



TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

BORIS KASHENTSEV
ESTIMATION OF DOMINANT SOUND SOURCE WITH THREE
MICROPHONE ARRAY

Master of Science thesis

Examiner: prof. Moncef Gabbouj
Examiner and topic approved by the
Faculty Council of the Faculty of
Computing and Electrical
Engineering
on 5th June 2013

ABSTRACT

BORIS KASHENTSEV: Estimation of dominant sound source with three microphone array

Tampere University of Technology

Master of Science Thesis, 47 pages

May 2015

Master's Degree Programme in Information Technology

Major: Multimedia

Examiner: Professor Moncef Gabbouj

Keywords: Time Delay Estimation, Direction of Arrival, Multimicrophone Array, Cross Correlation, Automated Sound Source Tracker

Several real-life applications require a system that would reliably locate and track a single speaker. This can be achieved by using visual or audio data. Processing of an incoming signal to obtain the location of a source is known as Direction of Arrival (DOA) estimation. The basic setting in audio based DOA estimation is a set of microphones situated in known locations. The signal is captured by each of the microphones, and the signals are analyzed by one of the following methods: steered beamformer based method; subspace based method; or time delay estimation based method.

The aim of this thesis is to review different classes of existing methods for DOA estimation and to create an application for visualizing the dominant sound source direction around a three-microphone array in real time. In practice, the objective is to enhance an algorithm for a DOA estimation proposed by Nokia Research Center. As visualization of dominant sound source creates a basis for many audio related applications, a practical example of such applications is developed.

The proposed algorithm is based on time delay estimation method and utilizes cross correlation. Several enhancements are developed to the initial algorithm to improve its performance. The proposed algorithm is evaluated by comparing it with one of the most common methods, general cross correlation with phase transform (GCC PHAT). The evaluation includes testing all algorithms on three types of signals: speech signal arriving from a stationary location, speech signal arriving from a moving source, and a transient signal. Additionally, using the proposed algorithm, a computer application with a video tracker is developed.

The results show that the initially proposed algorithm does not perform as well as GCC PHAT. The enhancements improve the algorithm performance notably, although they did not bring the efficiency of the algorithm to the level of GCC PHAT when processing speech signals. In case of transient signals, the enhanced algorithm was superior to GCC PHAT. The video tracker was able to successfully track the dominant sound source.

PREFACE

This Master's thesis was conducted in collaboration with Nokia Research Center and Department of Signal Processing of Tampere University of Technology. All experimental work was conducted in Nokia Research Center under supervision of PhD Kemal Ugur, PhD Mikko Tammi, Miikka Vilermo and Roope Järvinen. They assisted me with all technical matters and issues that I stumbled upon. I want to express my gratitude to them for sharing their knowledge and expertise.

I specially thank my supervisor professor Moncef Gabbouj for advice, patience and for never giving up on me, even after years of working. I'm grateful for examining my thesis and his constructive comments that helped me to finish this thesis.

I would love to thank my parents for their support over the years of my studies. Without them I would have never had chance to study abroad.

Finally, my gratitude goes to my girlfriend Jette who was expressing her support and motivated me during the course of the whole work.

In Tampere, Finland, on May 14 2015

BORIS KASHENTSEV

CONTENTS

1.	INTRODUCTION	1
2.	DIRECTION OF ARRIVAL ESTIMATION TECHNIQUES.....	4
2.1	Microphone array structure and conventions	4
2.2	Steered beamformer based methods.....	5
2.3	Subspace based direction of arrival estimation	8
2.4	Time delay estimate based methods.....	8
2.4.1	Time delay estimation	9
2.4.2	Source localization in two-dimensional space	11
3.	PROPOSED ALGORITHM	15
3.1	Assumptions	15
3.2	Basic algorithm	15
3.3	Algorithm enhancement	19
3.3.1	Adjustment of time delay array.....	19
3.3.2	Adjustment of subbands.....	21
3.3.3	Smoothing of DOA estimation	23
3.4	Automated sound source tracker	24
4.	RESULTS AND DISCUSSION	26
4.1	Time Delay Estimation.....	26
4.2	Direction of Arrival Estimation.....	35
4.3	Computational complexity	39
4.4	Automated sound source tracker	41
5.	CONCLUSION AND FUTURE WORK	43
	REFERENCES.....	45

LIST OF FIGURES

Figure 1.	<i>Uniform linear array with Far Field Source.</i>	5
Figure 2.	<i>A common broadband beamformer forms a linear combination of the sensor outputs.</i>	7
Figure 3.	<i>Block diagram of a generalized cross-correlator for time-delay of arrival estimation. x_i denote incoming signals, which might be filtered through H_i to obtain signals y_i.</i>	9
Figure 4.	<i>Localization in a 2-D plane. Circles represent microphones and triangle represents the sound source.</i>	11
Figure 5.	<i>A generated image of possible positions of a sound source for two known time delays of arrival.</i>	13
Figure 6.	<i>Comparison between possible hyperbolic sound source locations with a straight line.</i>	14
Figure 7.	<i>Setup of the used three microphone array. The microphones are located in equal distances from each other.</i>	16
Figure 8.	<i>Calculating the angle of the arriving sound.</i>	18
Figure 9.	<i>All possible angles using equidistant array of delays.</i>	20
Figure 10.	<i>Possible detectable angles using the optimal array of time delays.</i>	21
Figure 11.	<i>Normalized PSD of an average speech signal used during the experiment.</i>	22
Figure 12.	<i>Normalized PSDs for different subband arrays applied to the speech signal.</i>	23
Figure 13.	<i>Flow chart of the signals in the built automated sound source tracker.</i>	25
Figure 14.	<i>TDE of algorithms applied to a speech signal which originates from a static location in front of the microphone array.</i>	27
Figure 15.	<i>TDE of algorithms applied to speech signals originating from static sound sources.</i>	28
Figure 16.	<i>TDE of the proposed algorithms using cube root scaling applied to speech signals originating from static sound sources.</i>	29
Figure 17.	<i>TDE of applying algorithms to moving sound signals.</i>	29
Figure 18.	<i>TDE of applying algorithms to moving sound signals.</i>	30
Figure 19.	<i>TDE of the algorithms applied to transient signals.</i>	31
Figure 20.	<i>TDE results of the GCC PHAT algorithm the basic algorithm and the enhanced algorithm applied to transient signals, scaled with cube root.</i>	32
Figure 21.	<i>A part of the TDE result of the GCC PHAT algorithm in 3D applied to transient signals.</i>	33
Figure 22.	<i>A part of the TDE result of the basic algorithm in 3D applied to the transient signals and scaled with cube root.</i>	34
Figure 23.	<i>A part of the TDE result of the enhanced algorithm in 3D applied to the transient signal and scaled with cube root.</i>	34

Figure 24. Estimation of DOA angle for static sound source placed in front of the microphone array.....	36
Figure 25. Estimation of DOA angle for static sound source placed behind of the microphone array.....	37
Figure 26. Estimation of DOA angles for a moving sound source.....	38
Figure 27. Visualization of the DOA estimation without applying limitation of dominant sound source.....	41
Figure 28. Image taken by the camera of the video tracking system.....	42

LIST OF ABBREVIATIONS

DFT	Discrete Fourier transform
DOA	Direction of Arrival
DSB	Delay-and-sum beamformer
FIR	Finite impulse response
GCC	Generalized cross correlation
LCMV	Linearly constrained minimum-variance
MVB	Minimum-variance beamformer
PHAT	Phase transform
PSD	Power spectral density
SNR	Signal to noise ratio
TDE	Time delay estimation
ULA	Uniform linear array
USB	Universal serial bus

1. INTRODUCTION

For humans it is very natural to communicate through speech. Therefore, there are a lot of attempts to implement this type of communication between a human and a machine. The first step was to invent a microphone. However, signal picked up by a microphone contains many additional signals that for certain tasks are considered as noise. Humans are able to understand speech in presence of noise of the same power [1, pp. 383–385], in some cases even of greater power [2]. Computers, on the other hand, are unable to perform this task, and that started the race of developing methods for speech enhancement. One of the methods to achieve better signal was employment of multiple microphones. Currently, multiple microphone arrays are used in two categories of tasks: speech enhancement and positioning the location of a signal emitter.

Speech enhancement often attempts to solve problems related to presence of background noise and reverberation in a room [3]–[5]. Promising areas for speech enhancement using multiple microphones are teleconferencing, hearing aids, hands-free communication, cars and home entertainment systems. Applications used for hands-free communication in cars and home entertainment systems strongly depend on speech recognition, which depends on sufficient performance of speech enhancement techniques [6]. In case of cars, possible locations of speakers are limited, which makes it possible to use beamforming techniques to pick up sound from a specific direction [7]. However, only top car manufacturers can afford such systems inside of a car [1, p. 391]. Good example of using multiple microphones in home entertainment systems is Kinect motion controller by Microsoft [8].

Processing of an incoming signal to obtain the location of the sound source is known as Direction of Arrival (DOA) estimation. Estimating DOA is not limited to estimation of the location of a sound source. It is possible to locate sources emitting different types of energy (e.g., radio frequency, acoustic, ultrasonic, optical, infrared, seismic and thermal). For example, the Federal Communications Commission has authorized the E911 system (or E112 system in Europe), which requires cellular telephone providers to locate a cell phone user to tens of meters in an emergency situation. [9, p. 343]

Another example of serious area for DOA estimation is surveillance, especially underwater surveillance. However, in this case hydrophones are used instead of regular microphones. Some of the systems are even able to classify surfaces and underwater sources of sound, for example, surface vessels, swimmers, divers and unmanned underwater vehicles. [10]

DOA estimation of speech has many practical applications. Particular examples of applications where DOA estimation is especially useful are video conferencing and long-distance video classrooms. In a conference, participants want to see the person in the room who is speaking at each time. Especially in video classrooms, the speaker can also be moving around in the room. With existing video conferencing systems, viewing at the speaker is achieved by one of the following ways. First, multiple stationary cameras can be placed around the room to have different views on all participants. Secondly, the camera system can entail switches, which participants can use to steer the camera to their direction. Finally, another person can manually operate a camera.

The current systems are often costly and require additional manpower or hardware to function in a reliable and efficient manner. Thus, a new approach is needed to reliably and automatically track a single speaker. In general, the speaker can be localized based on either visual or acoustic signals. Visual tracking systems have been developed, for example by Wren et al. [11]. This method is, however, complex and has a high computational load requiring powerful computers to perform. One of the recent examples of tracking using video input has been demonstrated in the iCam system [12]. Presented system employs several complex and expensive video cameras. Thus, using acoustic signals is reasonable. A useful system would be a video tracker steering the camera towards the speaker automatically. This could be built based on DOA estimation, using a microphone array placed in the conference room or classroom. [13]

Further applications of DOA estimation include human computer interfaces, where communicating with the computer occurs through speech. These systems utilize methods such as superdirective beamforming for DOA estimation of speech. [14] Hearing aids that capture sound signals in the presence of background noise also employ adaptive beamforming [15], [16].

The ultimate principle of DOA estimation using spatially separated microphones is to process the phase difference of an audio signal detected by the individual microphones in an array. The audio signal arrives to the spatially separated microphones with a time difference, giving time delays when one of the microphones is set as a reference point. As the geometry of a microphone array is known, the time delays allow estimating the respective DOA of the signal. Three classes of DOA estimation methods exist: steered beamformer based methods, subspace based methods, and time-delay estimate methods. [1, pp. 158 – 164]

The aim of this thesis is to create an application for visualizing the main sound source direction around a microphone array in real time. In practice, the objective is to enhance an algorithm for DOA estimation proposed by Nokia Research Center. As visualization of dominant sound source creates a basis for many audio related applications, a practical example of such applications is developed and built using a three microphone array: a computer application with a video tracker.

The structure of the thesis is the following. First, different techniques of DOA estimation and principles of sound source localization in two-dimensional plane are discussed in Chapter 2. Chapter 3 presents initial and enhanced algorithms proposed in this thesis for DOA estimation, as well as a practical application of the proposed algorithm. Chapter 4 discusses results of the tests made with the proposed algorithm and the built practical system. Chapter 5 briefly states conclusions for this research and possible future work and enhancements.

2. DIRECTION OF ARRIVAL ESTIMATION TECHNIQUES

The basic setting in direction of arrival (DOA) estimation is a given set of acoustic sensors (microphones) situated in known locations. The goal is to estimate two or three-dimensional coordinates of the acoustic sound source. A single or multiple sound sources are assumed to be present in the system. The signal is captured by each of the acoustic sensors, and the signals are analyzed by one of the following methods: steered beamformer based method; subspace based method; or time delay based method.

Majority of DOA estimation algorithms apply narrowband beamforming techniques in order to obtain separate DOA estimates for different frequency bands. These separate estimates are later combined to extract one estimate based on statistical observations.

2.1 Microphone array structure and conventions

In order to estimate the DOA of a single source, the sensors should receive the same signal but at slightly different time instants. This is accomplished by spatially separating them. Basic structure of microphone placement shown in Figure 1 is called uniform linear array (ULA). It is used in this chapter to explain principles of conventional methods of DOA estimation.

Microphones are placed in a straight line with equal distance, d , between neighboring microphones. Distance between microphone array and the sound source is assumed to be greater than distance between neighboring microphones, which guarantees the same angle, θ , of sound signal arrival into the microphones [9, p. 345].

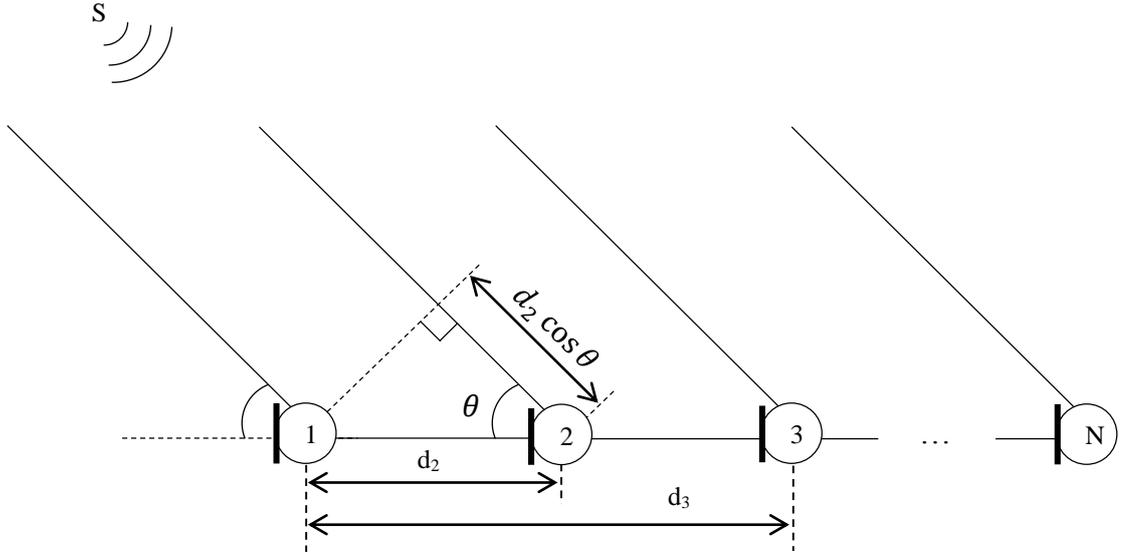


Figure 1. Uniform linear array with Far Field Source.

Traditionally microphone 1 is fixed as the reference microphone, and the signal received by the microphone is $s(t)$, without taking into account noises from the air. Then the signal received by microphone 2 would be the same signal $s(t)$ with a time delay or time advance of $\frac{d \cos \theta}{c}$, where c is the velocity of sound. Extending this idea to the rest of the microphone array, signals arrived to arbitrary microphone can be written as

$$x_i = s(t + \tau_i), \quad i = 2 \dots N, \quad (2.1)$$

where

$$\tau_i = \frac{d_i \cos \theta}{c}, \quad i = 2 \dots N, \quad (2.2)$$

and d_i is the distance between reference microphone and microphone i :

$$d_i = (i - 1)d, \quad i = 2 \dots N. \quad (2.3)$$

2.2 Steered beamformer based methods

The first class of DOA estimation methods contains the steered beamformer based methods. The signals from spatially separated array-sensors are joined by beamformers in a way that the array output accentuates signals from a specific viewing direction: if a signal is coming from the viewing direction, then the power of the array output is high; and following the same logic, the power of the array output is low in case signal is absent in the viewing direction. Therefore, the array is used to construct beamformers that inspect in all possible directions. [17]

Two major approaches can be distinguished among beamformer based methods: the delay-and-sum beamformer (DSB) and the linearly constrained minimum-variance (LCMV) beamforming. The DSB is the simplest type of beamformer that can be implemented, also most often referred to as a conventional beamformer. In a DSB time shifts are applied to the signals which have reached the microphones to compensate the delays in the arrival of the incoming signal to each microphone. After time-alignment these signals are summed together to create a single output signal. Additionally, filters might be applied to the array signals. That action is used to advance DSB. [1] In LCMV beamforming, the response of the beamformer is constrained so that signals from the viewing direction are passed with specified gain and phase. The weights are chosen to minimize either the output variance or power, depending on the response constraint. Thereby, the signal from the direction of interest is preserved, while noise and signals from other directions contribute little to the output. [18, pp. 14–15]

The capability of beamformers to enhance signals from a particular direction as well as to decrease signals from other directions is used in DOA estimation. A beamformer is constructed for each direction of interest and the power of the array output is computed. The direction that gives the largest power is taken as the estimated DOA of the incoming signal. In other words, when the power is plotted against the viewing direction, it shows a peak for each viewing direction from which a signal is detected.

There are two types of beamformers: narrowband and broadband beamformers. Classification depends on the bandwidth of the signals on which beamformers are used. Narrowband beamformers expect that the incoming signal captured by the beamformer has a narrow bandwidth centered at a particular frequency. To satisfy that condition, a signal might be bandpass filtered to convert it to narrowband signal. Additionally, the same bandpass filter has to be applied to all channels of the microphone array. It assures that relative phase information between channels is not altered. [18]

Figure 2 presents broadband beamformer scheme. That beamformer samples the propagating wave field in both space and time. The output at time k , $y(k)$, can be depicted as

$$y(k) = \sum_{l=1}^J \sum_{p=0}^{K-1} w_{l,p}^* x_l(k-p), \quad (2.4)$$

where $w_{l,p}$ is the p th weight of the filter applied to signal from l th microphone, x_l is the signal from l th microphone, $K-1$ is the maximum number of delays in each of the J sensor channels, and $*$ represents complex conjugate. [17]

In case of narrowband signals, equation (2.4) takes variable K equals 1 . For convenience equation in matrix form takes appearance of

$$y(k) = \mathbf{w}^H \mathbf{x}(k), \quad (2.5)$$

where $(^H)$ represents Hermitian (complex conjugate) transpose, and boldface is used to represent vector quantities. [17]

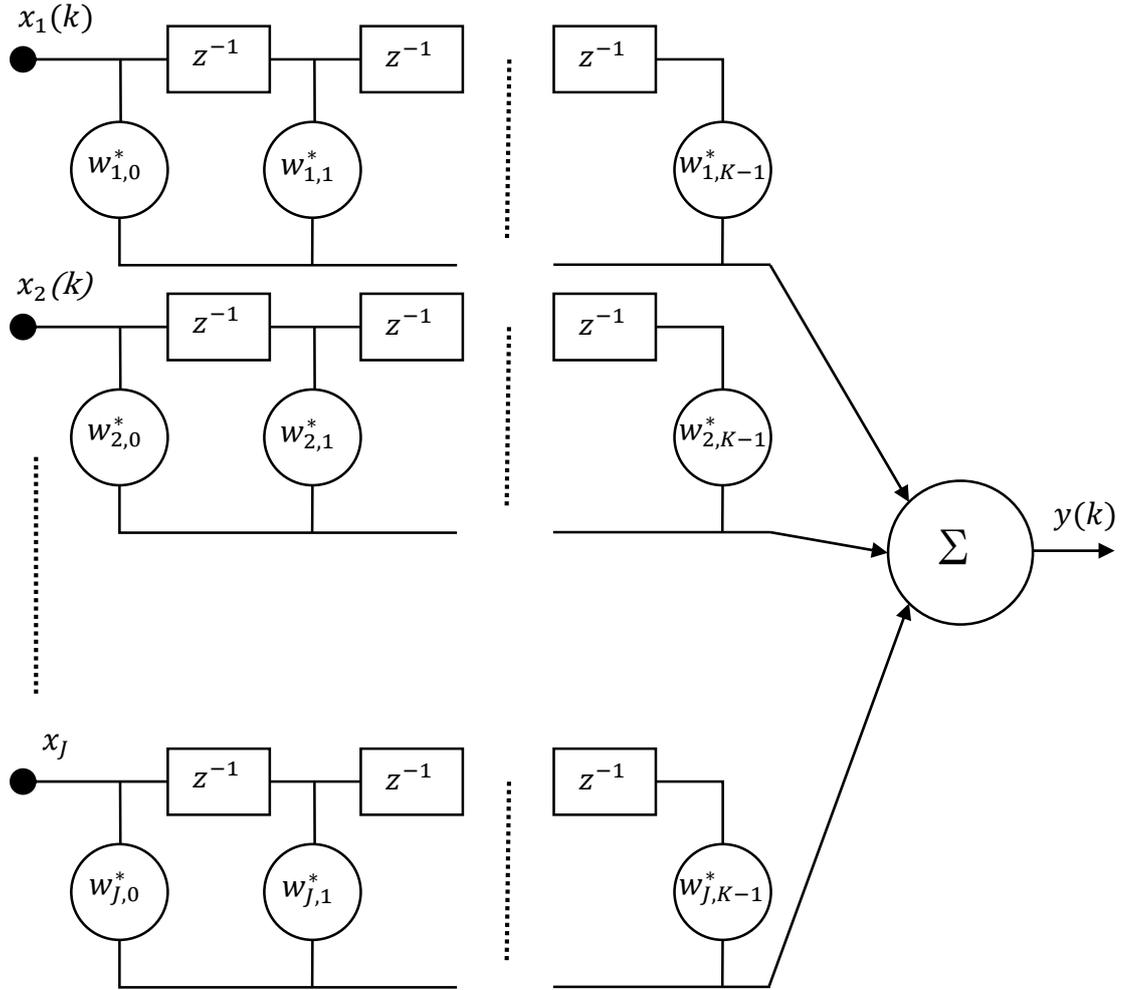


Figure 2. A common broadband beamformer forms a linear combination of the sensor outputs. Modified from [18].

The frequency response of a finite impulse response (FIR) filter with tap weights w_p^* , where $1 \leq p \leq J$ and tap delay of T seconds is given by

$$r(\omega) = \sum_{p=1}^J w_p^* e^{-j\omega T(p-1)}, \quad (2.6)$$

which can be expressed as

$$r(\omega) = \mathbf{w}^H \mathbf{d}(\omega), \quad (2.7)$$

where $r(\omega)$ represents the filter response to a complex sinusoid of frequency ω ; $\mathbf{w}^H = [w_1^* \ w_2^* \ \dots \ w_J^*]$ are weights of the filter; $\mathbf{d}(\omega) = [1 \ e^{j\omega T} \ e^{j\omega 2T} \ \dots \ e^{-j\omega T(J-1)}]^H$ is a vector describing the phase of the complex sinusoid at each tap in the FIR filter relative to the tap associated with w_1 .

Assume that ω_0 is a frequency of interest, therefore, according to the property of a beamformer, the desired frequency response is unity at ω_0 and zero elsewhere. A common solution to this problem is to choose \mathbf{w} as the vector $\mathbf{d}(\omega_0)$. This choice can be shown to be optimal in terms of minimizing the squared error between the actual response and desired response.[17]

The advantage of a steered beamformer based algorithm is ability to detect the directions of all sound sources that effect array with one set of computations. Therefore, this class of algorithms is suitable for detecting multiple sources. The computational load required for steered beamformer based methods is massive, thereby not being suitable for all applications. [19, p. 4]

2.3 Subspace based direction of arrival estimation

The second class of DOA estimation methods contains high-resolution subspace based methods. Subspace based methods divide the cross-correlation matrix of signals array into signal and noise subspaces applying eigenvalue decomposition. Additionally, these methods are widely used in the context of spectral estimation. These methods are able to differentiate multiple sources located close to each other. Subspace based methods handle that task better than the steered beamformer based methods because computational results give much sharper peaks at the correct locations. These methods works on a principle of exhaustive search over the set of possible source locations. [20], [21]

2.4 Time delay estimate based methods

The final class of methods is time delay estimation (TDE) based methods. In that class DOA estimation is completed in two steps. First, the time delay is estimated for each pair of microphones in the array. Second, the time delays acquired in previous step are combined with the knowledge of microphone array geometry to determine the best estimation of the DOA. [20]

TDE based methods have the advantage of lower computational load compared to other methods, because the need of extensive search over all possible directions of arrival is avoided. This makes TDE based methods the most efficient. Additionally, TDE based methods can be applied to broadband signals, unlike the other methods. However, TDE

based methods produce the most reliable results in case of a single sound source. [19, pp. 7–8]

2.4.1 Time delay estimation

Various techniques exist to compute pair-wise time delays [22]. General cross correlation (GCC) method and GCC with phase transform (PHAT) were chosen for demonstration purposes, as an example of TDE based methods. Calculation of time delay \hat{D} between two signals x_i in a microphone pair can be calculated based of the following principle, which is also presented in the schematic illustration (Figure 3).

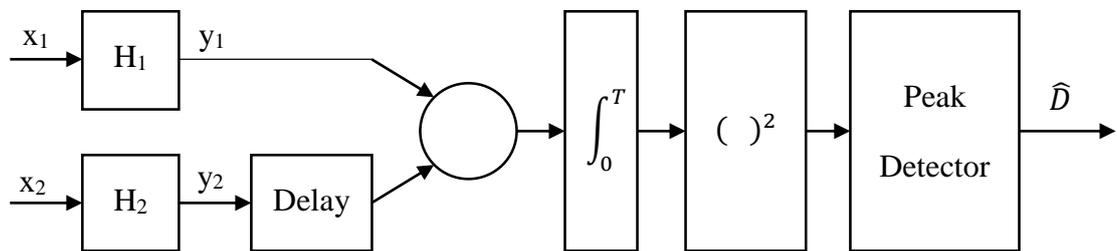


Figure 3. Block diagram of a generalized cross-correlator for time-delay of arrival estimation. x_i denote incoming signals, which might be filtered through H_i to obtain signals y_i . [23]

Signals x_1 and x_2 arrive to spatially separated microphones. The signals can be mathematically modelled as

$$x_1(t) = s(t) + n_1(t), \quad (2.8)$$

$$x_2(t) = \alpha s(t + D) + n_2(t), \quad (2.9)$$

where $s(t)$, $n_1(t)$, and $n_2(t)$ are real, jointly stationary random signals, α is a linear coefficient, explained by the fade property of a sound. Signal $s(t)$ is assumed to be uncorrelated with noises $n_1(t)$ and $n_2(t)$. In order to estimate time delay, it is required to find cross correlation between the signals x_1 and x_2 :

$$\hat{R}_{x_1 x_2}(\tau) = \frac{1}{T - \tau} \int_{\tau}^T x_1(t) x_2(t - \tau) dt, \quad (2.10)$$

where T represents the observation interval and argument τ that maximizes equation (2.10) provides an estimate of delay. [23]

The cross correlation between $x_1(t)$ and $x_2(t)$ is related to the cross power spectral density function by the well-known Fourier transform relationship:

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} G_{x_1x_2}(f)e^{j2\pi f\tau}df, \quad (2.11)$$

where

$$G_{x_1x_2}(f) = X_1(f)X_2^*(f), \quad (2.12)$$

where X_1 and X_2 are DFT of incoming signals and $(^*)$ denotes complex conjugate. Equation (2.11) is valid for cases when pre-filtering of signals is not required. Taking into account that signals $x_1(t)$ and $x_2(t)$ are going through filtering (filters H_1 and H_2), equation (2.11) will become:

$$R_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \psi_g(f)G_{x_1x_2}(f)e^{j2\pi f\tau}df, \quad (2.13)$$

where

$$\psi_g(f) = H_1(f)H_2^*(f) \quad (2.14)$$

denotes the general frequency weighting. [23]

Equation (2.11) is a representation of GCC approach. In order to get a representation of the GCC approach with PHAT, the frequency weighting function is written as [23]:

$$\psi_g(f) = \frac{1}{|G_{x_1x_2}(f)|}, \quad (2.15)$$

which yields

$$R_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \frac{G_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} e^{j2\pi f\tau}df. \quad (2.16)$$

Taking into account the fact that noise in (2.8) and (2.9) is uncorrelated (i.e., $G_{n_1n_2}(f) = 0$):

$$\begin{aligned} R_{y_1y_2}(\tau) &= \int_{-\infty}^{\infty} \frac{G_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} e^{j2\pi f\tau}df = \int_{-\infty}^{\infty} e^{j2\pi fD} e^{j2\pi f\tau}df \\ &= \delta(\tau - D). \end{aligned} \quad (2.17)$$

This means that, in an ideal situation, the result of cross correlation would be a delta function, which is supposed to provide only one optimal solution for the time delay of signal arrivals to microphones.

2.4.2 Source localization in two-dimensional space

After collecting information from the Peak Detector (Figure 3), the system acquires data about time delays between a signal reaching the reference microphone and other microphones of an array. The next step of the TDE based methods is to localize the sound source.

Time delays of arrival are estimated for each microphone i with respect to the first (reference) microphone: $\dot{d}_{i,1} = \dot{d}_i - \dot{d}_1$, for $i = 2, 3, \dots, N$, and \dot{d}_i are the time delay associated with microphone i . In Figure 4 there are N arbitrarily distributed microphones and one sound source.

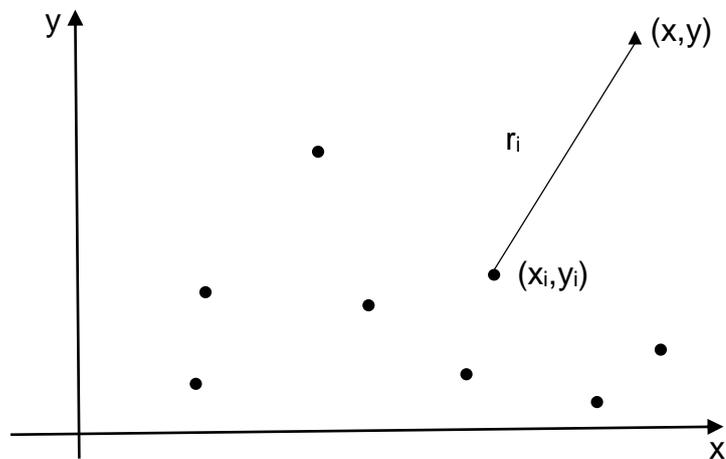


Figure 4. Localization in a 2-D plane. Circles represent microphones and triangle represents the sound source.

Coordinates of the sound source (x, y) are unknown. Coordinates of each microphone (x_i, y_i) are known. Therefore, the squared distance between the source and sensor i is:

$$\begin{aligned} r_i^2 &= (x_i - x)^2 + (y_i - y)^2 \\ &= (x_i^2 + y_i^2) - 2x_i x - 2y_i y + x^2 + y^2, i = 1, 2, \dots, N \end{aligned} \quad (2.18)$$

As before, c is the velocity of sound, also known as the signal propagation speed. Then

$$r_{i,1} = c\dot{d}_{i,1} = r_i - r_1 \quad (2.19)$$

define a set of nonlinear equations, the solution of which gives (x, y) . [24, p. 1906], [25, pp. 1317–1318]

Solving those nonlinear equations is difficult. There are several iterative solutions existing: using linearization by Taylor-series expansion [25], [26]; or rearranging equation (2.19) so that in the end there will be linear equation of three unknown variables x , y and r_1 [27]. Additionally, Chan and Ho proposed a simple and efficient estimator for

hyperbolic location. They proposed a technique in locating a source based on intersection of hyperbolic curves defined by the time differences of arrival of a signal received at a number of microphones. This estimator is noniterative and gives an explicit solution. [24]

However, this thesis proposes a different method of location estimation. For that reason it was decided to investigate the geometry of possible sound source locations for a single time delay of arrival, which was noticed to be hyperbolic.

A hyperbola is a set of all points in a plane, the locations of which are characterized by a constant difference of their distance from two fixed points. The two fixed points are called the focal points, or foci. In case of our problem, microphones represent the focal points of a hyperbolic curve. A hyperbola consists of two branches, and the sound source is located on one of the branches. The standard equation of a hyperbola centered at the origin is given by:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \quad (2.20)$$

when the transverse axis matches x axis. A transverse axis is an axis, which goes through the focal points of a hyperbola. The center of the hyperbola is the center of a section connecting focal points. a is the distance between the center of the hyperbola and a vertex, which is the intersection point of a hyperbola branch and transverse axis. Coefficient $b^2 = c^2 - a^2$, where c is the distance between the center of the hyperbola and one of the focus points.

As shown in Figure 5, two microphones would give an infinite array of possible sound source locations. For that reason, one more non-collinear microphone is required. An additional microphone delivers an extra hyperbola. The intersection of the two hyperbolas gives the true location of the sound source. In case of collinear microphones, the intersection(s) of the formed hyperbolas would leave ambiguity of the sound source location. Three dimensional sound source localization requires one more microphone. This thesis, however, focuses on the estimation of the sound source direction in a two-dimensional plane.

Instead of using the existing methods mentioned above ([24]–[27]) for locating a sound source, for the purposes of the thesis it has been decided to use the following approximation for a hyperbola: after a certain distance from a focus point, it is allowed to assume that branches become straight lines. That kind of approximation saves time during DOA estimation, which positively affects real-time execution. To illustrate the approximation, an application was developed that draws possible locations of a sound source according to a time delay between a signal reaching different microphones. Figure 5 shows one possible situation for two arbitrary time delays between microphones one – two and one – three.

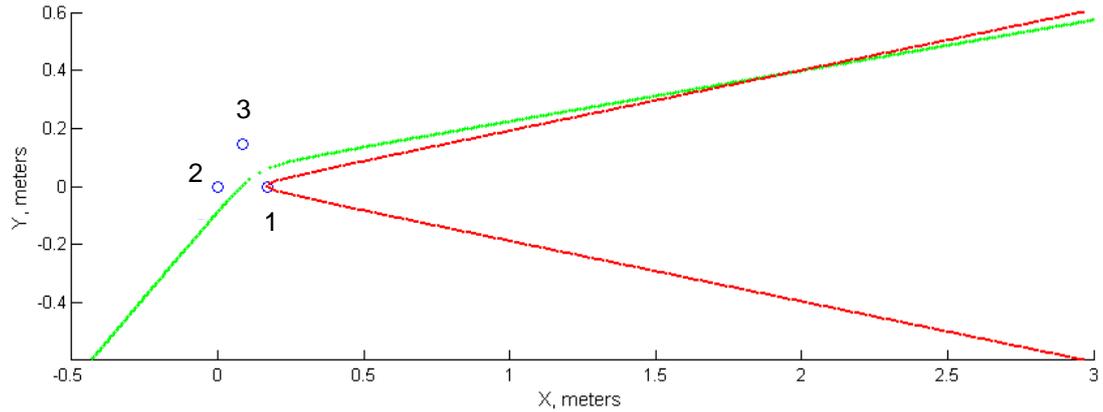
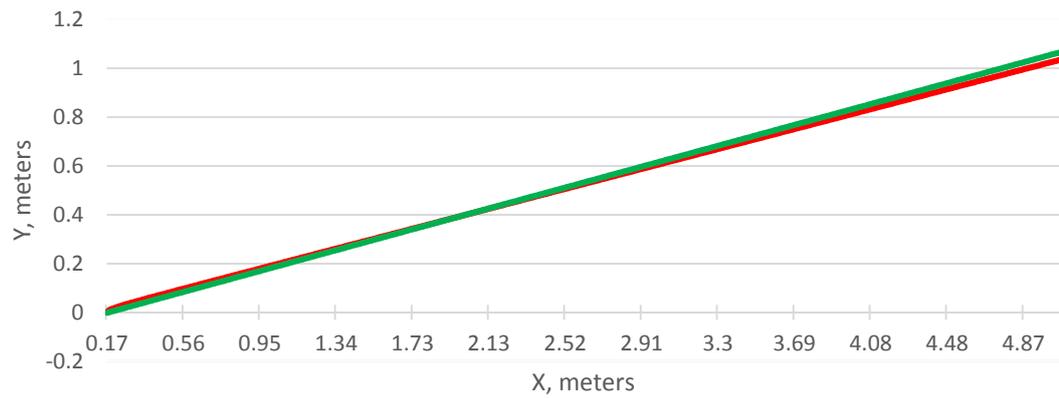
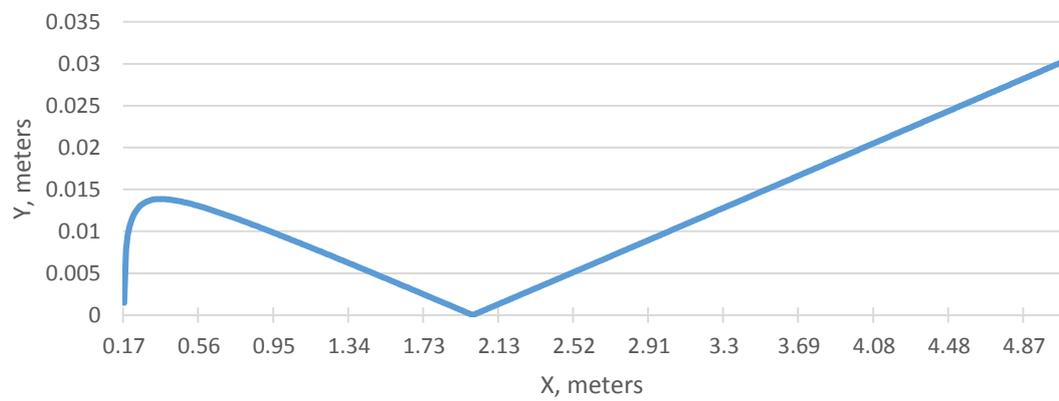


Figure 5. A generated image of possible positions of a sound source for two known time delays of arrival. Blue circles represent microphones; red points show possible locations of a sound source for one time delay between microphones one and two; green dots show possible locations for other time delay between microphones one and three.

To demonstrate this approximation, possible sound source locations were compared with a straight line (Figure 6). Figure 6 (a) clearly shows that those lines are almost undistinguishable. Figure 6 (b) shows better what is the actual distance between the possible locations of a sound source and a straight line. It is visible that even at the distance of 5 meters, displacement is only 3 cm in case this approximation is used. In other words, use of straight line instead of hyperbola is acceptable.



(a)



(b)

Figure 6. Comparison between possible hyperbolic sound source locations with a straight line. (a) Shows possible locations of a sound source for a particular time delay (red plot), and a straight line (green line) which goes through the location of one of the microphones and the possible location of sound source at a distance of around 2 meters. (b) Shows distance between the straight line and possible coordinates of the sound source.

3. PROPOSED ALGORITHM

In this chapter an algorithm for DOA estimation is presented. The work was developed in collaboration with Nokia Research Center, based on an earlier implementation. This implementation is referred in this thesis as the basic algorithm. In chapter 3.3 several improvements to the basic algorithm are presented with their justifications. The algorithm with improvements is denoted as the enhanced algorithm.

3.1 Assumptions

In the surrounding of the microphone array (in the room) multiple sound sources can be present, including noise sources contributing to the sound field. A dominant sound is defined here as the loudest sound.

The following conditions are assumed, under which the location of sound source is estimated:

1. Single sound source, infinitesimally small, omnidirectional source.
2. Reflections from the bottom of the plane and from the surrounding objects are negligible.
3. The dominant sound source to be located, is not assumed to be stationary during the data acquisition period.
4. Microphones are assumed to be both phase and amplitude matched and without self-noise.
5. The change in sound velocity due to change in pressure and temperature are neglected. The velocity of sound in air is taken as 343 m/s.

3.2 Basic algorithm

The basic algorithm is constructed on similar principles as presented by Wang et al. [28] in terms of utilizing the method of non-circular cross correlation in frequency domain. In current work, three microphones were placed in corners of an equilateral triangle, as illustrated in Figure 7. The direction of arrival of a sound is estimated independently for B frequency domain subbands. The objective is to find the direction of the perceptually dominating sound source for every subband.

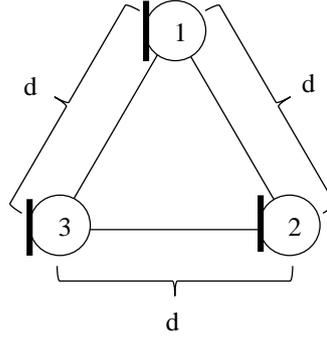


Figure 7. Setup of the used three microphone array. The microphones are located in equal distances from each other.

Signals from each input channel $k = 1, \dots, 3$ are transformed to frequency domain using discrete Fourier transform (DFT). Hamming windows with 50% overlap and effective length of 20 ms are used, as recommended by Paliwal et al.[29]. Before DFT, a number of zeroes equal to D_{max} are added to the end of the window, where D_{max} denotes the maximum time delay in samples between the microphones. In the microphone setup presented in Figure 7, the maximum delay is obtained as

$$D_{max} = \frac{dF_s}{c}, \quad (3.1)$$

where d is the distance between a pair of microphones, F_s is the sampling rate of signal and c is the speed of the sound in the air. The DFT gives the frequency domain representations $X_k(n)$ of all three channels, $k = 1, \dots, 3$, $n = 0, \dots, N - 1$. N is the total length of the window consisting of the Hamming window and the additional D_{max} zeroes.

The frequency domain representation is divided into B subbands:

$$X_k^b(n) = X_k(n_b + n), n = 0, \dots, n_{b+1} - n_b - 1, b = 0, \dots, B - 1 \quad (3.2)$$

where n_b is the first index of b th subband. The widths of the subbands follow the Bark scale.

For every subband the directional analysis is performed as follows. First the direction is estimated with two channels. The task is to find time delay τ_b that maximizes the correlation between two channels for subband b . The frequency domain representation of $X_k^b(n)$ can be shifted τ time domain samples using equation

$$X_{k,\tau}^b(n) = X_k^b(n) e^{-j \frac{2\pi n \tau}{N}}. \quad (3.3)$$

Now the optimal delay τ_b is obtained from

$$\max_{\tau_b} \text{Re}(X_{2,\tau_b}^b * X_3^b), \tau_b \in [-D_{max}, D_{max}], \quad (3.4)$$

where Re indicates the real part of the result and $*$ denotes combined transpose and complex conjugation operations. X_{2,τ_b}^b and X_3^b are considered vectors with length of $(n_{b+1} - n_b)$ samples. Resolution of one sample is generally suitable for the search of the delay. With the delay information a sum signal is created. It is constructed using following logic:

$$X_{sum}^b = \begin{cases} (X_{2,\tau_b}^b + X_3^b)/2 & \tau_b \geq 0 \\ (X_2^b + X_{3,\tau_b}^b)/2 & \tau_b < 0 \end{cases} \quad (3.5)$$

Equation (3.5) confirms that in the sum signal the content of the channel in which an event occurs first is added as such, whereas the channel in which the event occurs later is shifted to obtain the best match.

Shift τ_b indicates how much closer the sound source is to microphone 2 than microphone 3. The actual distance Δ_{23} can be calculated as

$$\Delta_{23} = \frac{c\tau_b}{F_s}. \quad (3.6)$$

Figure 8 presents a scheme of sound arrival to two microphones. From cosine laws follows:

$$(\Delta_{23} + b)^2 = d^2 + b^2 - 2db \cos \beta. \quad (3.7)$$

Since

$$\beta = \pi - \alpha_b \Rightarrow \cos \beta = -\cos \alpha_b, \quad (3.8)$$

substituting (3.8) to (3.7) gives:

$$\alpha_b = \pm \cos^{-1} \left(\frac{\Delta_{23}^2 + 2b\Delta_{23} - d^2}{2db} \right), \quad (3.9)$$

where d is the distance between microphones and b is the estimated distance between sound sources and nearest microphone. As discussed in the previous chapter, b can be set to a fixed value. For example, $b = 2$ meters was found to provide stable results. Notice that there are two alternatives for the direction of the arriving sound as the exact direction cannot be determined with only 2 microphones. The third microphone is utilized to determine which of the angles in (3.9) is correct.

The distances between microphone 1 and the two estimated sound sources are:

$$\begin{aligned}\delta_b^+ &= \sqrt{(h + b \sin(\hat{\alpha}_b))^2 + (d/2 + b \cos(\hat{\alpha}_b))^2} \\ \delta_b^- &= \sqrt{(h - b \sin(\hat{\alpha}_b))^2 + (d/2 + b \cos(\hat{\alpha}_b))^2}\end{aligned}\quad (3.10)$$

where h is the height of the equilateral triangle, and calculated as

$$h = \frac{\sqrt{3}}{4} d. \quad (3.11)$$

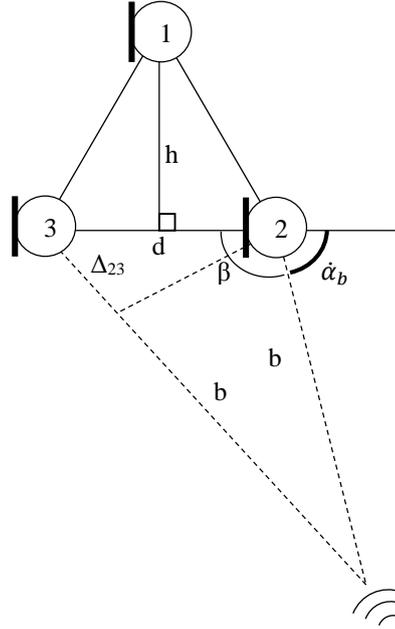


Figure 8. Calculating the angle of the arriving sound.

The distances in (3.10) equal to delays (in samples)

$$\begin{aligned}\tau_b^+ &= \frac{\delta_b^+ - b}{c} Fs \\ \tau_b^- &= \frac{\delta_b^- - b}{c} Fs\end{aligned}\quad (3.12)$$

Out of these two delays the one is selected which provides better correlation with the sum signal. The correlations are obtained as

$$\begin{aligned}c_b^+ &= Re(X_{sum, \tau_b^+}^b X_1^b)^* \\ c_b^- &= Re(X_{sum, \tau_b^-}^b X_1^b)^*\end{aligned}\quad (3.13)$$

Finally, the direction of the dominant sound source for subband b is:

$$\alpha_b = \begin{cases} |\hat{\alpha}_b| & c_b^+ \geq c_b^- \\ -|\hat{\alpha}_b| & c_b^+ < c_b^- \end{cases} \quad (3.14)$$

The same estimation is repeated for every subband.

3.3 Algorithm enhancement

Later, in chapter 4, it will be shown that basic algorithm is able to perform DOA estimation, nevertheless, results of the basic algorithm are not sufficient enough. During the course of development and testing it has been noted that several enhancements can improve results without increasing the complexity of the algorithm. These enhancements include adjustment of frequency plane division into subbands, calculation of the optimal time delay array and smoothing of the DOA estimation.

3.3.1 Adjustment of time delay array

In the basic algorithm, equation (3.4) is used to calculate the optimal delay of incoming signal to two microphones. In that equation, a correlation between two channels is calculated for different delays τ so that it would maximize the correlation. For this calculation, it was initially proposed to use one sample as the resolution of τ , meaning that the array of delays τ was equidistant. However, during the experiments it was noted that equidistant array was not the best choice. In order to illustrate that issue, equations (3.6) and (3.9) were combined:

$$\alpha_{all} = \pm \cos^{-1} \left(\frac{\left(\frac{\tau_{all}c}{F_s} + b \right)^2 - d^2 - b^2}{2db} \right), \tau_{all} \in [-D_{max}; D_{max}], \quad (3.15)$$

where, α_{all} are all possible angles in respect to τ_{all} , all possible time delays in samples for the current microphone setup. Using equation (3.15), all possible angles in respect to all possible time delays were plotted in Figure 9. It illustrates that the chosen array of delays would not cover the whole array of possible angles. For example, signal that comes from angles 0 and 180 degrees will be, most probably, associated with signals coming from angles 10 and 170 degrees. Additionally, concentration of points close to 0 and 180 is very small, which leads to certain angles being undetectable by the basic algorithm.

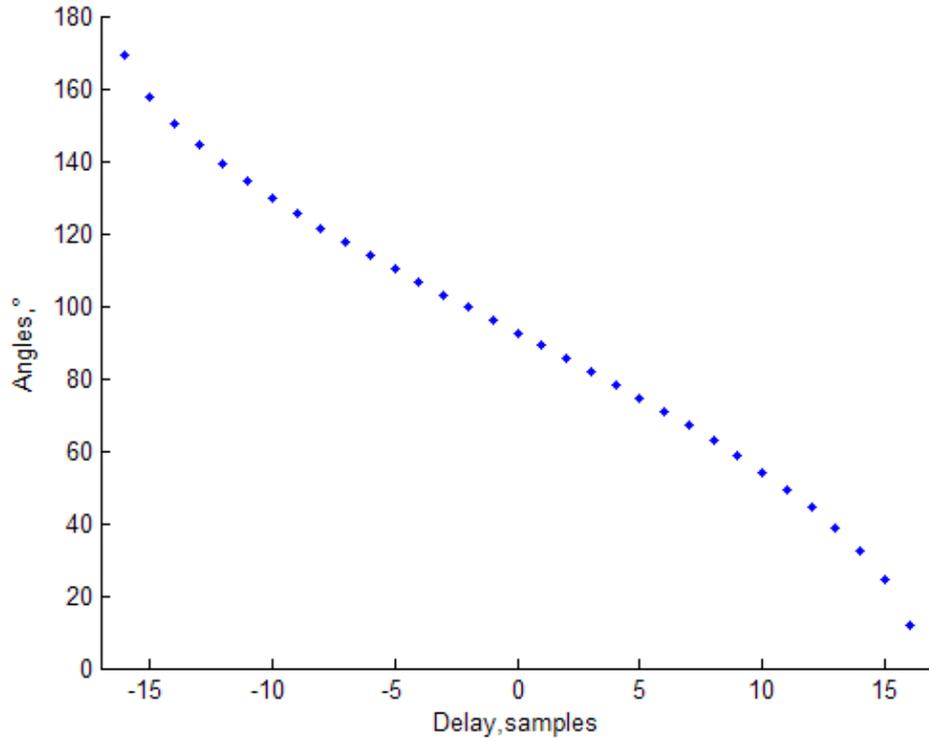


Figure 9. All possible angles using equidistant array of delays.

To resolve this problem, equations (3.6) – (3.8) were combined to determine the optimal array of delays τ_b in a way that, opposed to equidistant time delays, it would cover the whole range of possible angles α equidistantly:

$$\tau_b = \frac{(\pm\sqrt{d^2 + b^2 + 2db \cos \alpha} - b)F_s}{c} \quad (3.16)$$

Equation (3.16) is a function of angle α , which results in Figure 10. As can be noticed in the figure, the obtained array of delays now satisfies the requirement of covering the whole array of angles equidistantly.

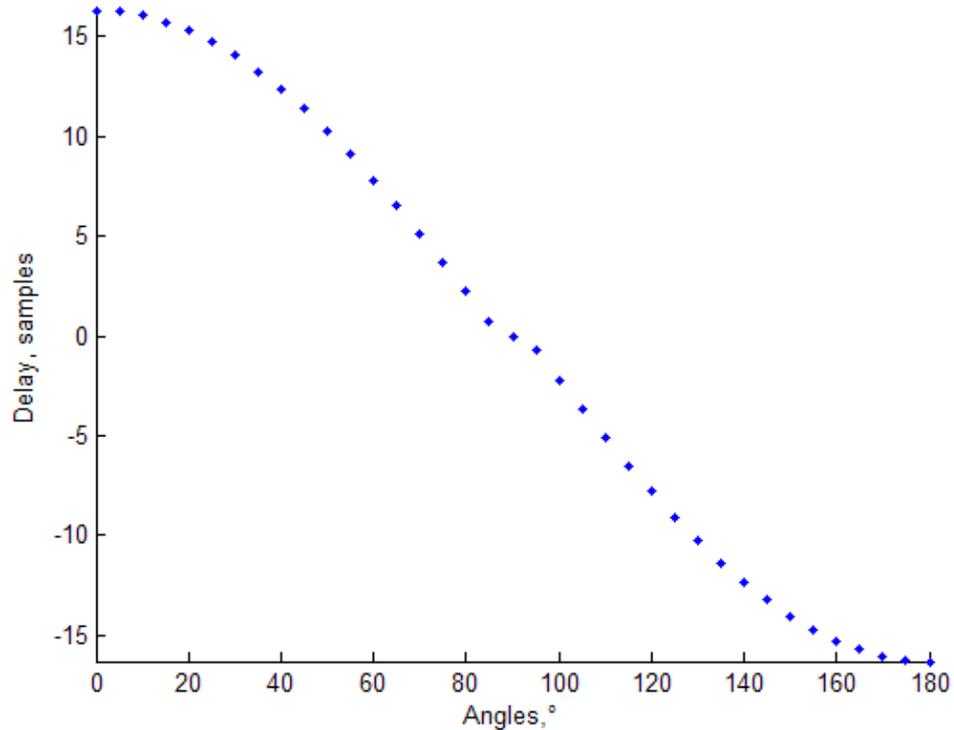


Figure 10. Possible detectable angles using the optimal array of time delays, which are calculated with equation (3.16).

3.3.2 Adjustment of subbands

The second enhancement concerns the width of subbands used for division of frequencies of an incoming audio signal. Figure 11 presents a normalized power spectral density (PSD) of a speech signal, which was used as a sample in the experiment. It is visible that the power of smaller frequencies is much higher comparing to high frequencies. There are some peaks in the high frequencies, but they can be explained by additional environmental noise.

Initial proposition in the basic algorithm was to use the Bark scale. The Bark scale divides a frequency plane into subbands so that frequencies that are perceived by human hearing as one frequency are divided into the same subband. Such approach can be justified by the need of making audio manipulations in a way that they would be undistinguishable for human hearing. One example of such audio manipulation is converting audio signal captured by multimicrophone setup to binaural audio signal, which could be achieved by using Head-Related Transfer Function [30, pp. 283–302].

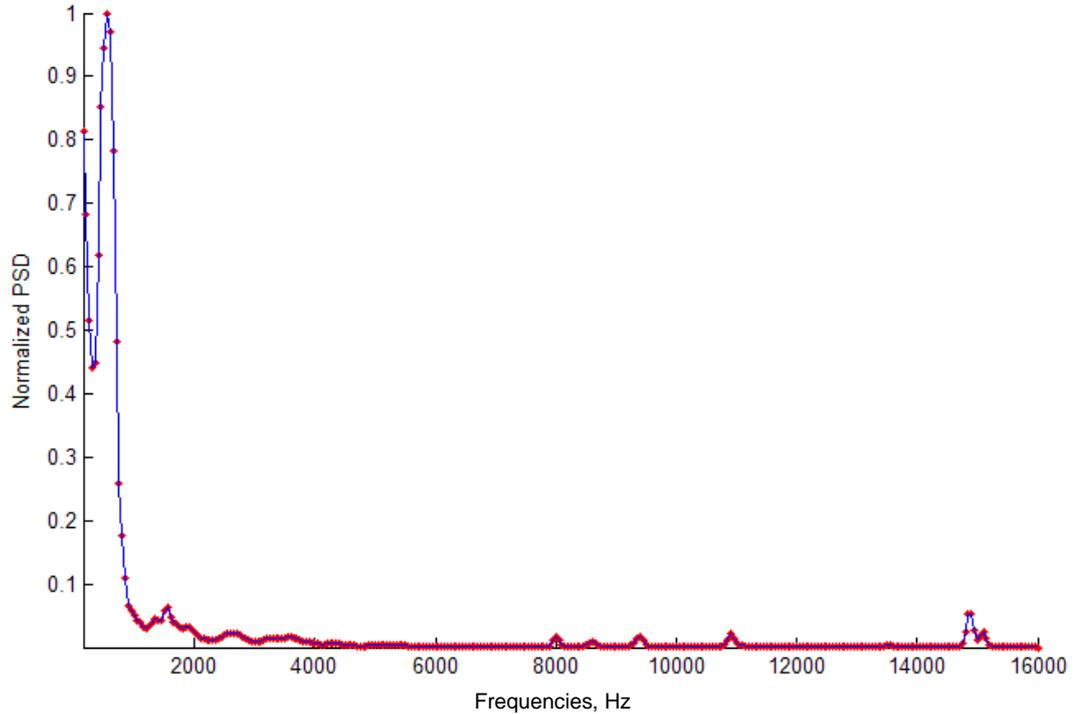


Figure 11. Normalized PSD of an average speech signal used during the experiment.

Overall using Bark scale for subband division gives sufficient results, producing correct time delays for different subbands. Nevertheless, it has been observed that subbands consisting of high frequencies do not produce large values of signal power, as calculated from the real part of equation (3.4):

$$\text{Re}(X_{2,\tau_b}^b * X_3^b), \tau_b \in [-D_{max}, D_{max}]. \quad (3.17)$$

Practically that leads to overlooking a portion of directional information.

To avoid such behavior, an array of subbands was created by splitting the entire frequency band into divisions that produce equal power values when processing an average speech or transient signal. Figure 12 displays the power magnitudes different subbands, when this division was applied to a speech signal.

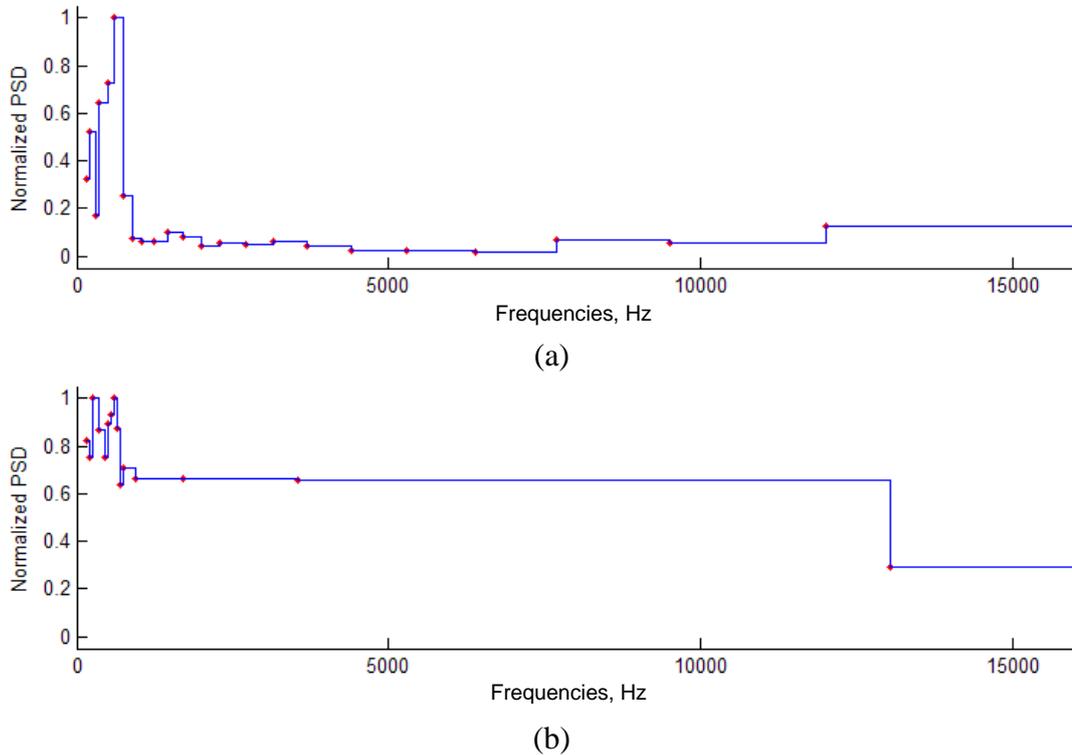


Figure 12. Normalized PSDs for different subband arrays applied to the speech signal. (a) The result of utilizing Bark scale; (b) the result of utilizing suggested scale. Each red point represents the beginning of a new subband; blue steps represent width of a subband.

3.3.3 Smoothing of DOA estimation

The last step of the algorithm returns direction of the dominant sound source of a particular frequency subband. The purpose of the following enhancement was to prepare that information for further visualization. This was executed by smoothing of received data. Two histograms were created: an angle-of-arrival histogram $H_{D,n}[\varphi]$ and a magnitude histogram $H_{M,n}[\varphi]$. $H_{D,n}[\varphi]$ is computed for the current time index n by counting the number of frequency subbands that have the angle φ as the assigned direction and normalized by the total number of frequency subbands. $H_{M,n}[\varphi]$ is computed for the current time index n by finding the frequency subbands, that have φ as direction of signal arrival, and then summing the corresponding values of power of the frequency subbands calculated by equation (3.17). It is advised to use decibel scale for $H_{M,n}[\varphi]$.

The changes in the histograms of the angle-of-arrival and magnitudes can be rapid from frame to frame, therefore angle-of-arrival histogram is slowed down using leaky integrator:

$$\langle H_{D,n}[\varphi] \rangle = \beta_H \cdot \langle H_{D,n-1}[\varphi] \rangle + (1 - \beta_H) \cdot H_{D,n}[\varphi], \quad (3.18)$$

where β_H is the forgetting factor and $\langle \rangle$ is a time-averaging operator. A good value for β_H is selected from range of 0.9 and 0.95. For the magnitude histogram similar formula is used:

$$\langle H_{M,n}[\varphi] \rangle = \beta_H \cdot \langle H_{M,n-1}[\varphi] \rangle + (1 - \beta_H) \cdot H_{M,n}[\varphi], \quad (3.19)$$

Finally, the two histograms are merged by using following equation:

$$\langle H_n[\varphi] \rangle = \alpha_H \cdot \langle H_{D,n}[\varphi] \rangle + (1 - \alpha_H) \cdot H_{M,n}[\varphi]. \quad (3.20)$$

It is worth noting, that it is better to assign value α_H in a way that the contribution of neither of the histograms will be eliminated in equation (3.20). In this thesis α_H value is assigned to 0.88.

Above explained enhancements give significant improvements to the result in cases when the algorithm is applied only on human speech signals. One disadvantage of smoothing directional information is possible loss of directional data of transient signals, such as claps or finger snaps. Therefore having an extra test for checking if the signal is a transient signal completes the algorithm. This test is easy to implement. The second enhancement split the frequency band into subbands with equal power values in case of a speech signal. The difference between a speech signal and a transient signal is that power of a transient signal is high for most frequencies. If power values of subbands with high frequencies are much higher than power values of subbands with low frequencies, it means that the signal is a transient signal. If the transient signal is detected, an additional visualization step is triggered, making the transient signal visible after the smoothing.

3.4 Automated sound source tracker

To evaluate the final algorithm and test it in real life situations, a video tracker system was built to follow the dominant sound source with a video camera. This system would be useful in applications such as video conferencing.

The challenge of this task was to build an equipment that would follow a dominant sound source mechanically and point a video camera viewing to the direction of the dominant sound source. The enhanced algorithm was used to develop an application for a desktop computer, and a video tracker system was built and connected with the computer application. The built system consisted of an Arduino microcontroller (a single-board microcontroller), a stepper motor, and a web camera. A generic web camera with 74 degree angle of view was used.

Reasons to use Arduino board were its ease and elegance of program designing. Arduino already has proven itself as a great instrument for different kind of projects ranging from simple school projects to extremely complicated projects [31]: Arduino projects can be stand-alone or communicate with software running on a computer. This microcontroller

is able to sense the environment by receiving input from a variety of sensors, as well as controlling its surrounding by controlling lights, motors and other actuators. [32], [33]

Communication with the computer was established using USB connection. Values of dominant sound source were pushed to microcontroller through serial port, and Arduino turned the stepper motor with attached web camera to the correct direction. Correct direction was assigned as the direction at which the dominant speech source is visible, i.e. within the viewing angle of the web camera. In case transient signals were present in the field of view of the camera, the area of estimated DOA of this signal was marked on the video taken by the web camera. Figure 13 shows how all elements of the system communicate with each other.

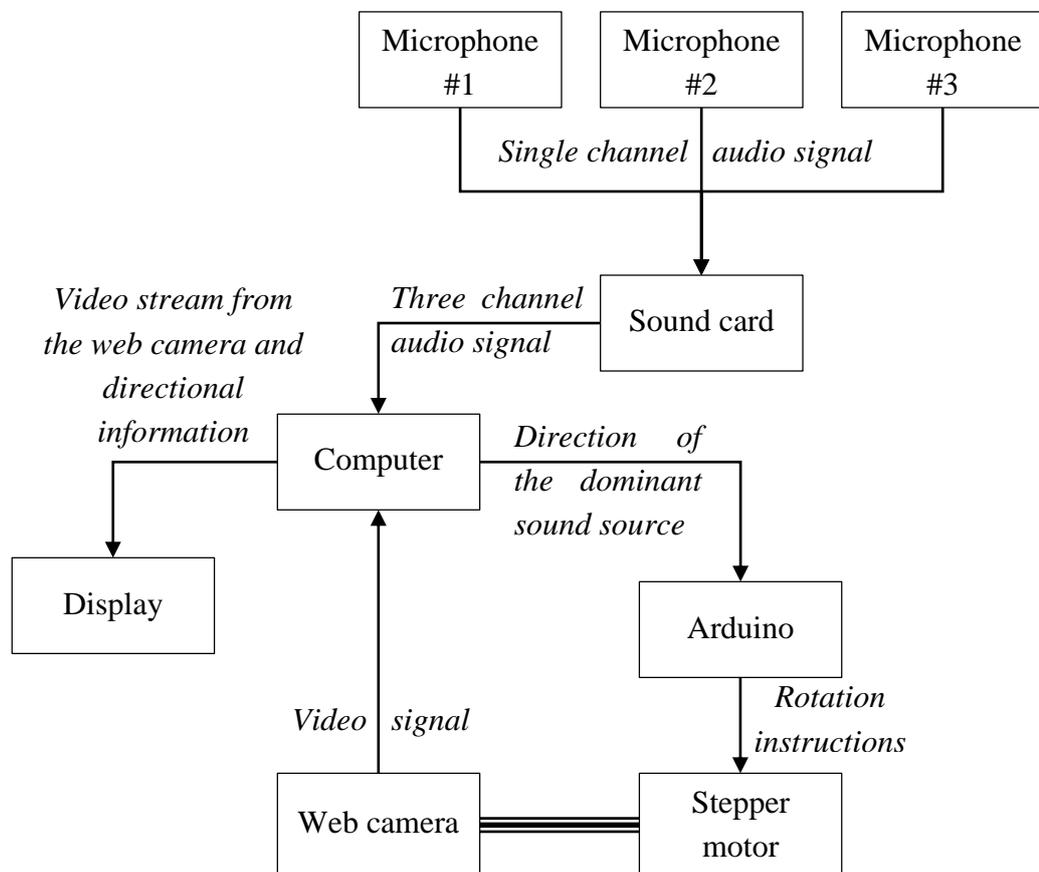


Figure 13. Flow chart of the signals in the built automated sound source tracker.

4. RESULTS AND DISCUSSION

In this chapter, the proposed algorithms are put to a performance test. To compare working abilities of the basic algorithm and its enhanced version, they were first compared with the GCC PHAT algorithm in time delay estimation task. Results of the time delay estimations are presented in chapter 4.1. The GCC PHAT algorithm was chosen for comparison, because it is one of the most commonly used TDE based algorithms, and GCC PHAT is considered to be the most robust method when the SNR is moderate [34].

After that, the basic algorithm and its enhanced version were compared for their ability to estimate the angle of a sound source (chapter 4.2). To compare the performance of the algorithms, they were tested on three types of signals: static location of a sound source, a moving sound source and transient signals. In most cases the tested signals had a SNR value of approximately 15 dB.

In chapter 4.3, the computational complexity of the basic, enhanced and GCC PHAT algorithms are assessed. Lastly, functioning of the built automated sound source tracker is demonstrated in chapter 4.4.

4.1 Time Delay Estimation

In order to compare results of the enhanced algorithm with that of the GCC PHAT algorithm, the values of time delays used with the enhanced algorithm had to be downscaled in order to match resolution of time delays used with GCC PHAT. In other words, knowing values of the signal sampling frequency and the distance between microphones, and applying those to equation (3.1), it gives the maximum time delay equal to 17 samples.

It is also worth noting, that results for the proposed basic algorithm and its enhanced version were scaled to decibel scale for displaying purposes. However, such scaling is not sufficient enough for transient signals. Therefore, cube root scaling was later applied to prove that the proposed algorithms are able to spot transient signals.

Figure 14 shows the performance of the three algorithms on a statically located speech source. The signal source was placed in front of the microphone array, therefore expected time delay was 0. The GCC PHAT algorithm gives a very clean result of delays. Result of the basic algorithm appears to be noisy. However, it is visible that the developed algorithms are able to highlight timeframes when the actual speech was present, as well as reflect the correct time delay, although with partial scattering. Result of the enhanced

algorithm shows less noise and more concentration around the expected time delay, with much higher peaks.

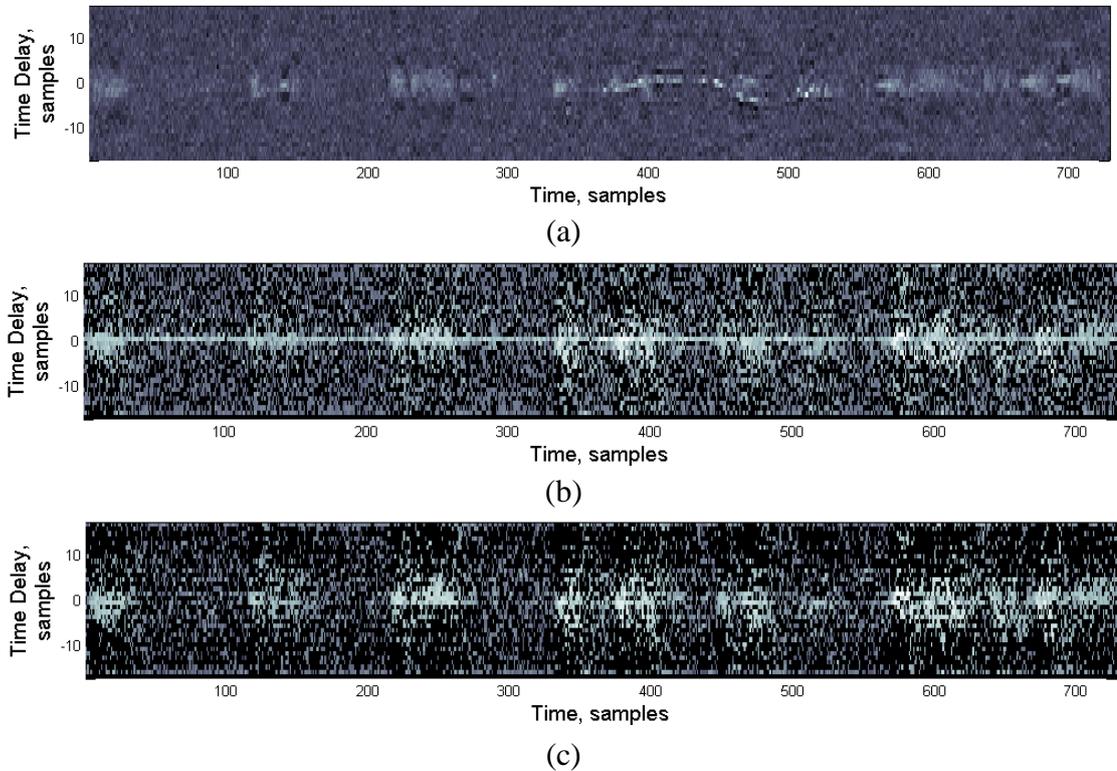


Figure 14. TDE of algorithms applied to a speech signal which originates from a static location in front of the microphone array. (a) GCC PHAT algorithm; (b) the basic algorithm; (c) the enhanced algorithm.

Similarly, Figure 15 presents results of applying algorithms to different signals coming from static sound sources. In this example, signals with different SNR values were tested. On the left column a signal was coming from the side of the microphone setup and had SNR value of 8 dB. The expected value for a time delay was -7 samples. It is visible that even PHAT gives a poor result, which can be justified by the low SNR. The proposed algorithms also give poor results. However, results of the enhanced algorithm have similar allocations as results of GCC PHAT considering expected time delay of -7 samples. On the right column, a signal was coming from the back of the microphone array, therefore expected time delay was 0 samples, and the SNR value was approximately 15 dB. Results are similar to ones presented in Figure 14.

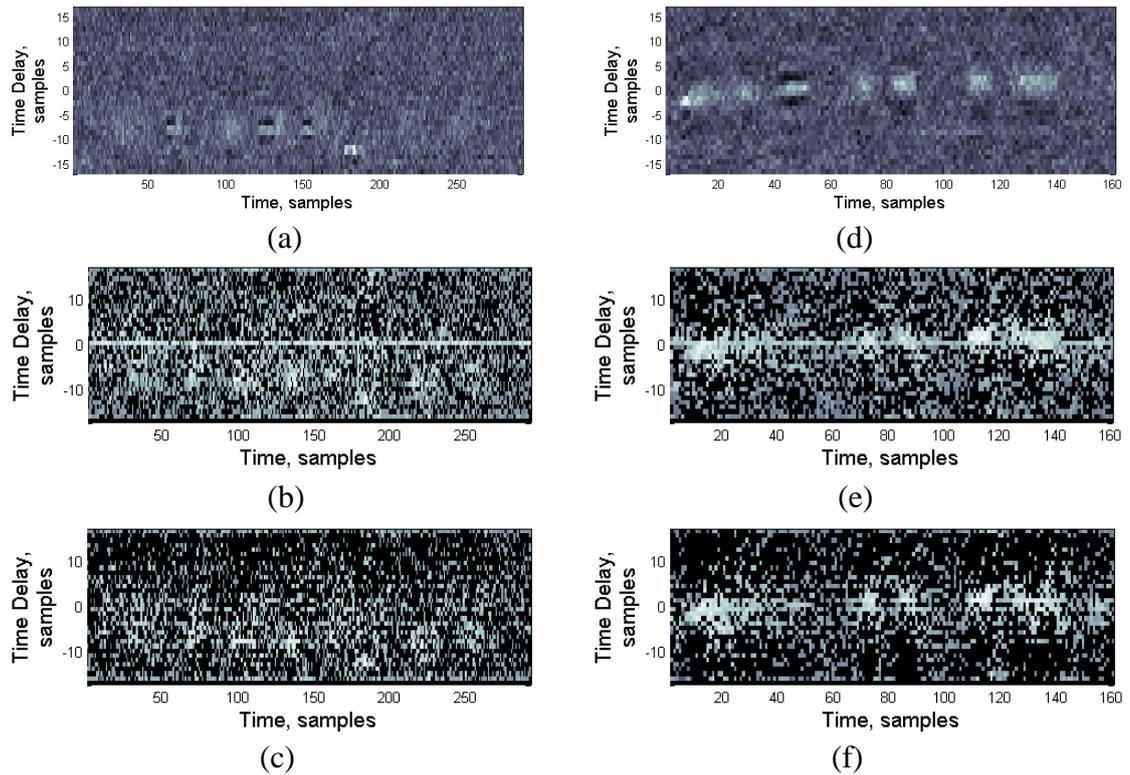


Figure 15. TDE of algorithms applied to speech signals originating from static sound sources. (a), (d) GCC PHAT algorithm; (b), (e) the basic algorithm; (c), (f) the enhanced algorithm. (a)-(c) Time delay estimation for a speech signal coming with a delay of -7 samples; (d)-(f) time delay estimation for a speech signal coming from the back of the microphone array.

Additionally, a different scaling was used for tests presented in Figure 15. Instead of decibel scaling, cube root scaling was applied (Figure 16). It is visible that the results became less noisy and time delay of the arriving signal is more distinguishable. However, later on, when angle of signal arrival was calculated, it was discovered that using decibel scale provides better DOA estimation. Therefore, decibel scale is still used to calculate angles of arrival, and cube root is merely used to visualize the difference between GCC PHAT, which does not require additional scaling.

Figure 17 and Figure 18 show results of time delay estimation for moving signals. Two experiments were conducted: speech signal source was moved clockwise around the microphone setup; and speech signal source traveled counterclockwise around the microphone setup. Similarly to static sound source experiments, results of proposed algorithms seem noisier. Nevertheless, it is visible that results of the enhanced algorithm are more precise, although far from the results of GCC PHAT algorithm.

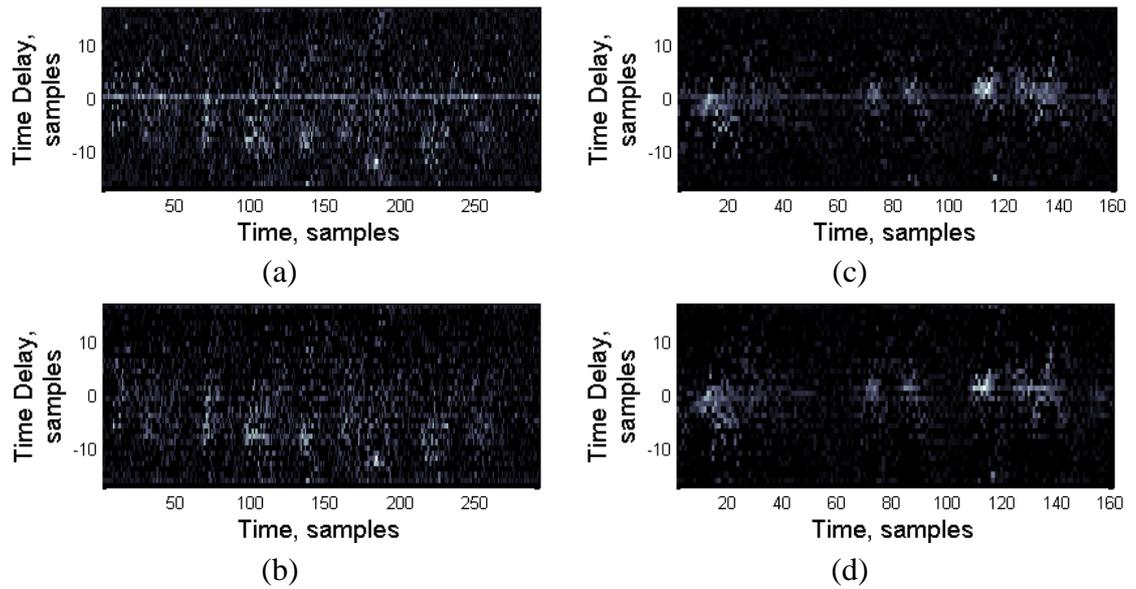


Figure 16. TDE of the proposed algorithms using cube root scaling applied to speech signals originating from static sound sources. (a), (c) The basic algorithm; (b), (d) the enhanced algorithm. The speech signal is coming with a delay of -7 samples (a),(b) or from the back of the microphone array (c), (d).

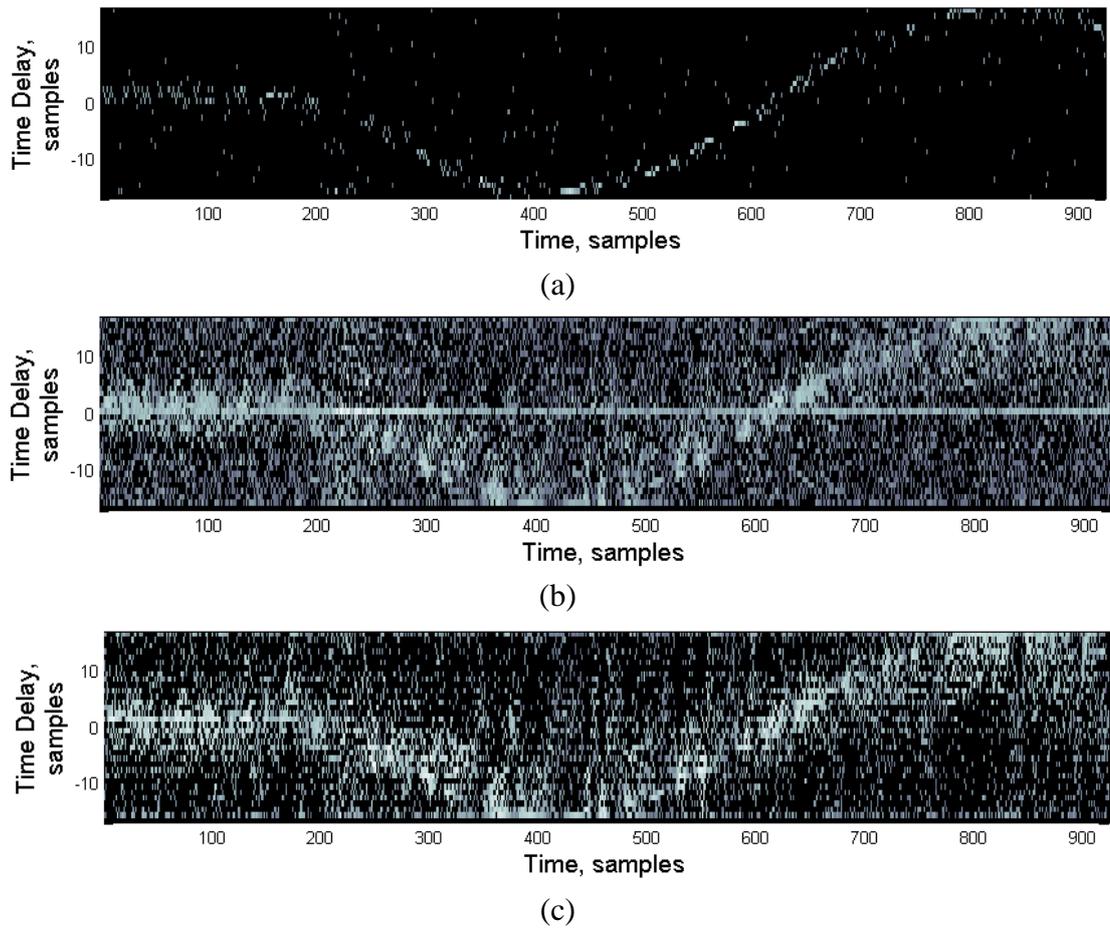


Figure 17. TDE of applying algorithms to moving sound signals. (a) GCC PHAT algorithm; (b) the basic algorithm; (c) the enhanced algorithm. The sound source is moving around the microphone array clockwise.

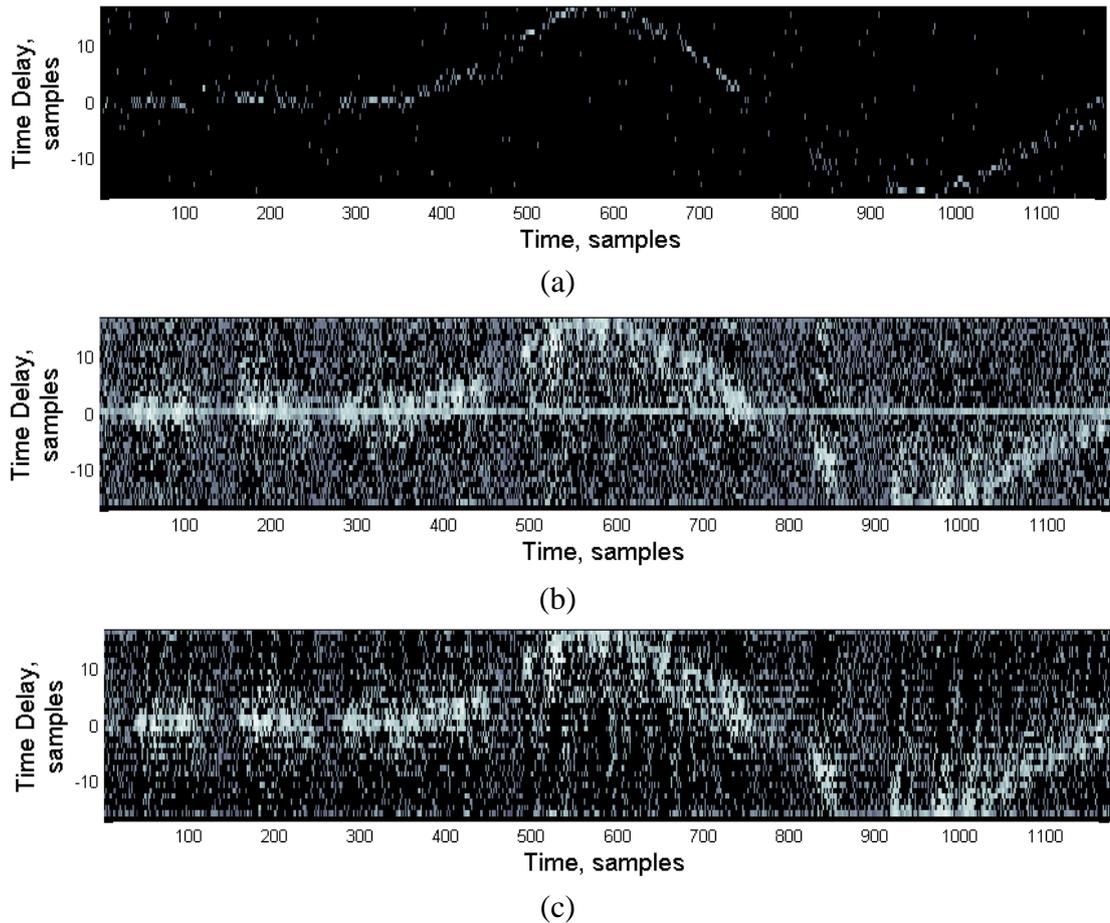


Figure 18. TDE of applying algorithms to moving sound signals. (a) GCC PHAT algorithm; (b) the basic algorithm; (c) the enhanced algorithm. The sound source is moving around the microphone array counterclockwise.

Results of handling transient signals are presented in Figure 19. As it was mentioned before, using decibel scale for transient signals does not properly visualize the true efficiency of the proposed algorithms. Hence, cube root was used, and the results are presented in Figure 20. The effect of using these different scales is visible by comparing Figure 19 (b) and Figure 20 (b) for the basic algorithm and Figure 19 (c) and Figure 20 (c) for the enhanced algorithm. With cube root scaling, peaks are very sharp and hardly any noise is seen.

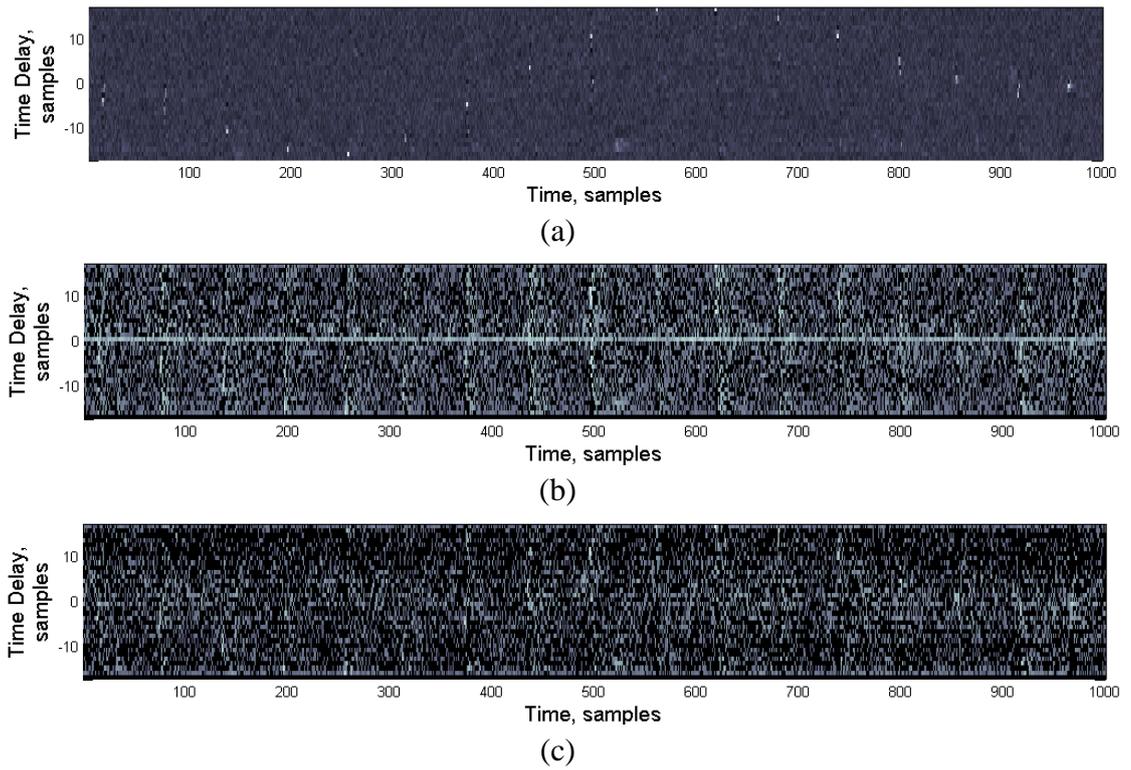


Figure 19. TDE of the algorithms applied to transient signals. (a) GCC PHAT algorithm; (b) the basic algorithm scaled to decibel scale; (c) the enhanced algorithm scaled to decibel scale.

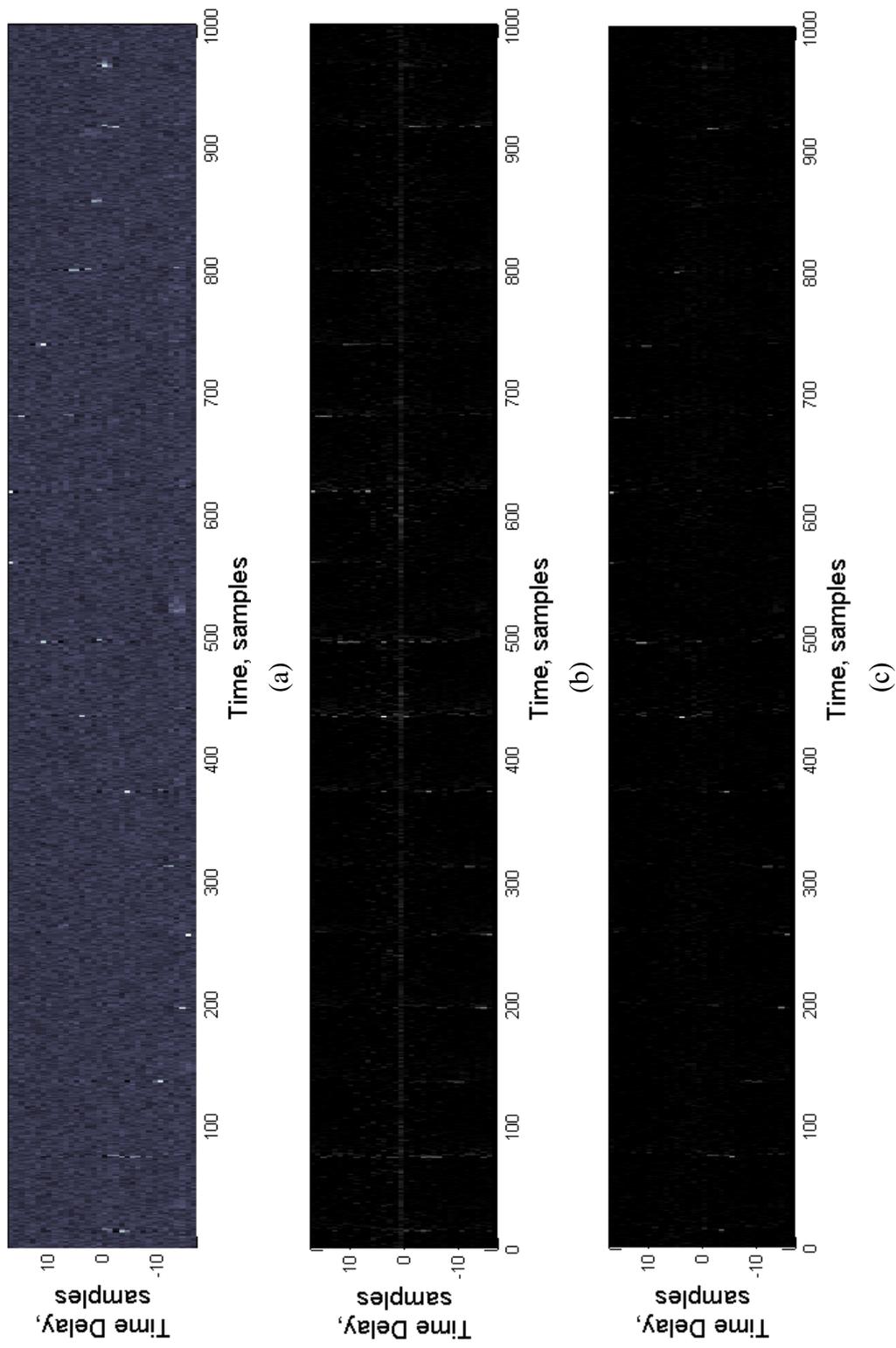


Figure 20. TDE results of the GCC PHAT algorithm (a); the basic algorithm (b) and the enhanced algorithm (c) applied to transient signals, scaled with cube root.

Having sharp peaks is desired, but in fact peaks in Figure 20 are so sharp that they are difficult to notice in this presentation format. Therefore, the same results are visualized in three-dimensions, the axis being time delay, time and normalized power (Figure 21- Figure 23). Only a part of TDE results are shown in these 3D figures in order to keep the figures clear. These parts correspond to the areas in Figure 20 from time point 350 to 1000 samples and time delay -4 to 17 samples.

Figure 21-Figure 23 show clear peaks at the time points and time delays of transient signal appearance. In ideal situation all other power values should be equal zero, identifying absence of any signal. However, in case of GCC PHAT (Figure 21) these power values are elevated to 0.3, while power values of the proposed algorithms are preserved close to zero. Yet Figure 22 shows elevation of power values around time delay of 0 samples over whole time period, which suggests presence of the signal coming from the front or the back of the microphone array. Although, as shown from results of the GCC PHAT and the enhanced algorithms there is no signal. Similar behavior has been observed in the previous test results of the basic algorithm as well.

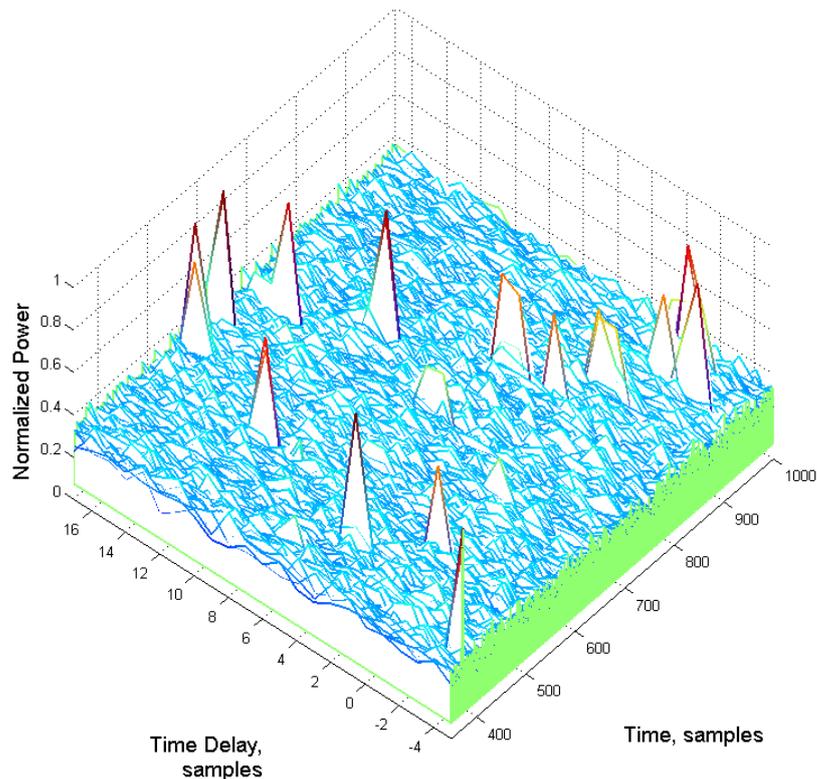


Figure 21. A part of the TDE result of the GCC PHAT algorithm in 3D applied to transient signals.

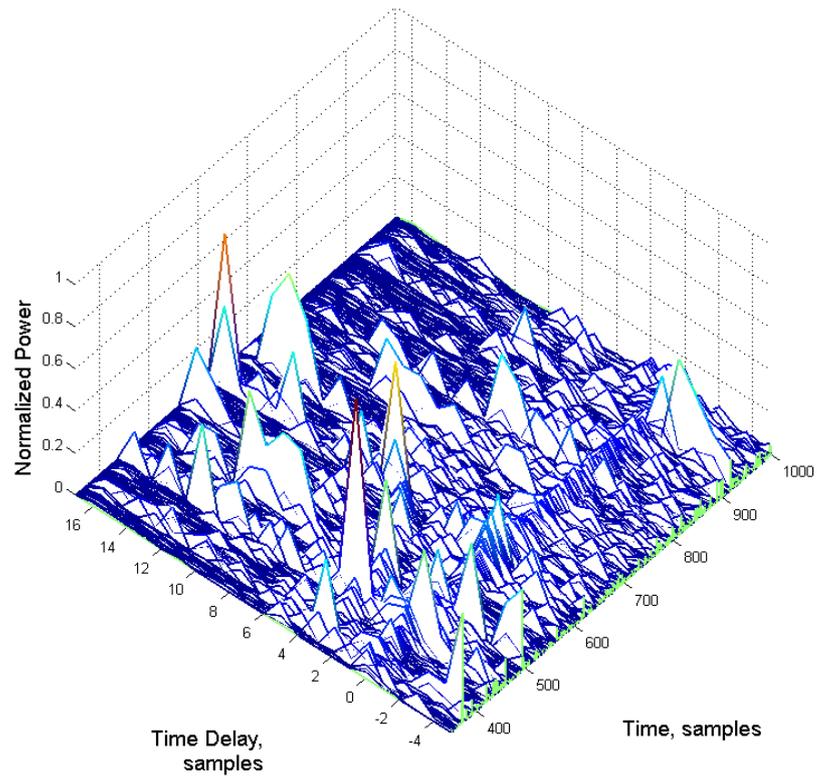


Figure 22. A part of the TDE result of the basic algorithm in 3D applied to the transient signals and scaled with cube root.

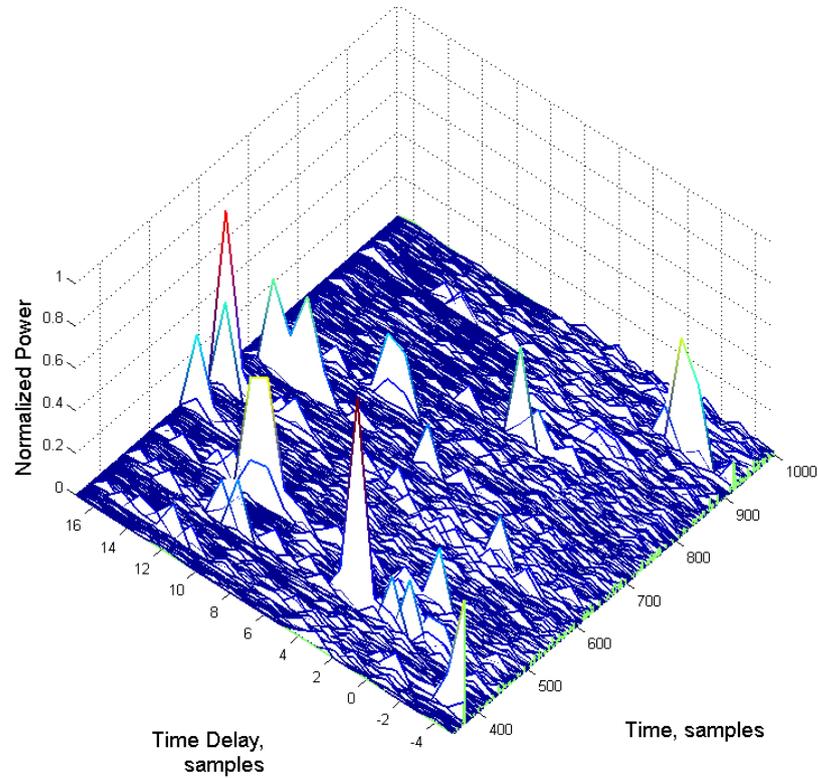


Figure 23. A part of the TDE result of the enhanced algorithm in 3D applied to the transient signal and scaled with cube root.

It is reasonable to conclude that the processing results of transient signals with the proposed algorithms, especially with the enhanced algorithm, exceed the result of GCC PHAT: the results are less noisy, and peaks are much sharper.

4.2 Direction of Arrival Estimation

The next phase was to present the efficiency of proposed algorithms by calculating directions of incoming speech signals. It was done in a similar way as all experimental data before: static speech signal source first, then moving speech signal source. Signals from Figure 14, Figure 15 (d) – (f) and Figure 18 were used for that purpose. Zero degrees is assigned as the direction in front of the microphone array. An angle value is increasing by moving in clockwise direction around the microphone array.

Figure 24 contains the direction estimation for the signal, the time delay of which was inspected in Figure 14. In Figure 24 and other following figures, odd panels should be compared between each other, same as even panels with each other. As before, the result of the enhanced algorithm seems better: there are few misestimations. A good example of misestimation occurs in the range from 210th sample to 280th sample. The basic algorithm points to the direction assigned as the back of the microphone array several times, while the enhanced algorithm is able to keep pointing to the correct direction constantly.

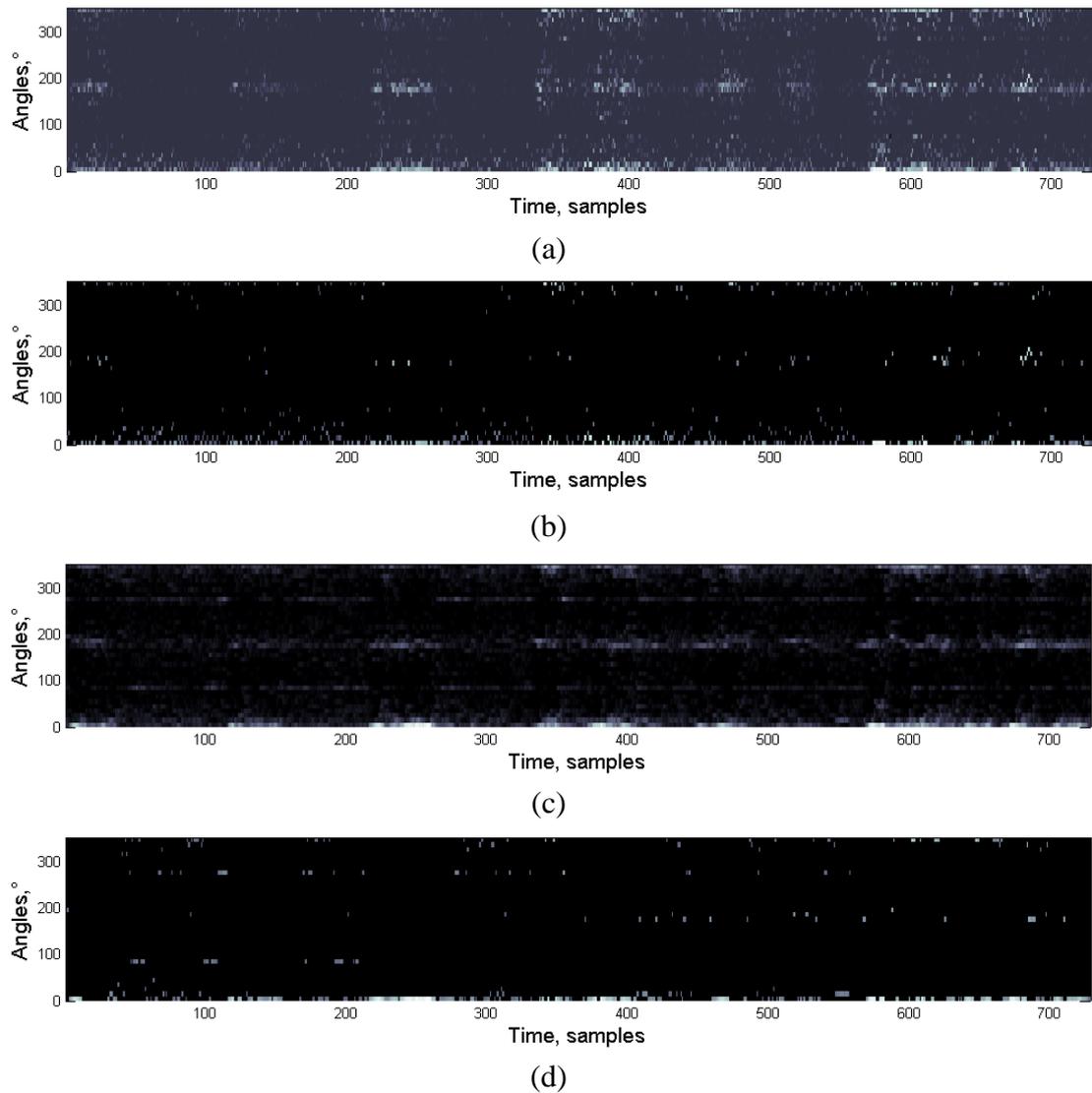


Figure 24. Estimation of DOA angle for static sound source placed in front of the microphone array. (a) Results of the basic algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (b) Results of the basic algorithm to estimate the dominant sound source direction. (c) Results of the enhanced algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (d) Results of the enhanced algorithm to estimate the dominant sound source direction.

Perhaps superiority of the enhanced algorithm to the basic algorithm is more evident in Figure 25. The DOA angle is estimated for the signal used in Figure 15 (right column), emitted from a static sound source. The result of the estimation of the dominant sound source looks less scattered around the expected angle.

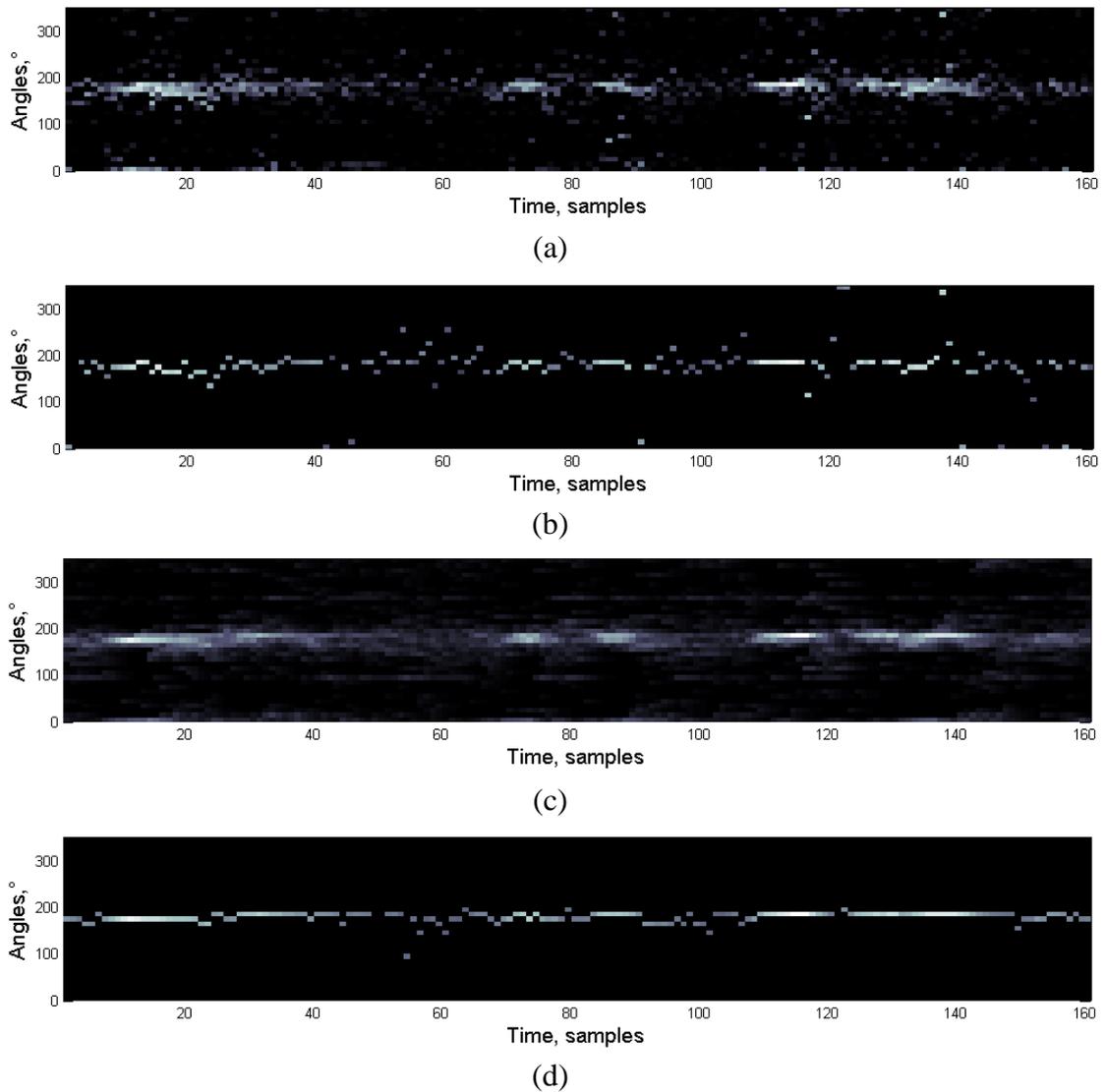


Figure 25. Estimation of DOA angle for static sound source placed behind of the microphone array. (a) Results of the basic algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (b) Results of the basic algorithm to estimate the dominant sound source direction. (c) Results of the enhanced algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (d) Results of the enhanced algorithm to estimate the dominant sound source direction.

To finish the comparison of efficiencies between the basic algorithm and the enhanced algorithm, the moving signal from Figure 17 (b), (d), (f) was used. Results of DOA angle estimation are presented in Figure 26. As expected from experiments presented before, the enhanced algorithm is performing better. Estimated directions are following the path of the dominant speech signal source precisely.

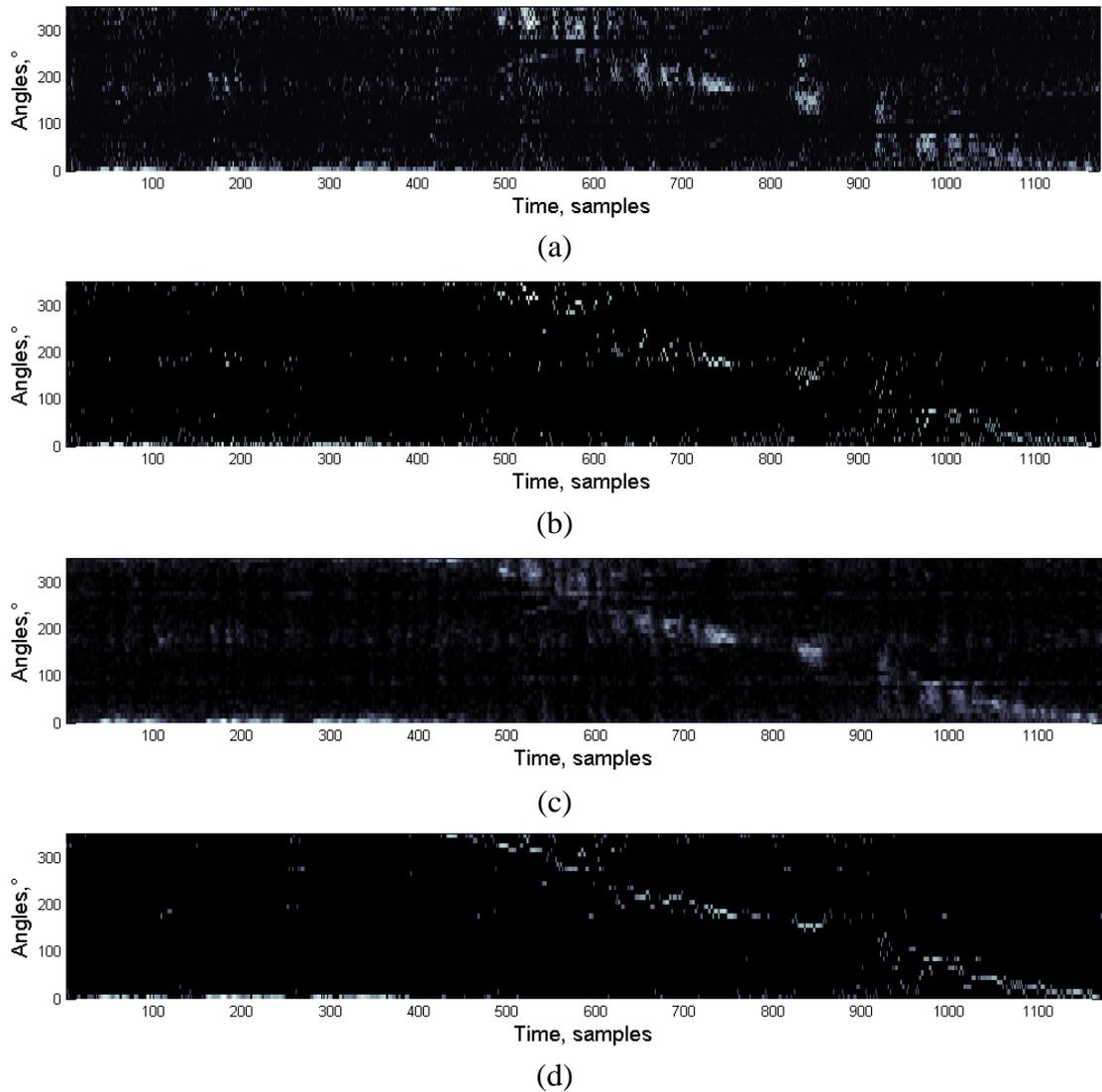


Figure 26. Estimation of DOA angles for a moving sound source. (a) Results of the basic algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (b) Results of the basic algorithm to estimate the dominant sound source direction. (c) Results of the enhanced algorithm without eliminating signals coming from directions other than direction of the dominant sound source. (d) Results of the enhanced algorithm to estimate the dominant sound source direction.

In conclusion of the experimental part, it is fair to say that estimation of the dominant sound source location with the proposed algorithm and especially its enhanced alternative is feasible. It might not top the GCC PHAT algorithm in estimating the direction of the dominant speech signal source, however, it gives particularly better results when applied to a transient signal.

4.3 Computational complexity

First, computational complexity of TDE task of the proposed algorithm is compared with that of the GCC PHAT algorithm, similarly as above. Second, computational complexity of DOA estimation task of the basic algorithm is compared with that of the enhanced algorithm. It will be done using “big-O” notation [35, p. 44] and list of frequencies requires to perform this evaluation are presented in Table 1.

Table 1. List of quantities required for computational complexity evaluation.

Variable	Description
N_{W1}	size of the signal array used in GCC PHAT, in practice equals to the length of Hamming window
N_{W2}	size of the signal array used in the proposed algorithms; it equals to the sum of the Hamming window length and D_{max}
N_D	size of the array of delays used in the proposed algorithms; the basic and the enhanced algorithms have different value of that variable
N_{SB}	number of the subbands used in the proposed algorithms; the basic and the enhanced algorithms have different value of that variable

Some useful observations about the quantities used for the evaluation are the following:

- size of the signal array used in GCC PHAT is less or equal than that of the proposed algorithms;
- size of the array of delay used in the basic algorithm is constant and equals 33, and that of the enhanced algorithm is arbitrary, however, in scope of this thesis, number of the time delays was chosen equal to 36;
- in general, number of the subbands used in the proposed algorithms is between 1 and $N_{W2}/2$ inclusive.

Table 2 list all operations that are executed by GCC PHAT for TDE after acquiring a single frame of the incoming signal. The respective information for the proposed algorithms is presented in Table 3.

Table 2. Computational complexities of operations included in TDE with the GCC PHAT algorithm.

Operation	Computational complexity
Fourier transformation of the incoming signals	$O(N_{W1} \log N_{W1})$ [36, p. 386]
Complex conjugate of the signals (equation (2.12))	$O(N_{W1})$
Denominator of the frequency weighting function (equation (2.15))	$O(N_{W1})$
Division of complex conjugate by frequency weighting function	$O(N_{W1})$
Inverse Fourier transformation of previous step result	$O(N_{W1} \log N_{W1})$

Table 3. Computational complexities of operations included in TDE with the proposed algorithms.

Operation	Computational complexity
Fourier transformation of the incoming signals	$O(N_{W_2} \log N_{W_2})$
Complex conjugate of the signals (part inside the brackets of equation (3.4) without shifting signals)	$O(N_{W_2})$
Shifting conjugated signals and acquiring real part of it	$O(N_{W_2} N_D)$
Looking for optimal delay	$O(N_D N_{SB})$

The final computational complexity of the GCC PHAT algorithm for TDE is sum of all its components in “big-O” notation, which gives $O(N_{W_1} \log N_{W_1})$. Similar analysis for the proposed algorithms results in $O(N_{W_2} \log N_{W_2} + N_{W_2} N_D)$. To get rid of the sum in the last formula it is required to estimate which summand is greater. Value of N_D in the basic algorithm is equal to 33, and that in the enhanced algorithm is 36 (depending on the precision demand). Value of $\log N_{W_2}$ is not more then 6.5. That means that final computational complexity of the proposed algorithms for TDE becomes $O(N_{W_2} N_D)$. Because signal array in case of the proposed algorithms was extended with additional zeroes, value of $\log N_{W_1}$ will be even less than 6.5. By comparing computational complexity of GCC PHAT and proposed algorithms, it is concluded that GCC PHAT produces result the TDE result with computational complexity of about 5 times smaller.

Table 4 shows the rest of the operations executed for DOA estimation with the proposed algorithms after TDE is complete. As it was defined above, N_{SB} is always smaller than N_{W_2} (maximum $N_{W_2}/2$), and thus “big-O” notation results in final computational complexity of $O(N_{W_2})$. This means that the basic and proposed algorithms have the same computational complexity when estimating angle of arrival from already known delay.

Table 4. Operations included in the rest of the proposed algorithms and their computational complexities.

Operation	Computational complexity
Calculating sum signal (equation (3.5))	$O(N_{SB})$
Calculation of the angles (equations (3.9)-(3.13))	$O(N_{W_2})$
Smoothing values of angles from the previous step (applied only for the enhanced algorithm)	$O(N_{SB})$

After determining computational complexity the proposed algorithm for both parts of DOA estimation, they can be summed and the computational complexity of the basic and enhanced algorithms can be compared. The sum, $O(N_{W_2} N_D + N_{W_2})$ is equal to $O(N_{W_2} N_D)$, which means that total computational complexity depends only on size of the signal array and size of array of delay. However, taking into account that size of signal array is the same for the basic and enhanced algorithms, it leaves only the size of array of delays as the differing quantity. As was said above, values of the size of array of delays are very similar to each other in the basic and enhanced algorithms (in this thesis, 33 and

36, respectively). For that reason, using the enhanced algorithm is completely justified, especially talking into account that performance of the enhanced algorithm exceeds that of the basic algorithm.

4.4 Automated sound source tracker

In the developed desktop application for automated sound source tracker, the DOA of a sound source was visualized using a function similar to wind rose (Figure 27). In this visualization method, the DOA was shown for all sound signals, not only the dominant sound source. The surrounding of the microphone array was divided so that DOA estimation had a resolution of 10° , creating 36 beams. The length of each beam corresponds to the volume of sound signals detected from the respective direction. As defined earlier, the dominant sound source is the loudest sound, therefore the direction of the longest beam indicates the direction of the dominant sound source. In case of transient signals, the beam in the respective DOA was shown as a brief highlight in different color.

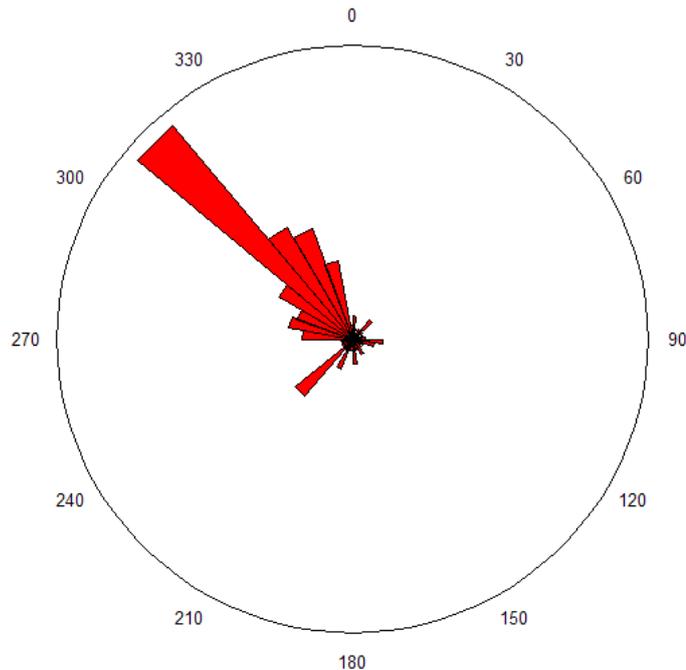


Figure 27. Visualization of the DOA estimation without applying limitation of dominant sound source. The volume of incoming signals from each 36 directions is shown as the length of the corresponding beam.

Sending the information of DOA angle of the dominant sound source to the Arduino board succeeded to turn the web camera to the direction of dominant speech signal, if the camera was not already pointing at the correct direction. In addition to the wind rose diagram, the desktop application showed video captured by the web camera. Figure 28 illustrates with one timeframe, what the video footage looked like in practice in the application, when speech signal was detected. Furthermore, the direction of the dominant sound source (person speaking) was marked with dashed red lines. Similarly, in case

transient signals were present in the field of view of the camera, the area of estimated DOA of this signal was briefly visualized on the video.



Figure 28. Image taken by the camera of the video tracking system. Red marks the direction, where dominant sound is coming from.

Similar sound source tracker was built by Garg et al. [37]. However, this team was primarily aiming for a non-expensive system, rather than a system that would track speaker in real time. Their system uses one microcontroller for audio processing and a rotating camera. In case of the system presented in this thesis, audio processing was completed on relatively fast computer, and only after that commands were sent to Arduino microcontroller just to inform it about a new angle of the dominant sound source. As result, the system tracks a sound source in real time (it is able to compute new values of angles in under 0,02 seconds), while system by Gang et al. tracks a speaker within 10 degrees of their location in less than 3 seconds. [37, p. 1680]

iCam system, which was mentioned in introduction, was tracking a speaker (in case of that system, a lecturer and audience of the lecture room) using only video signal. In the next iteration of this system, iCam2, audio processing was added for DOA estimation. [38] This system utilizes two pan/tilt/zoom cameras besides the microphone array, situated in the opposite sides of the lecture hall, one being close to the lecturer. Cost of implementation of this system is very high and it is used for the lecture recording/broadcasting purposes. For these reasons, it would be difficult to achieve similar results with the system presented in this thesis. The automated tracker which was built in this thesis, is meant to be used in closer distance to speaker. If the system is used in a lecture hall, it would most probably be able to perform, but requires shorter distance to speaker comparing to iCam2. Additionally, lecture halls usually have high reverberation and the built system was not tested in such an environment.

5. CONCLUSION AND FUTURE WORK

During the course of this Master's thesis different methods for DOA estimation were studied. An algorithm was proposed for DOA estimation that falls into the class of TDE based methods. A TDE based method was chosen among the classes because of its ease of implementation. The proposed algorithm is based on the GCC methods, with the difference of dividing the frequency plane into subbands using Bark scale. Along with the proposed algorithm, additional enhancements were implemented: time delays were recalculated; subbands were altered from Bark scale to optimal scale for speech signal; and additional smoothing was implemented for visualization purposes.

Time delays were recalculated in order to cover all possible directions of a signal arrival. The optimal subband division was found by modifying Bark scale so that subbands containing higher frequencies were able to have contribution in TDE, otherwise they were too insignificant comparing to subbands with lower frequencies. Lastly, smoothing was applied to eliminate short-term scattering of the DOA.

Efficiency of the proposed algorithms were compared to that of GCC PHAT by comparing TDE of all algorithms. Scattering around the true time delay of the arrived signal depends on the scaling function that was used during visualization for comparing purposes. Cube root scaling turned out to be the best to illustrate TDE efficiency of the proposed algorithms. In that case results were very close to results of GCC PHAT. However, results of the proposed algorithms were only superior when the algorithms were applied to transient signals. Decibel scale, which was used later for calculating DOA angles, gave more noisy impression. However, it did not affect the DOA angle estimation, due to the smoothening function of the enhanced algorithm. When comparing efficiency of the basic algorithm and enhanced algorithm, the results show that the made enhancements improve the performance of the algorithm, especially by decreasing scattering.

It was noticed that the best performance of DOA angle estimation is achieved on the stationary signal sources for both proposed algorithms. Nevertheless, both algorithms were capable of maintaining constant following of the moving signal source.

The efficiency of the proposed algorithms was correlated with SNR, as expected. Same behavior was observed using GCC PHAT. Good results were obtained when the algorithms were applied to signals with SNR value equal to 15 dB, and signals with SNR 8 dB did not give acceptable results. In this research relation of the algorithm performance to SNR was not conducted systematically, and this could be done in the future research.

The computational complexity of the basic and enhanced algorithms were comparable. However, the computational loads of the proposed algorithms were many times higher than the load of GCC PHAT.

The algorithm could be further enhanced by adding a function that would adapt the subbands depending on the current incoming signal type and signal distribution in the frequency plane. The automated signal source tracker may be improved by adding video processing to the system. Video processing would be able to detect movements, such as gestures and lip movements. However, video processing would be limited to the camera's angle of view. Therefore initial DOA would have to be estimated by audio processing. Alternatively, for video conferencing cameras with 360 degree view angle can be used to detect possible speakers, or separate people in the room; and that system can be combined with estimation of dominant sound source so that only the speaker would be shown to the participants of the video conference. Even so the idea is similar to the system already built in this thesis, video processing might add precision to DOA estimation.

Another interesting utilization of the algorithm that can be investigated in future is a mobile phone application that would allow the phone to recognize sound signals arriving from different directions. The basis for creating such mobile application is in understanding that future phones would accommodate three microphones. An application estimating DOA of an incoming sound signal could be used as a new way of communicating with a phone, e.g. unlocking the phone with consecutive finger snaps from different directions.

Such an application for mobile phones was shortly looked into during the thesis work alongside the computer application for automated video tracker. The enhanced algorithm was used to develop the mobile application for Windows operating system and results did not differ from results of the computer version, indicating that the algorithm works in a mobile platform. However, at the time of writing the thesis, no mobile phones existed with three microphones built in. Therefore, the mobile application was tested with already prerecorded signals. Taking into account that microphones on future mobile phones would be able to pick up sound according to assumptions used in this thesis, such as microphones being omnidirectional, the algorithm would most probably work. Nevertheless, real-life testing and implementing of the mobile application has to be left to future research, when suitable mobile phones with three built-in microphones exist.

REFERENCES

- [1] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*. Berlin Heidelberg: Springer Berlin Heidelberg, 2001, p. 402.
- [2] B. Langner and A. W. Black, "Improving the Understandability of Speech Synthesis by Modeling Speech in Noise," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, 2005, vol. 1, pp. 265–268.
- [3] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 2578–2581.
- [4] Q.-G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Commun.*, vol. 18, no. 4, pp. 317–334, Jun. 1996.
- [5] G. W. Elko, "Adaptive noise cancellation with directional microphones," in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.
- [6] M. L. Seltzer, "Microphone Array Processing For Robust Speech Recognition," Carnegie Mellon University, 2003.
- [7] S. Oh, V. Viswanathan, and P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array," in *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, vol. 1, pp. 281–284.
- [8] B. Mrazovac, M. Z. Bjelica, I. Papp, and N. Teslic, "Smart audio/video playback control based on presence detection and user localization in home environment," in *Proceedings - 2011 2nd Eastern European Regional Conference on the Engineering of Computer Based Systems, ECBS-EERC 2011*, 2011, pp. 44–53.
- [9] C.-E. Chen and K. Yao, *Classical and Modern Direction-of-Arrival Estimation Ch9 Source and Node Localization in Sensor Networks*. Elsevier, 2009, pp. 343–383.
- [10] A. Sutin, B. Bunin, A. Sedunov, N. Sedunov, L. Fillinger, M. Tsionskiy, and M. Bruno, "Stevens passive acoustic system for underwater surveillance," in *2010 International Waterside Security Conference, WSS 2010*, 2010, pp. 1–6.
- [11] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," vol. 19, no. 7, pp. 780–785, 1997.

- [12] C. Zhang, Y. Rui, L. W. He, and M. Wallick, "Hybrid speaker tracking in an automated lecture room," in *IEEE International Conference on Multimedia and Expo, ICME 2005*, 2005, vol. 2005, pp. 81–84.
- [13] Y. Huang, J. Benesty, and G. W. Elko, "Microphone Arrays for Video Camera Steering," in *Acoustic Signal Processing for Telecommunications*, S. L. Gay and J. Benesty, Eds. Kluwer Academic Publishers, 2000.
- [14] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing for Telecommunications*, S. L. Gay and J. Benesty, Eds. Kluwer Academic Publishers, 2000.
- [15] P. M. Peterson, N. I. Durlach, W. M. Rabinowitz, and P. M. Zurek, "Multimicrophone adaptive beamforming for interference reduction in hearing aids," *J. Rehabil. Res. Dev.*, vol. 24, no. 4, pp. 103–110, Jan. 1987.
- [16] K. Kokkinakis and P. C. Loizou, "Multi-microphone adaptive noise reduction strategies for coordinated stimulation in bilateral cochlear implant devices," *J. Acoust. Soc. Am.*, vol. 127, no. 5, pp. 3136–44, May 2010.
- [17] B. Van Veen and K. M. Buckley, "Beamforming: a Versatile Approach to Spatial Filtering," *IEEE ASSP Mag.*, pp. 4–23, 1988.
- [18] B. Van Veen and K. M. Buckley, "Beamforming Techniques for Spatial Filtering," in *The Digital Signal Processing Handbook*, 2nd ed., CRC Press LLC, 2009, pp. 1 – 22.
- [19] M. S. Brandstein, "A Framework for Speech Source Localization Using Sensor Arrays," Brown University, 1995.
- [20] J. Zhang, M. Walpola, D. Roelant, H. Zhu, and K. Yen, "Self-organization of unattended wireless acoustic sensor networks for ground target tracking," *Pervasive Mob. Comput.*, vol. 5, no. 2, pp. 148–164, Apr. 2009.
- [21] Y. Yuan, B. Zhang, D. Fan, and G. Tong, "DFT and PSD for estimating DOA with an active acoustic array," in *Proceedings of the IEEE International Conference on Automation and Logistics, ICAL 2008*, 2008, pp. 694–699.
- [22] S. Bjorklund and L. Ljung, "A review of time-delay estimation techniques," *42nd IEEE Int. Conf. Decis. Control (IEEE Cat. No.03CH37475)*, vol. 3, no. December, pp. 2502–2507, 2003.
- [23] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech Signal Proc.*, vol. 24, pp. 320–327, 1976.
- [24] Y. T. Chan and K. C. Ho, "A Simple and Efficient Estimator for Hyperbolic Location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, 1994.

- [25] G. Y. G. Yan, L. D. L. Daoben, and Z. Q. Z. Qishan, "The principle and high-speed algorithm for hyperbolic radiolocation in digital cellular networks," in *WCC 2000 - ICCT 2000. 2000 International Conference on Communication Technology Proceedings (Cat. No.00EX420)*, 2000, vol. 2, pp. 1314–1318.
- [26] W. H. Foy, "Position-Location Solutions by Taylor-Series Estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, no. 2, pp. 187–194, 1976.
- [27] B. Friedlander, "A passive localization algorithm and its accuracy analysis," *IEEE J. Ocean. Eng.*, vol. 12, no. 1, pp. 234–245, 1987.
- [28] S. Wang, D. Sen, and W. Lu, "Subband Analysis of Time Delay Estimation in STFT Domain," in *11th Australian International Conference on Speech Science & Technology*, 2006, pp. 211–215.
- [29] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," *4th Int. Conf. Signal Process. Commun. Syst. ICSPCS'2010 - Proc.*, pp. 26–29, 2010.
- [30] S. L. Gay and J. Benesty, Eds., *Acoustic Signal Processing for Telecommunication*. Springer US, 2000, p. 338.
- [31] M. Schwartz, *Arduino home automation projects : automate your home using the powerful Arduino platform*. Packt Publishing, 2014, p. 132.
- [32] R. Slyper, J. K. Hodgins, R. Slyper, and J. Hodgins, "Action capture with accelerometers," in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2008, pp. 193–199.
- [33] J. D. Brock, R. F. Bruce, and S. L. Reiser, "Using Arduino for introductory programming courses," *J. Comput. Sci. Coll.*, vol. 25, no. 2, 2009.
- [34] J. M. Perez-Lorenzo, R. Viciano-Abad, P. Reche-Lopez, F. Rivas, and J. Escolano, "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments," *Appl. Acoust.*, vol. 73, no. 8, pp. 698–712, Aug. 2012.
- [35] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction To Algorithms*. MIT Press, 2001, p. 1180.
- [36] L. R. Rabiner and B. Gold, *Theory and application of digital signal processing*. 1975, p. 777.
- [37] S. Garg, S. Tiwari, S. S. Chauhan, S. Singh, and S. Ahmad, "Rotating Camera Based on Speaker voice," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 2, no. 5, pp. 1674–1681, 2013.
- [38] C. Zhang, Y. Rui, J. Crawford, and L. He, "An Automated End-to-End Lecture Capture and Broadcasting System," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 4, no. 1, pp. 1–23, 2008.