Auditory scene analysis by time-delay analysis with three microphones

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

Włodzimierz Kasprzak
Paweł Przybysz
Institute of Control and Computation Eng. Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
W.Kasprzak@elka.pw.edu.pl, P.Przybysz.2@elka.pw.edu.pl

Abstract

We propose two methods for the disambiguation of results in time-delay based detection and localization of sound sources, when a triangle of microphones is applied for signal acquisition. A standard approach is to create histograms of time differences of arrival (TDOA) for each microphone pair in a triangular array and to create an averaged histogram. But each individual histogram is designed to detect unique orientation of source only within the local range of $[-\pi/2, \pi/2]$. *Hence, taking the average for different pairs is not appropriate and such method suffers from ambiguity of results in the full range of orientations:* $[0, 2\pi]$. *Our first proposition is a delay vector transformation method, that combines corresponding delay measurements into vectors and transforms them into a 2-D space in which a full-range orientation histogram can finally be established and analyzed. In our second method, individual orientation histograms obtained for pairs of microphones are analyzed first and for each detected source two competitive hypotheses are created. Due to a final clustering of the hypothesis set a unique orientation of each source can be estimated.*

1. Introduction

Sound source localization and computational auditory scene analysis (CASA) have potential applications such as tele-conference and distributed meeting systems, hands-free automatic voice recognition systems and mobile service robots [1], [2].

A standard approach for dominating-sound-source of multi-sound-source localization is to use an array of microphones [3]. Early methods concentrated on measuring the time differences of arrival (TDOA) to two sensors directly in the time domain, using the generalized cross-correlation or related methods [4], [5]. However, the resolution capability of this method is limited, which makes it impractical for use in some applications, like in multi-speaker localization.

Current most basic approach to find the direction of a speech source is to estimate the phase difference between two sensors in the frequency domain [6]. Several novel algorithms have been developed recently, such as TIFROM (TIme-Frequency Ratio Of Mixtures) [7], DEMIX [8] and uniform clustering [9], that try to overcome some weak points of basic algorithms. These improvements focus on making more efficient clustering in the attenuation-rate and delay-time spaces. In case of music and speech sources the specific signal characteristics can be explored, e.g. the harmonic signal structure leads to a clustering feature [10], [11].

Other recent research line is to provide proper microphone arrangements, e.g. an array of microphones [12] or multi-array sets [13], in order to implement a true 3-D localization ability. For example, in [13] the multi-array consisted of 8 omni-directional sensors organized in two squares, that were located on mutually orthogonal planes (e.g. two walls in a room). Two types of experiments have been conducted - with 4 microphone pairs (2 diagonal pairs in every square) or 12 microphone pairs (by taking the all possible pairs within a square for two squares). Average performance of detecting a single dominating source have been reported to be only slightly better for 12 pairs if compared to the 4-pair case. In fact, using a triangle of microphones (3 difference measurements for 3 pairs) we can already obtain robust results similar to a single square of 4 microphones.

With regard to a triangle of microphones the *histogram mapping* (HM) method was proposed in [14]. The HM method estimates histograms of phase differences of sensor signals for each microphone pair in a triangular array and averages them to make a combined histogram search. But each individual histogram is designed to detect unique orientation of source only within the local

range of $[-\pi/2, \pi/2]$. Hence, taking the average for different pairs is not appropriate and such method suffers from ambiguity of results in the full range of orientations: $[0, 2\pi]$. Additionally, as the estimation accuracy at one microphone pair differs from that at the other two pairs, some weighted averaging of the individual estimates obtained for these different pairs is needed.

Obviously, one can force the sources always to be "in front" of the microphones. This is even possible in stationary arrangements in in-door conditions. But with a mobile robot moving in an in-door or eventually also in an out-door environment the sound sources can really originate from all the places surrounding the robot.

In this paper, we propose improvements of the basic method of histogram mapping, based on signals coming from a triangle of microphones, as applied for the detection and localization of sound sources. In the first approach, called the *delay vector transformation* approach [15], a 3D time delay vector is created for a microphone triangle, and by observing obvious constraints it is transformed into a 1-D orientation histogram analysis. In the second approach, instead of averaging the three individual time-delay histograms, like it is done in the HM method, we apply orientation histogram analysis for each microphone pair separately, then generate from them source orientation hypotheses (two hypotheses for each detected source) and finally resolve the inherent hypothesis ambiguities by applying a clustering process.

We also propose to replace the standard time-delay (or phase-difference) histogram analysis by the orientation histogram analysis. It is known that the estimation of the TDOA at a microphone pair degrades when the direction of the speech source deviates from the broadside direction, e.g. [16], [11]. Thus, by dealing with the orientation histogram, we shall avoid this endpoint degradation problem.

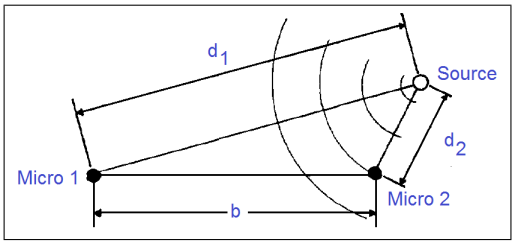
In next section, the principle of the TDOA-based source localization approaches are introduced. In section 3, the particular histogram mapping method (for a triangle of microphones) is described and the ambiguity problem of this method is identified. In section 4, the delay vector transformation method is proposed to avoid the ambiguity of the basic method. In section 5, the orientation ambiguity problem is solved by a so called *source voting* method. Individual orientation histograms obtained for pairs of microphones are analyzed first and for each detected source two competitive hypotheses are created. Due to a final clustering of the hypothesis set a unique orientation of each source can be estimated. Some experimental results, that verify the ability of both methods to resolve source detection ambiguity, complete the paper.

2. The TDOA-based source localization

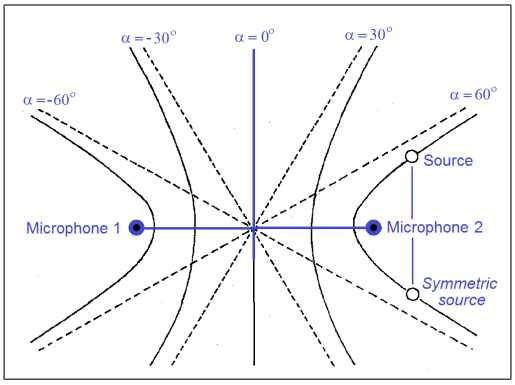
The basic principle of audio signal detection based on the time delay between two microphones is explained in Figure 1. For a single active sources, the peak value of the cross correlation of two sensor signals in the time domain occurs at the time delay:

$$D_{1,2} = (d_1 - d_2)/v, \tag{1}$$

where $d = d_1 - d_2$ is the path length difference and v is the propagation velocity of the source signal $s(t)$. Knowing v, the delay value $D_{1,2}$ defines a hyperbolic surface on which the source must lie [4]. For sources that are sufficiently distant from the microphone pair (with respect to the base distance between two microphones), one can approximately assume the orientation angle has reached the border value (e.g. 60^o or -60^o for source 1 in Figure 1).



(a)



(b)

Figure 1. The principle of TDOA-based source localization with two microphones: (a) one sound source, (b) two symmetric directions per source are detected, corresponding to a hyperbolic surface of possible locations

To be more specific, let us refer the experimental results given in [5]. The experimental setup included: the base distance, $b = 14$ cm and the sampling rate, $f_s = 64$ kHz. When no echo effects deteriorate the measurements, it was found that with source distances of $1 - 3$ m (i.e. $7b - 21b$) the average detection error was below 6^o. In the modern methods of multi-speaker detection and localization, the phase difference is measured in the frequency domain, rather than the time domain differences. Then the base distance is kept at $4 - 8$ cm and the sampling rate is $8 - 16$ kHz. Hence, the expected orientation approximation error will be below 6^o already at small source distances starting from 28 cm.

Unfortunately, there is an ambiguity inherited in this approach. From the principle in equation (1) and the Figure 1(b) it is visible, that two "symmetric" orientations are equally probable for a single source. Obviously, one can force the sources always to be "in front" of the microphones. This is even possible in stationary arrangements in in-door conditions. But with a mobile robot moving in an in-door- or eventually also in an out-door-environment, the sound sources can really originate from all the places surrounding the robot. Hence, in such applications we need to resolve this ambiguity problem.

3. The triangle of microphones

Let us assume a triangular microphone array as shown in Figure 2. There are three microphones located at the vertices of an equilateral triangle with size distance equal to d. There exist speech and sound signals s_k, ($k = 1, 2, ..., K$), each located at a different direction θ_k with respect to the microphone coordinate system. Here, K represents the number of sound sources that may be simultaneously active. For simplification, we restrict the discussion to a two sources case, without loss of generality.

3.1. Mixture of sources

Each mixture of sources, acquired by a particular microphone, is first transformed by a short-time Fourier transform (STFT). The value of d limits the interesting frequency bandwidth to $[0, f_{max}]$, so that the phase shift between every two microphones is less than π. Then the spectrogram image (i.e. the magnitudes of Fourier coefficients) of every mixed signal is virtually the same and the differences are only due to the phase information.

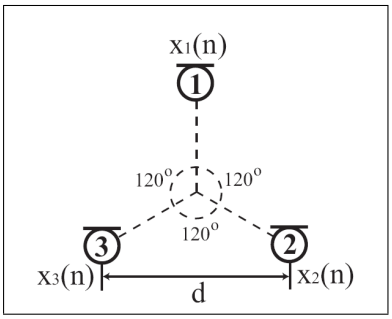


Figure 2. The arrangement of microphones.

3.2. WDO

One can assume, that different speech signals only rarely overlap in the time-frequency domain. In the ideal case, the W-disjoint orthogonality (WDO) property is satisfied [6]: a single-frequency bin of the input signal at any frame contains the spectral component of only one speech signal. Hence, in time-frequency domain the signals have the property of sparseness, i.e.

$$S_i\,(t,f) \cdot S_j\,(t,f) \approx 0 \quad , \quad \forall i \neq j, \forall (t,f) \qquad (2)$$

Assuming this WDO property, the many-source localization problem may be transformed into a set of narrowband, single-source problems.

3.3. Delay time calculation

For every pair of microphones (let us denote them here as 1 and 2) the anechoic mixing process can be expressed as

$$\begin{bmatrix} X_1\,(t,f) \\ X_2\,(t,f) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ e^{-j\frac{2\pi f \delta_1}{L}} & e^{-j\frac{2\pi f \delta_2}{L}} \end{bmatrix} \begin{bmatrix} S_1\,(t,f) \\ S_2\,(t,f) \end{bmatrix}, \qquad (3)$$

where δ_i (i =1,2) is the delay between two microphones, and L is the number of STFT points. Assuming that microphone no. 1 provides the reference data, under the condition of WDO, the mixing model can be simplified to

$$\begin{bmatrix} X_1\,(t,f) \\ X_2\,(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ e^{-j\frac{2\pi f \delta_i}{L}} \end{bmatrix} S_i\,(t,f) \qquad (4)$$

The delay δ_i is related to a phase function:

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

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

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

3.4. The histogram mapping method

In the histogram mapping method, the orientation distributions $\theta(i,j)$, one for each microphone pair (i,j), are mapped into a histogram (Figure 3).

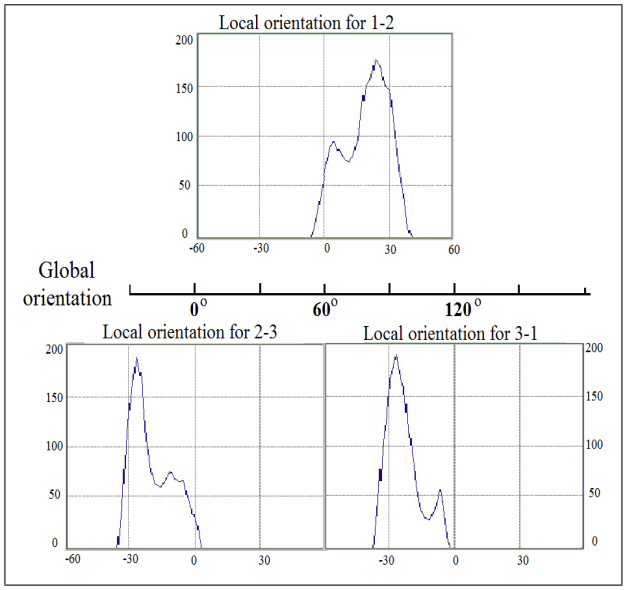


Figure 3. Example of histogram mapping with prior "front" constraint - three local orientation histograms.

The three individual histograms are eventually reversed and shifted in order to change the local orientations to the global orientation of the triangle (Figure 4).

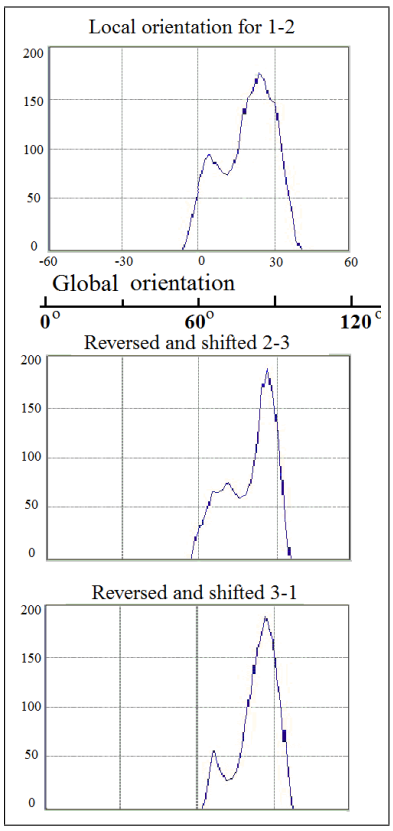


Figure 4. Example of histogram mapping with prior "front" constraint - histograms after order reversion and orientation shift.

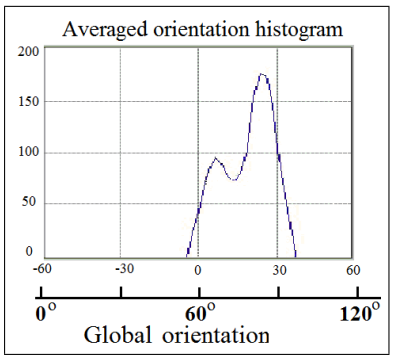


Figure 5. Example of histogram mapping - final averaged histogram.

Such normalized three individual histograms are added together and averaged (Figure 5). Eventually, the final histogram can be smoothed allowing to remove insignificant small peaks. Finally, the peaks are searched for in the averaged histogram. The number of peaks in the resultant histogram is equal to the number of active speakers, and the center of each peak is the estimated direction of a signal generated by an active source (e.g. a speaker).

In practice, this approach succeeds if the location of speakers can be in advance restricted to be in one of three areas: "in front" of microphone pair 1-2, or "in front" of microphone pair "2-3" or "in front" of microphone pair 3-1. Only then the ambiguity of individual histograms can be avoided, because in general any orientation solution established in the interval $[-90^o, 90^o]$ has an equally proper symmetric solution in the interval of $[90^o, 270^o]$. Without the above mentioned a priori knowledge the summation of histograms can not be applied. But this constitutes a severe constraint onto the use of all the three measurements. To be able to combine 3 histograms the sources should be located within of $[-30^o, 30^o]$ with respect to the local system of the "front" pair. If only two histograms need to be combined then this allowed interval is extended to $[-90^o, 30^o]$ or $[-30^o, 90^o]$. In the example in Figure 3 we know in advance that the sources, if any, are located "in front" of the microphone pair 1-2. Hence, the histogram for pair 1-2 is computed as usual, but the other two histograms have to be reversed, i.e. the order of microphones 2-3 changes to 3-2, and the order of 3-1 changes to 1-3. Due to this reversion the unknown sources are expected to be located "in front" of all the 3 microphone pairs.

3.5. The ambiguity in histogram mapping

If the time-delay (or orientation) histogram is computed for a single pair of microphones the useful interval of orientations is $[-90^o, 90^o]$ with respect to the normal-to-baseline vector. There is an inherent assumption that the source is "in front" of the two microphones. If the source is placed symmetrically (with respect to the baseline) "in the back part" of microphones, the same histogram data will appear.

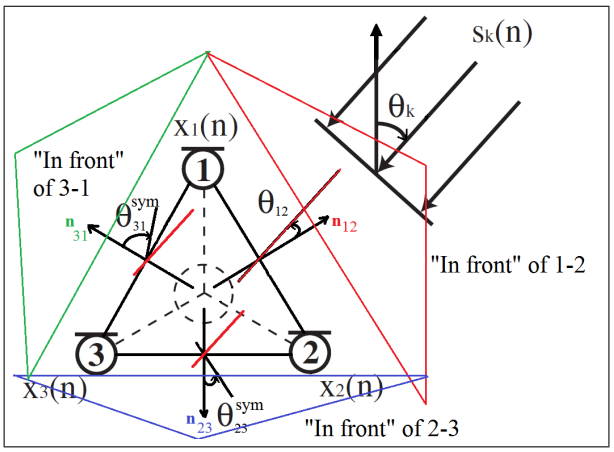


Figure 6. There are 3 regions representing locations "in front" of the corresponding microphone pair.

For the triangle microphone arrangement, in the standard approach we need to know in advance in which of the 3 "front" regions the source is located (Figure 6).

4. The delay vector transformation method

Let us define a TDOA vector, which consists of the TDOAs of the three microphone pairs, as:

$$\tau(\theta) = [\tau_{12}(\theta), \tau_{23}(\theta), \tau_{31}(\theta)]^T, \tag{7}$$

where the theoretical values of the particular components are given by

$$\tau_{12}(\theta) = \frac{d}{c}\sin\theta + \frac{2}{3}\pi \tag{8}$$

$$\tau_{23}(\theta) = \frac{d}{c}\sin\theta \tag{9}$$

$$\tau_{31}(\theta) = \frac{d}{c}\sin\theta - \frac{2}{3}\pi \tag{10}$$

It may be observed that for theoretical corresponding delay time values a linear equation is satisfied:

$$\tau_{12}(\theta) + \tau_{23}(\theta) + \tau_{31}(\theta) = 0 \tag{11}$$

There is also a constraint posed onto the second norm of the delay vector:

$$\tau_{12}^2(\theta) + \tau_{23}^2(\theta) + \tau_{31}^2(\theta) = \frac{3d^2}{2c^2} \tag{12}$$

This means that valid observations of vector $\tau(\theta)$, given in the 3D coordinate system spanned over delay times registered by microphone pairs, are located on a circle with radius $r(\theta) = \frac{d}{c}\sqrt{\frac{3}{2}}$ (Figure 7).

We can now define a mapping from this 3D space to a 2D space by using the following orthogonal transformation:

$$\boldsymbol{T} = [\boldsymbol{t_1}, \boldsymbol{t_2}, \boldsymbol{t_3}], \tag{13}$$

with the column vectors defined as:

$$t_1 = \left[-\sqrt{\frac{1}{6}}, \sqrt{\frac{2}{3}}, -\sqrt{\frac{1}{6}}\right] \tag{14}$$

$$t_2 = \left[-\sqrt{\frac{1}{2}}, 0, -\sqrt{\frac{1}{2}}\right] \tag{15}$$

$$t_3 = \left[\sqrt{\frac{1}{3}}, \sqrt{\frac{1}{3}}, \sqrt{\frac{1}{3}}\right] \tag{16}$$

Now every transformed delay vector takes the form:

$$\tau'(\theta) = T\tau(\theta) = \tag{17}$$

$$= \left[\frac{d}{c}\sqrt{\frac{3}{2}}\sin\theta, \frac{d}{c}\sqrt{\frac{3}{2}}\cos\theta, 0\right]^T. \tag{18}$$

Hence, the orientation is obtained as: $\theta = \arctan\left(\tau'_x/\tau'_y\right)$.

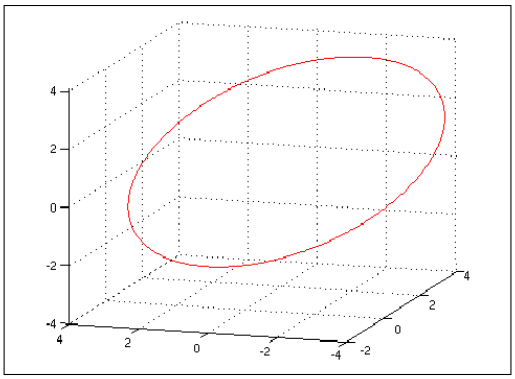


Figure 7. The theoretically correct delay vectors form a circle in the 3D delay space.

5. The source voting method

5.1. Selection of spectrogram data

It is already well recognized that particular spectrogram cell's provide delay data of different quality, as in practice the WDO principle is often violated. Here we propose to use a restrictive cell selection rule, considering two criteria:

1. select the local maxima along every column of the spectrogram (a distribution across frequencies) each frequency-indexed column,

2. select sufficiently large values along every row of the spectrogram - a maximum value in given row is first obtained and a threshold is set in proportion to this maximum.

The result of combining above two criteria is illustrated in Figure 8.

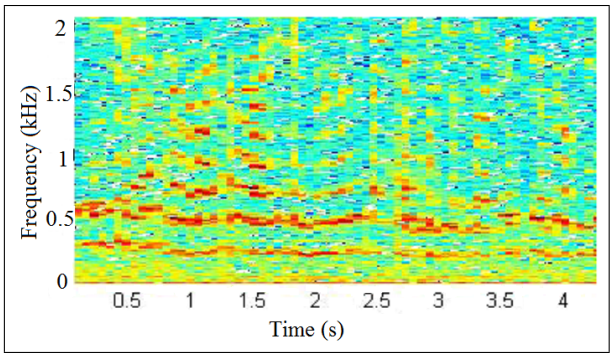
5.2. Orientation histogram

Another preliminary modification of a typical TDOA-based approach will be the use of an orientation histogram instead of analyzing directly the delay times or phase shifts. For the case of orientations, the histogram bins are linearly matching the angle scale, e.g. the difference of, say, 10 degrees corresponds to the same number of bins in cases when θ is nearly 90 degrees or near 0 degrees. In opposite, 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.

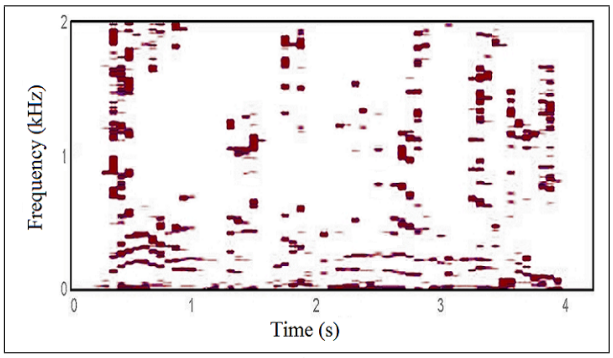
In fact, where two sources are located at the same distance of 2 m from the center of two microphones, we can write:

$$\theta(t, f) = \arcsin(\delta(t, f) \cdot c/d), \tag{19}$$

where c is the average speed of sound and d - the base distance between two microphones. The delay time $\delta(t, f)$ can be measured from the mixture spectrogram according

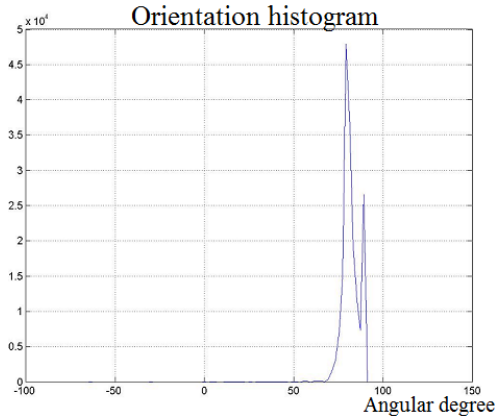


(a)

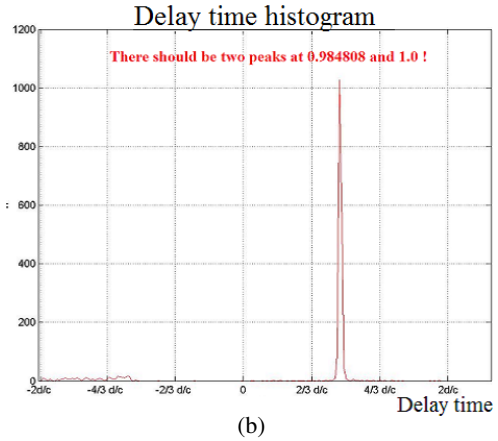


(b)

Figure 8. Illustration of the spectrogram selection step: a) example of a spectrogram of mixed signal, b) selected spectrogram elements that will be used for phase shift measurements.



(a)



(b)

Figure 9. A difficult case of two sources at orientations of 80 and 90 degrees: (a) the orientation histogram succeeds, (b) but the delay-time histogram fails

to equations (4) and (5). From (6) in turn we observe that the delay time is nonlinearly dependent on the orientation angle. We can write:

$$\delta(\theta) = \frac{d}{c}\sin(\theta) \tag{20}$$

Let us notice the Figure 9, 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.

5.3. The source localization algorithm

Let us consider again the problem illustrated in Figure 6. We assume that a single source is located at orientation θ, expressed in the global coordinate system (speaking more exactly - with respect to a vertical system axis). The histogram analysis performed independently for three microphone pairs will return single peaks inducing the orientations θ_{12}, θ_{23} and θ_{31}, specified in local coordinates of each microphone pair. In fact, every peak is ambiguous as it represents two locally symmetric solutions, θ_{ij} and θ_{ij}^{sym}:

$$\text{if } \theta_{ij} < 0 \text{ then } \theta_{ij}^{sym} = -180^o - \theta_{ij} < 0 \tag{21}$$

$$\text{if } \theta_{ij} > 0 \text{ then } \theta_{ij}^{sym} = 180^o - \theta_{ij} > 0 \tag{22}$$

for $i \neq j, i, j = 1, 2, 3$. Thus, six hypotheses for the unknown global θ can be created:

$$\theta(1) = \theta_{12} + 60^o, \text{or } \theta(2) = \theta_{12}^{sym} + 60^o \tag{23}$$

$$\theta(3) = \theta_{23} + 180^o, \text{or } \theta(4) = \theta_{23}^{sym} + 180^o \tag{24}$$

$$\theta(5) = \theta_{31} - 60^o, \text{or } \theta(6) = \theta_{31}^{sym} - 60^o \tag{25}$$

For example: $\theta(1) = 45^o$. Measurements at 1-2: location is "in front" - $\theta_{12} = -15^o$, $\rightarrow \theta_{12}^{sym} = 165^o \rightarrow \theta(2) = -125^o$.

Measurements at 2-3: location is "in the back" - $\theta_{23}^{sym} = -135^o$, $\rightarrow \theta_{23} = -45^o \rightarrow \theta(3) = 135^o$, $\theta(4) = 45^0$.

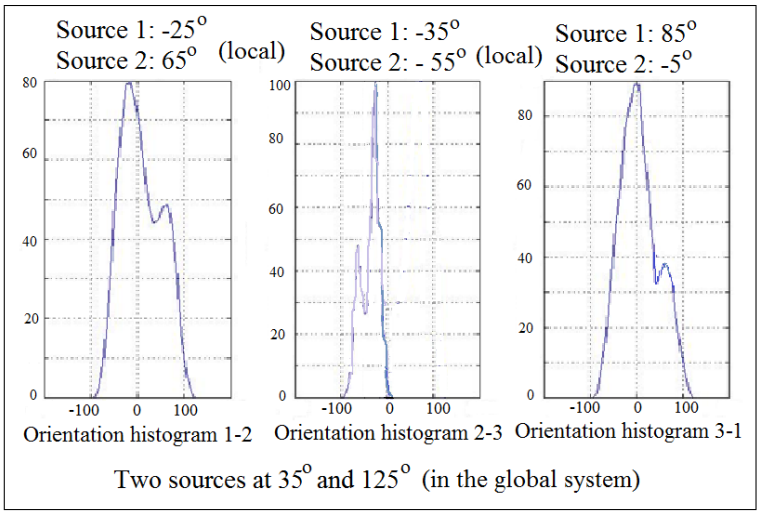


Figure 10. Three histograms obtained for pairs of microphones if two sources are active.

Measurements at 3-1: location is "in the back" - $\theta_{31}^{sym} = 105^o, \rightarrow \theta_{31} = 75^o, \rightarrow \theta(5) = 15^o$, $\theta(6) = 45^0$.

Hence, if the histogram analysis is perfectly done, we obtain 3 "votes" for the hypothesis $\theta = 45^o$ and single "votes" for 3 different other hypotheses: $-125^o, 135^o, 15^o$. The ambiguities that exist for each pair of microphones have been resolved by combining the results for three microphone pairs.

In general, all the orientation histogram peaks are detected first. Next, every "weak" peak is canceled when it is a "symmetric" version of some dominating peak. Hence, only strong peaks will remain that represent true source directions.

It has been validated in our experiments, that for two sources the triangle arrangement is sufficient in practice (Figure 10). The problem gets more complex if more sources are simultaneously active. Then the probability is growing, that a real orientation of one source is equal to a "symmetric" orientation of some other source. A possible solution is to assume some signal sparsity and to consider histograms created for shorter time intervals, i.e. to expect that a smaller number of sources is simultaneously active in every short-time interval.

5.4. Unique solution in the vector transformation method

The delay vector transformation method always solves the problem, while detecting source orientations in the interval $\theta \in [-\pi/2, \pi/2]$. Obviously, a symmetric solution $\theta^{sim} \in [\pi/2, 3/2\pi]$ is also possible. But here we can easily avoid the ambiguity problem if considering the signs of X and Y components in the arc tangent function (like for example in MATLAB is given as the atan2 function). We can use the location of given histogram peak (located along the red circle in Fig. 7) to make a one-to-one correspondence between the peak (a delay vector) and a direction angle within the entire interval of $[-\pi, \pi]$.

6. Results

The mixture signals have been generated by first acquiring each source signal by each microphone, and then adding appropriate signals for given microphones to simulate the mixture. Finally, a Gaussian white noise has been added to simulate the environment noise. The signal processing parameters were as follows: sampling frequency - 16 kHz, microphone distance - 0.08 m, window - Hamming, STFT frame - 1024 samples, STFT overlap - 512 samples, assumed wave speed - 340 m/s. The frame length was set in such a way that the bandwidth of frequencies up to 16 kHz was covered, which is the most appropriate bandwidth for analyzing a speech signal.

First, we evaluate the single source orientation quality. The following RMSE (root means square error) measure is computed from the results of experimental series:

$$RMSE = \sqrt{\sum_{i=1}^{N} (\bar{\theta}_i - \theta)^2 / N} \qquad (26)$$

In Table 1 we show that our "voting" method is able to detect the source and to estimate the orientation with similar high quality for the whole range of directions. Notice that the histogram resolution was set to 10^o, i.e. a range of orientations from -90^o to 90^o is digitized into 19 bins. The results of the second proposed approach - the vector transformation method - has been virtually the same. Both methods rely on the same pre-processed and selected data elements, representing reliable phase-difference measurements between pairs of microphones. This is the reason while the source estimation quality should be similar. The results are in fact the same as we use a relatively small number of histogram bins - this is motivated by the fact that our aim is not directly a high estimation quality but experimentally to verify the disambiguation ability of both methods.

Table 1. Source orientation estimation quality by the source voting method.

	Estimated θ for single source								
Real θ	Speaker no 1	Speaker no 2	Speaker no 3	Speaker no 4	Speaker no 5	Speaker no 6	Speaker no 7	Speaker no 8	RMSE
0	0	0	0	0	0	0	0	0	0
15	10	10	20	20	20	20	20	20	5
30	30	30	30	30	30	40	40	40	6,12
45	50	50	40	40	50	50	50	50	5
60	60	60	50	50	60	60	60	60	5
75	80	80	60	60	70	70	70	70	8,66
90	80	80	70	80	80	80	80	80	11,70
105	110	110	110	110	110	110	110	110	5
120	120	120	120	120	120	120	120	120	0
135	140	140	130	130	130	130	130	130	5
150	150	150	160	160	140	150	140	140	7,91
165	170	170	170	170	160	160	160	170	5
180	180	180	180	180	180	180	180	180	0
-165	-170	-170	-160	-160	-160	-160	-160	-160	5
-150	-150	-150	-150	-150	-150	-140	-150	-140	5
-135	-130	-130	-140	-140	-130	-130	-130	-130	5
-120	-120	-120	-130	-130	-120	-120	-120	-120	5
-105	-110	-110	-120	-110	-110	-110	-110	-110	7,07
-90	-80	-80	-110	-110	-80	-80	-80	-80	13,23
-75	-80	-80	-70	-70	-70	-70	-70	-70	5
-60	-60	-60	-60	-50	-60	-60	-60	-60	3,54
-45	-50	-50	-40	-40	-50	-50	-50	-50	5
-30	-30	-30	-20	-20	-30	-20	-30	-30	6,12
-15	-10	-10	-10	-10	-10	-10	-10	-10	5
								Aver.:	5.39

In Table 2 results of source estimation are shown when two speakers are simultaneously active. One speaker has been set at one of three directions: $0^o, -30^o, 30^o$, while the directions of the other speaker have varied from -165^o to 180^o with an interval of 15^o.

Table 2. Source orientation estimation quality if two sources are active.

Real θ		Errors of estimated θ		RMSE
0°	-165° : 180°	[-10 : 10]	[-10:10]	5.30
-30°	-165° : 180°	[-10 : 10]	[-10:10]	6.17
30°	-165° : 180°	[-10 : 10]	[-10:10]	6.01

7. Summary

Two methods has been proposed that independently resolve ambiguities in TDOA-based source detection and localization, when the signals are acquired by a triangle of microphones. We have shown that the previous histogram mapping method suffers from ambiguous detections due to symmetric localizations. The first improvement is achieved by changing the averaging of three TDOA histograms into the pre-analysis of individual orientation histograms, followed by a clustering of generated orientation hypotheses. The second improvement is achieved due to the delay vector transformation method. By experimental tests we confirmed that both proposed methods work properly for the full range of orientations.

Acknowledgements

The Authors gratefully acknowledge the implementation work support by Ai Kijima and Maciej Szumielewicz.

References

[1] D. Wang and G. E. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] C. Zhang, D. Florencio, D.E. Ba, and Z. Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Trans. on Multimedia*, 10(3):538–548, 2008.

[3] M. Brandstein and D. Ward. *Microphone Arrays*. Springer, Berlin Heidelberg New York, 2001.

[4] A.G. Piersol. Time Delay Estimation Using Phase Data. *IEEE Trans. on Acoustics Speech and Signal Processing*, ASSP-29(3):471-477, 1981.

[5] J-F. Wang, J.-C. Wang, B-W. Chen, and Z-W. Sun. A long-distance time domain sound localization. *UIC 2008*, LNCS 5061:616–625, Springer 2008.

[6] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52(7):1830–1847, 2004.

[7] F. Abrard and Y. Deville. Time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85:1389–1403, 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. on Signal Processing*, 58(1):121–133, 2010.

[9] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sane. K-hyperline clustering learning for sparse component analysis. *Signal Processing*, 89:1011–1022, 2009.

[10] H. Ouchi and N. Hamda. Separation of speech mixture by time-frequency masking utilizing sound harmonics. *Journal of Signal Processing, Research Institute of Signal Processing, Japan*, 13(4):331–334, 2009.

[11] W. Kasprzak, N. Ding, and N. Hamda. Relaxing the WDO assumption in blind extraction of speakers from speech mixtures. *Journal of Telecommunications and Information Technology, National Institute of Telecommunications, Poland*, 2010(4):50–58, 2010.

[12] A. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87:1833–1847, 2007.

[13] P.Svaizer, M.Matassoni, M.Omologo. Acoustic source location in a three-dimensional space using crosspower spectrum phase. *Proc. ICASSP-97*, 1:231-234, 1997.

[14] J. Huang, N. Ohnishi, and N. Sugie. A biomimetic system for localization and separation of multiple sound sources. *IEEE Trans. Instrum. Meas.*, 44:733–738, 1995.

[15] Y. Hioka, M. Matsuo, and N. Hamada. Multiple-speech-source localization using advanced histogram mapping method. *Acoust. Sci. & Tech., The Acoustical Society of Japan*, 30(2):143–146, 2009.

[16] S. Rickard. The DUET blind source separation algorithm. *Blind Speech Separation*, 217–237, Springer 2007.