MOKSLAS – LIETUVOS ATEITIS
SCIENCE – FUTURE OF LITHUANIA

ISSN 2029-2341 print / ISSN 2029-2252 online
2011 3(1): 43–49 doi: 10.3846/mla.2011.009

Signalų technologijos T 121
Signal Technologies T 121

Elektronika ir elektrotechnika
Electronics and Electrical Engineering

SPEAKER LOCALIZATION AND SPEECH SEPARATION IN TWO ECHOIC MIXTURES

Włodzimierz Kasprzak[1], **Ning Ding**[2], **Nozomu Hamada**[3]

[1]*Warsaw University of Technology, Poland,* [2,3]*Keio University, Japan*

E-mail: [1]*W.Kasprzak@ia.pw.edu.pl,* [2]*ding@sd.keio.ac.jp,* [3]*hamada@sd.keio.ac.jp*

Abstract. We are developing two crucial improvements on the time-frequency masking approach to the blind speech separation of underdetermined mixtures when processing anechoic and echoic mixtures. First, the proposed method copes with the usually large amount of delay estimation error that appears in a low frequency band. This step generates a restrictive mask for phase delays on the basis of local and global energy distribution analysis. This mask allows the selected cells to contribute to the orientation histogram. Second, the strong WDO assumption (disjoint orthogonal frequency domain) is relaxed by allowing some frequency bins to be shared by both sources. By detecting fundamental frequencies of speakers at instantaneous time points, mask creation is supported by exploring their harmonic frequencies. The proposed method is proved to be effective and reliable in conducting experiments with both simulated and real-life mixtures.

Keywords: blind source separation, histogram clustering, spectrogram analysis, speech reconstruction, time-frequency masking.

Introduction

Recently proposed time-frequency (T-F) methods of blind speech separation (Makino *et al.* 2007) for the underdetermined mixture case, known as DUET (Yilmaz *et al.* 2004; Rickard 2007), TIFROM (Abrard *et al.* 2005), DEMIX (Arberet *et al.* 2010), etc. explore differences in the locations of the speakers. They first make feature clustering or histogram analysis in the attenuation rate and time delay spaces to detect the number and characteristics of speakers. Second, the reconstruction of sources is performed by masking the spectrogram of a mixture with appropriate binary masks created in accordance with the estimated cluster centers (or histogram peaks) of the given feature. In DUET, a weighted two-dimensional (2-D) histogram is constructed expressing differences in amplitude and phase between time-frequency representations of two mixtures. A histogram peak is assumed to represent a source – uniquely characterized by relative attenuation and time delay.

Modifications to the basic T-F masking approach focus on increasing the reliability of delay information provided by spectrogram cells. In SAFIA (Aoki *et al.* 2001), the authors extend delay measurement to handle more than two observations. The DUET method uses a weighting scheme for features – weight grows with the growing frequency of the element. TIFROM, DEMIX and the uniform clustering approach (He *et al.* 2009) observe the stability of delay feature in a local neighborhood in spectrogram space and associate an appropriate confidence measure with such delay data. The HS method (Ouchi *et al.* 2009) explores other kind of information as a feature selection criterion. It uses a harmonic structure of a speech signal to support the feature clustering step. The proposed method estimates the fundamental frequency of the source given initial separation results. This information allows to select frequency components of the mixture which probably come from a single source.

T-F methods usually work well for anechoic mixtures and significantly different orientations of speakers' w.r.t the microphone set. Otherwise, they fail due to the following difficulties: 1) in echoic mixtures, the delays estimated (especially) for low frequencies are significantly disturbed; 2) even in anechoic mixtures, the WDO assumption (disjoint orthogonal of sources in the frequency domain) is not fully satisfied (especially in the low frequency band).

This paper proposes a more restrictive than before selection of the "true" phase delays based on local and global energy distribution analysis performed in the T-F domain (spectrogram). The second contribution is to relax with the WDO assumption by allowing frequency bins to be shared by more than one source. Then, the mask for the source extraction stage is created not only by applying the time delay criterion but also by exploring the harmonics of fundamental frequencies.

The remainder of this paper is organized as follows. First, the basic approach to time-frequency masking in blind source separation is introduced. Then, the first

modification, regarding histogram detection for the DOA feature, is described. Then, the second modification, regarding the creation of the WDO-based separation mask, is described. The experiments provided in the last section show how the proposed approach is validated and compared with the basic solution.

T-F masking Approach to BSS

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

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

where τ is the discrete time index, $h_{jk}(l)$ represents impulse response from source k at sensor j, N is the number of sources and M is the number of sensors.

We focus particularly on the situation where the number of sources N = 2 and the number of sensors M = 2 (Fig. 1). Typical time-frequency masking methods for BSS are based on the *difference of arrival time* (DOA) principle and the WDO (frequency-domain *disjoint orthogonal*) assumption.

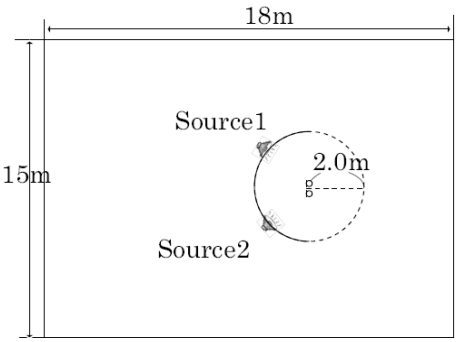


Fig. 1. The arrangement for the measurement of 2 mixtures of 2 sources

DOA histogram. Time domain signals, $x_j(\tau)$, sampled at frequency f are converted to the frequency domain into a time-series of vector signals $X_j(t, f)$ by applying L point STFT to consecutive signal frames:

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

where $w(r)$ is a window function, S is the size of window shift, t is the integer time frame index and f ($\in[0, L/2]$) is an integer index of the frequency bin.

The selection of an appropriate feature that differentiates between sources is essential in every approach to blind source separation. T-F masking approaches utilize the DOA prinicple – for different sources there must be a dfferent delay calculated from phase difference between observations. Assuming that microphone 1 is the reference point, the anechoic mixing process can be expressed as

$$\begin{pmatrix} X_1(t,f) \\ X_2(t,f) \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ e^{-j2\pi f\delta_1/L} & e^{-j2\pi f\delta_2/L} \end{bmatrix} \begin{pmatrix} S_1(t,f) \\ S_2(t,f) \end{pmatrix}, \qquad (3)$$

where δ_i (i =1,2) are time delays between two microphones for each source and L is the number of STFT points. Under the condition of WDO, the mixing model (3) can be simplified to

$$\begin{pmatrix} X_1(t,f^{(i)}) \\ X_2(t,f^{(i)}) \end{pmatrix} = \begin{pmatrix} 1 \\ e^{-j2\pi f^{(i)}\delta_1/L} \end{pmatrix} S_i(t,f^{(i)}), i = 1,2\,, \qquad (4)$$

where $f^{(1)}$, $f^{(2)}$ are disjoint frequency subsets, at which the components of sources 1 or 2 are of non-zero amplitude. Time delay δ corresponds to phase difference ϕ at frequency f as follows:

$$\delta(t,f) = \frac{L}{2\pi f}\varphi(t,f)\,. \qquad (5)$$

Phase difference $\phi(t,f)$ is obtained from mixture spectrograms as

$$\varphi(t,f) = \angle X_1(t,f) - \angle X_2(t,f)\,. \qquad (6)$$

Assuming *sparse* speech signal in time and frequency, to reconstruct original signals, time-frequency cells must be clustered into *two groups*. The *time delay* between the 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.

WDO assumption. The time-frequency masking approach to blind speech separation utilizes instantaneous mixtures at each time frame t and frequency bin f:

$$X_{j(t,f)} \approx \sum_{k=1}^{N} H_{jk}(f)\, S_k(t,f)\,, \qquad (7)$$

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. It is assumed that in the time-frequency domain, signals have the property of *sparseness* (also called WDO assumption) i. e.

$$S_1(t,f)\cdot S_2(t,f) \approx 0, \forall_{t,f}\,. \qquad (8)$$

Under this assumption, source reconstruction is possible up to a scaling coefficient even without the estimation of the frequency response of mixing matrix ($\mathbf{H}(f)=[H_{jk}(f)]$). Though delay data $\delta(t, f)$ are spread, the peaks can approximately estimate the direction of

sources. In the conventional method, clustering is given by drawing the *separation line in the middle* of two histogram peaks. Then, binary masks are generated according to the following decision rule:

$$M_1(t,f)=\begin{cases}1, & \text{if } |\delta(t,f)-\delta_1| < |\delta(t,f)-\delta_2|;\\ 0, & \text{otherwise;}\end{cases}$$
$$M_2(t,f)=\begin{cases}1, & \text{if } |\delta(t,f)-\delta_1| > |\delta(t,f)-\delta_2|;\\ 0, & \text{otherwise.}\end{cases} \quad (9)$$

Obviously, the quality of binary masks depends strongly on the clustering quality of the DOA feature and how well the WDO assumption is satisfied by the current mixture.

After *binary masks* over all the spectrogram's cell are created according to the histogram peaks of the clustering feature, spectral contributions to each expected source are filtered out from the specrum of the first mixture. Therefore, when using binary masks $M_i(t, f)$, separated spectral signals $\hat{S}_i(t, f)$ are given as

$$\hat{S}_i(t,f)=M_i(t,f)X_1(t,f). \quad (10)$$

By following inverse short-time Fourier Transform (ISTFT), the sources in time domain can finally by reconstructed.

DOA Feature and Histogram Detection

We can draw two important conclusions from our experimental analysis of speech mixtures:

1. A feature value should be linearly proportional to the orientation angle – then it can reliably be used to separate two sources very close to each other not only at directions around 0 degrees but also at those of around 90 degrees.
2. A highly selective scheme is needed to concentrate on relatively error-free feature information only.

Feature: orientation instead of delay time. Instead of delay time, typically applied as a clustering feature, we perform histogram analysis of the feature that is directly – *orientation angle* θ (t, f). In fact, for the arrangement given in Fig. 1, where two sources are located at the same distance of 2 m from the centre between microphones, we can accept that

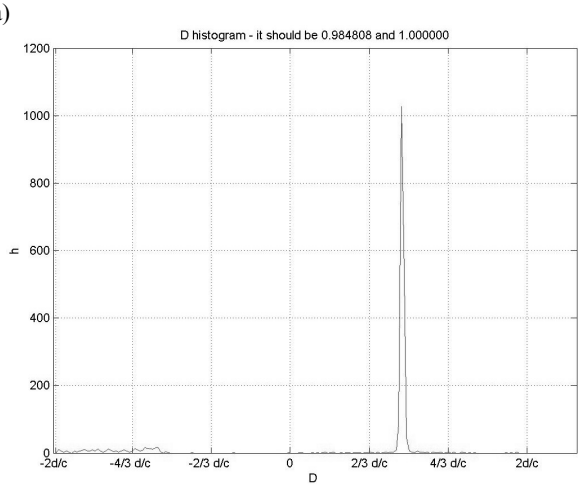
$$\theta(t,f)=\arcsin(\delta(t,f)\cdot c/d), \quad (11)$$

where c is the average speed of sound and d – the base distance between two microphones. Delay time $\delta(t, f)$ can be measured from the mixture spectrogram according to equations (5) and (6). From (11), in turn, we observe that delay time is nonlinearly dependent on the orientation angle. Thus, we can write that

$$\delta(\theta)=d\sin(\theta)/c. \quad (12)$$

Let us consider what happens if two sources are very close to each other. The most difficult case in T-F based speech separation is given when two sources are very close and nearly in-line with this base line of microphones. Assume that $\theta_1 = 80^o$ and $\theta_2 = 90^o$, with respect to the normal base line of microphones. Compare two histograms obtained in this case of different features – orientation $\theta(t, f)$ and time delay $\delta(t, f)$ (Fig. 2).

a)



b)

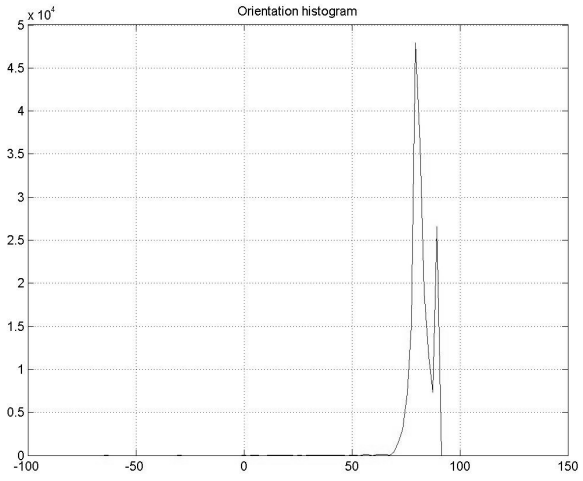


Fig. 2. Two histograms of different features in case two sources are located at 80^o and 90^o: (a) delay time, (b) orientation

Two clear local maxima are present in the orientation histogram (Fig. 2 b), because there are many bins between the values of 80 and 90 on the orientation scale, which is the same number of bins as that between 0 and 10 on this scale. However, in the time delay histogram (Fig. 2 a), orientations 80 and 90 correspond to very close

time delay values of 0.9848 and 1.000 [d/c]. It is virtually impossible to distinguish different histogram peaks for them.

Energy-based selection of a clustering feature. A restrictive feature selection procedure is developed, in which two criteria are jointly used, i.e. consider only information coming from

1) *local energy maxima* cells where the maxima are computed along each *frequency*-indexed column;
2) *near global* maximum cells along *time axis* for each frequency bin.

The high selectivity of both criteria is illustrated in Fig. 3.

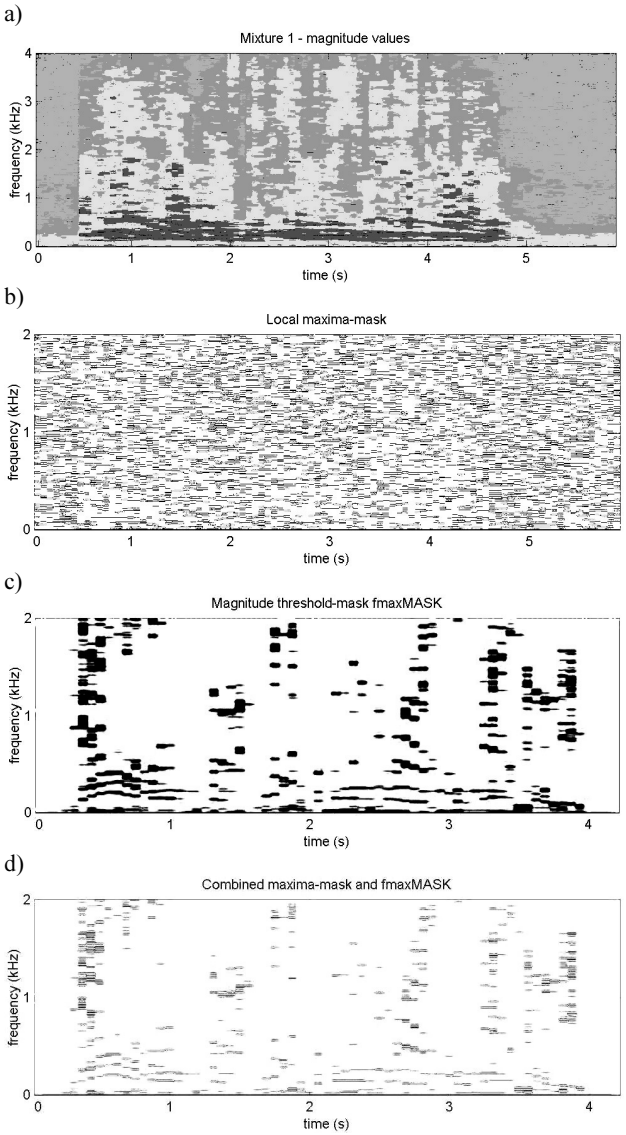


Fig. 3. The illustration of the feature selection procedure: (a) an input spectrogram, (b) the binary mask of local energy maxima (along frequency axis), (c) the binary mask of sufficient energy per frequency bin (along time axis), (d) the combined T-F cell selection mask

Source Mask

The next novelty is to explore fundamental frequencies of speakers, i.e. to detect and track them along in time and to influence by them the source mask creation. This step is illustrated in Fig. 4.

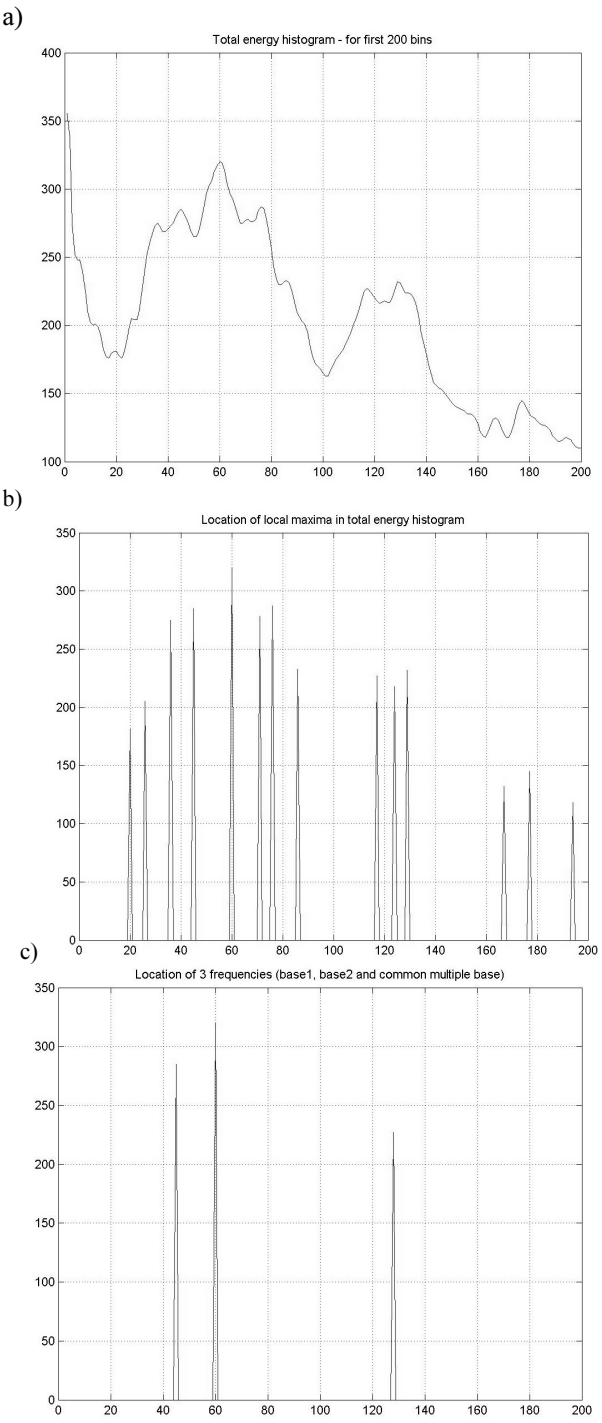


Fig. 4. The illustration of the fundamental frequency detection step: (a) energy distribution-per-frequency bin in some time window, (b) local energy maxima in (a), (c) selected fundamental frequencies of the first and second source and their first common frequency

Having detected the distribution of fundamental frequency $f0_i(t)$ for every expected source spectrogram $S_i(f, t)$, we can extract such spectrogram from given $X_1(f, t)$ more reliably than by using a binary mask only. Now, the mask no longer needs to be a binary one but the many-valued one. Spectrogram cells with delays matching the histogram peaks are classified to an appropriate source, i. e. an appropriate T-F source mask is filled with value 1 and the other with 0 at the given cell. Otherwise, the masks are filled with some values from interval [0, 1] computed by normalized frequency distance functions $Ai(t,f)$, i=1,2. The rule for creating T-F mask is:

$$M_1(t,f)=\begin{cases}1, & \text{if } (|\,\theta(t,f)-\theta_1\,|<\theta_{\max})\,;\\ 0, & \text{if } (M_2(t,f)=1)\,;\\ A1(t,f), & \text{otherwise}\,,\end{cases}$$
$$M_2(t,f)=\begin{cases}1, & \text{if } (|\,\theta(t,f)-\theta_2\,|<\theta_{\max})\,;\\ 0, & \text{if } (M_1(t,f)=1)\,;\\ A2(t,f), & \text{otherwise}\,.\end{cases} \quad (13)$$

Normalized frequency distance functions are

$$A1(t,f)=\frac{W1(t,f,f0_1(t))}{W1(t,f,f0_1(t))+W2(t,f,f0_2(t))}\,,$$
$$A2(t,f)=\frac{W2(t,f,f0_2(t))}{W1(t,f,f0_1(t))+W2(t,f,f0_2(t))}\,, \quad (14)$$

where $f0_i(t)$-s represent the fundamental frequency of source i in window t. Distance function $W(t, f, f0)$ gives weight in proportion to the distances of cell frequency f to the two nearest harmonic frequencies of the given source ($n_L f0_1$ and $n_H f0_2$) where

$$f\le n_L f0 \quad \text{and} \quad f\ge n_H f0 \quad \text{and} \quad H=L+1\,, \quad (15)$$

W($t,f, f0$) is a magnitude average in the above interval:

$$W(t,f,f0)=\frac{1}{2}\Big[\beta\cdot\mathrm{Mag}\big(X_1(t,n_L f0)\big)+ +(1-\beta)\cdot\mathrm{Mag}\big(X_1(t,n_H f0)\big)\Big]\,, \quad (16)$$

where

$$\beta=\big(f-n_L f0\big)\big/f0\,. \quad (17)$$

Results

The experimental setup (Fig. 1) is as follows:

- sampling frequency: f_s= 8 000 Hz;
- microphone distance: d = 40 mm;
- sound velocity: c = 340 m/s;
- window function: hamming;
- STFT frame length: L = 1 024;
- frame overlap: Δ = 512.

Orientation estimation. Tables 1, 2 and 3 show the results of two sources – men and women voices – located at different orientations in a circle with the radius of 2 [m] in front of a pair of microphones. The results in Table 1 were obtained for real acquired mixtures and present clustering a standard feature –time delay δ. It can be observed that the detection of orientations near the base line of microphones, i. e. 80°–90°, is at least very difficult if not impossible. The results in Tables 2 and 3 are from clustering directly the feature of orientation where Table 2 shows the results obtained for the same natural sources but for a simulated mixture pair, while Table 3 presents the results for real acquired mixtures. We observe how the detection of orientation has improved for orientations of 80°–90°.

Table 1. The estimated orientations θ_1 and θ_2 based on the *time delay* histogram for two *real* acquired mixtures

Women at / Men at	20°	30°	40°	50°	60°	70°	80°	90°
10°	13.5 21.9	14.9 29.4	13.5 40.8	13.5 46.4	13.5 54.8	12.1 67.4	12.1 73.2	12.1 78.4

Table 2. The estimated orientations θ_1 and θ_2 based on the feature of the *orientation* histogram for two *simulated* mixtures of real sources

Women at / Men at	20°	30°	40°	50°	60°	70°	80°	90°
10°	10.8 20.5	10.8 29.4	9.4 40.8	9.4 50.4	9.4 59.8	9.4 69.1	9.4 78.7	9.5 87.4

Table 3. The estimated orientations θ_1 and θ_2 based on the *orientation* histogram for two *real* acquired mixtures

Women at / Men at	20°	30°	40°	50°	60°	70°	80°	90°
10°	13 22	15 29	13 41	13 46	13 55	12 73	12 78	12 87

Source reconstruction quality. For total performance evaluation, we use the *WDO* coefficient (measure of *W*-disjoint orthogonal) computed from two other criteria – *PSR* (the Preserved-Signal Ratio) and *SIR* (the Signal-to-Interference Ratio) and defined as

$$WDO(D,I)=\frac{\|\,M_D(t,f)\,S_D(t,f)\,\|^2-\|\,M_D(t,f)\,S_I(t,f)\,\|^2}{\|\,S_D(t,f)\,\|^2}\,, \quad (18)$$
$$WDO(D,I)=PSR(D)-\frac{PSR(D)}{SIR(D,I)}\,, \quad \text{with}$$

$$PSR(D) = \frac{\| M_D(t,f)\, S_D(t,f) \|^2}{\| S_D(t,f) \|^2},$$
$$SIR(D,I) = \frac{\| M_D(t,f)\, S_D(t,f) \|^2}{\| M_D(t,f)\, S_I(t,f) \|^2}. \quad (19)$$

where $S_D(t, f)$ is a desired spectrogram, $M_D(t, f)$ is a spectrogram mask for a desired source and $S_I(t, f)$ is an interfering spectrogram. The interval of WDO values is $0 \le$ WDO ≤ 1. The ideal extraction of the desired source means that WDO(D, I) = 1.

The results in Table 4 clearly document that a binary spectrogram mask does not allow a proper extraction of speech sources from real echoic mixtures. WDO coefficients have low values in the range of [0.26, 0.66].

Table 4. WDO(1,2) and WDO(2,1) coefficients with *binary* spectrogram masks for histogram peaks in Table 1 for *real* acquired mixtures

Women at / Men at	20°	30°	40°	50°	60°	70°
10°	0.391 0.269	0.464 0.329	0.661 0.522	0.598 0.401	0.502 0.296	0.498 0.271

Table 5. WDO(1,2) and WDO(2,1) coefficients with *multi-valued* spectrogram masks and for histogram peaks in Table 2, for *simulated* mixtures

Women at / Men at	20°	30°	40°	50°	60°	70°	80°	90°
10°	0.907 0.964	0.921 0.962	0.942 0.952	0.931 0.958	0.928 0.958	0.925 0.960	0.931 0.956	0.934 0.956

The results provided in Tables 5 and 6 have been achieved by applying the multi-valued mask for source extraction. Table 5 shows WDO coefficients in case of simulated mixtures, whereas Table 6 – for real acquired mixtures and for difficult orientations tending towards 90°. The results are significantly better than in the case of the binary mask. With the multi-valued mask even for orientations of 80° or 90°, where the binary mask failed, sufficiently good source extraction is possible.

Table 6. WDO(1,2) and WDO(2,1) coefficients for *multi-valued* spectrogram masks and for *real* acquired mixtures

Women at / Men at	60°	70°	80°	90°
50°	0.926 0.904	0.910 0.889	0.900 0.874	0.880 0.858
60°	–	0.907 0.893	0.902 0.880	0.880 0.857
70°	–	–	0.816 0.756	0.687 0.586
80°	–	–	–	0.449 0.271

6. Conclusions

Two major improvements on time-frequency masking approaches to blind speech separation have been proposed and tested for a two-microphone case. They are based on observation that a strict WDO assumption (disjoint orthogonality of sources in the frequency doimain) is practically not fully satisfied. In the proposed method, the creation of an orientation histogram is efficiently performed by considering phase-difference data on reliable cells only. Hence, we combine an energy local maximum criterion along the frequency axis (for every time frame) with relative energy threshold along the time axis (for each particular frequency bin). Conversion from a binary spectrogram mask to a multi-valued mask, applied for source extraction, constitutes the second major improvement. The orientation peaks are responsible only for selecting spectrogram cells with nearly perfect match. Otherwise, harmonic frequencies are applied as a new selection criterion.

Acknowledgments

W. Kasprzak has been supported by the Polish Ministry of Science and Higher Education within the grant N N514 1287 33.

References

Abrard, F.; Deville, Y. 2005. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources, *Signal Processing* 85: 1389–1403. doi:10.1016/j.sigpro.2005.02.010

Aoki, M.; Okamoto, M.; Aoki, S.; Matsui, H.; Sakurai, T.; Kaneda, Y. 2001. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, *Acoust. Sci. & Tech* 22(2): 149–157. doi:10.1250/ast.22.149

Arberet, S.; Gribonval, R.; Bimbot, F. 2010. A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture, *IEEE Transactions on Signal Processing* 58(1): 121–133. doi:10.1109/TSP.2009.2030854

He, Z.; Cichocki, A.; Li, Y.; Xie, S.; Sanei, S. 2009. K-hyperline clustering learning for sparse component analysis, *Signal Processing* 89: 1011–1022. doi:10.1016/j.sigpro.2008.12.005

Makino, S.; Lee, T-W.; Sawada, H. (Eds.) 2007. *Blind Speech Separation*. Springer-Verlag, Berlin etc.

Ouchi, H.; Hamada, N. 2009. Separation of Speech Mixture by Time-Frequency Masking Utilizing Sound Harmonics, *Journal of Signal Processing* 13(4): 331–334.

Rickard, S. 2007. The DUET Blind Source Separation Algorithm, in Makino, S., *et al*. (Eds.). *Blind Speech Separation.* Springer, 2007, 217–237.

Yilmaz, O.; Rickard, S. 2004. Blind Separation of Speech Mixtures via Time-Frequency Masking, *IEEE Trans. on Signal Processing* 52(7): 1830–1847. doi:10.1109/TSP.2004.828896

KALBĖTOJO APTIKIMAS IR ŠNEKOS IŠSKYRIMAS DVIEJŲ SIGNALŲ MIŠINIUOSE SU AIDU

W. Kasprzak, N. Ding, N. Hamada

Santrauka

Straipsnyje nagrinėjamas aklasis signalų šaltinių išskyrimas apdorojant signalų mišinius su aido efektu ar be jo. Detaliai pristatomi matematiškai bei eksperimentų su dirbtiniais ir realiais šnekos duomenimis rezultatais pagrindžiami du esminiai šio metodo patobulinimai. Pirmasis patobulinimas leidžia sumažinti vėlinimo žemuose dažniuose įtaką šnekos signalo išskyrimo klaidai. Antrasis patobulinimas, paremtas kalbėtojo pagrindinio dažnio sekimu, leidžia algoritmui išnaudoti tas pačias dažnių sritis skirtingiems signalų šaltiniams išskirti.

Reikšminiai žodžiai: šaltinių aklasis atskyrimas, histogramos klasterizavimas, spektrogramos analizė, kalbos rekonstravimas, maskavimas laiko ir dažnių skalėje.