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 (spectrogram) cell contains at most one source signal component. DUET[2] and SAFIA[3] are based on WDO assumption and explore the differences in the directions of the speakers. These algorithms perform feature clustering or histogram analysis in the parameter plain of attenuation rate and the time delay between sensors' observations to characterize the sources. Reconstruction of source signals can then be performed by masking the spectrogram of a mixture. The DUET for two microphones uses a weighted, two-dimensional histogram to express the differences between the T-F representations of two mixtures in terms of amplitude and phase. The resulting histogram peaks are assumed to represent the respective sources.

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.

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 , $\cdots$,s_N are convolutively mixed and observed by M sensors in discrete time domain,

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

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:

$$X_j[k,l] = \sum_{r=-L/2}^{L/2-1} x_j(r+kS) win(r) e^{-i2\pi lr}, \qquad (2)$$

© 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 $l(0 \sim \frac{L}{2})$ is the integer index of frequency bin.

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

$$X_j[k,l] \approx \sum_{i=1}^{N} H_{ji}(l) S_i[k,l], \tag{3}$$

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 δ without spatial aliasing is calculated by:

$$\delta[k,l] = \frac{L}{2\pi f_s l} \phi[k,l], \tag{4}$$

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

$$\phi[k,l] = \angle X_1[k,l] - \angle X_2[k,l]. \tag{5}$$

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. PROPOSED METHOD

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

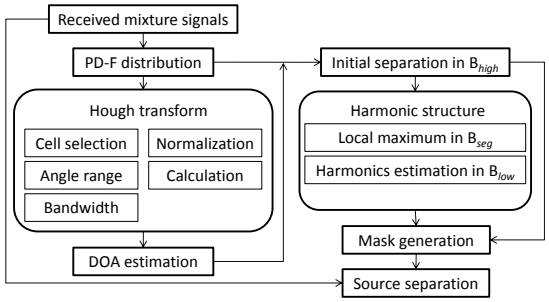
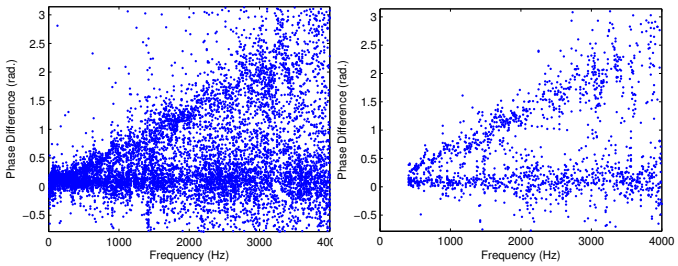


Fig. 1. Flow of proposed method

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:
(i) $B_{low} := \{l | l < l_1\}$,
(ii) $B_{high} := \{l | l > l_1\}$,
(iii) $B_{seg} := \{l | l_1 < l < l_2\} \subset B_{high}$,
where $l_1 = \lfloor (f_1 \cdot L/f_s) \rfloor$, $l_2 = \lfloor (f_2 \cdot L/f_s) \rfloor$, $f_1 = 400Hz$, $f_2 = 1kHz$, and $\lfloor \rfloor$ is the Gauss floor function which maps a real number to the largest previous integer.



(a) Whole T-F cells (b) Selected T-F cells

Fig. 2. PD-F distribution. $f_s = 8kHz$.

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 ($l \in B_{low}$) is too small for accurate estimation, we restrict $l \in B_{high}$.
(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

$$\gamma[k,l] = \frac{|X_1[k,l]|}{A(l)} \geq Th_1, \tag{6}$$

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 Ω_1. An example of PD-F distribution of Ω_1 is shown in Fig. 2(b).

For the analysis, all vectors in Ω_1 are normalized by

$$[y(l), z_k(l)]^T := [l/(L/2), \phi[k,l]/\pi]^T. \tag{7}$$

3.1.2. Angle range

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

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

$$[y(l), z_k(l)]^T \rightarrow [y_n, z_n]^T, \gamma[k,l] \rightarrow \gamma_n, [k,l] \in \Omega_1 \tag{8}$$

The Hough transform is calculated by

$$\rho_n(\alpha) = y_n \cdot \cos\alpha + z_n \cdot \sin\alpha, \quad |\alpha| < \alpha_{limit} \tag{9}$$

where ρ 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 α for $\rho = 0$. However, to consider phase estimation error at each frequency, the interval of $|\rho_n(\alpha)| \leq \varepsilon(\alpha)$, where $\varepsilon(\alpha) = \varepsilon_0 \cos\alpha$, at each α is combined into a unit rectangular cell for the Hough voting procedure. Through experimentation, we set $\varepsilon_0 = 0.03$. The Intersection Value (denoted by IV) at α with weight γ_n is calculated as

$$IV(\alpha) = \sum_{\Omega_1} \gamma_n, \quad if\ |\rho_n(\alpha)| \leq \varepsilon(\alpha). \tag{10}$$

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

In practice, $IV(\alpha)$ is evaluated at sampled α values, such as integer values within the interval $[-\alpha_{limit}, \alpha_{limit}]$. The DOA estimation is performed as follows: The α_1 which maximizes $IV(\alpha)$ gives the first source direction θ_1 using the relationship α and θ. Next, DOA θ_2 is obtained using α_2 at which $IV(\alpha_2)$ is the local maximum taking sub-maximum value, and θ_2 is more than 10 degrees apart from the estimated DOA θ_1.

3.2. Source separation

Because PD-F data for two sources are closely mixed in the B_{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_{high} into two groups as an initial separation, and estimate the fundamental frequency in B_{seg}. The binary separation mask in B_{low} is then generated using harmonic relationship. The following sections describe our procedure in detail.

3.2.1. Initial separation in B_{high}

The binary mask $\tilde{M}_i[k,l]$ $(i = 1,2)$ in B_{high} is defined as

$$\tilde{M}_i[k,l] = \begin{cases} 1,\ if\ \ i = \underset{c=(1,2)}{\arg\min} |\phi[k,l] - \alpha_c \cdot l|,\ l \in B_{high} \\ 0,\ otherwise. \end{cases} \tag{11}$$

The initially separated signals $\tilde{S}_i[k,l]$ are obtained by

$$\tilde{S}_i[k,l] = \tilde{M}_i[k,l] \cdot X_1[k,l]. \tag{12}$$

3.2.2. Local maximum in B_{seg}

To generate individual mask in B_{low}, the observed amplitude spectrum $|X_1[k,l]|$ in $l \in B_{low}$ is compared with the initially separated spectra $\tilde{S}_1[k,l]$ and $\tilde{S}_2[k,l]$ in $l \in B_{seg}$.

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

(1) Sufficient amplitude, where $|\tilde{S}_i[k,l]| / \max_v |\tilde{S}_i[k,v]| > Th_2$. In later experiments, $Th_2 = 0.2$ is adopted.

(2) Its amplitude takes the maximum among adjacent several frequency bins. Because the fundamental frequency of human voice is greater than $80Hz$, $l > \lfloor L/f_s \cdot 80 \rfloor = 10$, which means it is only possible to have one harmonic frequency within at least 10 adjacent bins.

Under these conditions, the frequency bins of local maxima are obtained. We denote the obtained local maximum frequencies of $|\tilde{S}_i[k,l]|$ are $b_{i1}(k), b_{i2}(k), \cdots$, and the number of local maxima in B_{seg} is $q_i(k)$.

3.2.3. Harmonics estimation in B_{low}

We define the frequency difference between adjacent local maxima $\Delta d_i(k)$ as

$$\Delta d_i(k) = b_{i2}(k) - b_{i1}(k),\ q_i(k) \geq 2 \tag{13}$$

When $q_i(k) = 0$ or 1, we regard that there is no harmonic characteristics in the source $\tilde{S}_i[k,l]$ at the frame k. The estimated harmonics $g_{in}(k)$ in B_{low} is

$$g_{in}(k) = b_{i1}(k) - \Delta d_i(k) \cdot n, \tag{14}$$

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

There is a special situation in which both $q_1(k)$ and $q_2(k) = 0$ or 1. In this case, the harmonics at the latest frame is used as follows:

$$g_{in}(k) = g_{in}(k - v), \tag{15}$$

for the smallest $v > 0$ with $q_i(k - v) \geq 2$.

3.2.4. Mask generation and separation

We assume that the spectral bandwidths at harmonics in B_{low} are the same, amount of 5 adjacent cells (i.e. $40Hz$). Thus, the mask in B_{low} is defined

$$\bar{M}_i[k,l] = \begin{cases} 1,\ if\ \ g_{in}(k) - 2 < l < g_{in}(k) + 2\ \ and \\ \qquad q_i(k) \geq 2,\ l \in B_{low},\ n = 1,2,3,\cdots \\ 0,\ otherwise. \end{cases} \tag{16}$$

The final mask is represented by

$$M_i[k,l] = \tilde{M}_i[k,l] + \bar{M}_i[k,l]. \tag{17}$$

The separated signals are got by applying inverse STFT to

$$\hat{S}_i[k,l] = M_i[k,l] X_1[k,l]. \tag{18}$$

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°.

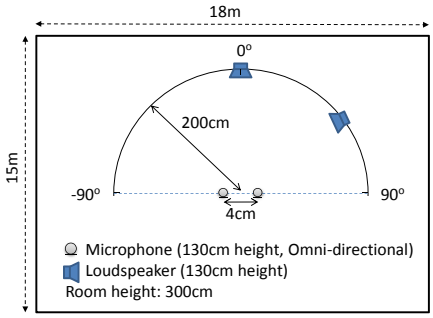


Fig. 3. Experimental setup

Table 1. Experiment parameters

Source Signal Duration	5 s
Sampling Frequency	8 kHz
Sound Velocity	340 m/s
Window	Hamming
STFT Frame Length	1024
Frame Overlap	512

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 sources's directions, but the proposed method can estimate DOA more accurate 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].

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

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 $0.06 \sim 0.13$.

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.

Table 2. DOA estimation results

Source direction	0^0 & 10^0	0^0 & 60^0
Conventional method	3.7^0 & *Fail*	3.7^0 & 64.3^0
Proposed method	3.1^0 & 14.1^0	2.1^0 & 63.0^0

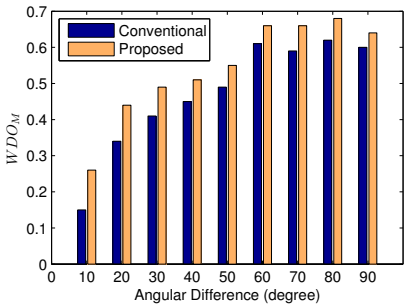


Fig. 4. Experimental results. One source was held fixed at 0°, while the other source was changed from 0° to 90° with a 10° increment. The horizontal axis also indicates the true direction of the second source.

5. CONCLUSION

The method combining the DOA estimation via Hough tranform 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

[1] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. on Speech Audio Process*, Vol. 8, No. 3, pp.320-327, 2000.

[2] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. On signal processing*, Vol. 52, No. 7, pp.1830-1847, 2004.

[3] M. Aoki, M. Okamoto S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech*, Vol. 22, No. 2, pp.149-157, 2001.

[4] Frederic Abrard, and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, Vol. 85, pp.1389-1403, 2005.

[5] S. Arberet, R. Gribonval, and F. 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.

[6] S. Marchand and A. Vialard, "The Hough transform for binaural source localization," *Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-08)*, Como, Italy, Sepetember1-4, 2009.