions recently in robot auditory systems. For instance, speaker's DOA tracking by microphone array mounted on mobile robot is the problem of moving sources and moving sensors.

BSS and DOA Problems:

The underlying BSS and DOA estimation problems addressed in this chapter are listed as follows:

a. Use of a pair of microphones

b. Multiple simultaneously uttered speech signals under the assumption that the number of sources is known a priori

c. Underdetermined cases, where the sources outnumber the sensors

d. The inter-sensor distance is bounded so as to avoid spatial aliasing (for instance, less than 4 cm spacing for an 8 kHz sampling rate)

http://dx.doi.org/10.5772/56013

While stereophonic sensor is the simplest sensor array, the study of how to improve the separation performance and to obtain accurate DOA by a pair of microphones is meaningful because any complex array configuration can be considered as an integration of these.

The rest of this chapter is organized as follows. In section 2, problems of underlying BSS and DOA estimation are described in detail. The proposed BSS method based on a frame-wise scheme is introduced in section 3. Section 4 describes a DOA estimation algorithm by using T-F cell selection and the kernel density estimator. The last section concludes this chapter.

2. Problem descriptions

2.1. Observation model

Source mixing models in time domain and its T–F domain description are described as follows. All discrete time signals are sampled version of analog signals with sampling frequency f_S. Suppose N source signals $s_1(t)$, $s_2(t)$, $\cdots$, $s_N(t)$ are mixed by time-invariant convolution and the observed signals $x_1(t)$, $x_2(t)$, $\cdots$, $x_M(t)$ at M sensors with omni-directive characteristic are described as:

$$x_m(t) = \sum_{i=1}^{N} \sum_{\tau} h_{mi}(\tau) s_i(t-\tau), \tag{1}$$

where $h_{mi}(\tau)$ represents the impulse response from i-th source to m-th sensor. Observed signals $x_m(t)$ (m=1~M) are converted into T–F domain signals $X_m[k, l]$ by using L-point windowed STFT as written by

$$X_m[k,l] = \sum_{r=-L/2}^{L/2-1} x_m(r+kS)\, win(r)\, e^{-j\frac{2\pi l}{L}r}, \qquad k = 0 \sim K, l = 0 \sim \frac{L}{2} \tag{2}$$

where r is dummy variable in convolution sum operation, $win(r)$ is a window and S is the window shift length. Here, we apply half window size overlapping transformation, namely $S = L/2$ in (2). Transformed T–F mixture model of Eq.(1) can be described by the instantaneous mixtures at each time frame index k and frequency bin l.

$$X_m[k,l] = \sum_{i=1}^{N} H_{mi}[l] S_i[k,l] \tag{3}$$

where $H_{mi}[l]$ is the frequency response (DFT) of $h_{mi}(t)$, $S_i[k, l]$ is the windowed STFT representation of i-th source signal $s_i(t)$, and the point $[k, l]$ is called "*T-F cell*" in this chapter.

Assuming an anechoic mixing, the source signals which we want to recover are alternatively redefined as the observed signals at the first mixture $x_1[k, l]$. In this case, the following mixing models in the T–F domain are henceforth considered without loss of generality.

$$\begin{aligned} X_1[k,l] &= \sum_{i=1}^{N} S_i[k,l], && \text{a} \\ X_m[k,l] &= \sum_{i=1}^{N} H_{mi}[l] S_i[k,l] \qquad m = 2 \sim M && \text{b} \end{aligned} \tag{4}$$

where $S_i[k, l]$ and $H_{mi}[l]$ are different from $\mathsf{S}_i[k, l]$ and $\mathcal{H}_{mi}[l]$ in (3), $S_i[k, l]$ is the i-th source signal observed at the first sensor (m=1), and $H_{mi}[l]$ eventually represents the DFT domain operation of the transfer function with relative attention and delay between m-th and the first sensors.

From then on, consider the mixture of two sources $S_1[k, l]$ and $S_2[k, l]$ which are received at a pair of microphones. Their mixture system (4a) and (4b) can thus be expressed as

$$\begin{bmatrix} X_1[k,l] \\ X_2[k,l] \end{bmatrix} = \begin{bmatrix} 1, & 1 \\ H_{21}[l], & H_{22}[l] \end{bmatrix} \begin{bmatrix} S_1[k,l] \\ S_2[k,l] \end{bmatrix} \tag{5}$$

2.2. Basic assumptions

As stated in Section 1, the WDO is commonly supposed in sparseness-based separation approaches. At first, we denote the T-F domain Ω on which $S_1[k, l]$ and $S_2[k, l]$ are defined

$$\Omega := \{[k,l], k = 0 \sim K, l \in B\} \tag{6}$$

where $B := [l_1, L/2]$ is the frequency band after deleting lower frequency components which do not exist in actual speech signals, and $l_1 = \lfloor f_1 L / f_s \rfloor$ means the Gauss floor function, and f_1 is the analog lowest frequency of speech components such as 80Hz in later experiments.

Next, define the T-F supports $\Omega_i (i = 1, 2)$ of $S_i[k, l] (i = 1, 2)$ by

$$\Omega_i := \{[k,l] \| S_i[k,l] \| > \varepsilon\} \quad i = 1,2 \tag{7}$$

where $\varepsilon (>0)$ is a sufficiently small value. Although, in theory, the support of $S_i[k, l] (i = 1, 2)$ is defined by the condition $| S_i[k, l] | \neq 0$, Eq. (7) gives a set of components of actual signals except noise-like ones satisfying $| S_i[k, l] | < \varepsilon$.

http://dx.doi.org/10.5772/56013

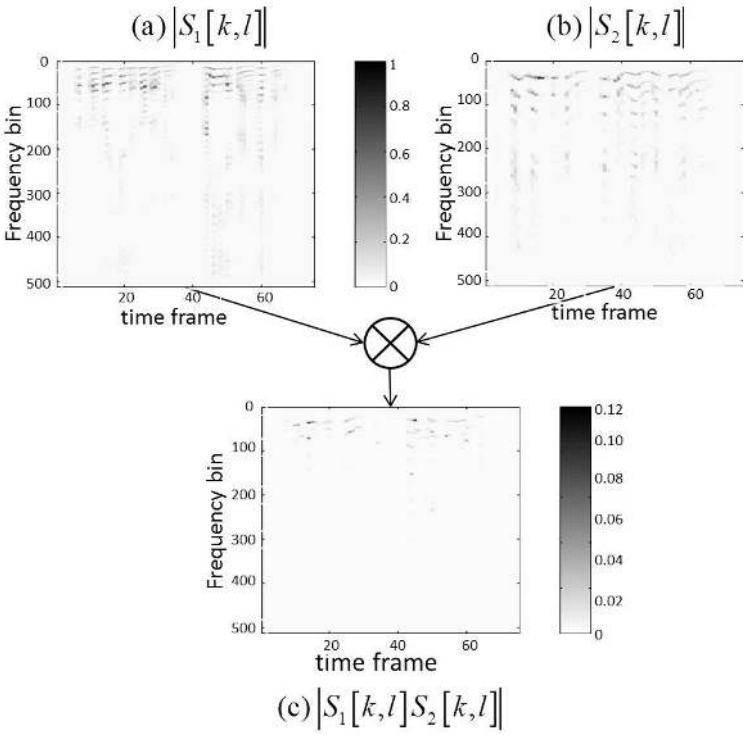


Figure 1. WDO property of two real speech signals

We may consequently express the WDO assumption between two source signals $s_1(t)$ and $s_2(t)$ by the disjoint condition

$$\Omega_1 \cap \Omega_2 = \phi(empty\ set) \tag{8}$$

This can equivalently be represented as follow.

$$S_1[k,l]S_2[k,l] = 0 \quad \text{at any } [k,l] \tag{9}$$

The verification of above WDO condition for actual speech signals is performed in Fig. 1 where (a) and (b) show spectrograms of two speech signals in the T-F domain, and (c) shows their multiplication in which we see rarely overlapping between two spectrograms.

Obviously, the supports of $X_1[k, l]$ and $X_2[k, l]$ are coincident and it, denoted by Ω_X, can be given as

$$\Omega_X = \Omega_1 \cup \Omega_2 \tag{10}$$

In addition the following null component domain, denoted by Ω_N, is also introduced as

$$\Omega_N = \bar{\Omega}_X = \overline{\Omega_1 \cup \Omega_2} \qquad \bar{X} : \text{complementary set of } X \tag{11}$$

The WDO condition (8) accordingly derives that the T–F domain representation of the mixed signal $X_1[k, l]$, given by Eq.(5), can be decomposed into the following three parts with no overlapping in Ω.

$$X_1[k,l] = \begin{cases} S_1[k,l] & [k,l] \in \Omega_1 \\ S_2[k,l] & [k,l] \in \Omega_2 \\ 0 & [k,l] \in \Omega_N \end{cases} \tag{12}$$

2.3. Source separation

Under the WDO assumption expressed in (12), the binary masking in the T–F domain is performed as follow:

Clustering the T-F cells in the support Ω_X of the mixture $X_1[k, l]$ into two sub-regions Ω_1 and Ω_2, the separated source estimates in T-F domain, $\hat{S}_1[k, l]$ and $\hat{S}_2[k, l]$, are obtained by applying the masks

$$M_i[k,l] = \begin{cases} 1 & [k,l] \in \Omega_i \\ 0 & otherwise. \end{cases} \quad (i = 1,2) \tag{13}$$

on $X_1[k, l]$ as follows.

$$\hat{S}_i[k,l] = M_i[k,l] X_1[k,l] \quad (i = 1,2) \tag{14}$$

Clustering features

The separation task is to classify T–F cells composing the support Ω_X of $X_1[k, l]$ into either Ω_1 or Ω_2. A pair of $X_1[k, l]$ and $X_2[k, l]$ is used to characterize a T-F cell $[k, l]$ at which spatial features are introduced, and the clustering process is performed in the estimated feature space.

Effective features must be the signal level or attenuation ratio defined by

$$\alpha[k,l] := \left| \frac{X_1[k,l]}{X_2[k,l]} \right| \tag{15}$$

and the arrival time difference defined by the frequency-normalized phase difference (PD) between $X_1[k, l]$ and $X_2[k, l]$ as

$$\delta[k,l] := \frac{L}{2\pi f_s l} \phi[k,l] \tag{16}$$

where $\phi[k, l]$ is the PD as defined by

$$\phi[k,l] = \angle X_1[k,l] - \angle X_2[k,l] \tag{17}$$

Other features used for characterizing T-F cells are listed in [13]as well as the attenuation ratio modifications. It is noted that the attenuation ratio would not give distinctive difference for short distance microphone array. In our experimental setup, for example, the distance between microphones is 4cm in order to avoid spatial aliasing at 8kHz sampling rate.

Clustering scheme

For given features at T-F cells in Ω_X, clustering of these is the next step. In DUET where a pair of microphones is used, the two dimensional histogram of feature vectors $\{\alpha[k, l], \delta[k, l]\}^T$ within a time interval, such as for several seconds, is generated and the clustering is performed by finding the maximum peaks which are corresponding to sources. When the attenuation feature is omitted the clustering problem is solely performed based on time delay histogram distribution. The dimension of feature space will be higher for array configuration with many microphones than two. For these cases more sophisticated clustering scheme such as k-means algorithm or EM algorithm [19] should be adopted.

Inverse STFT

The final stage of the separation process is to obtain time domain separated signals $\hat{s}_i(t)$ $(i=1, 2)$ by applying the inverse STFT.

3. Sound source separation

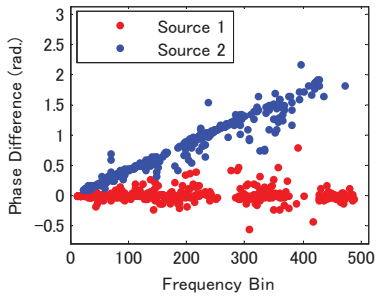
3.1. Phase–difference vs. frequency data

As a T-F cell's feature depending on the spatial location difference of sources, our strategies exploit a frame-wise, namely, a time sequence of phase difference of observations versus

frequency (PD-F) distribution. In a k-th frame, the point plot of the PD-F is defined as a collection of two-dimensional vectors at k-th frame $p_k(l)$as

$$\boldsymbol{p}_k(l) := \left\{ l, \phi\left[k,l\right] \right\}^T, \quad l \in B \quad k \in [1,K] \tag{18}$$

An example of PD-F in (l, ϕ)-plane and its time sequence for the mixture of two speech signals are shown in Fig.2 (a) and (b) respectively.



(a) PD-F plot

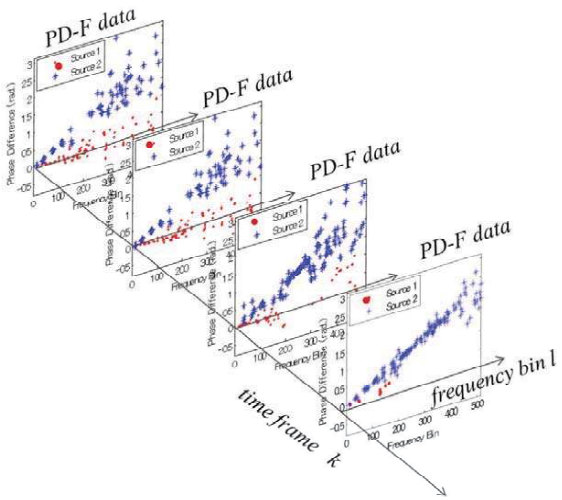


Figure 2. PD-F and time sequence of PD-F (Blue and red points respectively corrrespond invividual source components)

The relationship between the gradient β of a vector in PD-F plane defined in Eq.(18) and the source direction θ is: [33]

$$\beta = \left(\frac{2\pi}{L}\right)\frac{f_s d}{c}\sin\theta \tag{19}$$

where d is the distance between the sensors, c is the sound velocity, and θ is the direction of source. Here $\theta=0$ corresponds to the broadside direction and the term $(d/c)\sin\theta$ represents the wave arriving delay between microphones. For example, the dot distribution in Fig.2 (a) concentrates along two lines corresponding to two source directions. By determining the gradients of these lines two directions of sources are estimated from the relationship of (19).

The conventionally utilized features associating with delay time at each T-F cell can be estimated from the frequency normalization of PD-F dot corresponding to individual T-F cells. Unlike the conventional delay-like features PD-F dots keep a linear dot distribution on the plane and it is effectively utilized in both following source separation and direction finding methods.

3.2. Frame categorization

Fig. 3(a) shows two simultaneously uttered speech signals. In the figure four frame time points k_1 - k_4 indicated by the red rectangular parts are shown as the following four types of source signal activity states:

Frame $k=k_1$; No source signal is active (Non Source Active:NSA)

Frame $k=k_2$; Only the first source is active (Single Source SSA)

Frame $k=k_3$; Only the second source is active (Single Source SSA)

Frame $k=k_4$; Both sources are active (Double Source Active:DSA)

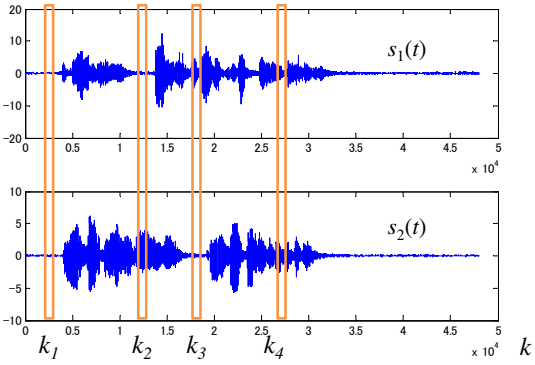
Here we may define three sets of time-frame indeces as follows:

The whole set of time-frames, denoted by $\boldsymbol{K}$:={1,.....,K}, is categorized into three sets with no overlapping.

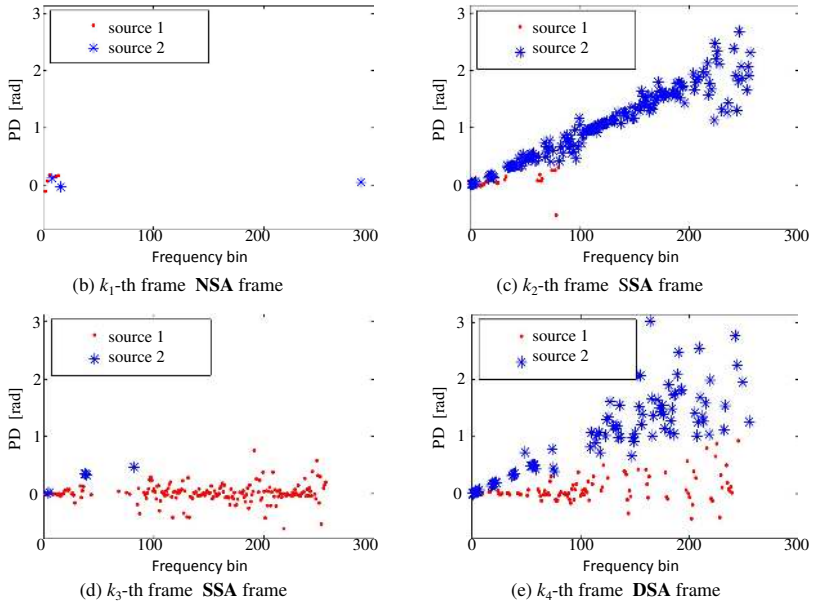
$$\boldsymbol{K} = \boldsymbol{K}_{NSA} \cup \boldsymbol{K}_{SSA} \cup \boldsymbol{K}_{DSA} \tag{20}$$

In addition, we define the following Sourse Active(SA) frame index set.

$$\boldsymbol{K}_{SA} = \boldsymbol{K}_{SSA} \cup \boldsymbol{K}_{DSA} \tag{21}$$



(a) Two speech signals



(b) k_1-th frame **NSA** frame

(c) k_2-th frame **SSA** frame

(d) k_3-th frame **SSA** frame

(e) k_4-th frame **DSA** frame

Figure 3. Frame categorization (NSA, SSA, DSA)

Above frame categorization suggests the source separation algorithm consisting of the following two parts:

- Assign each T-F component at SSA frame to either source by identifying the direction.
- Apply separation algorithm solely to DSA frames

The detail of these will be described in the next section.

http://dx.doi.org/10.5772/56013

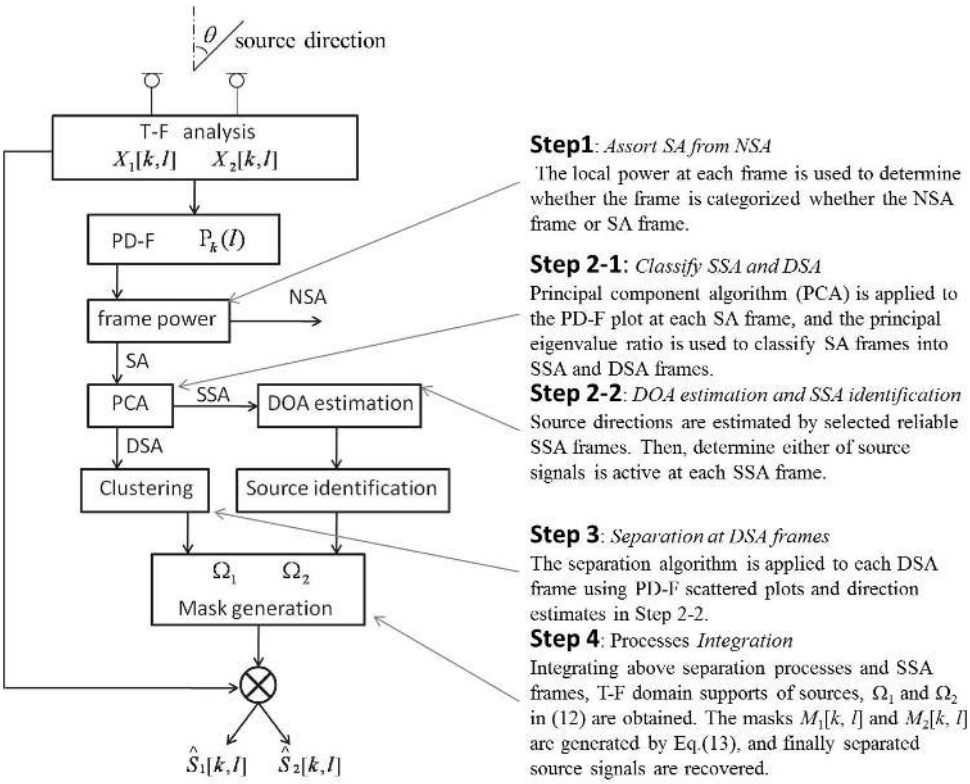


Figure 4. Flow of source separation method

3.3. Source separation algorithm

Outline of the method

The outline of the separation method using PD-F plot is shown in Fig.4 and summarized.

Step1: *Discriminate SA from NSA*

The following average power at a frame is employed to check the presence of speech signal at the frame.

$$E(k) := \frac{1}{L/2 - l_1 + 1} \sum_{l \in B} \left| X_1[k,l] \right|^2 \tag{22}$$

Here, the threshold operation is valid for basic voice activity detection as follow.

$$\boldsymbol{K}_{SA} = \left\{ k \middle| \, E(k) > Th_{SA} \right\} \tag{23}$$

where Th_{SA} is determined by a pre-experiment of noise level estimate during no utterance. In later experiments, we applied the following formula.

$$Th_{SA} = E_0 + 2\sigma_E \tag{24}$$

where E_0 is the average noise power estimate and σ_E is the standard deviation estimate given by respectively.

$$E_0 := \frac{1}{\left| \boldsymbol{k}_{NSA} \right|} \sum_{k \in \boldsymbol{k}_{NSA}} \left| X_1[k,l] \right|^2 \tag{25}$$

$$\sigma_E := \sqrt{\frac{1}{\left| \boldsymbol{k}_{NSA} \right|} \sum_{k \in \boldsymbol{k}_{NSA}} \left(E(k) - E_0 \right)^2} \tag{26}$$

Step 2-1: *Classify SA into SSA and DSA*

At each $k \in K_{SA}$ PCA is applied to the set of vectors $p_k(l)$ by computing the following 2 × 2 covariance matrix.

$$\boldsymbol{R}_k := \frac{1}{L/2 - l_1 + 1} \sum_{l \in B} \boldsymbol{p}_k(l) \, \boldsymbol{p}_k^{\mathrm{T}}(l) = \begin{bmatrix} R_{11}(k) & R_{12}(k) \\ R_{21}(k) & R_{22}(k) \end{bmatrix}. \tag{27}$$

Denoting the eigenvalues of R_k by $\lambda_1(k)$ and $\lambda_2(k)$ (assume $\lambda_1(k) \geq \lambda_2(k)$), the ratio of the principal eigenvalues defined by

$$r(k) := \frac{\lambda_2(k)}{\lambda_1(k)}. \tag{28}$$

is introduced to discriminate the SSA frames from the DSA. As shown in Fig.3 (c), (d), PD-F vector distribution at a SSA frame tends to concentrate around the first principal axis. This observation leads to the following discrimination of SSA from DSA frames and the estimation of the source directions.

The following criterion is applied to detect SSA frames.

$$\boldsymbol{K}_{SSA} = \left\{k \middle| r(k) < Th_{SSA}\right\} \tag{29}$$

where Th_{SSA} is determined experimentally.

Step 2-2:*DOA estimation and SSA identification*

Define the normalized eigenvector of the first principal eigenvalue as

$$\boldsymbol{e}_1(k) := \begin{bmatrix} \cos\beta(k) \\ \sin\beta(k) \end{bmatrix} \tag{30}$$

where $\beta(k)$ is the gradient of the principal axes at k-th frame. The histogram of the set

$$\left\{\beta(k), k \in \boldsymbol{K}_{SSA}\right\} \tag{31}$$

will have two peaks which are corresponding two source directions θ_1 and θ_2 caluculated by Eq. (19). By clustering the set of θ into two groups according to the distance from θ_1 and θ_2, each SSA frame in K_{SSA} is classified into each one of the sources from the direction θ_1 and θ_2.

Double Source Active (DSA)

For given set of DSA frames K_{DSA}, the clustering of the vectors $p_k(l),\ l \in B$ into two sets is the problem. Before describing this separation algorithm, three frequency bands, denoted by B_{high}, B_{low}, and B_{mid}, are introduced to use in the following separation algorithm.

Frequency Bands

The following three frequency bands are defined respectively.

$B_{high} := \{l \mid l_2 < l < L/2\},\ B_{low} := \{l \mid l_1 < l < l_2\},\ B_{mid} := \{l \mid l_2 < l < l_3\}$

where $l_i = \lfloor f_i L / f_s \rfloor,\ (i = 2, 3)$, f_2 is set 400Hz, and f_3 is set 1kHz in later experiments.

The idea of source separation at DSA frames utilizing these bands is divided into two parts according to above frequency bands.

1. The first scheme, called initial separation, is applied to the T-F cells in B_{high} based on the directions of sources which have been estimated at the SSA frames previously.
2. The clustering in B_{low} is performed utilizing a harmonic structure relationship between the spectral components in B_{low} and that of B_{mid}. The harmonic structure in B_{mid} can be obtained by the initial separation results in B_{high}.

1. *Initial separation*

Denote the source directions estimated in SSA frames by θ_1 and θ_2, and their corresponding gradients in PD-F plane are β_1 and β_2 as defined in Eq.(31). The points on these two lines can be expressed as

$$\phi[k,l] = \beta_i\, l \quad (i = 1,2) \tag{32}$$

At $k \in K_{DSA}$, the nearest neighbor rule gives the binary mask $\widetilde{M}_i[k, l]$ in B_{high} which is defined as

$$\tilde{M}_i[k,l] = \begin{cases} 1, & if\ i = \arg\min_c \left|\phi[k,l] - \beta_c\, l\right|, l \in B_{high} \\ 0, & otherwise. \end{cases} \tag{33}$$

As a result, the separated individual signals $\tilde{S}_i[k, l]$ (i = 1, 2) are represented by

$$\tilde{S}_i[k,l] = \tilde{M}_i[k,l] X_1[k,l], \quad l \in B_{high} \tag{34}$$

2. *Separation in B_{low}*

Local maximum points in B_{mid}

The final task for separation process is to generate individual mask applied to B_{low}. In this final separation process, the observed amplitude spectrum given by $|X_1[k, l]|$ with $l \in B_{low}$ is compared with the initially separated spectra $\tilde{S}_1[k, l]$ and $\tilde{S}_2[k, l]$ with $l \in B_{mid}$ in terms of harmonic relationships. At first, with the help of local maximum frequencies of $|\tilde{S}_i[k, l]|$, harmonic structure in B_{mid} is estimated for each separation spectra. We denote the obtained local maximum frequencies of $|\tilde{S}_i[k, l]|$ are $b_{i1}(k)$, $b_{i2}(k)$, ・ ・ ・, and the number of local maxima in B_{mid} is $q_i(k)$.

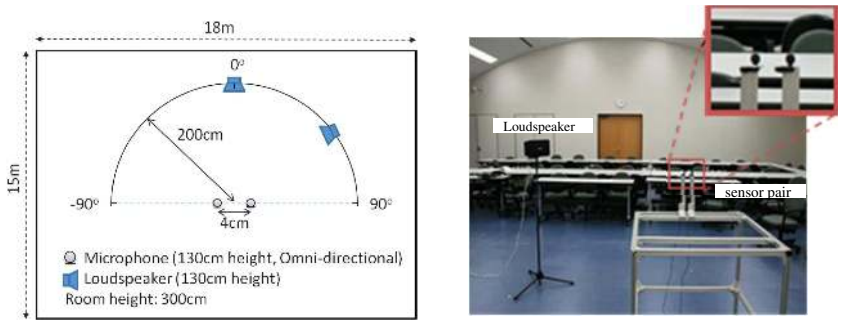
Harmonics estimation

The distance of adjacent harmonics $\Delta d_i(k)$ is defined as

$$\Delta d_i(k) = b_{i2}(k) - b_{i1}(k),\ q_i(k) > 2 \tag{35}$$

When $q_i(k) = 0$ or 1, we regard that there is no harmonic in the frame k. The estimated harmonics in low frequency band $g_{in}(k)$ is

$$g_{in}(k) = b_{l1}(k) - \Delta d_i(k) n, \tag{36}$$



(a) Experimental environments (b) Experimental setup

Figure 5. Experiments

where $n=1, 2, 3, \cdots$, $g_{in}(k) \in B_{low}$, and $g_{in}(k)$ means the harmonic structure of source i at frame k.

Massk generation

We assume that the bandwidth of each harmonics is the same, and use 5 adjacent cells as bandwidth in T–F domain. The mask in B_{low} is defined

$$\overline{M}_i[k,l] = \begin{cases} 1, if\ g_{in}(k) - 2 < l < g_{in}(k) + 2, and \\ \qquad\qquad q_i(k) \geq 2, l \in B_{low}, n = 1,2,3,\cdots \\ 0, otherwise. \end{cases} \tag{37}$$

The integrated mask combining Eq. (33) and Eq. (37) is represented by

$$M_i[k,l] = \tilde{M}_i[k,l] + \overline{M}_i[k,l]. \tag{38}$$

Finally, the separated signals are obtained as shown in Eq.(14).

3.3. Experiments

Experimental condition

Some real life experiments are performed in a conference room to evaluate the separation methods. Fig.5(a),(b) show the experimental environments and the setup. The experimental parameters are show in Tab.1. One source was placed at the broadside (θ=0°) and the location of the other source is varied from 0° to 80° at intervals of every 10°.

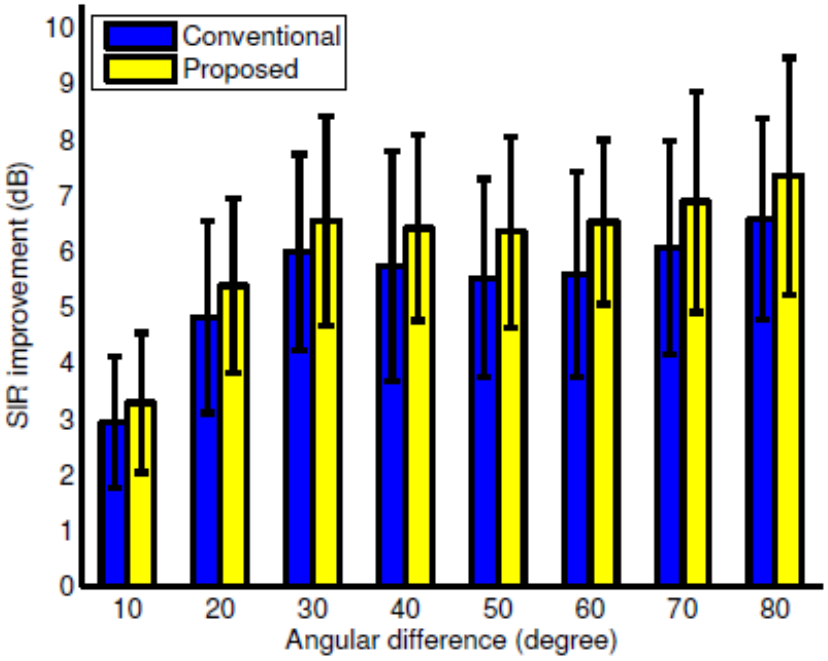


Figure 6. Experimental results (SIR)

Fig. 6 shows the average signal-to-interference ratio (SIR) improvement brought by the proposed and the conventional DUET method. The SIR improvement at the first sensor is defined as follows.

$$\text{SIR}_i \text{ improvement} = \text{Output SIR}_i - \text{Input SIR}_i \tag{39}$$

Where

$$\text{Input SIR}_i = 10\log_{10}\frac{\| s_i(t)\|}{\| s_j(t)\|}, \quad \text{Output SIR}_i = 10\log_{10}\frac{\| y_{ii}(t)\|}{\| y_{ij}(t)\|}$$

The proposed frame-wise PD-F approach exceeds the conventional method in terms of SIR improvement. The average improvement in our experiments is 6.22dB over the DUET. The most significant contribution in SIR improvement is made by the separation process in DSA frame which is 4.28dB. [31]

http://dx.doi.org/10.5772/56013

Source signal duration	5s speech signals
Sampling Frequency	8 kHz
Sound Velocity	340 m/s
Window	Hamming
STFT Frame Length	1024 sample
Frame Overlap	512 sample

Table 1. Experimental Parameters

4. DOA estimation

The DOA estimation method discussed in this section is based on the following three novel approaches.

1. Inspired by the ideas of TIme-Frequency Ratio Of Mixtures (TIFROM)-like assumptions, a novel reliability index is introduced. The selected cells with higher reliability are solely utilized for DOA estimation.
2. A statistical error propagation model relating PD-F and the consequent DOA is introduced. The model leads to a probability density function (PDF) of the DOA, and hence the DOA estimation problem is reduced to finding the most probable points of the PDF.
3. Source directions are determined using the kernel density estimator by utilizing the proposed bandwidth control strategy.

DOA information

Under the assumption of anechoic mixing with no-attenuation model and WDO in Eq. (5), the ratio between two observations $X_m[k, l]$ (m=1,2) is represented by

$$\frac{X_2[k,l]}{X_1[k,l]} = \frac{H_{2n}[l]}{H_{1n}[l]} = \exp\left[j\frac{2\pi f_s l}{L} \times \frac{d}{c}\sin\theta \right], \tag{40}$$

where θ is the direction of source which is dominant at [k,l]. The phase difference (PD) $\phi[k,l]$ between two observations $X_m[k, l]$ (m=1,2) defined by Eq. (17) is related to the angle θ as follows.

$$\phi[k,l] = \frac{2\pi f_s l d}{Lc}\sin\theta = \Delta\omega T l \sin\theta, \tag{41}$$

where $T = d/c$ is the maximum delay time between sensors, and $\Delta\omega = 2\pi f_S / L$ is the unit frequency width in L-point STFT. From Eqs. (16) and (41), the TDOA normalized by T, denoted by $\tau[k, l]$, can be represented as follows.

$$\tau\left[k,l\right]=\sin\theta=\frac{\phi\left[k,l\right]}{T\Delta\omega l} \tag{42}$$

4.1. Reliable T–F cell selection

As stated in 2.2, the following selection processes are applied only to the T-F cells in the support Ω_X of $X_1[k, l]$ as in 2.2. This eventually reduces the computation time by eliminating noise-like T-F components.

Since the PD estimation by (17) is subjected to unavoidable error, the success of DOA estimation is generally expected if reliable PD data are selected to use and outliers are eliminated. Likewise in [24], the following assumption is employed. When a source is dominant in a set of cells, all delays in it will take almost the same value; hence, the delay (42) and obviously the PD data (17) in this set are expected to be reliable. Conventionally, the confidence measure is obtained from the results of applying the principal component analysis to a set of steering vectors in individual horizontal and vertical T-F regions. Unlike this approach, the normalized delays $\tau[k, l]$ given by Eq.(42) are used to evaluate the attribute consistency of the T-F cells. According to the above assumption, two types of T-F regions around a cell $[k, l]$ are considered: a temporal neighborhood $\Gamma_t[k, l]$ and a frequency neighborhood $\Gamma_f[k, l]$,

$$\Gamma_t\left[k,l\right]:=\left\{\left[k+y,l\right]\middle|\,|y|\le Y\right\},\Gamma_f\left[k,l\right]:=\left\{\left[k,l+z\right]\middle|\,|z|\le Z\right\}, \tag{43}$$

where integers Y and Z determine the numbers of cells in these regions, as denoted by $|\Gamma_t[k, l]| := 2Y + 1$ and $|\Gamma_f[k, l]| := 2Z + 1$.

For each $\Gamma_t[k, l]$ and $\Gamma_f[k, l]$, the standard deviations of the normalized delays $\sigma_{\Gamma t}[k, l]$ and $\sigma_{\Gamma f}[k, l]$ are calculated by

$$\sigma_\Gamma\left[k,l\right]=\frac{1}{|\Gamma|}\sum_{\left[p,q\right]\in\Gamma}\left(\delta\left[p,q\right]-\mu_\Gamma\left[k,l\right]\right)^2 \tag{44}$$

$$\mu_\Gamma\left[k,l\right]=\frac{1}{|\Gamma|}\sum_{\left[p,q\right]\in\Gamma}\delta\left[p,q\right],\quad\Gamma=\Gamma_t,\Gamma_f. \tag{45}$$

Now, the reliability index $\eta[k, l]$ is calculated by

$$\eta\left[k,l\right]=\exp\left\{-\min\left(\sigma_{\Gamma_t}\left[k,l\right],\sigma_{\Gamma_f}\left[k,l\right]\right)\right\} \tag{46}$$

where $\eta[k, l]$ is a normalized index satisfying $0 < \eta \leq 1$. When at least $\sigma_{\Gamma t}[k, l]$ or $\sigma_{\Gamma f}[k, l]$ at $[k, l]$ is sufficiently small, $\eta[k, l]$ approaches unity, thereby the corresponding delay value $\delta[k, l]$ is considered to be reliable. We observed the tendency that the PD error decreases as the reliability index increases. Then, the cell group is selected with reliability index $\eta[k, l] > \eta_{th}$ for subsequent DOA estimation. In this paper, η_{th} is set to 0.96. The reason for using this value and related remarks are given in later.

For each selected reliable T-F cell, the direction θ is computed using Eq.(41). Here the set of computed directions is denoted as follows:

$$\left\{ \theta_i^{[l_i]} \middle| \, i = 1, 2, \ldots, I \right\}, \tag{47}$$

where i is the numbering integer of the selected cells, I is the total number of data, and l_i is the frequency bin at which the i-th cell is located.

DOA error distribution model

Consider a T-F cell at which the n-th source dominates and is located in the unknown direction θ_n. From Eq. (41), the theoretical PD at the cell is given by

$$\phi_n[l] = \Delta\omega T l \sin\theta_n = B_n l, \tag{48}$$

where $B_n = \Delta\omega T \sin\theta_n$. The frame index k is omitted because k is not essential in this section. In the l-th frequency bin, the observed $\phi_n[l]$ is distributed around its mean value $B_n l$,

$$\phi_n[l] = B_n I + \Delta\phi[l], \tag{49}$$

where $\Delta\phi[l]$ is a random variable representing the PD estimation error. Then, assume that the random variable $\Delta\phi[l]$ is an independent identical Gaussian distribution with zero mean and constant variance σ_ϕ^2, that is, $N(o, \sigma_\phi^2)$. The constant variance means that $\Delta\phi[l]$ is independent of the frequency bin l; this assumption is represented as follows:

$$\Delta\phi[l] \sim N(0, \sigma^2{}_\phi). \tag{50}$$

Fig. 7 (a) illustrates Gaussian error distribution at l-th frequency bin in PD-F plane in two-source case. The Gaussian distribution assumption is motivated from the simplicity of theoretical manipulation. From these error distribution model the problem is to estimate the probability distribution of the direction θ as shown in Fig.7(b).

Now, the following proposition can be proved.

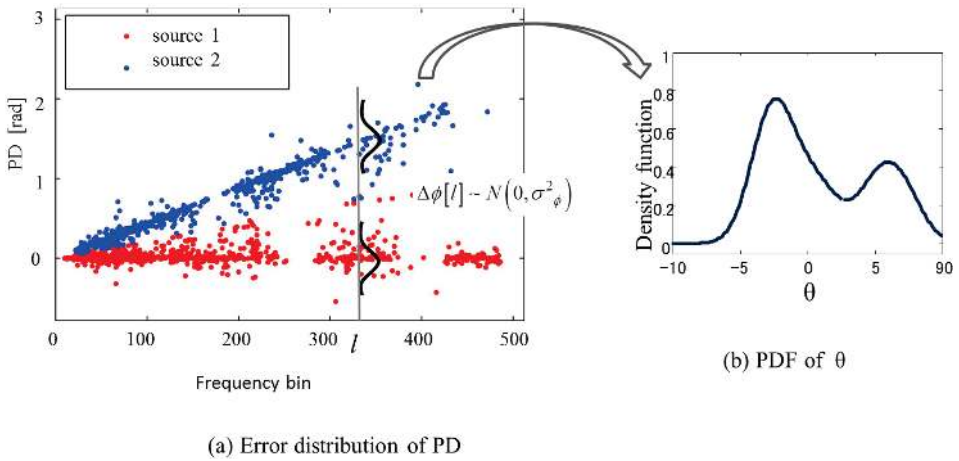


Figure 7. PD error distribution and Kernel density estimation

Proposition: If the random variable $\Delta\phi[l]$ is given by (50) and σ_{ϕ} is sufficiently small,

the PDF of $\theta_n^{[l]}$ is given by

$$\theta_n^{[l]} \sim N\left(\theta_n, \sigma^2_{\theta_n}[l]\right), \tag{51}$$

$$\sigma_{\theta n}[l] = \frac{1}{T\Delta\omega l\cos\theta_n}\sigma_{\phi}. \tag{52}$$

This proposition can be proved by the linearized incremental analysis between $\phi[l]$ and $\theta^{[l]}$. The DOA error distribution model is shown in Fig. 8.

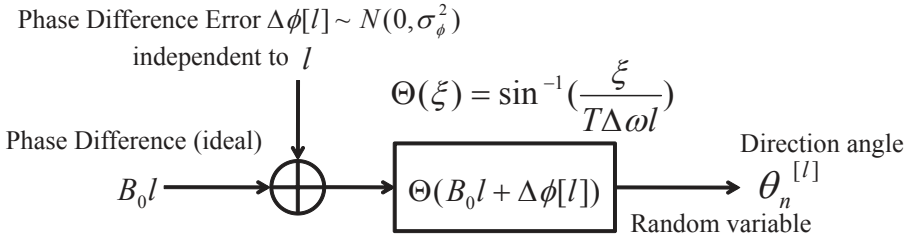


Figure 8. PD error and DOA estimation error distributions

4.2. DOA estimation using kernel density estimator

The kernel density estimation algorithm known as Parzen window in machine learning [32] is useful for statistical estimation even for a multiple-source problem. The algorithm provides an estimate PDF of $\theta[l]$ by using the observed samples (47). The maximum PDF point or the mode of the PDF can be considered as the optimal estimate of θ_n in the sense of the most probable value. The kernel density estimator approach yields an approximate estimation of the PDF of $\theta[l]$.

It is necessary to generalize the theoretical investigation noted above multisource and multi-frequency cases. The theoretical PDF formulation of θ in the case of multiple sources should be a Gaussian mixture with the same number of local modes (local peaks), each of which corresponds to an individual source. For the selected reliable data in Eq. (47), the kernel density estimator is applied to estimate the multi-modal PDF as follows:

$$\hat{p}(\theta) = \frac{1}{I}\sum_{i=1}^{I} \frac{1}{\varepsilon\left[l_i\right]} K\left(\frac{\theta - \theta_i^{\left[l_i\right]}}{\varepsilon\left[l_i\right]}\right), \tag{53}$$

where $K(\theta)$ is a kernel function, for which a Gaussian function is adopted in this study. $\varepsilon[l]$ is the bandwidth of the kernel. The idea behind applying the kernel density estimator is to reflect the theoretical result represented by the above proposition for the determination of the bandwidth. Since the variance of $\theta[l]$ depends on l and θ_n as indicated in Eq. (52), the bandwidth is determined as the form of

$$\varepsilon\left[I_i\right] = \frac{1}{T\Delta\omega l_i \cos\theta_i^{\left[I_i\right]}}\hbar. \tag{54}$$

where $\hbar$ is the control parameter and the observed $\theta[l_i]$ is substituted in place of a real unknown θ_n in Eq. (52). Accordingly, the dependence of the bandwidth on θ_n is indirectly controlled. The control parameter $\hbar$ is predetermined experimentally. Fig. 9 shows three examples of estimated PDFs for a two-source case with different $\hbar$. Finally, by finding the same number of local modes (peaks) as the number of pre-assigned source numbers, the source directions are estimated.

4.3. Experiments

Some experiments were conducted by the same setup and parameters as shown in Tab. 1. The first experiment is the case of two sources one of which is placed at the broad side (near 0 degree) as shown in Fig.10 (a). The results are shown in Fig.10 (b) and (c). While the proposed method gives a non-biased estimation, the estimates of the conventional method [27] tend to be biased for the cases of non-symmetric source positions with respect to the broadside. The second experiments for underdetermined case of three sources were performed. In this case

three sources were symmetrically located at the closer locations { -23, 4, 23 degrees} and far apart locations{ -42, 4, 42 degrees}. Fig.11 (a) and (b) show the results of the conventional method [27] and the proposed. In the “far apart” case both methods can estimate the source directions well. However, for the “closer” case, the proposed method provides less biased estimates than [27]. From the additional experimental results with diffuse noise presented in [33] and [34] it is proved the proposed cell selection method provides noise robust estimation better than the conventional.

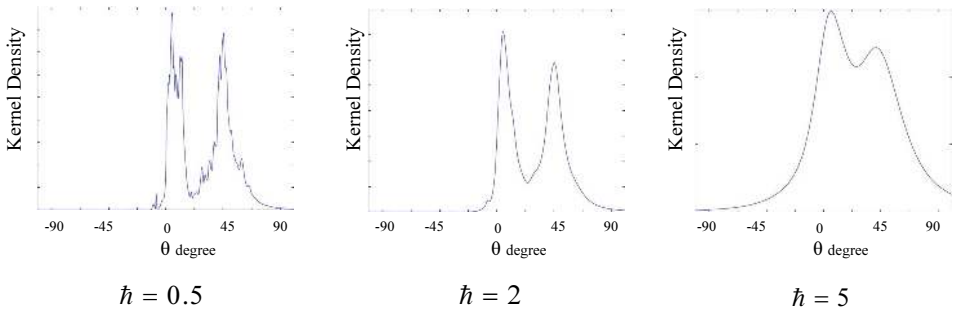


Figure 9. Estimated PDF and $\hbar$

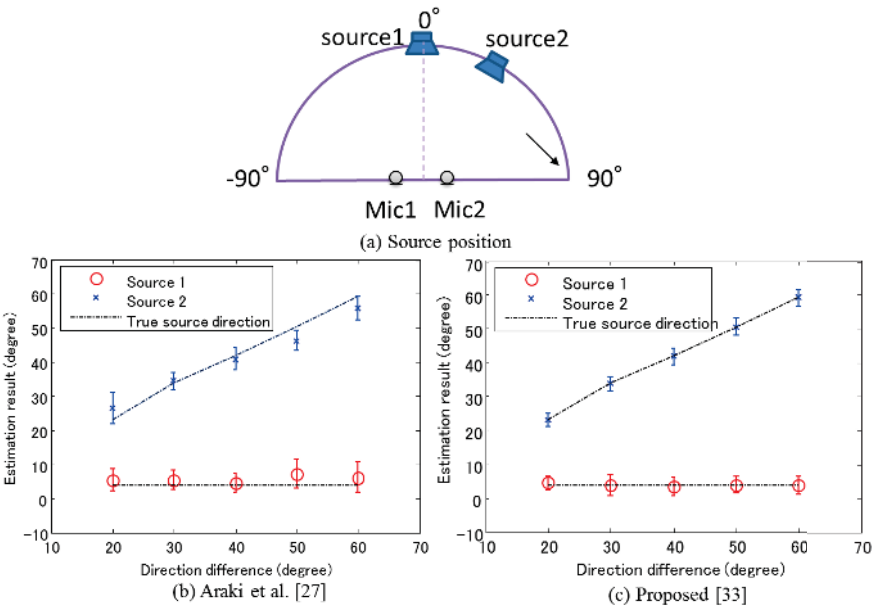


Figure 10. DOA estimation results for two sources

http://dx.doi.org/10.5772/56013

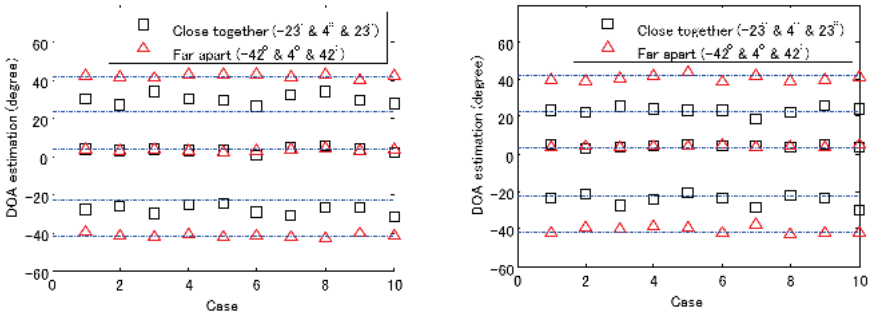


Figure 11. DOA estimation results for three sources

5. Conclusions

This monograph summarizes speech segregation and speaker's direction estimation methods which are based on sparseness of T-F components of speech signals. Throughout the discussion we are interested in underdetermined source-sensor conditions. At first recent progresses on BSS and DOA estimation algorithms associated with T-F sparse representation are reviewed. Then we focus on presenting an author's solution of BSS problems exploiting a series of phase difference versus frequency data. In the algorithm time frame classification concerning source active states is performed, and actual separation procedure is solely applied to the mixing frames.

The latter half of this chapter treats DOA estimation algorithm in a pair of microphones.

The basic error propagating mechanism is introduced and then the kernel density estimator is applied. The method provides a robust and non-biased DOA estimation and it develops theory for arbitrary microphone array configuration. [35]

One of recent human machine speech communication research on segregation and localization is associated with robot auditory system where the tracking of moving sources and sensors have to be considered.[36] For coping with these cases the particle filter and adaptive array processing have been attractive, and further efforts will be made.

Acknowledgements

The authors would like to appreciate Professor Wlodzimierz Kasprzak of Warsaw University of Technology for his valuable suggestions, and all members of speech signal processing group of Hamada Laboratory in Keio University for their great help.

Author details

Nozomu Hamada and Ning Ding

Keio University, System Design Engineering, Faculty of Science and Technology, Japan

References

[1] Divenyi P., editor. Speech Separation by Humans and Machines. Kluwer Academic Publishers; 2005.

[2] Benesty J., Chen J., Huang Y. Microphone Array Signal Processing. Springer; 2008.

[3] Makino S, Lee TW, Sawada H., editors. Blind Speech Separation. Springer; 2007.

[4] Hyvarinen A., Karhumen J., Oja E. Independent Component Analysis. John Wiley & Sons, Inc.; 2001.

[5] Saruwatari S., Takatani T., Shikano K. SIMO-Model-Based Blind Source Separation - Principle and its Applications. In: Makino S et al. (ed.) Blind Speech Separation. Springer; 2007. p149-168.

[6] Sawada H., Araki S., Makino S. Frequency-domain Blind Source Separation. In: Makino S et al. (ed.) Blind Speech Separation. Springer; 2007. p47-78.

[7] Choi S., Lyu Y., Berthommier F., Glotin H., Cichoki A. Blind separation of delayed and superimposed acoustic sources: learning algorithms an experimental study. Proc. IEEE Int. Conference on Speech Processing (ICsP), Seoul 1999.

[8] Choi S., Hong H., Glotin H., Berthommier F. Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network. Neurocomputing 2002; 49 (1) 299-314.

[9] Huang J., Ohnishi N., Sugie N. A biomimetic system for localization and separation of multiple sound sources. IEEE Trans. on Instrumentation and Measurement 1995; 44(3) 733-738.

[10] Aoki M, Okamoto M, Aoki A, Matsui H, Sakurai T, Kaneda Y. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. Acoust. Sci.& Tech 2001; 22(2) 149-157.

[11] Yilmaz O, Rickard S. Blind Separation of Speech Mixtures via Time- Frequency Masking. IEEE Trans. On signal processing 2004; 52(7) 1830-1847.

[12] Rickard S. The DUET Blind Source Separation Algorithm. In: Makino S et al. (ed.) Blind Speech Separation. Springer; 2007. p217-241.

[13] Araki S., Sawada H., Murai R., Makino S. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal Processing 2007; 87( ) 1833-1847.

[14] Sawada H., Araki S., Murai R., Makino S. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. IEEE Trans. on Audio, Speech, and Language Processing 2007 15(5) 1592-1604.

[15] Plumbley MD., Blumensath T., Daudet L., Gribonval R., Davies ME. Sparse representations in audio and music From coding to source separation. Proceedings of the IEEE 2010; 98(6) 995–1005.

[16] Nakatani T., Okuno H. Harmonic sound stream segregation using localization and its application to speech to speech stream segregation. Speech Communication 1999; 27 209-222.

[17] Parsons TW. Separation of speech from interfering speech by means of harmonic selection. Journal of the Acoustical Society of America 1976; 60(4) 911-918.

[18] Rickard S, Balan R., Rosca J. Real-time time frequency based blind source separation. ICA2001 2001; 651-656

[19] Izumi Y., Ono N., Sagayama S. Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics 2007; 147–150.

[20] Benesty J., Chen J., Huang Y. Direction of Arrival and Time-Difference-of-Arrival Estimation. chapter 9 in Microphone Array Signal Processing, Springer, 2008.

[21] Claudio EDD., Parisi R. Multi-Source Localization Strategies. In: (ed.) Microphone Arrays. Springer-Verlag; 2001. p181–201.

[22] Knapp CH., Carter GC. The generalized correlation method for estimation of time delays. IEEE Trans. on Acoust. Speech Signal Process. 1976; ASSP24 320–327. .

[23] Schmidt RO. Multiple emitter location and signal parameter estimation. IEEE Trans. on Antennas and Propagation. 1986; 34 276–280.

[24] Abrard F., Deville Y. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. Signal Processing 2005; 85 1389-1403.

[25] Arberet S., Gribonval R., Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," IEEE Trans. on Signal Processing, Vol. 58, No. 1, pp. 121-133, Jan. 2010.

[26] Berdugo B., Rosenhouse J., Azhari H. Speaker's direction finding using estimated time delays in the frequency domain. Signal Processing, 2002; 82 19–30.

[27] Araki S., Sawada H., Mukai R., Makino S. DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors. Journal of Signal Processing Systems, 2009; 63 265–275.

[28] Nesta F., Svaizer P., Omologo M. Cumulative state coherence transform for a robust two-channel multiple source localization. Proc. of ICA 2009; 290–297.

[29] Glotin H., Berthommier FB., Tessier E. A CASA-Labeling Model using the Localization Cue for Robust Cocktail-party Speech Recognition. Sixth European Conference on Speech Communication and Technology 1999; 22

[30] Tessier E., Berthommier F., Glotin H., Choi S. A CASA front-end using the localization cue for segregation and Then Cocktail-Party Speech Recognit6ion. Proc. IEEE Int. Conference on Speech Processing (ICsP) 1999; Seoul

[31] Ding N., Yoshida M., Ono J., Hamada N. Blind Source Separation Using Sequential Phase Difference versus Frequency Distortion. Journal of Signal Processing 2011; 15(5) 375-385.

[32] Duda R., Hart PE., Stork DG. Pattern Classification. John Wiley & Sons 2001.

[33] DING N., Hamada N. DOA Estimation of Multiple Speech Source from a Stereophonic Mixture in Underdetermined Case", IEICE Trans. Fundamentals, Vol.E95-A, No.4, Apr. 2012

[34] Ding N. Blind Source Separation and Direction Estimation for StereophonicMixtures of Multiple Speech Signals Based on Time-Frequency Sparseness. PhD thesis. Keio University Yokohama; 2012

[35] Fujimoto K., Ding N., Hamada N. Multiple Sources' Direction Finding by using Reliable Component on Phase Difference Manifold and Kernel Density Estimator. IEEE Proc. ICASSP 2012; Kyoto

[36] Valin JM., Michaud F., Rouat J. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robotics and Autonomous System 2007; 55 216—228.

[37] Daobilige Su, Masashi Sekikawa, and Nozomu Hamada, Novel scheme of real-time direction finding and tracking of multiple speakers by robot-embedded microphone array, 1st Int. Con. on Robot Intelligence Tech. RiTA, 2012 Korea