Invited paper

Relaxing the WDO Assumption in Blind Extraction of Speakers from Speech Mixtures

Włodzimierz Kasprzak[a], Ning Ding[b], and Nozomu Hamada[b]

[a] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[b] *Signal Processing Lab., School of Integrated Design Engineering, Keio University, Kanagawa, Japan*

Abstract—The time-frequency masking approach in blind speech extraction consists of two main steps: feature clustering in a space spanned over delay-time and attenuation rate, and spectrogram masking in order to reconstruct the sources. Usually a binary mask is generated under the strong W-disjoint orthogonal (WDO) assumption (disjoint orthogonal representations in the frequency domain). In practice, this assumption is most often violated leading to weak quality of reconstructed sources. In this paper we propose the WDO to be relaxed by allowing some frequency bins to be shared by both sources. As we detect instantaneous fundamental frequencies the mask creation is supported by exploring a harmonic structure of speech. The proposed method is proved to be effective and reliable in experiments with both simulated and real acquired mixtures.

Keywords—blind source extraction, harmonic frequencies, histogram clustering, spectrogram analysis, speech reconstruction, time-frequency masking, W-disjoint orthogonal.

1. Introduction

Blind source separation (BSS) is an approach for estimating source signals by using only mixed signals observed at many input channels [1], [2], [3]. The source reconstruction is performed blindly, without possessing information on each source, such as its location and active time, and having no knowledge about the mixing matrix. Many methods have been proposed for BSS problems, among them the most popular approaches are: independent component analysis (ICA) [2], multichannel blind deconvolution (MBD) [3] and time-frequency masking (TFM) [4]. ICA and MBD rely on statistical independence of the speech sources and that sources are mixed instantaneously or by FIR filters. However, it is difficult for ICA or MBD to solve the underdetermined case in which the source number is greater than the microphone number.

Time-frequency (T-F) masking methods are based on the assumption called *W-disjoint orthogonal* (WDO) (e.g., the DUET method [4], [5]). This assumes a sparse representation of speech in the frequency domain: although the observed signal is a mixture of several sources, most part of its time-frequency (spectrogram) cells contain one of the source signals' component only. Some other assumptions are also proposed in the method known as SAFIA [6], that performs sound source segregation based on estimating the incident angle of each frequency component. The T-F masking methods firstly make histogram (or cluster) analysis in the attenuation- and time delay-space, in order to detect the number of speakers and their characteristics, and secondly they perform source reconstruction via spectrogram masking. These methods work well for anechoic mixtures and significantly different orientations of speakers w.r.t. the microphone set.

Several novel algorithms have been developed recently, such as time-frequency ratio of mixtures (TIFROM) [7], DEMIX [8] and uniform clustering [9], that try to overcome some weak points of basic T-F masking algorithms. These improvements focus on making more efficient clustering in the 2-D attenuation rate- and delay-time space. Other recent research topic line is to provide proper microphone arrangements for T-F masking, e.g., an array of microphones or a triangle of microphones [10], [11]. Some background knowledge about speech signals can also help. In the HS method [12] it is proposed to use harmonic structure as the clustering feature.

Today the T-F masking methods work quite well for anechoic mixtures. However still there are some drawbacks of them. One of them is the error of phase-difference estimation, which is especially large in the low frequency band. It seems that in this range of frequencies, say up to 1 kHz, the assumptions of WDO is not satisfied. But one can not simply filter out these frequency components or skip it during source reconstruction, as they carry crucial information about signal's energy. The problem gets even more complicated if echoic mixtures are processed.

This paper proposes several improvements to both steps in T-F masking, the feature clustering and source reconstruction, that are relaxing the strict WDO assumption.

The paper is organized as follows. In Section 2, the BSS problem is briefly introduced and a basic T-F masking approach is defined. In Section 3 an analysis of the time-delay feature across the whole frequency spectrum is performed. The observation of large errors of time-delay estimation in the low frequency band leads to the first improvement - the use of a restrictive mask based on local and global energy distribution analysis (Section 4). The second problem is to improve the spectrogram masks for source reconstruction. A novel method for multi-valued mask generation is proposed (Section 5). Experimental results verify that crucial improvements in both histogram analysis and source reconstruction has been achieved (Section 6).

2. T-F Masking

In the following, we introduce the BSS problem and main processing stages of a T-F masking approach:

- extraction of spectrogram features and their clustering,
- spectrogram mask generation and source reconstruction.

In experiments we focus particularly on a situation where the number of sources $N = 2$, and the number of sensors $M = 2$.

2.1. The BSS Problem

In discrete time domain, suppose that sources $s_1, \ldots, s_N$ are convolved and mixed. This is observed at M sensors

$$x_j(\tau) = \sum_{k=1}^{N} \sum_{l} h_{jk}(l) s_k(\tau - l), \quad j = 1, \ldots, M, \tag{1}$$

where: $h_{jk}(l)$ represents the impulse response from source k at sensor j, N is the number of sources, and M is the number of sensors.
The time domain signals $x_j(\tau)$ sampled at frequency f_s are converted to frequency domain into a time-series of vector signals $X_j(t,f)$ by applying a L point STFT to consecutive signal frames:

$$X_j(t,f) = \sum_{r=-L/2}^{L/2-1} x_j(r+tS) win(r) e^{-i2\pi fr}, \tag{2}$$

where: $win(r)$ is a window function, S is the window shift size, t is the integer time frame index, and f is the integer $(0 \sim \frac{L}{2})$ frequency bin.
The time-frequency approach to blind speech separation utilizes instantaneous mixtures at each time frame t and frequency bin f:

$$X_j(t,f) \approx \sum_{k=1}^{N} H_{jk}(f) S_k(t,f), \tag{3}$$

where: $H_{jk}(f)$ is the frequency response of the mixing system, and $S_k(t,f)$ is a frequency domain representation of the k-th source signal.
In time-frequency domain, signals have the property of sparseness. In mathematical form, this is described as:

$$S_1\ (t,f) \cdot S_2\ (t,f) \approx 0, \quad \forall (t,f). \tag{4}$$

2.2. Spectral Feature Clustering

Currently most T-F masking algorithms utilize two features:

- the delay time calculated from the phase difference between observations,
- the attenuation rate between observations.

We limit our interest only to the delay-time. Due to our experimental setup, where all sources are located at the same distance from the microphone center, the attenuation rate provides no cues for separating among sources.

2.2.1. Delay Time Calculation

The anechoic mixing process can be expressed as

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ e^{-j\frac{2\pi f \delta_1}{L}} & e^{-j\frac{2\pi f \delta_2}{L}} \end{bmatrix} \begin{bmatrix} S_1\ (t,f) \\ S_2\ (t,f) \end{bmatrix}, \tag{5}$$

where: δ_i (i =1,2) is the delay between two microphones, and L is the number of STFT points.
Assuming that microphone 1 is the reference point, under the condition of WDO, the mixing model can be simplified to

$$\begin{bmatrix} X_1\ (t,f) \\ X_2\ (t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ e^{-j\frac{2\pi f \delta_i}{L}} \end{bmatrix} S_i\ (t,f). \tag{6}$$

The delay δ_i is obtained using a phase correlation function [6]:

$$\delta(t,f) = \frac{L}{2\pi f} \phi(t,f), \tag{7}$$

where $\phi(t,f)$ is the phase difference,

$$\phi(t,f) = \angle X_1(t,f) - \angle X_2(t,f). \tag{8}$$

2.2.2. Delay Time Histogram

Since speech signal has sparsity property against both time and frequency, to reconstruct the original signals, time-frequency cells must be clustered into two groups. The delay between observed signals can be an effective feature. Using the estimated delays and creating their histogram, we shall be able to detect two histogram peaks, δ_1 and δ_2, corresponding to two sources.

2.3. Spectrogram Masking for Source Reconstruction

Source reconstruction is performed by binary mask detection for the spectrogram's cell, for each expected source, due to some specific feature, followed by an inverse short time Fourier transform (ISTFT). The binary mask approach depends strongly on the clustering quality of given feature. Though the delay data $\delta(t,f)$ are spread, the peaks can approximately estimate the direction of sources.
In conventional method the clustering is given by drawing the separation line at the middle of two histogram peaks. Then the binary masks are generated by

$$M_1(t,f) = \begin{cases} 1 & \text{if } |\delta(t,f) - \delta_1| < |\delta(t,f) - \delta_2|, \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$M_2(t,f) = \begin{cases} 1 & \text{if } |\delta(t,f) - \delta_1| > |\delta(t,f) - \delta_2|. \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Therefore, the speech mixture signal can be separated by binary masks $M_i(t,f)$, and the separated signals $\hat{S}_i(t,f)$ are given by the following:

$$\hat{S}_i(t,f) = M_i(t,f) X_j(t,f). \tag{11}$$

Finally, by using the ISTFT, the separated signals are transformed in time domain.

3. Analysis: the WDO Assumption

3.1. Experimental Set-Up

Some experiments are performed in a conference room to certify our methods. The geometrical arrangement and parameters are shown in Fig. 1, and other parameters are shown in Table 1.

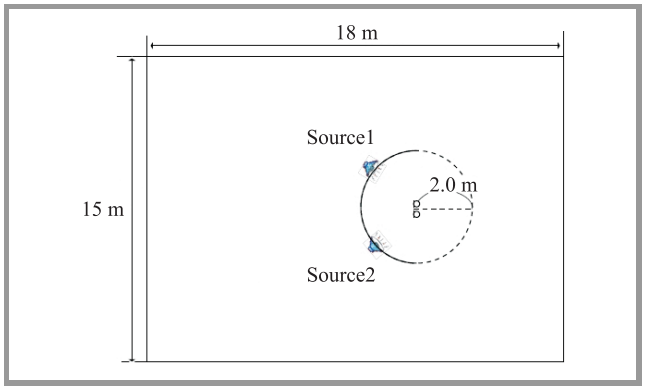

Fig. 1. The arrangement for signal acquisition.

Table 1
Experimental parameter setup

Sampling frequency	$f_s = 8000$ Hz
Microphone distance	$d = 40$ mm
Sound velocity	$c = 340$ m/s
Window type	Hamming
STFT frame length	$L = 1024$
Frame overlap	$\Delta = 512$

We use sources and mixtures coming from the *ASJ continuous speech* corpus [13], available for research work. One source is located at orientation expressed with respect to the normal line of the base line of two microphones. The normal line corresponds to the direction of 0 degree. The first source can be oriented as follows: starting from 0° it can take next orientations with 10-degree increments up to 80 degrees. The other sources can be located from 10° to 90° with 10 degree-increments. Hence the smallest possible orientation difference between two sources is 10 degrees, whereas the largest one is 90 degrees.

3.2. Phase Difference Errors

Although in principle the time-frequency masking based on time delay between microphones is a good method for the BSS problem, in real circumstances there appear large errors of phase difference estimation. For example, the calculated delay time derived from phase difference between two microphone signals should be less than d/c, where d is the distance between microphones, c is the sound velocity, but the estimated delays for lower frequencies obviously often violate this restriction. Fig. 2 shows examples of the time delays as a function of individual frequency bins $\{\tau = f(freq)\}$ for real and simulated mixtures.

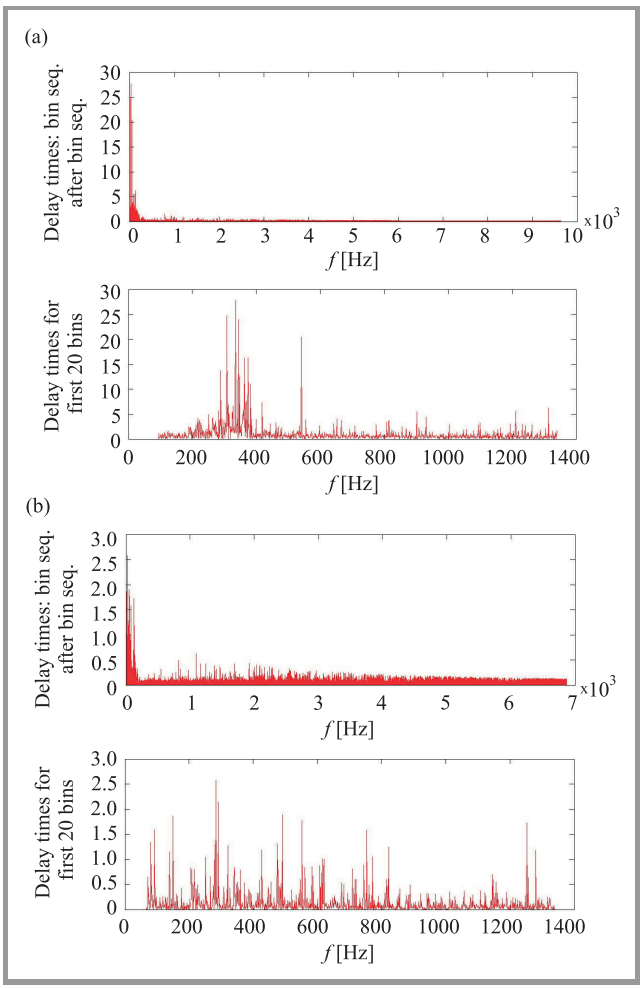

Fig. 2. Time delays as a function of frequency: (a) in real mixtures and (b) simulated mixtures.

Due to large errors, both histogram analysis and cell clustering by the use of delay values will be very difficult in the lower frequency band.

3.3. Cut-Off Frequency

One way to cope with the large estimation error in low frequency band is deleting these lower frequency components when generating binary mask. If we set the cut off frequency very low, the separated signal will still contain the error components. On the other hand, if we set the cut off frequency very high, it will affect the tone quality of separated signals.
The selection criterion for the cut-off frequency is keeping the tone quality of separated signals, at the same time, eliminating error components as much as possible. As demonstrated in reference [12] and by experiments, the cut off frequency need to be set around 400 Hz. In this paper instead of applying a general and simple cut-off we will individually examine each cell according to some energy criteria (Section 5).

4. Orientation Histogram Generation and Analysis

4.1. 1-D or 2-D Histogram?

With similar magnitude of sources no significant differences in attenuation rate appear. In Fig. 3a a histogram of symmetric attenuation values (A – 1/A) is shown, computed for already restricted, selected cells. For simulated mixtures there is only one clear maximum, at 0, that corresponds to attenuation ratio, A = 1. The reason for this observation is easy explained as the sources are of normalized amplitude and both mixtures are approximately in the same amplitude range. Hence in the ideal case of simulated data there is no gain of using attenuation ratio. Our element selection rules (that will be explained later) are sufficient to get clear 1-D direction histograms.

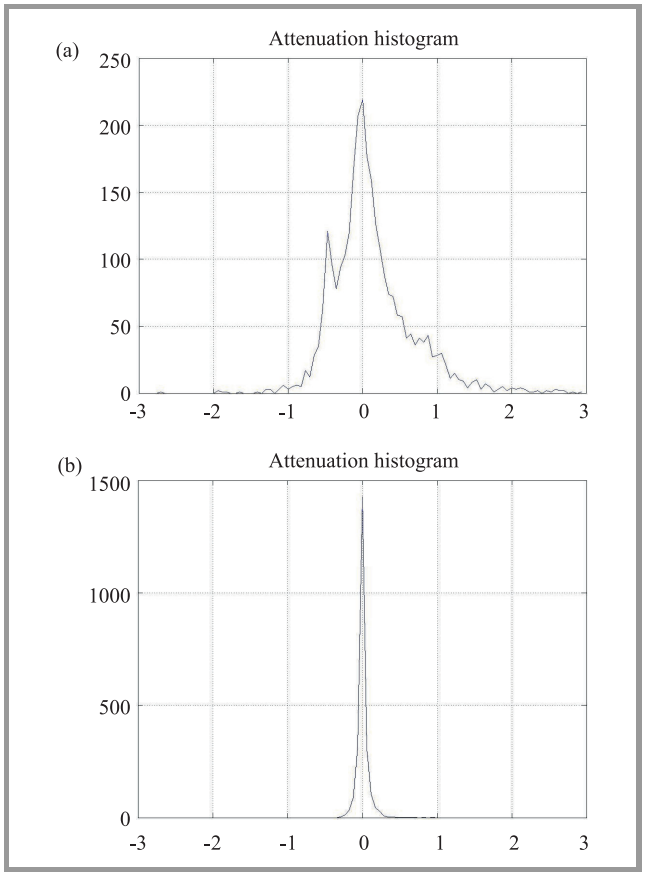

Fig. 3. The 1-D histogram of symmetric attenuation ratio for (a) simulated mixtures and (b) real acquired mixtures.

But is the attenuation ratio helpful for echoic mixtures? In Fig. 3b the attenuation histogram is now deteriorated and it shows a second local maximum, around –0.5. But there is no correspondence of this maximum to any of the sources. By checking where these values come from we conclude that they are the result of a significantly delayed echo, which still is a mixture of sources.

To summarize this discussion: attenuation ratio can eventually help if the mixed signals are of significantly different amplitudes. But these can not be echoic mixtures.

4.2. Orientation Instead of Delay Time

In our approach we compute a histogram of orientation angles instead of delay times. For this feature the histogram bins are linearly matching the angle scale, e.g., the difference of, say, 10 degrees corresponds to the same number of bins when θ is nearly 90 degrees or near 0 degrees. But in the histogram of delay times, the linear decomposition of histogram bins in the time space will correspond to a nonlinear scale in the orientation space, due to the mapping by the sin() function.
In fact, for the arrangement given in Fig. 1, where two sources are located at the same distance of 2 m from the center of two microphones, we can write:

$$\theta(t,f) = \arcsin(\delta(t,f)\cdot c/d), \tag{12}$$

where: c is the average speed of sound and d – the base distance between two microphones.

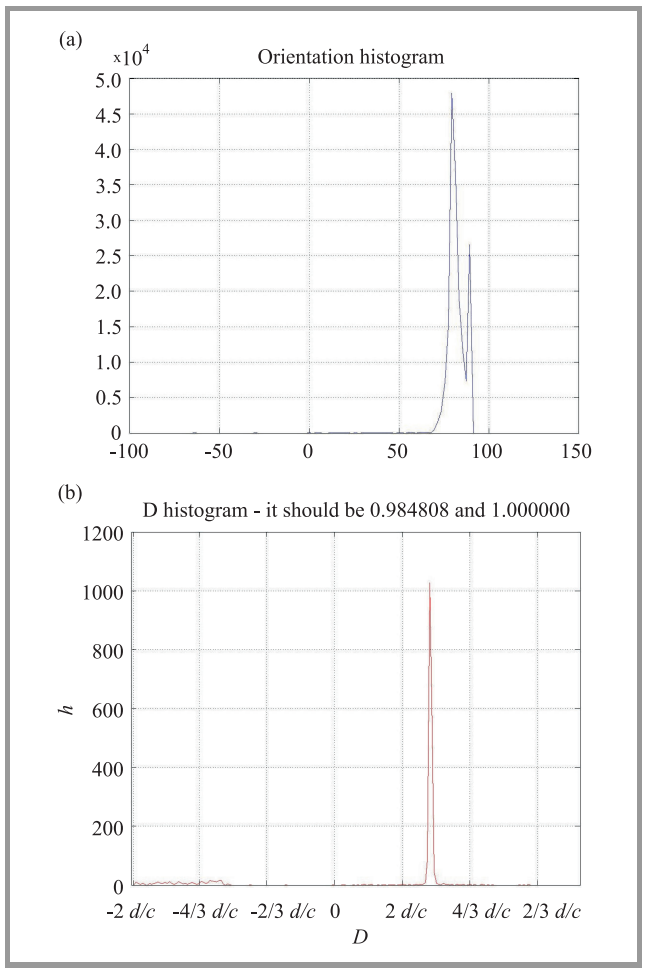

Fig. 4. A difficult separation case of two sources at orientations of 80 and 90 degrees with respect to the normal to base line of microphones: (a) the orientation histogram succeeds, whereas (b) the delay-time histogram fails

The delay time $\delta(t,f)$ can be measured from the mixture spectrogram according to Eqs. (7) and (8). From Eq. (12) in turn we observe that the delay time is nonlinearly dependent on the orientation angle. We can write:

$$\delta(\theta) = \frac{d}{c}\sin(\theta). \tag{13}$$

Let us observe the Fig. 4, which illustrates the most difficult case in T-F based speech separation when both sources are oriented very closely and at 80 and 90 degrees with respect to the normal to base line of microphones, i.e., nearly in-line with this base line. Still two clear local maxima are present in the orientation histogram, but not in the time delay histogram. In the latter case the time delays are nearly the same and they fall into a single histogram bin.

4.3. Confidence of Time Delays versus Energy-Based Selection

It is already well recognized that particular F-T cell's provide features with different quality or error, as in practice the WDO principle is often violated. The recently proposed methods, called TIFROM and DEMIX, use a "confidence measure" to select elements of the T-F signal (mixture) representation, which are with high probability "produced" by a single source only. The "confidence" is based on multiple PCA analysis in the attenuation-delay space for samples coming from the local neighborhood (say 3×3) of given element in the T-F space. The principal PCA-based axis is determined for each T-F cell and a confidence value is established that reflects the eigenvalue related to such principal eigenvector. The confidence value plays the role of a weight and allows to generate a weighted histogram. In our experiments, where both sources have similar amplitude, this approach performs worse.

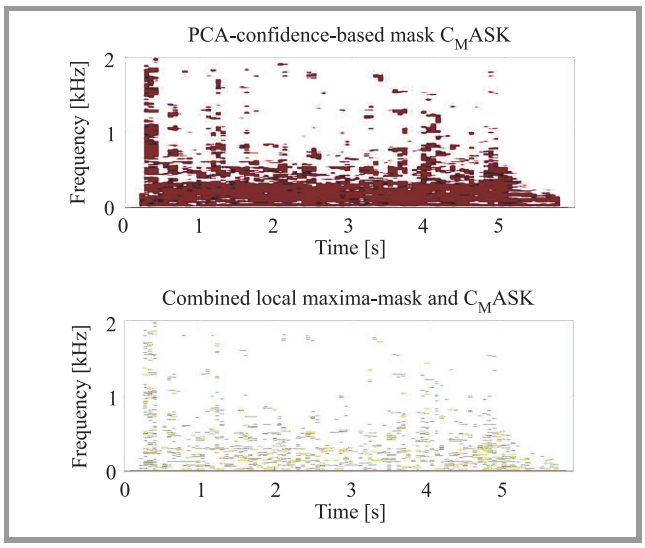

Fig. 5. Selection mask for spectrogram cells based on confidence value (top), compared with energy-based selection mask (bottom) (for real mixtures).

As it is seen in Fig. 5, if we follow the DEMIX approach and allow only highly confident elements (with confidence value > 90, we still enable most of the high energy, low-frequency elements to contribute to our direction histogram. As we have already shown, the delay information at low frequency bins is deteriorated by large errors. Applying our selection scheme we are able to concentrate on the relatively error-free information. This is further validated by results provided in Section 6.

4.4. Energy-Based Selection Criteria

Instead of computationally expensive approaches in TIFROM or DEMIX, our proposition is to use a restrictive cell selection rule, considering two criteria.

1. The local maxima along each frequency-indexed column (Fig. 6).
2. Near global maximum cells along the time axis for each frequency bin (Fig. 7).

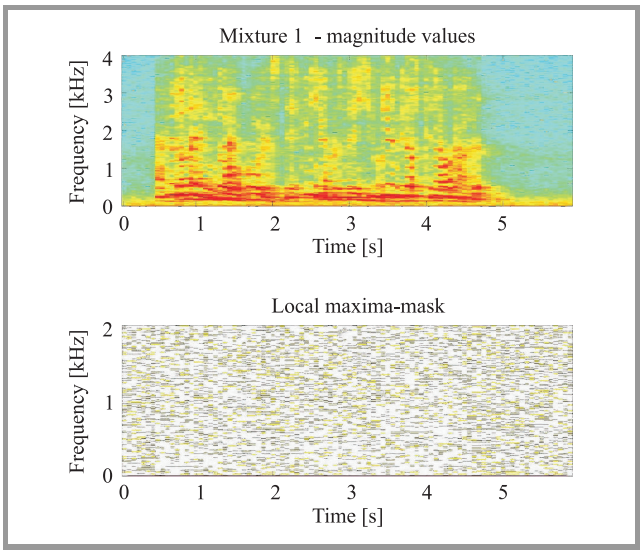

Fig. 6. The spectrogram of first mixture and a local maximum-based cell selection mask.

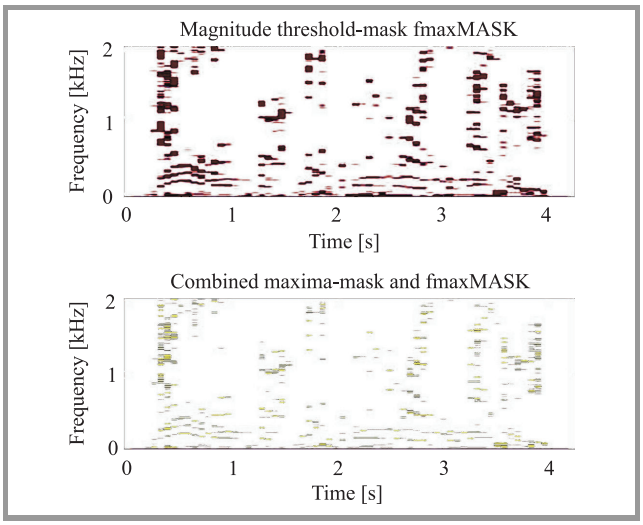

Fig. 7. Global-maximum-based cell selection mask (top) and the combined local- and global-based selection mask.

5. Many-Valued Masking

5.1. Classification Rule

The simple criterion for T-F mask generation, presented in Section 2, which classifies a spectrogram cell according to the smallest distance of its feature value to a histogram peak of this feature, is not a proper solution. We propose a more restrictive criterion that allows a feature peak-based classification only for cells where the distances-to-peak are relatively small. Then the appropriate mask is filled with value 1 and the other with 0 at given cell. If the feature value differs to much from all the detected histogram peaks, the masks are filled with some values from the interval [0, 1], computed by some frequency distance functions: $A_i(t,f), i = 1,2$. The rule for creation of the two spectrogram masks is as follows:

$$M_1(t,f) = \begin{cases} 1 & \text{if } |\theta(t,f) - \theta_1| < \theta_{max} \\ 0 & \text{if } M_2(t,f) = 1 \\ A_1(t,f) & \text{otherwise} \end{cases}, \quad (14)$$

$$M_2(t,f) = \begin{cases} 1 & \text{if } |\theta(t,f) - \theta_2| < \theta_{max} \\ 0 & \text{if } M_1(t,f) = 1 \\ A_2(t,f) & \text{otherwise} \end{cases}. \quad (15)$$

The normalized frequency distance functions are:

$$A_1(t,f) = \frac{W_1(t,f,f0_1(t))}{W_1(.) + W_2(.)}, \quad (16)$$

$$A_2(t,f) = \frac{W_2(t,f,f0_2(t))}{W_1(.) + W_2(.)}, \quad (17)$$

where the $f0_i(t) - s$ represents the fundamental frequency of source i in window t. The distance function $W_i(t,f,f0_i)$ gives a weight in proportion to two distances of cell's frequency f to the two nearest harmonic frequencies of given source $(n_L f0, n_H f0)$.

5.2. Harmonic Frequencies

The next results illustrate processing steps for the detection of two fundamental frequencies and their common multiple frequency. Even in a general overview of the total energy distribution along frequency bins we can already distinguish local maxima that corresponds to fundamental frequencies of both speakers and to magnifications around common multiple frequency (Fig. 8).

As the fundamental frequency can change during the speech the energy measurements are repeated every several consecutive frames. At first the gradient function is computed from the energy function along frequency axis (Fig. 9). Then the clearly visible local maxima peaks are detected and their harmonic structure is analyzed in order to select the fundamental frequencies and their common multiple frequency (Fig. 10).

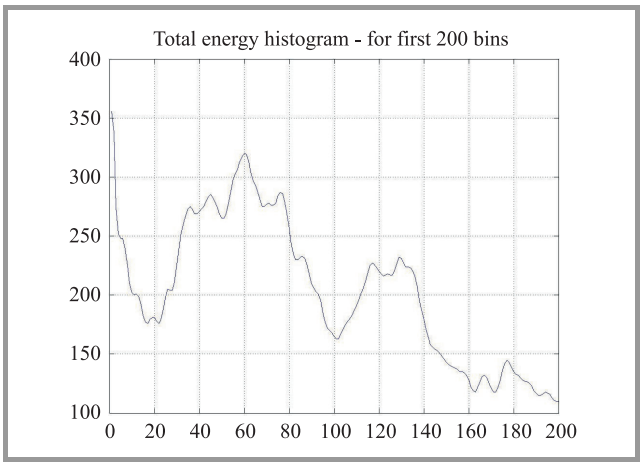

Fig. 8. Total energy distribution per frequency bins for real mixtures.

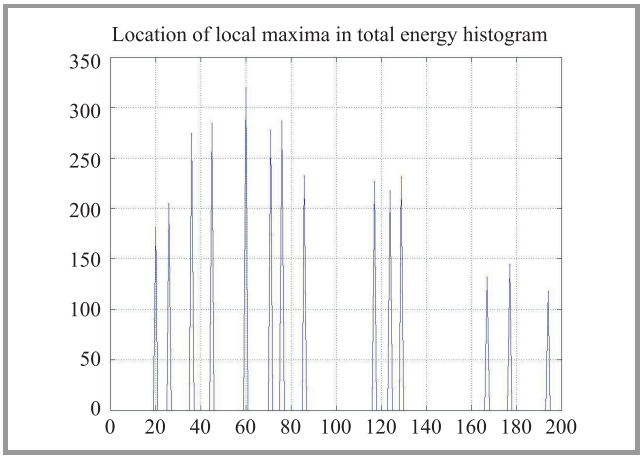

Fig. 9. The energy gradient along the frequency axis computed for real mixtures.

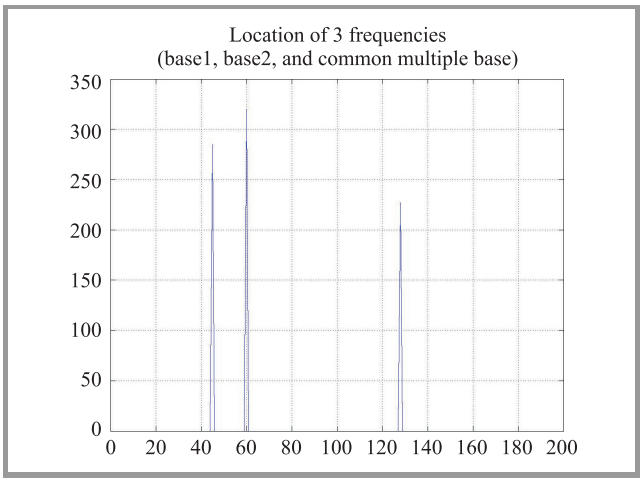

Fig. 10. The locations of two fundamental frequencies and a common multiple frequency for real mixtures.

6. Results

6.1. Histogram Analysis – Experiments

In experiments it turned out that the most difficult case is to distinguish between orientations of 80–90 degrees. This was the reason while we prefer to use direction feature clustering instead of the time delay one. The direction histogram for real data is not a simple mixture of two Gaussians, centered at speaker directions, as the interference is so that the second signal seems to "generate symmetric peaks" around the center of the first signal (Fig. 11). This interference effect could also be responsible for lower histogram peaks of the second source.

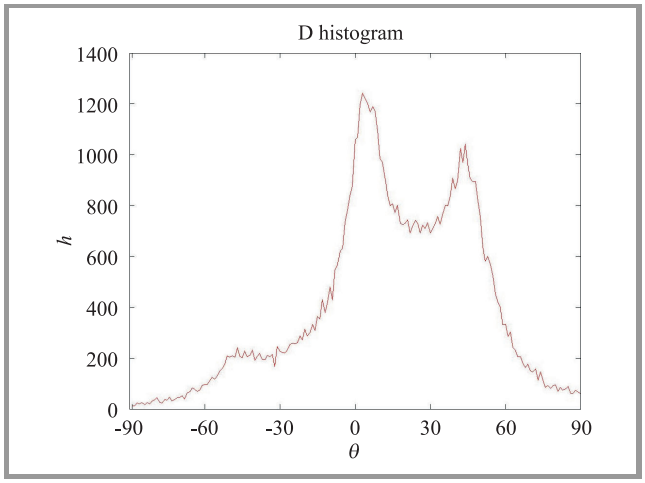

Fig. 11. Example of orientation histograms obtained for mixtures of two sources.

Table 2
The estimated orientations θ_1 and θ_2 (in top and bottom rows) based on the orientation histogram for two real acquired mixtures

s_2 at:	20°	30°	40°	50°	60°	70°	80°	90°
s_1 at	13	15	13	13	13	12	12	12
10°	22	29	41	46	55	73	78	87

Table 2 shows how well the clustering can be done for the whole representative range of orientation angles $[0°, 90°]$. The detection of orientations for both speakers has improved, especially in the most difficult range of orientations: $[80°, 90°]$.

6.2. Source Reconstruction

The performance of the spectrogram masking step will be evaluated in terms of the WDO coefficient (measure of W-disjoint orthogonality) [4]. This coefficient is computed for given useful destination source and interference signal. Related criteria are: the preserved-signal ratio (PSR) and the signal-to-interference ratio (SIR). The definitions are as follows:

$$WDO(d,i) = \frac{||M_d(t,f)S_d(t,f)||^2 - ||M_d(t,f)S_i(t,f)||^2}{||S_d(t,f)||^2} = PSR - \frac{PSR}{SIR}, \quad (18)$$

$$PSR = \frac{||M_d(t,f)S_d(t,f)||^2}{||S_d(t,f)||^2}, \quad (19)$$

$$SIR = \frac{||M_d(t,f)S_d(t,f)||^2}{||M_d(t,f)S_i(t,f)||^2}, \quad (20)$$

where: $S_d(t,f)$ is the desired source signal, $M_d(t,f)$ is the spectrogram mask for source d, and $S_i(t,f)$ is the interfering signal. The range of WDO values is: $0 \leq WDO \leq 1$. For ideal source reconstruction it would be $WDO = 1$.

The results in Table 3 clearly illustrate the statement that a binary spectrogram mask does not allow a proper extraction of speech sources from real echoic mixtures. The WDO coefficients have low values within the range of $[0.26, 0.66]$.

Table 3
The WDO(1,2) and WDO(2,1) coefficients (in top and bottom rows) for source reconstruction with ordinary binary spectrogram masks (according to Eqs. (9)–(10))

s_2 at:	20°	30°	40°	50°	60°	70°	80°	90°
s_1 at	0.39	0.46	0.66	0.59	0.50	0.49	0.29	0.26
10°	0.27	0.33	0.52	0.40	0.30	0.27	0.28	0.26

The results in Table 4 have been achieved by applying the multi-valued mask for source extraction, proposed in this paper. Here we focus on most difficult situations, when the sources are close to each other and their orientations w.r.t. the microphones are ending towards 90°. The results are significantly better than in the binary mask case. With the multi-valued mask a sufficiently good source extraction is

Table 4
The WDO(1,2) and WDO(2,1) coefficients (in top and bottom rows) for source reconstruction with multi-valued spectrogram masks (according to Eqs. (14)–(17))

s_2 at: / s_1 at	60°	70°	80°	90°
50°	0.92	0.91	0.90	0.88
	0.90	0.89	0.87	0.85
60°	–	0.90	0.90	0.88
	–	0.89	0.88	0.85
70°	–	–	0.81	0.68
	–	–	0.75	0.58
80°	–	–	–	0.45
	–	–	–	0.27

possible even for orientations in the range of 80° (and to some part even to 90°), where the binary mask definitely failed.

7. Conclusion

This paper introduces several improvements to the time-frequency masking approach to blind speech separation, that relax the strict DOA assumption. After experiments with anechoic (simulated) mixtures and echoic (real) mixtures of speech sources, acquired by two microphones, we worked out methods that improve two steps of such conventional approach – orientation histogram analysis and T-F mask creation. The creation of an orientation histogram is efficiently performed by considering the phase-difference data of reliable cells only. For this we combine an energy local maximum criterion along the frequency axis (for every time frame) with near global maxima intervals along the time axis (for each particular frequency bin).
Next improvement is due to the use of many-valued spectrogram mask. Thus we relax the strict WDO assumption, that seems not to hold perfectly in practice. The clustering feature is now only responsible for selecting cells with obviously perfect behavior. Otherwise the harmonic frequencies are applied as a new selection criterion. The WDO coefficients for the reconstructed sources document a significant improvement, even for close sources and for large orientation angles – reaching nearly 90°.

Acknowledgments

Włodzimierz Kasprzak was supported by the Polish Ministry of Science and Higher Education within the grant N N514 1287 33.

References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Berlin: Springer, 1997.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing.* Chichester: Wiley, 2003.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. Sig. Proces.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] S. Rickard, "The DUET blind source separation algorithm", in [1], pp. 217–237.

[6] 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, pp. 149–157, no. 2, 2001.

[7] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources", *Sig. Proces.*, vol. 85, pp. 1389-1403, 2005.

[8] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture", *IEEE Trans. Sig. Proces.*, vol. 58, no. 1, pp. 121–133, 2010.

[9] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sanei, "K-hyperline clustering learning for sparse component analysis", *Sig. Proces.*, vol. 89, pp. 1011–1022, 2009.

[10] S. Makino, H. Sawada, R.Mukai, and S. Araki, "Blind source separation of convolutive mixture of speech in frequency domain", *IEICE Trans. Fundament.*, vol. 88, no. 7, pp. 1830–1847, 2004.

[11] A. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors", *Sig. Proces.*, vol. 87, pp. 1833–1847, 2007.

[12] H. Ouchi and N. Hamada, "Separation of speech mixture by time-frequency masking utilizing sound harmonics, *J. Sig. Proces.*, vol. 13, no. 4, pp. 331–334, 2009.

[13] T. Kobayasi, S. Itahashi, S. Hayamizu, and T. Takezawa, "ASJ continuous speech corpus for research, *J. Acoust. Soc. Japan*, vol. 48, no. 12, pp. 888–893, 1992 (in Japanese).

Włodzimierz Kasprzak received the German Dr.-Ing. in pattern recognition (1996) from University of Erlangen-Nuremberg, and the Polish Ph.D. in computer science (1987) and D.Sc. in automation control (2001) from Warsaw University of Technology (WUT). From 1997, he is with the Institute of Control and Computation Engineering at WUT, currently appointed as Associate Professor. His main interests are in pattern recognition and artificial intelligence, and their applications in image and speech analysis. Among others, he has conducted research at the University of Erlangen-Nuremberg (1989–1991), the Bavarian Research Center FORWISS, Erlangen, Germany (1992–1995), and the RIKEN Institute, Wako-shi, Japan (1995–1996). He is Member of the IAPR and Polish TPO societies.
e-mail: W.Kasprzak@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warszawa, Poland

Ning Ding received the B.Sc. and the M.Sc. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2005 and 2008, respectively. He is currently working toward the Ph.D. degree in signal processing at Keio University. His main research interest is the speech signal processing, Direction of Arrival estimation and speech separation.
e-mail: ding@hamada.sd.keio.ac.jp
Signal Processing Lab.
School of Integrated Design Engineering
Keio University
3-14-1 Hiyoshi, Kohokuku
Yokohama 223-8522, Japan

Nozomu Hamada received the B.Sc., the M.Sc. and Ph.D. degrees in electrical engineering from Keio University. In 1974, he became an Instructor in electrical engineering at Keio University. He has been a Professor there in the Department of System Deign Engineering since 1991. He was the visiting researcher in Australian National University in 1982. His research interests include circuit theory, stability theory of dynamical system, design of one or multi-dimensional digital filters, and image processing. One of his resent research fields is a realization of human interface system using microphone array. The main topics in this study are acquisition of audio signal from spatially distributed sound sources and the separation of multiple speech signals. He is the author of "Linear Circuits, Systems and Signal Processing" (Chapter 5) (Marcell Dekker Inc. 1990), "Two-Dimensional Signal and Image Processing" (SICE, 1996), "Introduction of Modern Control Systems" (Corona Pub. Inc. 1997). He was the chair of IEEE Signal Processing Society in Japan Chapter (2004) and editorial board in Journal of Signal Processing. He was guest editor of special issues relevant to multi-dimensional signal processing: its application and realization technique in the IEICE (2000), Signal Processing (2002, 2003, 2006, 2007).

e-mail: hamada@sd.keio.ac.jp
Signal Processing Lab.
School of Integrated Design Engineering
Keio University
3-14-1 Hiyoshi, Kohokuku
Yokohama 223-8522, Japan