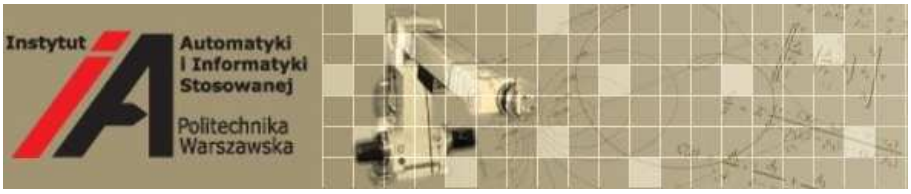


Blind localization and separation of two speakers based on two mixtures

Włodzimierz KASPRZAK

Institute of Control and Computation Engineering

Warsaw University of Technology, Poland

Ning DING and **Nozomu HAMADA**

Signal Processing Lab., School of Integrated Design Engineering,

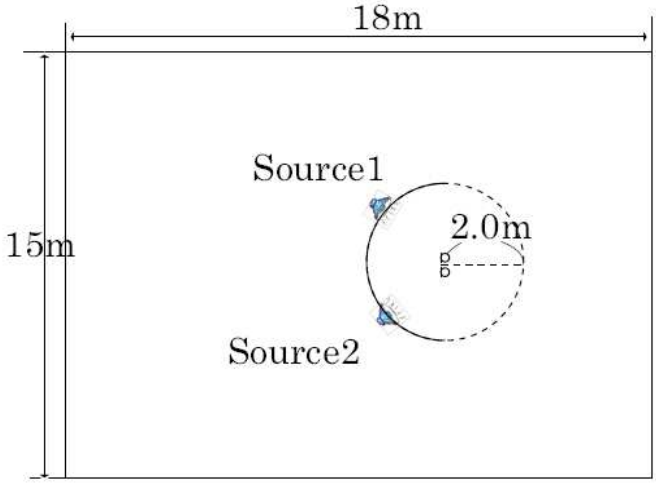
Keio University, Japan

INTRODUCTION

Problem: blindly to separate speech mixtures having two microphones.

Approach:

1. assume orientation and/or location difference of sources,
2. assume WDO (W disjoint orthogonally) of sources.



INTRODUCTION

Existing methods: DUET, TIFROM, DEMIX, etc.,

1. histogram analysis in the attenuation-time delay-space and
2. time-frequency masking controlled by estimated delay peaks to reconstruct the sources.

Problems:

1. well for anechoic mixtures but not echoic mixtures;
2. WDO (W disjoint orthogonal) assumption not satisfied (especially in the low frequency band)

Our solution:

1. a restrictive mask for phase delays on the basis of local and global energy distribution analysis in T-F domain (spectrogram),
2. the WDO assumption is relaxed by allowing some frequency bins to be shared by both sources. The mask creation is supported by exploring harmonics of fundamental frequencies.

1. THE BSS PROBLEM

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

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

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:

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

1. THE BSS PROBLEM

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

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

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

where $H_{jk}(f)$ is the frequency response, and $S_k(t,f)$ is a frequency domain time-series source signal.

It is assumed that in time-frequency domain, signals have the property of **sparseness**, i.e.:

$$S_1\ (t,f) \cdot S_2\ (t,f) \approx 0 \quad \forall (t,f)$$

2. T-F MASKING APPROACHES

Binary mask detection

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 STFT.

The binary mask approach depends strongly on the clustering quality of given feature, so the selection of an appropriate feature is essential in every T-F masking approach to blind source separation.

T-F masking approaches utilize the **delay** calculated from the **phase difference** between observations:

- **DUET**: a power weighted two-dimensional (2-D) histogram constructed from the ratio of the time-frequency representations of the mixtures, which is shown to have one peak for each source with peak location corresponding to the relative attenuation and delay mixing parameters.

2. T-F MASKING APPROACHES

- **SAFIA** : differences in the amplitude and phase between channels are calculated as in DUET. These features are used to select frequency components of the signal that comes from the desired direction and to reconstruct these components as the desired source signal.
- **MENUET**: a normalization and clustering of the level ratios and phase differences between multiple observations.
- In the **HS** method it is proposed to use harmonic structure as the clustering feature. To estimate the harmonic structure, the proposed method estimates the fundamental frequency using the initial separation results.

.

3. THE 2 MIXTURE APPROACH

We focus particularly on a situation where the number of sources $N = 2$, and the number of sensors $M = 2$.
Our approach can be classified as a time-frequency masking method based on ***difference of arrival time*** DOA.

1) Delay calculation

The anechoic mixing process can be expressed as

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

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

3. THE 2 MIXTURE APPROACH

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

The delay δ_i is obtained using a phase correlation function:

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

where ϕ(t,f) is the phase difference,

$$\phi(t,f) = \angle X_1(t,f) - \angle X_2(t,f)$$

3. THE 2 MIXTURE APPROACH

2) Delay histogram

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

3. THE 2 MIXTURE APPROACH

3) T-F masks for source reconstruction

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 binary masks are generated by

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

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

3. THE 2 MIXTURE APPROACH

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:

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

4) ISFFT

Finally, by using the Inverse Short Time Fourier Transform (ISTFT), the separated signals are transformed in time domain.

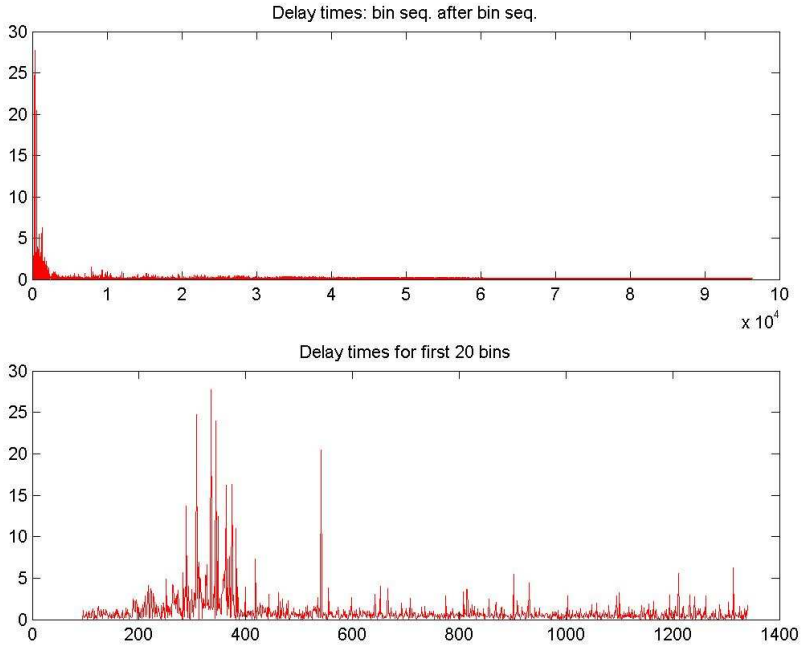
4. ANALYSIS

1) Quality of data

- Phase difference errors seem unavoidable.
- Under echoic mixtures – how to detect the number of sources?
- Do we need a cut-off frequency ?

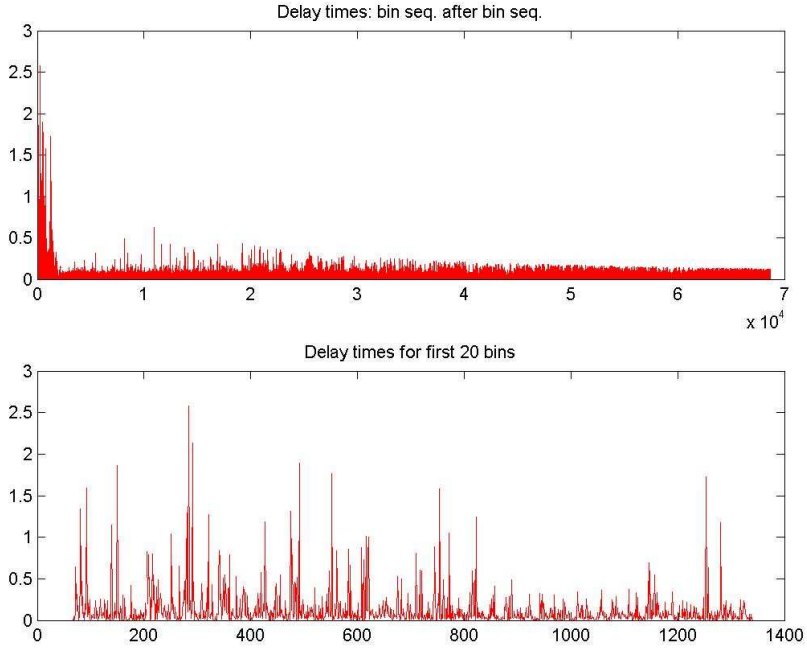
4. ANALYSIS

Delay times (real mixture)



4. ANALYSIS

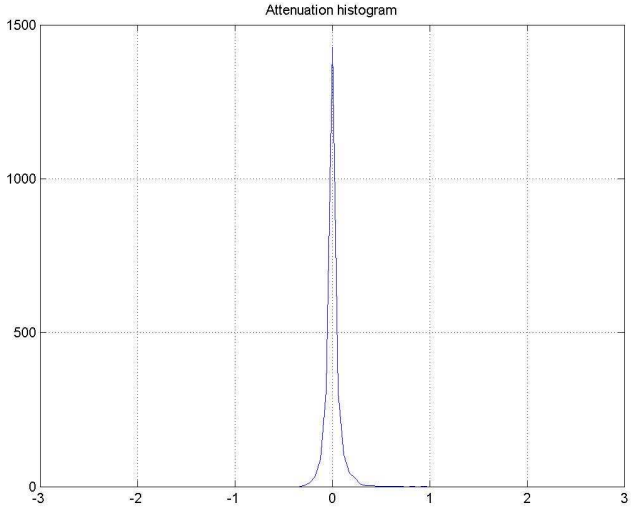
Delay times (simulated mixture)

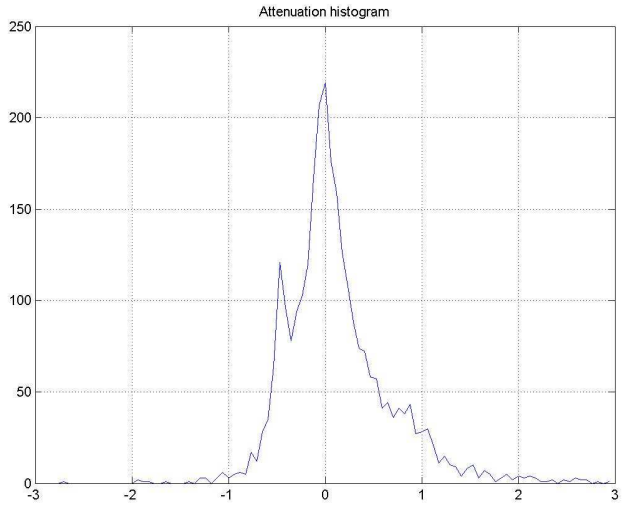


4. ANALYSIS

2) 1-D or 2-D histogram (DUET)

Histograms of the attenuation ratio





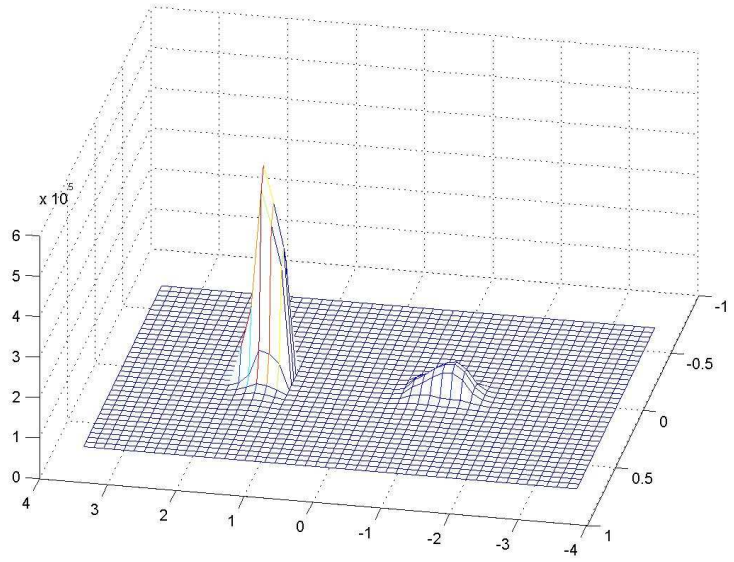
simulated (normalized amplitude)

real data (second peak for echo)

4. ANALYSIS

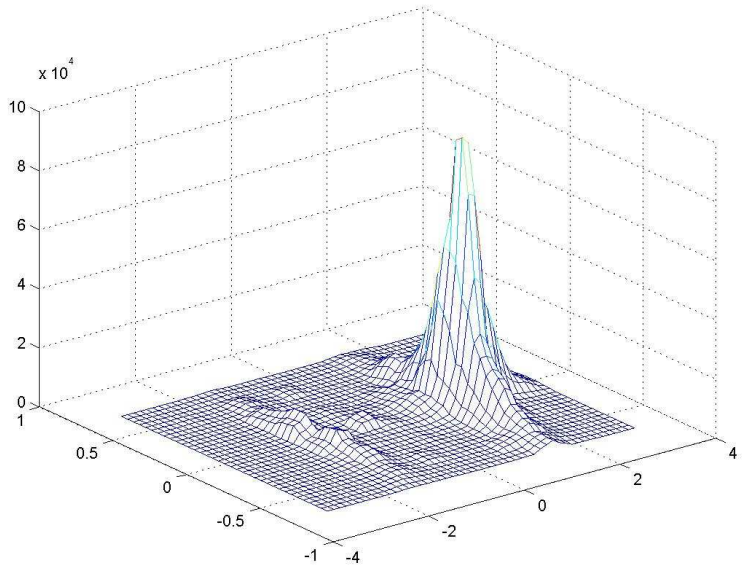
Test: two sources at 50^o and 60^o

2-D histogram in the DUET method: the second peak is an error.



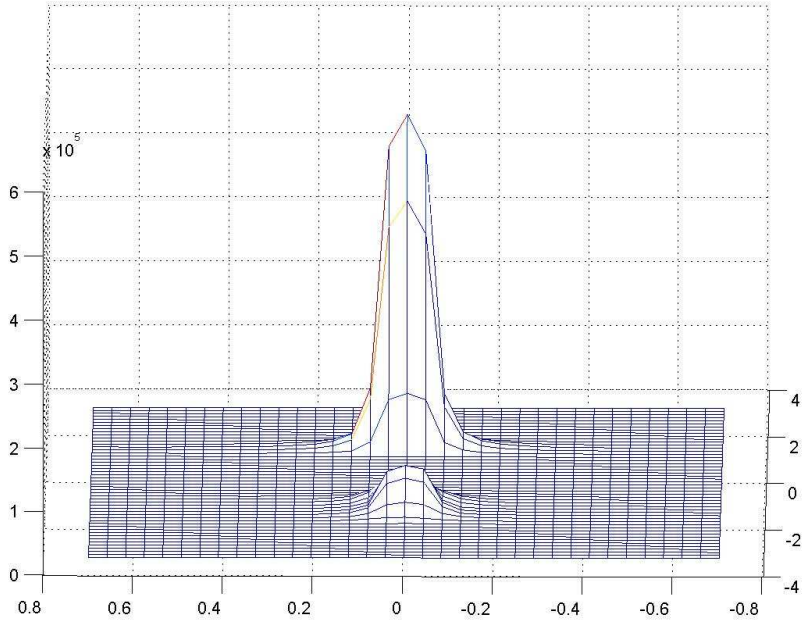
(a) for simulated mixture

4. ANALYSIS



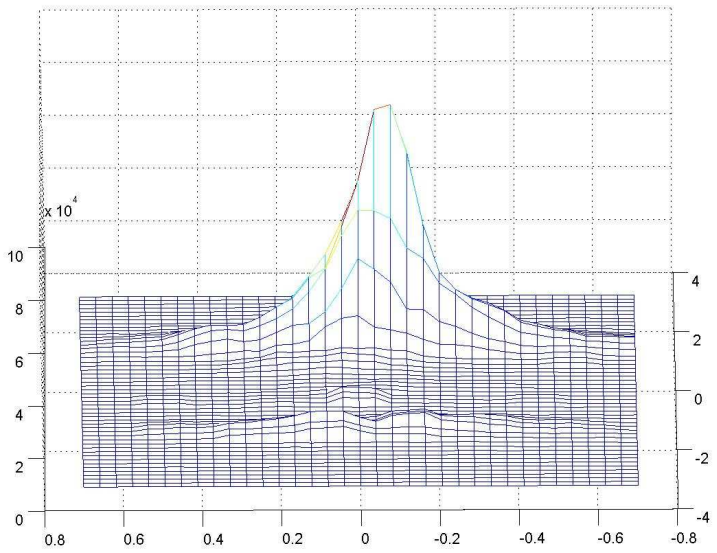
(b) for real mixtures

4. ANALYSIS



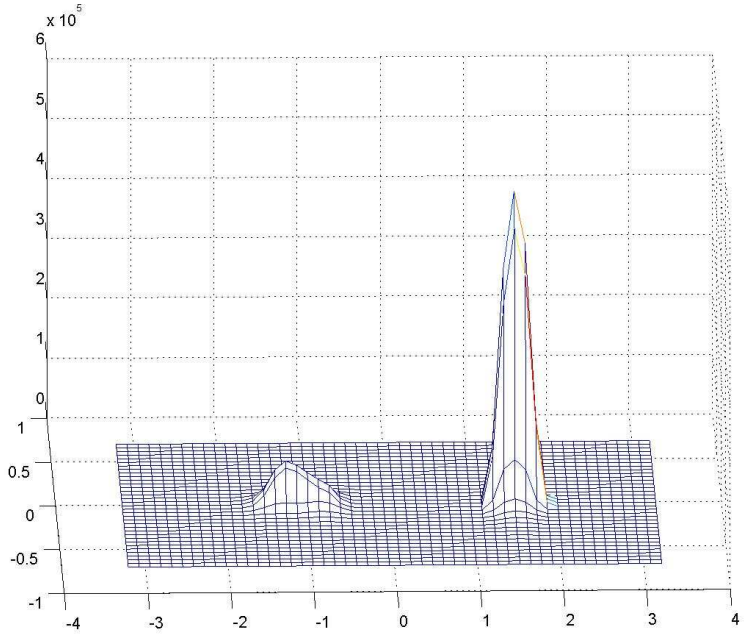
(c) projection onto attenuation axis – simulated case

4. ANALYSIS



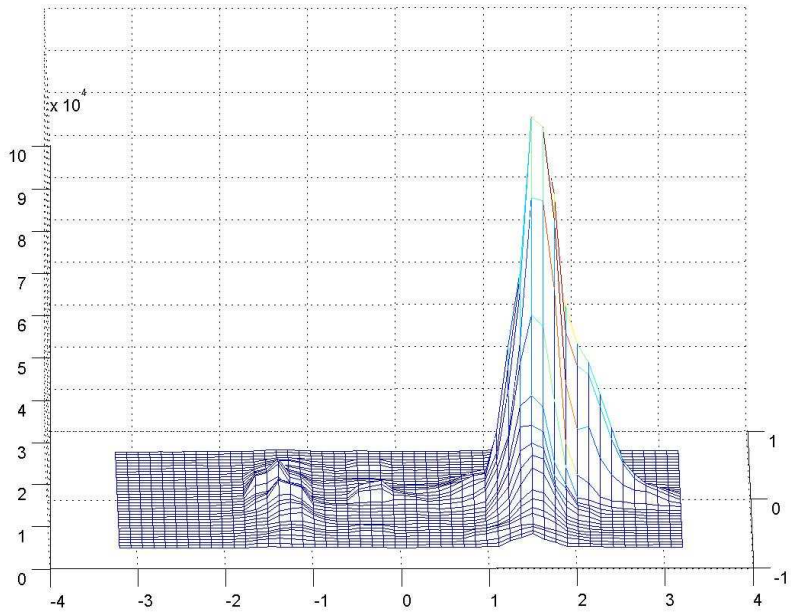
(d) projection onto attenuation axis – real case

4. ANALYSIS



(e) projection onto delay axis – simulated case

4. ANALYSIS



(f) projection onto delay axis – real case

4. ANALYSIS

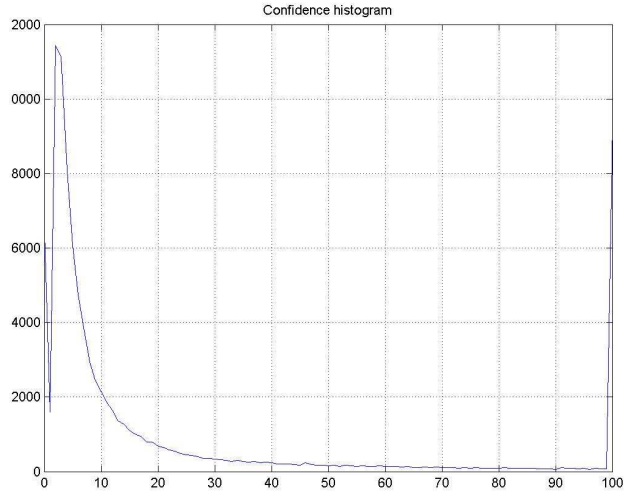
Conclusion: before extending the 1-D to 2-D histogram provide a better differentiation along the first axis (delay time or ?)

3) **Confidence measure** (TIFROM, DEMIX)

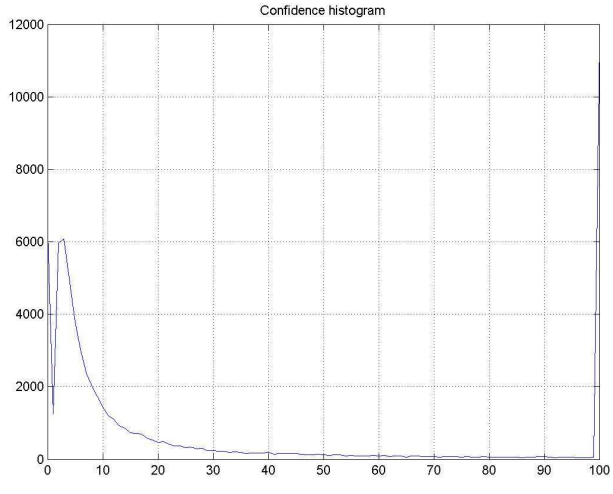
TIFROM and DEMIX use a "confidence measure" to select elements of the T-F signal (mixture) representation, which are with high probability "produced" by a single source only. The "confidence" is based on multiple PCA analysis in the attenuation-delay space for samples coming from the local neighborhood (say 3x3) 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.

4. ANALYSIS

In DEMIX only highly confident elements are considered (with confidence value > 90).



(a) for simulated data

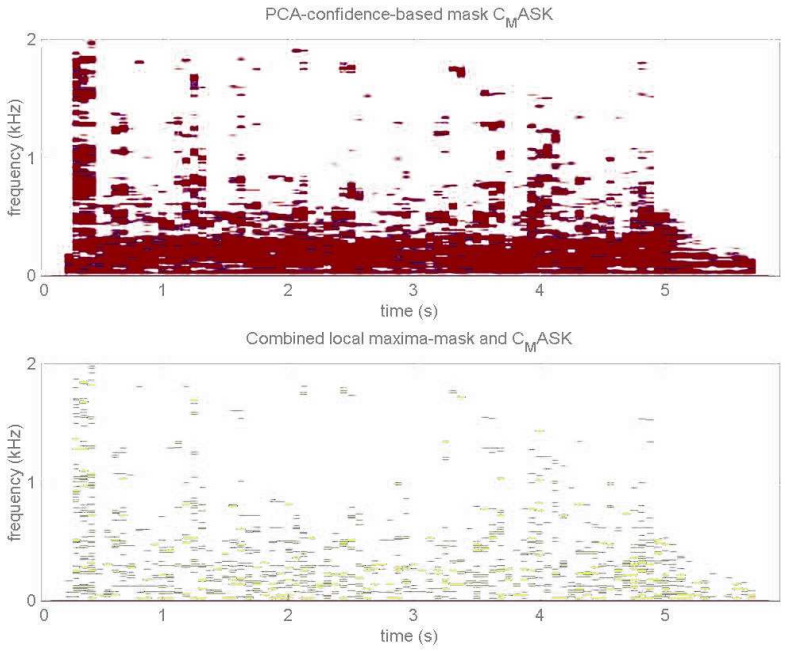


(b) for real data

4. ANALYSIS

But still high energy, low-frequency elements remain (top) – the delay information at low frequency bins is deteriorated by large errors.

Conclusion: a selection scheme is needed to concentrate on the relatively error-free information (bottom).



5. SOLUTION AND EXPERIMENTS

1) **Orientation** instead of delay time histogram

$$\text{Phase difference} : \phi(k,l) = \arg \frac{X_1(k,l)}{X_2(k,l)}$$

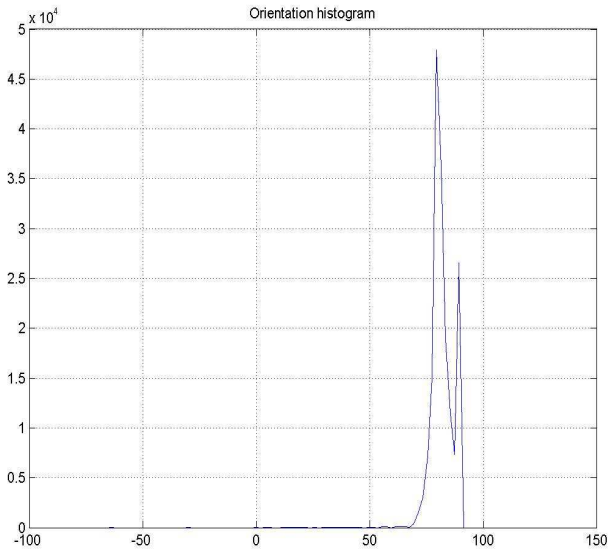
$$\text{Delay time} : \tau(k,l) = \phi(k,l) \frac{L}{2\pi f_s}$$

$$\tau(\theta) = \frac{d}{c} \sin(\theta)$$

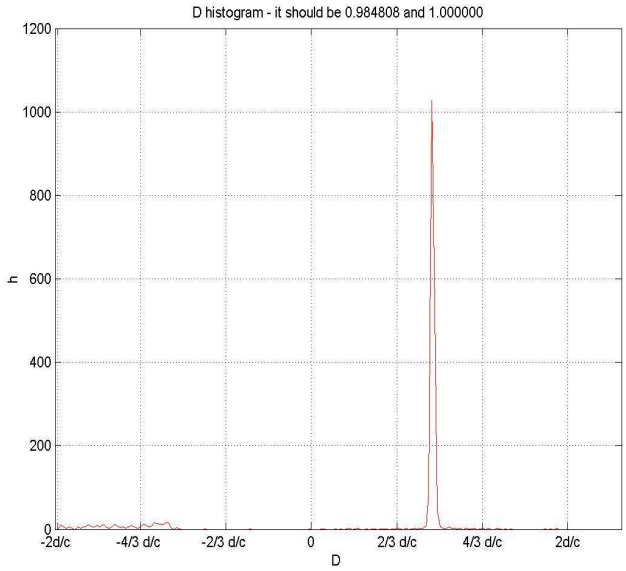
$$\text{Orientation angle} : \theta(k,l) = \arcsin(\tau(k,l)\dot{c}/d)$$

A difficult case in T-F based speech separation: 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.

5. SOLUTION AND EXPERIMENTS



(a) orientation is able to separate



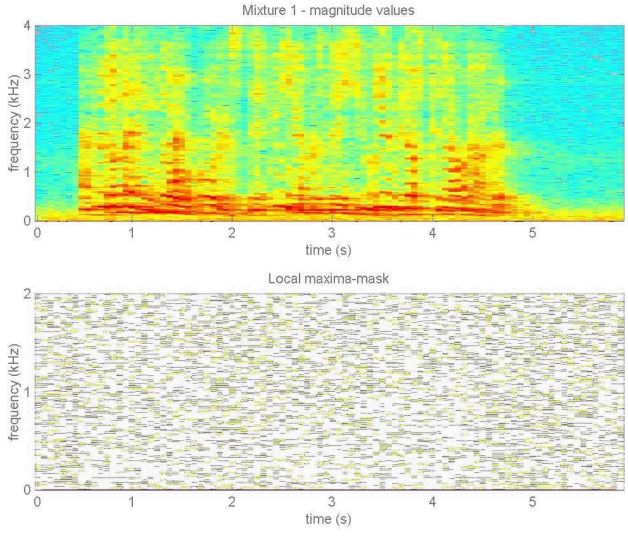
(b) delay – no separation possible

5. SOLUTION AND EXPERIMENTS

2) Energy-based selection of histogram information

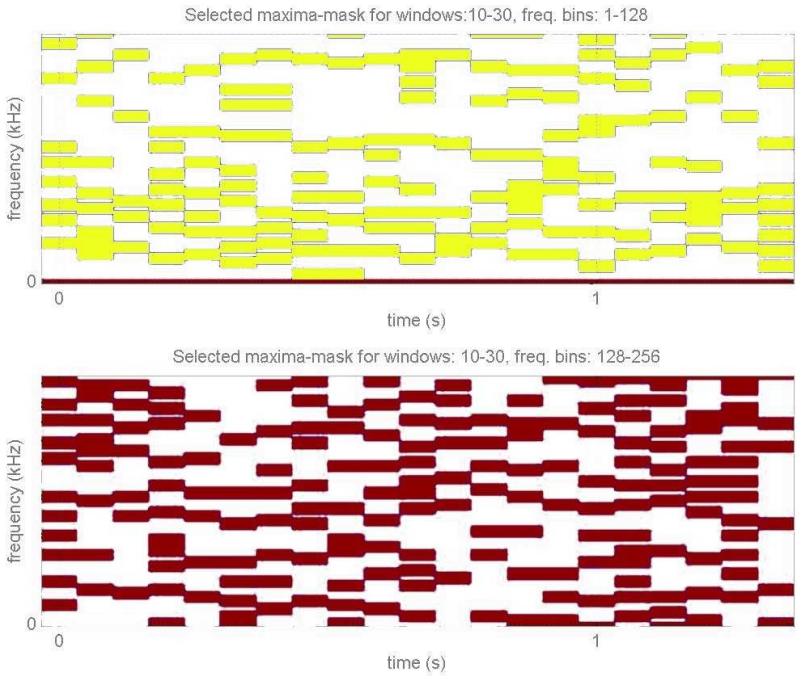
A restrictive cell selection procedure, in which two criteria are jointly used:

1. **local maxima** along each **frequency**-indexed column,
2. **near global** maximum cells along the **time axis** for each frequency bin.

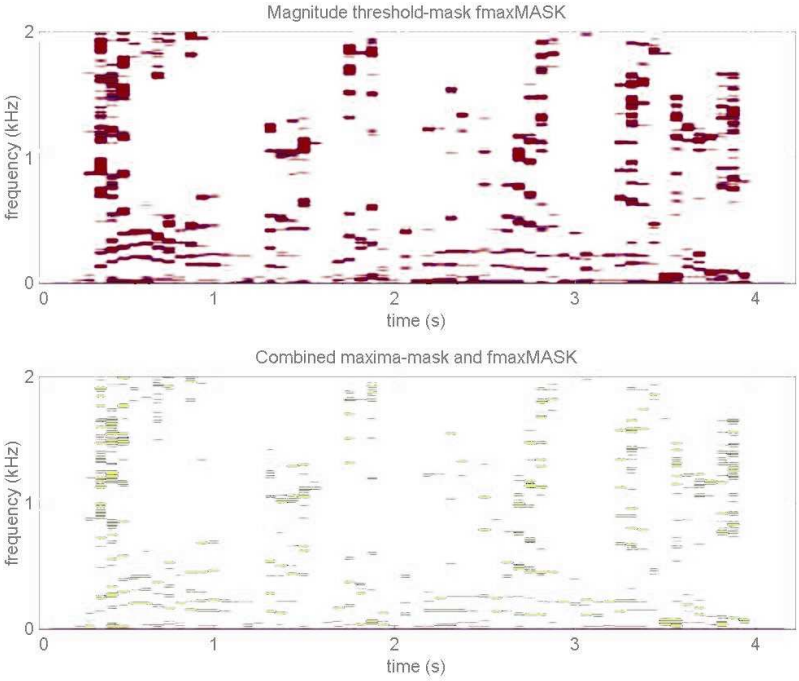


5. SOLUTION AND EXPERIMENTS

Local maximum-mask:

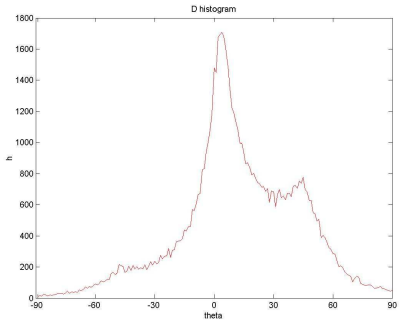


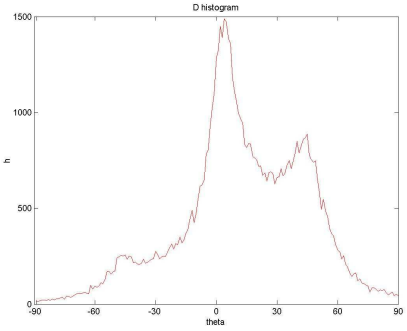
5. SOLUTION AND EXPERIMENTS

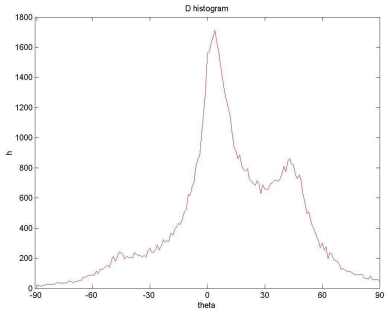


5. SOLUTION AND EXPERIMENTS

Histograms for selected T-F cells:



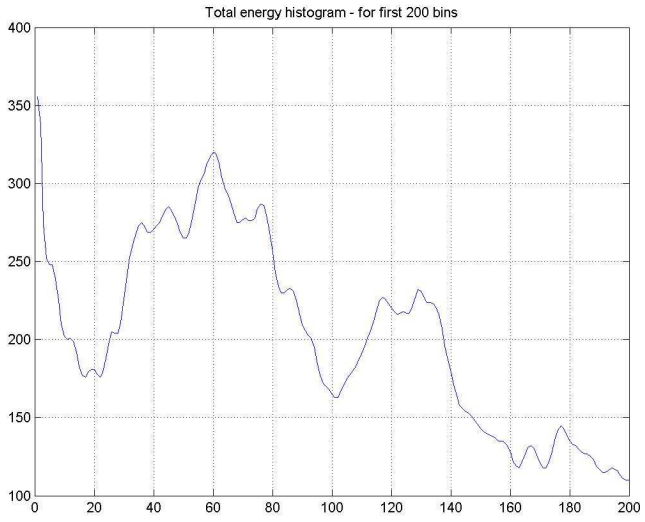


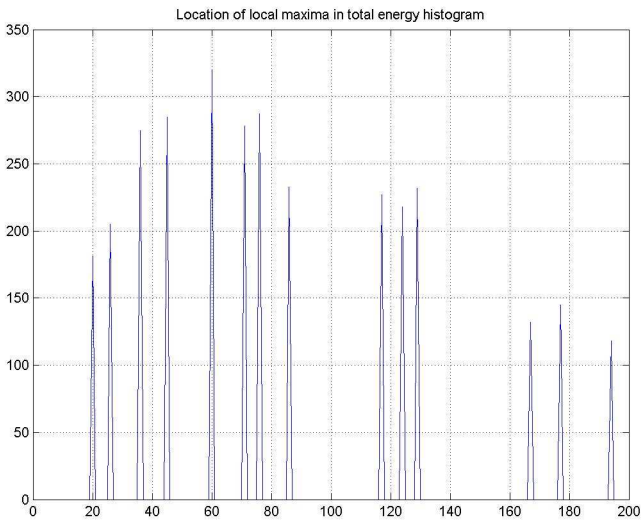


5. SOLUTION AND EXPERIMENTS

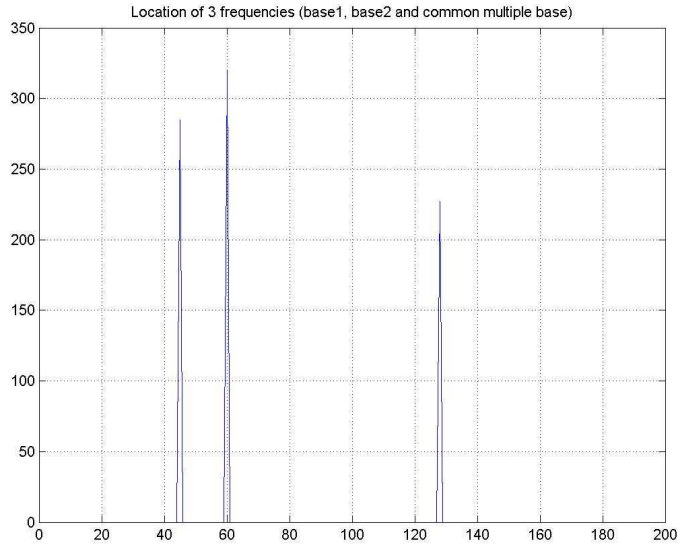
3) **Source mask** generation

Fundamental frequencies of speakers:





5. SOLUTION AND EXPERIMENTS



5. SOLUTION AND EXPERIMENTS

Modification of masking: 3-valued masks

$$M_1(t,f) = \begin{cases} 1 & if(d1 >= max\_distance)\&|f - f1\_harmonic| < max\_freq\_distance \\ 0.5 & \text{if } d1 \text{ and } f1\_harmonic \text{ do not match} \\ & \text{and neither d2 nor f2\_harmonic} \\ 0 & \text{if second mask} M_2(t,f) = 1 \end{cases}$$

$$M_2(t,f) = \begin{cases} 1 & if(d2 >= max\_distance)\&|f - f2\_harmonic| < max\_freq\_distance \\ 0.5 & \text{if d2 and f2\_harmonic do not match} \\ & \text{and neither d1 nor f1\_harmonic} \\ 0 & \text{if first mask} M_1(t,f) = 1 \end{cases}$$

6. RESULTS

Experimental setup:

Sampling frequency	$f_0 = 8000Hz$
Microphone distance	d = 40mm
Sound velocity	c = 340m/s
Window type	Hamming
STFT frame length	$L = 1024$
Frame overlap	$\Delta = 512$

Real acquired mixtures

1st: pair: 10 deg - 20 deg

theta1 = 13.5; theta2 = 21.9

6. RESULTS

corrBetweenRecons = (in perfect case should be 0.0) 0.2129

corrRecAndS = in perfect case should be

[1 0]

[0 1]

0.6856 0.2946

0.3982 0.6699

2nd: 10 deg - 30 deg

theta1 = 14.9; theta2 = 29.4

corrRecAndS =

0.7394 0.2746

0.3512 0.6801

6. RESULTS

3th: 10 deg - 40 deg
theta1 = 13.5; theta2 = 40.8
corrBetweenRecons = 0.1408
corrRecAndS =
0.8104 0.1486
0.2883 0.8109

4th : 10 deg - 50 deg
theta1 = 13.5; theta2 = 46.4
corrBetweenRecons = 0.1498
corrRecAndS =
0.7607 0.1633
0.3812 0.7820
---

6. RESULTS

5th: 10 deg - 60 deg
theta1 = 13.5; theta2 = 54.8
corrBetweenRecons = 0.1084
corrRecAndS =
0.6623 0.1605
0.4550 0.7506

6th: 10 deg - 70 deg
theta1 = 12.1; theta2 = 73.2
corrBetweenRecons = 0.0978
corrRecAndS =
0.6435 0.1457
0.4870 0.7580

6. RESULTS

Simulated mixtures

1st: 10 deg – 20 deg

theta1 = 10.8; theta2 = 20.5

corrRecAndS = correlation between reconstructed signals and sources

	s1	s2
s1:	0.9760	0.0694
s2:	0.0106	0.9747

2nd: 10 deg – 30 deg

theta1 = 10.8; theta2 = 29.4

corrBetweenRecons = 0.0325

corrRecAndS =

0.9778	0.0563
0.0142	0.9759

6. RESULTS

3th: 10 deg - 40 deg

theta1 = 9.4; theta2 = 40.8

corrBetweenRecons = 0.0308

corrRecAndS =

0.9793	0.0376
0.0255	0.9778

4th: 10deg - 50 deg

theta1 = 9.4; theta2 = 50.4

corrBetweenRecons = 0.0319

corrRecAndS1 =

0.9790	0.0480
0.0178	0.9756

6. RESULTS

5th: 10 deg - 60 deg

theta1 = 9.4; theta2 = 59.8

corrBetweenRecons = 0.0322

corrRecAndS =

0.9787 0.0505

0.0166 0.9748

6th: 10 deg - 70 deg

theta1 = 9.4; theta2 = 69.1

corrBetweenRecons = 0.0335

corrRecAndS =

0.9786 0.0538

0.0148 0.9745

6. RESULTS

7th: 10 deg - 80 deg

theta1 = 9.4; theta2 = 78.7

corrBetweenRecons = 0.0319

corrRecAndS =

0.9788 0.0478

0.0182 0.9746

8th: 10 deg - 90 deg

theta1 = 9.4; theta2 = 90.00

corrBetweenRecons = 0.0319

corrRecAndS =

0.9790 0.0452

0.0194 0.9755

---

6. RESULTS

For a total performance evaluation, we use WDO (measure of W-disjoint orthogonal). It is computed from two other criteria:

- PSR (the Preserved-Signal Ratio) and
- SIR (the Signal-to-Interference Ratio) defined as,

$$WDO = \frac{||M(t,f)S_d(t,f)||^2 - ||M(t,f)S_i(t,f)||^2}{||S_d(t,f)||^2} = PSR - \frac{PSR}{SIR}$$

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

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

6. RESULTS

Where S_d(t,f) is the desired signal, M(t,f) is the binary mask, and S_i(t,f) is the interfering signal. Usually, $0 \leq WDO \leq 1$. For ideal separation: WDO=1.

WDO

Women at: Men at	60°	70°	8°	90°
50 °	0.9264	0.9109	0.9004	0.8808
60 °		0.9073	0.9020	0.8807
70 °			0.8166	0.6876
80 °				0.4491

7. CONCLUSION

- Several improvements to the Time-Frequency masking approach to blind speech separation, that **relax** the strict **WDO** assumption (in the two microphone-case).

- 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 **relative energy threshold** along the time axis (for each particular frequency bin).

- The **delay** feature is only responsible for selecting cells with obviously **perfect** behaviour. Otherwise the **harmonic frequencies** are applied as a new selection criterion.

7. CONCLUSION

REFERENCES

1. Makino S, Te-Won Lee, Sawada H.: Blind Speech Separation, Springer-Verlag, 1997. Rickard S.: The DUET Blind Source Separation Algorithm, pages 217-237.
2. 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.
3. 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.
4. Arberet S, Gribonval R, Bimbot F. A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture, IEEE Transactions on Signal Processing, 2008/2009.