SPEAKER LOCALIZATION AND SPEECH SEPARATION USING PHASE DIFFERENCE VERSUS FREQUENCY DISTRIBUTION

Ning Ding and Nozomu Hamada

Signal Processing Lab., School of Integrated Design Engineering, Keio University
Address: 3-14-1 Hiyoshi, Yokohama 223-8522, Japan
Phone:+(81)45-566-1738, FAX:+(81)45-566-1720
Email: ding@hamada.sd.keio.ac.jp, hamada@sd.keio.ac.jp

ABSTRACT

This paper proposes a novel sparse source separation method using a pair of microphones. The method is based on time-frequency (T-F) decomposition, applies the weighted Hough transform to the Phase Difference (PD) versus Frequency (PD-F) distribution of received mixture signals, and estimates source directions. Then, the estimated source directions and harmonic structure are used to separate the mixture signals. The effectiveness of the proposed method is shown through experiments in real acoustic circumstances.

1. INTRODUCTION

Blind source separation (BSS) aims to estimate source signals by using only mixed signals without any priori information about the source position, mixing process, or circumstances. Various approaches have been proposed to resolve BSS problems in speech signals; the most popular approaches are independent component analysis (ICA)[1] and time-frequency (T-F) masking method[2][3]. ICA relies on statistical independence of speech sources; therefore it is difficult to rely on ICA to solve an underdetermined case in which the number of sources $N$ is greater than the number of sensors $M$.

The T-F masking method is applied to the signals transformed from time domain to T-F domain by Short Time Fourier Transform (STFT), and is applicable to the underdetermined BSS cases. Usually, T-F masking methods are based on an assumption known as the W-disjoint orthogonality (WDO) of speech signals in which a sparse representation of speech in the T-F domain is expected. Although the observed signal is a mixture of several sources, its T-F distribution of received mixture signals, and estimates source directions. Then, the estimated source directions and harmonic structure are used to separate the mixture signals. The effectiveness of the proposed method is shown through experiments in real acoustic circumstances.

1) A modified Hough transform in the PD-F distribution is introduced to estimate DOA.
2) PD-F distribution and harmonic structure are combined for source separation.

In Section 2 of this paper, we briefly review the BSS problem. Section 3 discusses our proposed method in detail, and Section 4 includes results of experiments performed to verify our method. Section 5 presents our conclusion.

2. BSS PROBLEM DESCRIPTION

2.1. Observation model

Assume that sources $S_1, \ldots, S_N$ are convolutively mixed and observed by $M$ sensors in discrete time domain,

$$x_j(\tau) = \sum_{i=1}^{N} \sum_{l} h_{ji}(l)s_i(\tau - l), \quad j = 1, \ldots, M, \quad (1)$$

where $h_{ji}(l)$ represents the impulse response from source $i$ to sensor $j$, $N$ is the number of sources, and $M$ is the number of sensors.

Discrete time domain signals $x_j(\tau)$ sampled at frequency $f_s$ are converted into T-F domain signals $X_j[k, l]$ using an $L$-point STFT:

$$X_j[k, l] = \sum_{r=-L/2}^{L/2-1} x_j(r + kS)\text{win}(r)e^{-j2\pi lr}, \quad (2)$$

Several new T-F masking and source direction estimation methods proposed recently, such as TIFROM[4] and DEMIX[5], modify the DUET to increase feature reliability given by T-F cells. These methods observe the stability of mixing parameters in a local neighborhood and focus on creating efficient clustering in a two-dimensional space. These modifications are intended to overcome drawbacks in the previous attenuation ratio- and delay-based clustering algorithm. Amplitude difference between the sensors is negligible for small distance sensor configuration. In addition, an error in phase-difference estimation, particularly in the low-frequency band, would cause a large delay error.

This paper proposes a novel T-F masking method based on PD-F distribution to obtain Direction Of Arrival (DOA) information and separate mixtures in speech signals. The novelty of this paper compared with previous studies can be summarized as follows.

© EURASIP, 2011 - ISSN 2076-1465
where win(r) is the window, S is the window shift size, k(0 \sim K) is the integer index of time frame, and l(0 \sim \frac{L}{f_s}) is the integer index of frequency bin.

The T-F masking approach utilizes instantaneous mixtures at each time frame k and frequency bin l:

\[ X_j[k, l] \approx \sum_{i=1}^{N} H_{ji}(l) S_i[k, l], \]

where \( H_{ji}(l) \) is the frequency response of \( h_{ji}(l) \), and \( S_i[k, l] \) is the i-th T-F domain source signal. This paper focuses on experiments performed for \( N = 2 \) and \( M = 2 \).

2.2. Conventional clustering feature and the limit

The conventional histogram mapping method obtains source direction information by estimating the delay between two observations. The delay \( \delta \) without spatial aliasing is calculated by:

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

where \( \phi[k, l] \) is the PD between \( X_1[k, l] \) and \( X_2[k, l] \):

\[ \phi[k, l] = \angle X_1[k, l] - \angle X_2[k, l]. \]

Although the conventional delay histogram-based clustering method is a good approach for speech separation, it fails where

\[ \{\text{ii}\} \quad B_{\text{low}} := \{l | l < l_1\}, \]

\[ \{\text{ii}\} \quad B_{\text{high}} := \{l | l > l_1\}, \]

\[ \{\text{ii}\} \quad B_{\text{seg}} := \{l | l_1 < l < l_2\} \subset B_{\text{high}}, \]

where \( l_1 = [(f_1 \cdot L/f_s)\], l_2 = [(f_2 \cdot L/f_s)\], f_1 = 400Hz, f_2 = 1kHz, and \( \lfloor \cdot \rfloor \) is the Gauss floor function which maps a real number to the largest previous integer.

3. PROPOSED METHOD

The flow of the proposed method is shown in Fig. 1.

An example of PD-F data set defined by vectors \{l, \phi[k, l]\} in a two-dimensional plane is shown in Fig. 2(a).

In our proposed method, the following three frequency bands are used:

\[ \{\text{i}\} \quad B_{\text{low}} := \{l | l < l_1\}, \]

\[ \{\text{ii}\} \quad B_{\text{high}} := \{l | l > l_1\}, \]

\[ \{\text{iii}\} \quad B_{\text{seg}} := \{l | l_1 < l < l_2\} \subset B_{\text{high}}, \]

where \( l_1 = [(f_1 \cdot L/f_s)\], l_2 = [(f_2 \cdot L/f_s)\], f_1 = 400Hz, f_2 = 1kHz, and \( \lfloor \cdot \rfloor \) is the Gauss floor function which maps a real number to the largest previous integer.

\[ \angle X_1[k, l] = \angle X_2[k, l]. \]

Although the conventional delay histogram-based clustering method is a good approach for speech separation, it fails to estimate DOA and to separate T-F cells properly due to the estimation error in \( \delta[k, l] \).

3.1. DOA estimation by weighted Hough transform with bandwidth

Because DOA estimation corresponds to finding a linear phase relationship in PD-F distribution, we apply the Hough transform as a line extraction technique.

3.1.1. Cell selection and normalization

We select a set of T-F cells, which satisfied the following two conditions:

1. Because the PD in the low-frequency band \( l \in B_{\text{low}} \) is too small for accurate estimation, we restrict \( l \in B_{\text{high}} \).

2. Define the maximum amplitude value \( A(l) \) at each frequency bin by \( A(l) = \max_{k \in [0, K]} |X_1[k, l]| \), and then select the T-F cells \( [k, l] \) satisfying

\[ \gamma[k, l] = \frac{|X_1[k, l]|}{A(l)} \geq Th_1, \]

where \( \gamma[k, l] \) is used as the weight factor in Hough transform, and \( Th_1 = 0.5 \) is set by experiments. We denote all selected cells \( [k, l] \) as \( \Omega_1 \). An example of PD-F distribution of \( \Omega_1 \) is shown in Fig. 2(b).

For the analysis, all vectors in \( \Omega_1 \) are normalized by

\[ [y(l), z_k(l)]^T := [l/(L/2), \phi[k, l]/\pi]^T. \]

3.1.2. Angle range

The gradient of a line from the origin in the normalized PD-F plane, denoted by \( \alpha \), corresponds to the actual DOA \( \theta \) (degree) using the equation \( \theta = arcsin\left(\frac{f_2}{2\pi f_s} \cdot \tan\alpha\right) \). In addition, the theoretical limitation of \( \alpha \) is given by \( |\alpha| \leq \arctan\left(\frac{f_2}{2\pi f_s}\right) \), where \( f_2 \) is the distance between sensors and \( c \) is the sound velocity. From this inequality, \( \alpha \) is restricted within the interval \( |\alpha| < \alpha_{\text{limit}} \).

3.1.3. Hough transform calculation

By transforming the two-dimensional grid index \( [k, l] \in \Omega_1 \) into an arbitrary one-dimensional alignment integer index \( n \), we obtain the corresponding formula

\[ [y(l), z_k(l)]^T \rightarrow [y_n, z_n]^T, \quad \gamma[k, l] \rightarrow \gamma_n, [k, l] \in \Omega_1 \]

The Hough transform is calculated by

![Fig. 1. Flow of proposed method](image-url)
\[ \rho_n(\alpha) = y_n \cdot \cos \alpha + z_n \cdot \sin \alpha, \quad |\alpha| < \alpha_{\text{limit}} \] (9)

where \( \rho \) is the shortest distance from the origin to the line.

### 3.1.4. DOA estimation with bandwidth

In theory, the DOA corresponds to the gradient angle \( \alpha \) for \( \rho = 0 \). However, to consider phase estimation error at each frequency, the interval of \( |\rho_n(\alpha)| \leq \epsilon(\alpha) \), where \( \epsilon(\alpha) = \epsilon_0 \cos \alpha \), at each \( \alpha \) is combined into a unit rectangular cell for the Hough voting procedure. Through experimentation, we set \( \epsilon_0 = 0.03 \). The Intersection Value (denoted by \( IV \)) at \( \alpha \) with weight \( \gamma_n \) is calculated as

\[ IV(\alpha) = \sum_{\Omega_1} \gamma_n, \quad \text{if} \quad |\rho_n(\alpha)| \leq \epsilon(\alpha). \] (10)

Our approach is different from [6] in this regard.

In practice, \( IV(\alpha) \) is evaluated at sampled \( \alpha \) values, such as integer values within the interval \([-\alpha_{\text{limit}}, \alpha_{\text{limit}}]\]. The DOA estimation is performed as follows: The \( \alpha_1 \) which maximizes \( IV(\alpha) \) gives the first source direction \( \theta_1 \) using the relationship \( \alpha \) and \( \theta \). Next, DOA \( \theta_2 \) is obtained using \( \alpha_2 \) at which \( IV(\alpha_2) \) is the local maximum taking sub-maximum value, and \( \theta_2 \) is more than 10 degrees apart from the estimated DOA \( \theta_1 \).

### 3.2. Source separation

Because PD-F data for two sources are closely mixed in the \( B_{\text{low}} \) region, it is difficult to cluster PD-F dots into each respective source. Therefore, we cluster the PD-F distribution in the band \( B_{\text{high}} \) into two groups as an initial separation, and estimate the fundamental frequency in \( B_{\text{seg}} \). The binary separation mask in \( B_{\text{low}} \) is then generated using harmonic relationship. The following sections describe our procedure in detail.

#### 3.2.1. Initial separation in \( B_{\text{high}} \)

The binary mask \( M_i[k,l] \) (i = 1, 2) in \( B_{\text{high}} \) is defined as

\[ M_i[k,l] = \begin{cases} 1, & \text{if } i = \arg\min_{e = \{1,2\}} |\phi[k,l] - \alpha \cdot l|, \quad l \in B_{\text{high}} \\ 0, & \text{otherwise}, \quad (11) \end{cases} \]

The initially separated signals \( \hat{S}_i[k,l] \) are obtained by

\[ \hat{S}_i[k,l] = M_i[k,l] \cdot X_i[k,l]. \] (12)

#### 3.2.2. Local maximum in \( B_{\text{seg}} \)

To generate individual mask in \( B_{\text{low}} \), the observed amplitude spectrum \( |X_l[k,l]| \) in \( l \in B_{\text{low}} \) is compared with the initially separated spectra \( \hat{S}_1[k,l] \) and \( \hat{S}_2[k,l] \) in \( l \in B_{\text{seg}} \).

With the help of local maximum frequencies of \( \hat{S}_i[k,l] \), harmonic structure in \( B_{\text{seg}} \) is estimated. We select the local maximum frequencies of \( \hat{S}_i[k,l] \) satisfying the following two conditions:

1. Sufficient amplitude, where \( |\hat{S}_i[k,l]| / \max_v |\hat{S}_i[k,v]| > T_{h2} \).
2. In later experiments, \( T_{h2} = 0.2 \) is adopted.

### 4. EXPERIMENTS

#### 4.1. Experimental condition

Some experiments are performed in a conference room to evaluate our methods. The experimental setup is shown in Fig.3, and the experimental parameters are shown in Tab.1. One source is placed at the broadside (0°) and the location of the other source is varied from 0° to 90° at intervals of every 10°.
4.2. Experimental results

We compared the experimental results with the conventional method[2]. The separation results are based on the DOA estimation to some extent, and some results are shown in Tab.2.

From the table we can see that in some cases, the proposed method can estimate sources’ directions but the conventional method can not. In other cases, both the proposed method and the conventional method can estimate source’s directions, but the proposed method can estimate DOA more accurately than conventional method. It is evident that our proposed method can estimate the DOA with more reliability and accuracy than the conventional method.

The separation performance is evaluated using $WDO_M$ (measure of W-Disjoint Orthogonality) as in [2].

$$WDO_M = \frac{||\{M[k,l]|S_D[k,l]\}||^2 - ||\{M[k,l]|S_I[k,l]\}||^2}{||S_D[k,l]||^2}, \quad (19)$$

where $S_D[k,l]$ is the desired signal, $M[k,l]$ is the binary mask, and $S_I[k,l]$ is the interfering signal.

Fig.4 shows the separation results. The value of $WDO_M$ of our proposed method exceeds that of the conventional method about 0.06 to 0.13.

In general, the separation performance will raise with the increasing of angular difference. When two sources are closely located, the PD in low frequency band are mixed, and the conventional method can not separate the cells properly. But by the harmonic structure estimation of proposed method, we can overcome this drawback. That is why the improvement of the proposed method is relatively large in small angular difference.

5. CONCLUSION

The method combining the DOA estimation via Hough transform and the T-F masking utilizing harmonic structure is proposed. Comparing to the DUET algorithm, it improves the average WDO nearly 0.1 for two-source two-sensor case. The extension to arbitrary sensor configuration is one of the future issues.

6. REFERENCES