Mikko Parviainen

Sound source separation in real environments using two sensors

Master of Science Thesis

The subject was approved by the department council on 8th May 2002.
Examiners: Prof. Jaakko Astola
MSc Tuomas Virtanen
Preface

This work is the outcome of the project which was offered to me while working as a research assistant in the institute of signal processing at Tampere university of technology. Audio research group (ARG) has provided an inspiring working atmosphere and great facilities enabling research work. I would like to thank Mr. Anssi Klapuri who offered me the possibility to take my first steps in signal processing.

I wish to thank professor Jaakko Astola for his valuable advice and comments. Especially I would like to express my gratitude to Mr. Tuomas Virtanen for his advice. Without his dedicated attitude to the guidance, the writing process would have been almost mission impossible. Thank you Tuomas!

Tampere, May 2003

Mikko Parviainen
Insinööriinkatu 66 B 48
33720 Tampere
e-mail: mikko.p.parviainen@tut.fi
Contents

1 Introduction 7
  1.1 System overview .................................................. 8
  1.2 Organization of this thesis ........................................ 9

2 Spatial hearing 11
  2.1 Role of outer ear and head in spatial hearing ............... 13
    2.1.1 Head-related transfer function ......................... 13
    2.1.2 Interaural time difference and interaural intensity difference .... 15
  2.2 Localization ability ............................................ 16
    2.2.1 Localization of one sound source ....................... 16
    2.2.2 Lateralization ............................................... 19
    2.2.3 Localization of multiple sound sources ............... 20
  2.3 Signal separation ability of human hearing (Cocktail party effect) .... 22
    2.3.1 Selective attention ......................................... 23
    2.3.2 Effects of spatial information – physiological views and experiments on sound separation .... 26
    2.3.3 Auditory scene analysis .................................... 27

3 Computational models of spatial hearing and human sound source separation 32
  3.1 Sound localization models in horizontal plane .............. 33
  3.2 Binaural models applied to solve cocktail-party problem ... 37
  3.3 Auditory scene analysis models applied to solve cocktail-party problem 40
  3.4 Other sound source separation schemes ...................... 47

4 Mid-level representation 51
  4.1 Desirable properties of mid-level representations ........... 52
  4.2 Sinusoids + noise modeling .................................... 53
  4.3 Deterministic + stochastic model as a mid-level representation ... 54
  4.4 Parameter extraction of sinusoidal modeling ................ 55
    4.4.1 Peak picking ................................................. 55
    4.4.2 Cross-correlation method .................................. 56
    4.4.3 Peak picking and Cross-correlation – properties and comparison .. 58
  4.5 Transient modeling .............................................. 61
  4.6 Transient detection ............................................. 62
Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Electrical Engineering
Institute of Signal Processing

Parviainen, Mikko:
Sound source separation in real environments using two sensors
Master of science thesis, 109 pages, 7 enclosure pages
Examiners: Prof. Jaakko Astola, MSc Tuomas Virtanen
Financial support: Tampere University of Technology, Institute of Signal Processing
May 2003
Keywords: Sound source separation, direction-of-arrival estimation, spatial hearing, mid-level representation, sinusoidal modeling, transient detection

In this work a sound source separation system operating in real-world acoustic environments is proposed. The separation is based on the estimation of the spatial origin of the sound source. Sound source separation in general refers to signal processing techniques, the primary goal of which is to isolate one or several sound sources from a mixture signal which contains desired sound sources and undesired sound sources. Human being is able to perform this task tremendously well. The task is often referred to as cocktail-party problem induced by the fact that one is able to have a conversation with a person in such difficult acoustic conditions.

The selection of the method for sound source separation is constrained by the measurement configuration which consists of two normal high-quality microphones. The system in this work is based on the primary hypothesis that the spatial origin of the sound source is a major factor affecting the separation ability of the system. The overall system consists of two separate parts. The first part is for finding the direction-of-arrival (DOA) of the strongest sound source in the environment. The second part performs the separation of the sound source using DOA and a mid-level representation of the signal.

The simulations performed with the system showed that the separation is possible using the selected approach. Both subsystems operated in the desired manner with most of the signals types. Demonstration signals can be found at

Tiivistelmä

TAMPEREEN TEKNILLINEN YLIOPISTO
Sähköteknikko-osasto
Signaalinkäsittelyn laitos

Parviainen, Mikko:
Aänilähteen erottelu luonnollisissa ympäristöissä käyttäen kahta sensoria
Diplomityö, 109 sivua, 7 liitesivua
Tarkastajat: Prof. Jaakko Astola, DI Tuomas Virtanen
Rahoittajat: Tampereen teknillinen yliopisto, signaalinkäsittelyn laitos
Toukokuu 2003
Avainsanat: aänilähteiden erottelu, tulosuunnan määrittely, suuntakuulo, välitason esitysmuoto, sinimalli, transienttien havainnointi


List of acronyms and symbols

ASA  auditory scene analysis
BMLD  binaural masking level difference
BSS  blind source separation
CASA  computational auditory scene analysis
DOA  direction-of-arrival
DFT  discrete-time Fourier transform
HRIR  head-related impulse response
HRTF  head-related transfer function
IID  interaural intensity difference
ITD  interaural time difference
LMS  least mean square
MAA  minimum audible angle
MLD  masking level difference
MSE  mean-square error
SNR  signal-to-noise ratio
STFT  short-time Fourier transform

\( \mathbb{E}\{\cdot\} \)  expectation value operator
\( \Delta_{\text{min}}\{\cdot\} \)  minimum change operator
\( \text{var}\{\cdot\} \)  variance operator
Chapter 1

Introduction

Signal separation - The decomposition of a complex sound into separate parts as it would appear to human listener [18] -Daniel PW Ellis

In this thesis the ultimate goal is to search for a technique to build a system capable of sound source separation in auditory circumstances corresponding everyday life conditions. As an example of sound source separation consider listening to a person in some noisy room in which the radio is playing music with very loud volume, if the listener prefers to listen only the person or the radio, one of these sources is separated from the mixture in which both sources are present.

The concept of signal separation, however, has a more general meaning than the one reflected by the example. It is also used in communication signal processing. For instance, a communication link may need some sort of separation unit to recover the desired signal from the received noisy mixture. A radar must be able to detect the previously transmitted original pulse. However, in this thesis the term separation refers to the separation of audio signals. More precisely, the frequency range of signals is assumed to be suitable for human hearing system.

Partly due to the fact that the ultimate goal in all signal separation, regardless of the application, is virtually the same, a wide range of possible solutions are available. In fact, some of the solutions have clearly adapted ideas from completely different areas. However, in general the ideas developed for a certain application are not directly applicable in some other. Instead, the approaches in different research areas provide merely a source of ideas for new solutions.

The sound source separation has received a lot of attention over the years. Solely in audio signal processing, several drastically differing models have been proposed. One of the motivating factors is to increase the knowledge of human hearing. But there are also numerous applications directed to everyday life of people and to other scientific purposes in which a system capable of separating the desired sound source is useful.

For a man-made separation system the information by which it should be able to perform the separation is a sum signal, or mixture, the constituents of which are signals arising from different sources. In many cases the mixture consists not only of “clean” source signals but also background noise. From this viewpoint it is
easy to understand that picking only one of the signals in the mixture introduces a very challenging task. Despite the fact that area of sound source separation has been studied for decades, still, the currently employed schemes are merely rather theoretically oriented. This is to say, very few systems have been proposed that are intended to operate in the same kind of situations as in the example above.

It must be pointed out that for some of the existing schemes for sound source separation, the goals are somewhat different from the situation as above. In the example, the goal is merely to amplify one of the sound sources and attenuate the other rather than completely isolate one signal from the mixture. This aspect basically divides the research area of separation into two quite drastically differing parts.

Including the differences in the ultimate goals, major differences on which the algorithms are based can be recognized within each part. Human hearing system has a tremendous ability to separate sound sources in rather challenging auditory conditions. The first class of separation methods can be viewed to consist of systems that try to model this ability. The second class of methods initially arise from communication signal processing. The methods of the first class aim at both goals: “the amplification goal” and “the isolation goal”. In the second class, the latter goal is considered.

In this thesis the primary target is to search for a solution aiming at sound source separation with a specified microphone setup. Two conventional microphones attached to a rod 10 cm apart. The rod is at the top of a microphone stand adjusted to 1 m in height. This receiver is placed in three rooms the auditory properties of which differ from each other. Using the signals recorded by the microphones, the system tries to separate the strongest source from mixture signals. The aspect in this sound source separation system is in trying to amplify the desired signal merely than isolate the source of interest completely. The ideas for the system are explored in all the areas of signal separation.

1.1 System overview

The main hypothesis in this work is that the spatial information related to sound sources is the main cue based on which the separation can performed in real-world environments. The hypothesis, as well as, the solutions for sound source separation arise from the previous research done in the area, meaning, no fundamentally new ideas are proposed but the existing knowledge and approaches are utilized and ideas from different research areas are merged. The current knowledge and some proposed approaches are discussed in Chapter 2 and Chapter 3. The selection of the approach to meet the primary goal of sound source separation is made based on the fact that the receiver unit and the conditions are basically fixed in advance. The recording setup is explained in detail in Chapter 6.

The sound source separation system in this thesis consists of the following main steps: (1) The locations of sound sources are determined, (2) input mixture is modeled
using sinusoids + transients modeling, (3) the components of the modeled mixture are grouped based on the location information, and finally, (4) the components regarded as arising from the desired sound source are synthesized. This process is illustrated in Figure 1.1.

The location information in this thesis is restricted to determining the direction-of-arrival of a sound source in horizontal plane. Due to the fact that the system is utilized in real-world environments, the location of sound sources should be determined in three planes: horizontal plane, median plane and frontal plane (see Figure 1.2). The estimation of the horizontal angle $\varphi$ and the elevation angle $\delta$ uniquely define DOA in three dimensional space. Yet, to determine the unique location of a sound source the distance $r$ has to be estimated.

This is the initial evaluation of sound source separation based on the architecture in Figure 1.1. $\delta$ and $r$ are neglected at this phase. Thus, sound sources are assumed to exist exactly on horizontal plane, and yet, only one sound source is placed at certain horizontal angle. The estimation of horizontal angle is based on the time delay between the left and right channel signals.

Signals are modeled using a mid-level representation. The mid-level representation that is utilized enables signal separation by presenting it in a form which, in turn, enables the usage of the grouping criterion. The grouping in this system is based on DOA estimate, that is, time delay between left and right channel. Using the estimated time delay, the elements of the mid-level representation are grouped to sources which can be synthesized separately.

Despite the fact that sound source separation is the ultimate goal, using direction-of-arrival as a primary cue, the DOA subsystem has to be selected carefully. The work done for this thesis is thus directed to two areas, (1) choosing and implementing DOA subsystem and (2) employing signal modeling techniques that are believed to perform the sound source separation in the desired manner in given circumstances.

1.2 Organization of this thesis

The solutions selected in this thesis are largely based on the findings in the area of human hearing. The knowledge concerning human hearing stressing the issues essential to this work are discussed in Chapter 2. The physical factors as well as the
current knowledge in the area of psychoacoustics are presented. The aspect in this chapter is in stressing the mechanisms that are believed to be involved in the sound source localization ability and sound source separation ability in human auditory system. Additionally, the methods by which the knowledge is achieved are referred to in each section. A lot of useful terms are introduced and defined, which may help while studying the referred literature.

The systems modeling human hearing are presented in Chapter 3. This chapter follows the guidelines of the previous chapter. These two chapters are related in such a manner that the former presents the knowledge of an ability of human hearing and the latter presents the modeling techniques of the ability.

Chapter 4 introduces signal modeling techniques relevant to this work. However, the chapter provides also a general view of the techniques. The methods utilized in this work are presented in a way they are implemented in the system.

In Chapter 5 the overall system is described thoroughly. It is divided into two parts. In the first part the DOA subsystem is described. The second part is dedicated to the separation subsystem.

Since the system is directed to operate in real-world environments, appropriate measurement data is needed to perform the final evaluation of the methods. Chapter 6 is dedicated to the description of the audio measurements. The signals, the environments, the hardware and the procedures are discussed. It is maybe useful to consult this chapter while evaluating the results since some important issues that were observed during the measurements are pointed out.

The discussion on the results in Chapter 7 is two-fold. The performance of DOA subsystem is described in various auditory conditions. The second part consists of the evaluation of how well the system is able to separate sound sources based on the information provided by DOA subsystem.

Chapter 8 concludes this thesis by summarizing the observations concerning DOA subsystem and the overall system. Some ideas concerning the further development are suggested.
Chapter 2

Spatial hearing

In order to understand the solutions that try to solve the problem discussed in this thesis a short review into human auditory system is needed. The aim is not in trying to exhaustively explain how the human auditory system works but the focus is on the ability known as spatial hearing.

Human auditory system can be roughly categorized as in Figure 2.1 “Signal processing of human auditory system” embraces all the signal processing, that is, the alteration process of the original source signal, or merely the alteration process of the sound waves which they undergo while propagating from the spatial location of the sound source to left and right ear. The alteration arises from: (1) the space in which a sound is heard. (2) Human body affects the propagation of sound waves in a certain manner. (3) Auditory periphery does the final processing on that path. Left branch includes physiology of human hearing which is a research area concentrated on studying the actual physical structures and their role, covering outer ear, middle ear and inner ear. Other parts of this branch includes, for instance, the effects of the auditory environment. Figure 2.2 illustrates this chain of alterations.

“Psychology of human auditory system” refers to high-level processes that take place on the auditory cortex of human brain. The research that is conducted in this area is referred to as psychoacoustics. Alterations caused by this part that enable and assist spatial hearing are not known exactly. However, the existence of some processes and structures has been successfully proved.

Both parts of the auditory system have a significant role in spatial hearing. It is even hard to say which part is more significant.

To explain the meaning of spatial hearing the definitions of two terms is needed. Two core concepts of spatial hearing are presented in Table 2.1. A stick hitting
drum membrane is an example of a sound event. Music played via a loudspeaker produces multiple sequential sound events. Often auditory events occur as a result of sound events. Drum-sound would cause a listener to perceive an auditory event resulting from sound event. A phenomenon called tinnitus, instead, is an auditory event that is not a result of any sound event.  

Let us get back to the definition of spatial hearing. “The concept of spatial hearing embraces the relationships between the locations of auditory events and other parameters - particularly those of sound events, but also others such as those related to the physiology of the brain.” [5] p. 2-3. The definition is complete but difficult to understand without any background knowledge. A more relaxed definition or explanation might be: “Spatial hearing is about locating sound sources and sensing the space in which the sound sources are, or as they are perceived to be placed”. The last part of this definition refers to the fact that the perception of location of sound sources and true location of sound sources do not necessarily coincide.

Spatial hearing has an important role in analyzing and recognizing the auditory environment, and yet, recognizing and isolating objects in environments. Also it assists the navigation ability of human being. Despite the fact that vision is much more accurate and faster, for instance, in 3D-localization of sound events and in the tasks mentioned above, spatial hearing provides at least supplementary information thus helping, for instance, to resolve possible contradictions. Still, 3D-localization can be based solely on spatial hearing [32]. It must be pointed out that this type of “deaf-3D-localization” by using visual information is actually a resultant effect of experience and vision. For instance a person seeing a door closing expects a certain type of auditory event from that direction.

For the blind spatial hearing is the only way to form a picture of the space. It helps even in localization of objects that are not emitting any sound via reflections. The navigation ability of bats is almost solely based on the ability to produce a high-frequency sound and listening the reflections from objects (“auditory radar”). The ability of forming 3D-picture via spatial hearing improves when a subject stays longer in that particular space. It is also more sophisticated among older persons. Thus adaptation and learning are involved in performance of spatial hearing.

Finally, let us define some terms related to spatial hearing because (1) the reader will certainly come across these terms in the literature of the topic, and (2) in order to understand the derivation of the names for the methods and for the theories ap-
After this lengthy introduction to spatial hearing, the focus in the following sections is on the localization ability and on the separation ability. At first, physical facts that are involved in this task are presented. After that the actual localization ability and the ability to separate sound sources are described.

2.1 Role of outer ear and head in spatial hearing

Outer ear consists of pinnae and ear canal. It must be pointed out that sometimes outer ear is refers to only pinnae.

It is self-evident, by experience, that a sound source located at horizonthal plane let us say $\varphi = 90^\circ$ (see Figure 1.2) sound is perceived louder in left than in right ear. The difference is due to two facts (1) sound intensity level decreases $\propto 1/r^2$ where $r$ is the distance between source and observation point. In this case the increase in $r$ is approximately 20 cm corresponding the path difference between ears. (2) Head casts a shadow that attenuates the sound waves reaching right ear canal. The latter is significant at frequencies higher than 200 Hz. The amplification rises fast with frequency, but saturates at frequencies above 6 kHz. At the maximum, sound is amplified nearly twice compared to free-field [32].

With the help of pinna a person is able to judge whether the origin of sound source is either in front or back. It creates the border, or spans, “zero-plane” called frontal-plane in Figure 1.2. From pattern recognition viewpoint, frontal-plane acts as a decision boundary of a simple classifier the output of which is either class “in front” or “behind”. That is, it is able to determine, at this rough level, the location of a sound source. Again this ability is functional at high frequency region.

Ear canal is the other part of the outer ear structure. It has no effect on spatial hearing. It may be viewed as a tube the other end of which is closed. Based on the knowledge of basic physics, such a tube is an amplifier. The ear canal amplifies the frequency regions 1.5 - 7.5 kHz and 13 - 16 kHz. Whereas the region of 7.5 - 13 kHz is attenuated. These regions are found in studies using subjects and conducting several measurements and listening tests. Detailed examples of such measurements and tests can be found for instance in [5].

2.1.1 Head-related transfer function

The alterations by the physical attributes of human body were described above enabling us to discuss about the actual feature extraction that human auditory

<table>
<thead>
<tr>
<th>concept</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>monaural hearing</td>
<td>hearing with one ear or with two ears when a sound waves reaching both ears are equal in intensity and there is no time delay between the waves fronts</td>
</tr>
<tr>
<td>binaural hearing</td>
<td>hearing with two ears</td>
</tr>
<tr>
<td>localization</td>
<td>determining the physical location of a sound source</td>
</tr>
<tr>
<td>laterilization</td>
<td>determining the location of a sound “inside” the listener’s head</td>
</tr>
</tbody>
</table>
system uses to map sound events, or auditory events, to particular points in space. The modifications to sound waves propagating from the vicinity of a person to a point where they enter to ear canal are described with head-related transfer function (HRTF). This is a modeling technique that takes into account the effects of head, torso and pinnae by a filter the response of which is different for each direction-of-arrival of sound waves. This is to say that sound waves arriving from location $x_1 = (r, \varphi, \delta) = (1.0\,m, -90^\circ, 45^\circ)$ are filtered with a different filter the waves arriving from $x_2 = (1.0\,m, -90^\circ, -45^\circ)$. At least the coefficients of the corresponding filter are changed. The impulse response of the particular filter is referred to as head-related impulse response (HRIR). It can be stated that physiological attributes of human body form a direction-dependent filter. HRTFs are specific for each person. From modeling viewpoint this means that determining the HRTF of a single person is absolutely insufficient. In order to obtain a model of human hearing system pre-stage, hundreds of subjects are needed. In the course of time numerous HRTF measurements have been conducted and certain trends are visible in the responses. However, better general model of human hearing pre-stage is achieved by dummy head or head-and-torso simulator. The latter is illustrated in Figure 2.3 with three extra microphones which should be removed while measuring HRTFs because they would affect the total response. Measurements are basically easy to conduct with this equipment; sound source (loudspeaker) produces stimuli and the microphones at the ears canals of the simulator measure the filtered signal. A vast project with KEMAR dummy head was conducted by Bill Gardner and Keith Martin. The data and processing tools are available (<http://sound.media.mit.edu/KEMAR.html>). From the response figures presented in [5] and especially from the responses measured simultaneously from left and right ear for an experimental subject in [32], it is easy to see that the response varies depending on source position. The biggest changes in responses are detected as a source moves along perpendicular to median plane. The shape of the response changes, but more importantly, great changes in intensity levels are measured. As the source moves along median plane, or along planes parallel
to median plane the changes are not so drastic and detected intensity differences are also more moderate (in median plane for dummy head the intensity difference is zero).

2.1.2 Interaural time difference and interaural intensity difference

By studying HRIRs and HRTFs it may be stated that a sound source for which the following holds: \( \varphi \neq 0^\circ \) or \( \varphi \neq 180^\circ \) and \( \delta = [0^\circ, 360^\circ] \) basically detect two of differences in the sound waves should be detected; time difference for longer propagation path and intensity difference due to attenuation for longer path. In the literature these differences are referred to as interaural time difference (ITD) and interaural intensity difference (IID).

ITD is approximated by Blauert in [5, p. 75] for three different cases. The following estimates for time difference \( \Delta t \) are for the case in which a sound source on horizontal plane.

\[ \Delta t \approx \frac{D}{2c}(\varphi + \sin \varphi) \]  \hspace{1cm} (2.1)

\[ \Delta t \approx \frac{D}{c} \cdot \varphi \] \hspace{1cm} (2.2)

\[ \Delta t \approx \frac{D}{c} \left[ \left( n + \frac{1}{2} \right) \cos \epsilon + \frac{1}{2}(\varphi + \epsilon) - \sqrt{n^2 + n + \frac{1}{2} - \left( n + \frac{1}{2} \right) \sin \varphi} \right], \]  \hspace{1cm} (2.3)

where \( D \) is distance between ears, \( c \) propagation velocity of sound waves and \( \varphi \) is horizontal angle. Three different cases are presented to take into account the effect of distance \( r \). (1) Equation (2.1) is valid for the case in which the distance \( r \) between the sound source and the listener satisfies \( r \gg D/2 \), (2) Equation (2.2) assumes that sources are near head and fulfilling \( \sin \varphi \leq D/2r \) and (3) Equation (2.3) assumes \( \sin \varphi > D/2r \).

The equations with appropriate illustration can be found in [5, p. 75-76] with two distinctions. (A) Time delay \( \Delta t \) instead of path difference \( \Delta s \), and (B) \( \approx \) -notation is used instead of equality to point out that there are several errors included in evaluating the true path difference starting from the fact that human head is not a circle as assumed by the equations above. The boundary conditions for \( r \) are not specified. However, accurate boundaries are not possible to determine. However, the equations provide a sufficient model for most purposes.

The equations above highlight an important issue concerning the fact that \( \Delta t \) is not resulting only from the path difference measured by vector which is set to start from one ear and to end at the other and the direction of which is same as the wave-fronts from source. Diffraction or bending of wave-fronts introduces a significant increment in the resultant \( \Delta t \). The phenomenon is known as Huyghnesin principle. According to this principle, or law, we know that this phenomenon operates at only waves of
certain frequencies. In this case the range of operation consists of those frequencies
the wave length of which is comparable to dimensions of human head. Wave length
$\lambda$ is defined as $\lambda = c/f$ where $f$ is the frequency of sound waves.

The diffraction of wave-fronts affects not only ITD. As the wave-fronts are able
to bypass the object, which is human head in this case, changes in intensity level
are insignificant meaning not perceived by human. This is to say that IID's are
significant only at high frequencies, namely, starting around 500 Hz \cite{14}. The waves
of that frequency and above are more or less blocked by head resulting in significant
intensity difference.

2.2 Localization ability

In general the term localization covers estimation of three parameters defining the
location of the sound source in space. That is, location is specified uniquely by
$(r, \varphi, \delta)$ as it is illustrated by Figure 1.2. Additionally, the term localization also
covers the estimation of uncertainty of each parameter. The uncertainty is referred
to as localization blur \cite{5} which is the smallest change in any of the parameters $r$, $\varphi$ or $\delta$ detected by human.

In the previous section, the means and tools that human auditory system has for
localization were discussed. The focus is from now on in studying the actual ability,
or to be exact, on its performance. In the course of time it has been discovered that
the topic should be handled by dividing it in different cases. This is due to the fact
that the localization performance of human auditory system varies so drastically in
various conditions. The cases and the related phenomena are:

1. Localization of one sound source
   - Estimation of direction - specified by $\varphi$ and $\delta$
   - Estimation of distance - specified by $r$

2. Lateralization
   - Performance
   - Pseudofonic localization

3. Localization of multiple sound sources
   - Multiple locations with one physical source
   - Multiple physical sources
   - Summing localization
   - Precedence effect

2.2.1 Localization of one sound source

First of all, while studying the localization ability we come across several difficulties
that are not directly related to the performance of itself. Because the research must
be made via listening tests it has to be remembered while interpreting the results
that each subject is an individual behaving at least a bit different from each other.
For instance, while asking a person about the direction of a sound source, one person
is able to address more precise location than the other. Additionally, the precision
may vary between test sessions even if the same person is under testing. Signal content is also involved. The list of errors and variables is practically innumerable.

Localization in horizontal plane

Despite the problems related to the tests, some conclusions can be made. Blauert summarizes the human ability in one sound source configuration by presenting the results of listening tests by Preibisch-Effenberger and Haustein & Schirmer in [5]. Detailed descriptions can be found in [48] and in [28]. Figure 2.4 reflects the results found in these studies. By virtue of the figure it may be stated that a sound source placed either in front or behind, the listener can locate a sound source more accurately compared to the case where the source is placed on either side of the listener. This is generally true for any type of signal [5].

As it was stated earlier, signal content is one factor that affects the localization blur. Still, no other general conclusions than the fact that signal content is involved can be stated when it comes to localization blur because the listening tests done so far vary a great deal. For instance, it can not be claimed that a certain signal type results in smaller blur than some other.

Another factor that seems to affect the localization blur is the duration of a stimulus. This makes sense because the amount of available information by which subjects deduce the location of a source is increased.

Figure 2.5 presents how frequency of a sound affects angular resolution (or blur in parameter \( \varphi \)). The threshold values of this parameter are often referred to as minimum audible angle (MAA) [44]. It is obtained with sinusoidal signals for different horizontal angles. Two key issues which are worth noting from the set of curves: (1) at angles near “in front” the localization blur is smaller than for those that are on the left and (2) the collapse in performance at frequency region near 1500 Hz.

The latter observation actually proves the well-known fact concerning the role of IID and ITD. Interaural time difference is working as the frequency is below 1500 Hz and interaural intensity difference provides the means for localization at higher frequencies [44]. The notch in the performance can be explained as follows: the localization performance based on IID begins to improve and the localization performance based on ITD begins to fall drastically as that boundary is exceeded.
Let us discuss why the performance change occurs at this particular frequency. It can be explained by plain physical facts. The corresponding wavelength $\lambda$ is then approximately equal compared to the difference between ears ($\approx 20$ cm). This results in a phase ambiguity that the auditory system is basically not able to resolve without additional information. The additional information might be, for instance, movement of head and thus producing more estimates of the location. The information maybe be provided by other senses too, for instance, vision.

However, the performance is restored beyond this boundary. This is due to the fact that high frequencies and corresponding sound waves are not able to bend around the human head which results in intensity difference at ear drums. From this basis it is tempting to deduce the role of these two cues of localization ability.

Localization of vertically located sound sources

The estimation of vertical angles is poor compared to the horizontal angles. Figure 2.6 illustrates the estimation of parameter $\delta$ and related $\Delta_{\min}\{\delta = \delta_o\}$. The signal content in the experiments the results of which are illustrated was speech which is a broadband signal. One would expect good performance with this type of signals, thus, the results are somewhat surprising.

According to Blauert, in the median plane the location of an auditory event may not be related to the location of sound event at all [5 3]. Instead, the frequency has the key role. This phenomenon is especially apparent with narrowband signals of less than $2/3$ octaves in bandwidth. As the center frequency of that kind of a signal is raised, the location of auditory event is changing in the median plane. Figure 2.7 illustrates this behavior. Signal content in the experiments is noise with limited bandwidth. One expects a good performance with noise-like signals which carry a lot of information compared to sinusoidal signals.
Ability to estimate the distance of a sound source

The same type of phenomenon can be observed while studying the distance estimation ability as for the angle estimation ability in the median plane. In the experiments the main factor affecting to the estimation of distance is not the actual location but the sound intensity level of the sound source. The precision of distance estimation is illustrated in Figure 2.8. The signal content is impulse-like sounds with intensity level of 70 phon.

The ability to estimate distances is also affected by signal content. More specifically, the familiarity of a sound is found to be one of the main factors that contributes to the performance. The summaries of listening test (by Gardner [26] and by von Békésy [59]) and general discussion on the subject can be found in [5] and [32].

2.2.2 Lateralization

Lateralization is another tremendous ability of human hearing. It may be spelled out as an estimation of apparent location of a sound source inside human head. The phenomenon is studied via listening tests with headphones. If a sound played for each ear differ only by left ear signal being delayed compared to the right ear signal or vice versa, several stages from our model in Figure 2.2 are ignored and thus several cues making spatial hearing hearing to work are not available.

Despite the lack of auditory attributes affecting the signal, a human being is capable of localization. The phenomenon is enabled by either IID or ITD, or both. However, spatial hearing has to suffer in some way due to dropping the cues out of signal. For instance, 3D-localization of sounds has turned into 1D-localization. It is worth pointing out that the studies on the lateralization ability enables us to present spatial hearing models as in Figure 2.2.

To evaluate the performance of the localization ability is difficult in general. The evaluation of lateralization is even harder because subjects indicate the perceived location of a sound source by verbal means. However, listening tests have been conducted by Toole & Sayers who studied the effect of ITD [56]. Additionally, Sayers arranged a test to study the effect of IID [50]. In both studies a subjective scale was used. Subjects were asked to indicate the location of a sound source on the imaginary

![Figure 2.6](https://example.com/figure2.6.png)

Figure 2.6. The performance in the median plane at different $\delta$ values and the related blur. Signal content is speech presented by a person familiar to the listener. The experiments are conducted by Damaske [17]. The figure is taken from [5], p. 44]
axis connecting ears. For instance, after a stimulus was played a subject indicated its origin by expressing: “I noticed the sound source at location between 4-5”.

It is difficult to state hardly anything on these tests when it comes to measures like precision or robustness. But more importantly they present the existence of the well-known ability using scientific means. Blauert summarizes the results of Toole & Sayers and Sayers in [5].

2.2.3 Localization of multiple sound sources

This topic is actually a generalization of the sound source case. More precisely, less assumptions of the listening conditions are made. All the issues embraced by this topic are not even tried to be presented here but a few phenomena that are of great importance in location estimation are discussed. An exhaustive presentation of the topic can be found in [5].

As pointed out, the topic covers also the cases in which there is only one physical sound present but several auditory events are perceived by the listener. For instance a large empty storage with concrete walls may “produce” multiple sound sources out of one physical sound source.

In natural listening conditions two significant and easily perceivable phenomena are worth pointing out because they reflect directly the performance of hearing. Let us consider an everyday case, say, listening music via stereo equipment that contains two loudspeakers and signal source. Figure 2.9 illustrates the case. Listener is in front of speakers LC and RC, and the distance is the same to each speaker. If the signal content is the same in both channels of the source signal and the path from the sources to the ears is approximately the same, the detected sound source location, or the location of the auditory event is perceived at point C. Let us assume that the

![Figure 2.7](image1.png) The center frequency of a signal is the main contributor in location estimation in the case of narrowband signals. The figure is taken from [5, p. 45].

![Figure 2.8](image2.png) Distance estimation in anechoic chamber. $r$ estimate of an auditory event is mainly affected by sound pressure level not by the location of a sound event. The experiments are conducted by Gardner [26]. The figure is taken from [5, p. 47].
listener moves towards left of the current position, while holding the perpendicular distance to the point C the same. In this case the auditory event is detected as well to the left of the previous position C. The phenomenon results naturally from the fact that signal path is shorter from LC than from RC, time difference is experienced by spatial hearing system.

If the the level of RC is increased slowly, finally the virtual sound source that moved towards LC in the previous experiment due to time difference is now re-centered to point C\(^1\). Blauert calls this behavior “summing localization”\(^2\) induced by the fact that two sound events are summed into one auditory event. It has been discovered that summing localization occurs with time differences 630 μs - 1 ms \(^5\).

Another effect related to the listening signals differing in time domain is observed with impulse sounds. If time difference is more than 1-1.5 ms the auditory event is, however, detected at LC. Researchers refer that type of behavior with as “law of first wave-front”, “precedence effect” or “Haas effect”\(^3\). This effect has been largely explored, and it is still a popular topic. The precedence effect is more significant than one may at first think. Moore in \(^4\) suggests that it plays an important role in localization ability in everyday environments. According to Moore, the capability of human brain would experience an information flood if the precedence effect was “turned off”. Thus the precedence effect is a moderator that decreases the amount of data to such a level that is possible to process in given time. In a large storage room, for instance, the localization of sound sources would probably be significantly harder without precedence effect.

As time difference between sounds is further increased to 30-40 ms the listener begins to observe changes in timbre\(^4\) but still only one auditory events is detected. This is often perceived in large spaces. The change in timbre may be drastic compared to the original sound. The space is said to be reverberant and the phenomenon reverberation. Reverberation is commonly considered as an annoying feature of a room.

Time differences exceeding the limit resulting in reverberation, human auditory system detects two auditory events instead of one. The latter auditory event has a very common name, that is, echo.

Detailed discussion including illustration that reflect the results of listening tests on

---

1. The phenomenon has been utilized in stereo equipments for long time ago and its known as balance tuning.
2. The term originally defined by H. Warncke 1941
3. Haas’s discovery, 1951
4. common expression of timbre: “color of sound”
the phenomena can be found in the texts by Blauert and Moore.

2.3 Signal separation ability of human hearing (Cocktail party effect)

In the previous section the localization ability was discussed. This section is focused on presenting a phenomenon which is, again, hard to put into a single sentence simultaneously preserving completeness and understandability. Thus let us give a brief explanation by an example. Consider a situation that a person is having a conversation with another in a large room in which several other talkers are present. Additionally, the space may be reverberant which is to say that actually the number of detected auditory events may be much larger compared to the number of sound events. Still the conversation of these two persons is in most cases successful when it comes to extraction of the message.

As pointed out earlier, reverberant auditory conditions may be viewed as if there were multiple sound sources present in the space while only one physical source is present. Let us illustrate the problem addressed in this section. The individual sound sources are called objects. The sum of the objects is also an object called “Mixture”. The problem addressed in this section may be formulated as a restoring process of the original objects with help of “Mixture” and some additional information. The additional information in this case is provided by human hearing. Figure 2.10 illustrates formally the process. $o_1$, $o_2$ and $o_3$ are the objects and $\hat{o}_1$, $\hat{o}_2$ and $\hat{o}_3$ are the restored objects, respectively. The different notation is used to illustrate that the process is not an ideal one but, for instance, $\hat{o}_1$ might be contain components that actually belong to $\hat{o}_2$. The reverberant conditions may be formulated in light of the figure as $O = \{o_1, o_2, \ldots, o_m\} \mapsto \hat{O} = \{\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_n\} : n \geq m$.

Cocktail-party effect is discussed in numerous publications. In fact, the vast variety of material causes problems because different issues are stressed and thus not necessarily all important phenomena are pointed out. However, we try to make a review of the main issues belonging to this topic. The discussion in this section is organized by presenting the earliest findings and evolving to the most recent theories. Note that the theories to be presented are mostly deriving from the research on speech recognition of human. Thus one should remember while evaluating the theory that
(1) it is highly probable that the separation ability is affected by signal content. Thus a theory derived from listening test using speech signals does not necessarily apply to an arbitrary signal. (2) The concept of recognition is not so vast as the concept of separation because the synthesis of sounds is often connected to the latter.

The outline of this section is the following:

1. Selective attention – psychological views and experiments on sound separation
   - Listening two simultaneous messages
   - Two simultaneous messages and spatial information
   - Conclusions on selective attention

2. Effects of spatial information – physiological views and experiments on sound separation
   - Role of binaural cues (IID and ITD) in separation
   - Binaural detection – Masking level differences (MLD)

3. Auditory scene analysis – modern view on source separation
   - Perceptual grouping and auditory streams
   - Grouping cues – psychological aspect

### 2.3.1 Selective attention

The use of term “cocktail-party effect” to illustrate unmixing of the desired sounds is inspired by above described example of two talkers discussion in a noisy environment. However, in literature, it has achieved a more general role; it is often used as to cover also the separation model suggested by several modern theories that can be stated to belong to auditory scene analysis. The term was introduced by Cherry as early as in 1950s from the basis of his experiments in which the concentration on single talker speech in the presence of several other talkers was studied. The experiments are described in [12] and in [13]. These references are considered as first relevant publications in studying the phenomenon.

**Listening two simultaneous messages**

Cherry presented a list of cues which he believed to be involved in human separation process of speech signals. The cues that he suggested are: The spatial origin of talkers, visually acquired information (lip reading, gestures), different speaking voices (male or female speaker, mean speaking rate, different accent), transition probabilities (utilization of speech content/context, syntax of language, voice dynamics). In the listening tests it was discovered that attention to a message content was still possible if all the other cues except transition probabilities were canceled. This was achieved by recording the messages to be extracted using the same talker for each presented message.

The interpretation of the last cue can be illustrated with the following example: for instance, after hearing the word “computer” one is expects to hear the words as “keyboard”, “screen” or “programming” rather than “ice cream”, that is, contextually correlated words are expected.
Cherry conducted a set of two-simultaneous-talker experiments that were mainly inclined to study the significance of the last cue on his list. It was concluded based on the experiments that human separation system relies strongly on context and longer term processing. The robustness of recognition is, at least, better as the context is available. For instance, words that sound almost the same are sometimes incorrectly reproduced in this type of listening tests. For instance, a subject may claim hearing the word “bar” instead of “bath”.

Cherry’s the most famous experiment describes the selective attention ability. The subjects were asked to attend to only one of the simultaneous messages. It was found that subjects had no difficulties in extracting the desired message. On the other hand they could not say much about the other message. Message content or even the spoken language were almost perfectly ignored. Switch from male speaker to female speaker or switch from speech to a pure tone was noticed in general. However, yet another tremendous ability was discovered; the subjects were able switch their object of attention over a short delay.

Detailed descriptions of all Cherry’s tests discussed here can be found in [12] and in [13]. An excellent summary of the tests including general discussion is presented by Arons in [2].

Two simultaneous messages and spatial information

In another set of experiments two-loudspeaker configuration was used, thus, providing spatial information to subjects. Also in this experiment the ability to extract one message from two simultaneous messages was tested. Each message was produced by a different loudspeaker. The study was conducted by Spieth and it is presented in [54].

It was noticed that the more spatially apart from each other the loudspeakers were, the more easily the content of a message was extracted. For instance, the configuration: loudspeaker 1 at $\varphi = -90^\circ$ and loudspeaker 2 at $\varphi = 90^\circ$ outperformed the configuration: $-10^\circ$ and $10^\circ$.

The messages had a limited vocabulary, and the content of the messages was relatively simple implying high transition probabilities not only at word-to-word intervals but also at phrase-to-phrase intervals. In the experiments it was discovered that the messages fused partly. For instance, the subjects picked some individual words from the irrelevant message to the desired message. The fusion of two messages was detected thus by getting incorrect answer to the question.

In further experiments by Spieth the desired signal was filtered by using low- and high-pass filters cut-off frequency set to 1.6 kHz. Filtering affects basically the transition probabilities by decreasing the probability between the words in two message streams. For instance, even if the same word occurred simultaneously in the messages they sound different.

Spieth pointed out two major observations based on his study: (1) transition probabilities are a significant cue for the separation of speech signals and (2) spatial cues
provide supplementary information for the separation. However, these findings raise also new questions: is the performance improvement the consequence of manipulation done to transition probabilities. Or is it just resulting from the fact that the total amount of information is changed. And yet, is the total amount of information increased or decreased. A plausible suggestion for the interpretation of the experiment is that the filtering decreases the amount of irrelevant information. As well one might suggest that the filtering decreases the total amount of information to a level that human is able to process at a time.

Conclusions on selective listening

A lot of psychologically oriented theories on selective attention have been proposed by several researches. It has been proposed that human information processing capability is limited at certain stage in the brain. This view has received wide consensus among the researchers. The term bottleneck is used to describe the limiting factor in the process of attention. The location of this bottleneck is the main difference between the theories. We conclude this discussion by presenting the principles of one of these theories.

The early work on selective attention is summarized by Broadbent in [10]. Additionally, he conducted a series of experiments of his own. He proposed the general model of information processing (Figure 2.11). “Sensory register” is considered as a storing device capable of storing a large amount of information. “Selective filter” selects information stream which is routed to “Short-term memory” which presents the bottleneck in this model by being able to store only a moderate amount of information at a time.

The model presents basically the same fact that has been proposed by Cherry and Spieth, that is, only one message can be attended at a time. However, Broadbent stated that human is able to concentrate on two messages if the information content of two simultaneous messages is sufficiently low. Another important coinciding result was achieved concerning the significance of spatial information. He concluded that spatial information is useful in the case of ignoring the other message whereas in two-message extraction the significance is lesser. [2]

Also Broadbent presented a list of cues affecting the attention ability. The outstanding difference to the research done until, is that he stressed the significance of neural systems and regarded the role of sensory systems as supplementary. Further discussion on Broadbent’s views can be found in [2].

5. Refers to the level at which the phenomenon occurs as viewed by psychologists
2.3.2 Effects of spatial information – physiological views and experiments on sound separation

Role of IID and ITD in separation

The motivation behind this topic lies in the assumption that binaural cues are the major factor not only in localization but also in separation. The topic is discussed here based on Moore’s book (44), in which, the topic is presented based on the experiments by Kubovy et. al (34). In their experiments a signal consisting of eight sinusoids was presented to subjects via headphones. The phase of each sinusoid was adjusted to zero, that is, the location of auditory event was at $\phi = 0$. In the first experiment it was discovered that if any of components was shifted in time by varying its phase in one ear relative the corresponding component in the other ear, the subjects reported that a new auditory event was detected at some different location. While the sound presented in one ear only such a phenomenon can not occur. In the second experiment they modified the components such a manner that a certain phase shift sequence was introduced in each component of one ear signal whereas the signal presented to other ear was untouched. The subjects reported to hear melody in test signal. Again, if one ear was plugged the melody was not detected. The frequencies for signal components were selected to correspond to the frequencies of musical notes.

Kubovy’s experiments and the example above can be interpreted as an ability (1) to extract the binaural cues; IID and ITD, (2) to localize auditory events, and (3) based on the difference in location of auditory events conclude the number of sound events. Note that in the example only ITD was provided for the subjects.

Let us make a quick review into modeling of this detection via spatial cues even if the spatial hearing model will be discussed later more thoroughly. Most separation systems based on binaural processing, suggest that there is “a field” or “spatial map” in auditory cortex in which a certain point corresponds to an auditory event at a certain location. In the example above this would imply that two separate auditory events would have occurred at two separate locations on that field. Whether such a structure truly exists in physical form – neural activity at certain point in auditory cortex can be mapped to a certain location in space. However some research has been done on guinea pigs by Hartung and Sterbing (27), in which, according to Bodden (7), the precedence effect and monaural cues were studied at the neural level.

Masking level differences

To illustrate the situation what is about in masking level difference (MLD) let us focus on Figure 2.12. Two types of signals are used: a single sinusoid and white noise. Additionally, only two observations are used, that is, the signals are from the same source. This is a very important notification since each sample of noise signal is different. A tone, however, can be always reproduced with a sufficient accuracy. The amplitudes or the phases of the test signals are modified. Phase shifts of 0 and $\pi$ are used only. In the experiment (a) the signals are adjusted such that white noise masks the tone. Let this tone level be $L_0$ dB. In the experiment (b) the signal
played to one ear is inverted, and it turns out that the tone is audible again despite
the masker. Now the tone can be adjusted to a lower level $L_\pi$ dB. The difference
$(L_0 - L_\pi)$ dB is masking level difference (MLD), also known as binaural masking level
difference (BMLD). In the experiment (c) one ear is plugged which causes re-masking
of the tone by noise. Whereas in the experiment (d) it becomes audible again once
the noise generator is activated. Note that in the experiments (c) and (d) the signal
levels were preserved the same as in the experiment (b).

According to Moore the phenomenon is observed also with other types of signals
than just speech signals [44]. It has been observed in several tests concerning MLDs
that the requirements for the signals are maybe more relaxed than in the experiment.
The binaural detection seems to improve once there is some difference in the signals’,
in phase or in level. The difference, in turn, implies the different spatial locations
of the sources. The maximum MLDs have been detected in circumstances as in the
example. In this kind of situations human auditory system is not able to detect
phase differences and thus not able to accurate localization of the sources. This
is an important notification from modeling point of view because some separation
models assume that the localization of a sound source occurs first, and after that
the separation of the source can be made. The model, or imitation, may be wrong,
in this sense, but as a separation technique it has been proved to be working, and
in this sense the model is valid. Taking into account the findings presented above,
it can be stated that cocktail-party effect can be partly explained by the process
observed in studying MLDs.

2.3.3 Auditory scene analysis

More psychoacoustically oriented view of separation process is offered by Bregman
in his famous book “Auditory scene analysis”. Auditory scene analysis (ASA) is not
restricted to sound source separation but it is analyzing human hearing behavior
in general. However, sound source separation has been studied exhaustively by the
ASA researchers.

![Figure 2.12. The significance of binaural data to MLDs. The figure is taken from [44], pp. 236](image)
At first it is worth stressing ASA researchers handle the phenomena related to hearing completely differently compared to the previous research. In general, the phenomena of hearing are explained with high-level processes. This is an important reminder from modeling viewpoint because the systems based on ASA see their ultimate goal differently compared to the systems based binaural techniques. Thus the separation models based on ASA are often referred to as sound source segregation and the use of terms as cocktail-party effect are avoided. However, broadly speaking the goals are the same, that is, to modify the signal mixture in such a manner that certain source stands out of the mixture. ASA from human hearing viewpoint is discussed thoroughly in [44]. Good summaries can be found in [2], [16] and [32]. For the most demanding purposes one should acquire Bregman’s book ([9]) that defines the whole concept.

Perceptual grouping and auditory streams

To illustrate what is all about in ASA let us use Bregman’s example that he used to enlighten the topic. Think of a signal mixture recorded from a cocktail-party. It would be nice to be able to paint the spectrogram of the signal in a way that regions belonging to the same sound source are of the same color. Painting of the regions of the spectrogram is formally referred to as perceptual grouping [44].

In addition to perceptual grouping, ASA introduces new concepts comparable to sound events and auditory events. The concepts and their meaning are presented Table 2.3.

<table>
<thead>
<tr>
<th>concept</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>physical entity that gives rise to acoustic pressure waves</td>
</tr>
<tr>
<td>stream</td>
<td>percept of a group of successive and/or simultaneous sound elements as a coherent whole, appearing to emanate from a single source</td>
</tr>
</tbody>
</table>

Grouping cues – psychological aspect

To understand further steps let us consider a few concepts deriving from psychology. One of the basic concepts of ASA is perception. Ears, for instance, produce sensation as any other sensory organ in human body. Perception is stimulated by the sum of distinct sensations. As an example, let us consider a well-known example from image processing books illustrating the figure-ground phenomenon. Colloquially, “figure” is what you see and “ground” is everything else. In Figure 2.13 human “wants” to see a triangle while only three incomplete circles are organized in a certain manner.
Gestalt psychology is a research area which is focused on studying why human sees a complete form or pattern instead of the individual parts that are constituents of the form or the pattern. However, it has been established that some theories derived from the research done on vision apply to other sensory systems in human. A few principles that have been established by the Gestaltists deriving from research done on visual perception are enumerated below. The list is summarized from [32] and [2]. The term elements refers to the constituents of the perceived object. The difference between element and object is important to recognize while studying the list.

- Principle of proximity
  - Two objects that are close to each other (e.g. time, place, etc.) are probably classified into a single object than into two different objects

- Principle of similarity
  - Similar properties group elements together whereas dissimilar group elements apart

- Principle of closure
  - Forms, shapes and figures that are enclosed are perceived as single objects

- Principle of continuity
  - Elements that appear to follow the same “trajectory” (time, place, including other attributes) tend to be grouped into single object

- Principle of common motion
  - Constituent objects or elements that move together tend to be grouped into a single object

**Grouping**

The concept of grouping is divided into two categories by ASA researchers. (1) Temporal grouping, or sequential grouping – grouping of sound streams that do not overlap in time and (2) spectral grouping, or simultaneous grouping – grouping of sounds that do not overlap in frequency. Let us consider these two grouping schemes and see how they are related to basic principles of Gestalt psychology. This discussion derives from [8].

The factors that are involved in temporal grouping of two tones are presented in Table 2.4. “Factor” refers to an attribute the existence of which increases the perceptual distance between two tones. It is worth stressing that the findings are the result of tests in which only one sound source is active at a time. This is considerably easier situation compared to simultaneous grouping in which the sounds overlap in time. The factors, or the attributes, that are believed to be involved in spectral grouping are presented in Table 2.6.

The grouping is involved in perception of auditory environments. A change in an attribute affecting the grouping results in change in perception. Some of the changes can be understood by the principles of Gestalt psychology. Whereas some can be understood as perceiving a change in spatial origin and thus the sounds are grouped

---

6. This is not a definition of Gestalt psychology but rather a colloquial expression of the topics covered by the research area.
CHAPTER 2. SPATIAL HEARING

Table 2.4. Factors that increase perceptual separation in temporal grouping process

<table>
<thead>
<tr>
<th>factor</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency separation</td>
<td>the more two tones are apart in frequency the bigger the perceptual distance between the streams</td>
</tr>
<tr>
<td>separation in time</td>
<td>increasing temporal gap between streams</td>
</tr>
<tr>
<td>differences in timbre</td>
<td></td>
</tr>
<tr>
<td>differences in spatial origin</td>
<td></td>
</tr>
<tr>
<td>differences in loudness</td>
<td>weak effect, but may “assist” other factors leading to separation</td>
</tr>
<tr>
<td>abruptness of the transition</td>
<td>the changes in continuous sounds are often perceived</td>
</tr>
<tr>
<td>from one sound to another</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5. The effects on perception resulting from temporal grouping

<table>
<thead>
<tr>
<th>effect</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>perception of melody: the melody is formed</td>
<td>two sound streams produce a separate melody instead of single</td>
</tr>
<tr>
<td>within an auditory stream</td>
<td></td>
</tr>
<tr>
<td>rhythm emerges primarily within segregated</td>
<td></td>
</tr>
<tr>
<td>streams</td>
<td></td>
</tr>
<tr>
<td>finer judgments of timing within a perceived</td>
<td></td>
</tr>
<tr>
<td>stream than across streams</td>
<td></td>
</tr>
<tr>
<td>continuity of speech is lost if the pitch</td>
<td></td>
</tr>
<tr>
<td>of speech is altered</td>
<td></td>
</tr>
<tr>
<td>perceived location can be affected by its</td>
<td></td>
</tr>
<tr>
<td>sequential grouping</td>
<td></td>
</tr>
<tr>
<td>perceived loudness can be affected</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.6. Factors that increase perceptual separation in simultaneous grouping process

<table>
<thead>
<tr>
<th>factor</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>different harmonic relations among subsets</td>
<td>fundamental frequencies can be the same</td>
</tr>
<tr>
<td>of partials</td>
<td></td>
</tr>
<tr>
<td>different spatial origin</td>
<td></td>
</tr>
<tr>
<td>frequency separation</td>
<td></td>
</tr>
<tr>
<td>independent changes in loudness</td>
<td></td>
</tr>
<tr>
<td>asynchrony of attack of frequency</td>
<td>also asynchrony in shutoff</td>
</tr>
<tr>
<td>components</td>
<td></td>
</tr>
<tr>
<td>asynchronous changes in frequency</td>
<td>may have an effect</td>
</tr>
</tbody>
</table>

Table 2.7. The effects on perception resulting from simultaneous grouping

<table>
<thead>
<tr>
<th>effect</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>each global sound has its own properties</td>
<td>pitch, loudness, timbre and perceived location</td>
</tr>
<tr>
<td>isolation of a region in spectrum</td>
<td>organizing parts of the spectrum into separate groups reduces their perceptual interaction</td>
</tr>
</tbody>
</table>

differently. A few effects that the temporal grouping process has on perception are presented in Table 2.5. Increasing some of the factors that help to perform spectral grouping also causes changes in perceptions. These are listed in Table 2.7.

As for sequential grouping, apparently simultaneous grouping results in perceiving also higher level attributes of sound streams, but now as there is basically no historical information as in sequential grouping the factors of Table 2.6 are resulting from primitive properties for each sound (e.g. frequency).

The final grouping of streams in auditory environment is more complicated than the resultant effect of grouping schemes. Actually, two grouping schemes seem to compete in certain situations. Figure 2.14 is an example of one possible result of two grouping schemes working together.

Another examples is so called “old-plus-new” principle. It states that human hearing
Figure 2.14. Single component belonging to complex source B tends to group to source A when both sources are active. This is due to the fact that A and B have the same frequency component.

Figure 2.15. “Old-plus-new” strategy. Two sounds, A and B, are heard instead of “A with something new on it”. A is expected to remain unchanged.

tries to interpret auditory environment as it was, reluctant to “see” any changes in the streams which have already been detected. In Figure 2.15, new component B is introduced to existing sound A. Human detects two streams or sources instead of one.
Chapter 3

Computational models of spatial hearing and human sound source separation

In the Chapter 2 human auditory system was reviewed, more specifically, spatial hearing. In this chapter the focus is on computational modeling of the phenomena related to spatial hearing.

Note that basically all the models do not even try to copy all the abilities of human auditory system. Instead, they pick one ability, assume certain signal content and environment (often considered as ideal by generating artificial stimuli), and try to maximize the performance of the ability in these conditions. For instance, the models which address to the localization ability, the problem solving strategies often assume that sound sources are in horizontal plane. Other angles or distance are often neglected, or are not of interest. In spite of the fact that several things are fixed in advance and only only one ability is modeled, the current models appear to have a number of shortcomings, thus, further development is certainly needed. Early models for sound localization as well as models for sound separation are presented. The outline of the chapter is the following;

1. Sound source localization models
   - Jeffress model
   - Colburn model
   - Lindemann model
   - Gaik model

2. Binaural models applied to solve cocktail-party problem
   - Binaural dereverberation system
   - Binaural dereverberation – using binaural model as a “filter”
   - Binaural dereverberation – sub-band multi-channel processing approach
   - Cocktail-party processor – Bodden model

3. Auditory scene analysis (ASA) based models applied to solve cocktail-party problem
   - Parsons model – early model for speech separation
   - CASA – data-driven architecture
   - CASA – prediction-driven architecture

4. Other sound source separation schemes
   - blind source separation (BSS)
   - neural models
3.1 Sound localization models in horizontal plane

In general, models of spatial hearing or binaural hearing models, as they are often referred to, aim at responding different stimuli similar to human. As the ultimate goal the systems aim at repeating the phenomena described in Chapter 2. The models presented so far focus on localization in horizontal plane, that is, they can not give any information of elevation angle or distance. In general, they try to find computational equivalent of human hearing system that, in an ideal case, responds to any type of stimulus as human auditory system does.

The earliest models date back to 1940s, and still, the basic structure of the early models can be recognized in state-of-the-art systems. Does that reflect the maturity of the early models or the lack of new ideas in the area, can only be guessed. However, the amount of completely different approaches is surprisingly low. Still, significant progress has been achieved in modeling the constituent structures compared to the earlier systems.

This is a short review into the models that have received wide consensus among the researchers. The viewpoint is in describing the models that serve our goals because a more general description would turn into writing a book of the subject.

Jeffress model

The earliest model in the area is presented by Jeffress in [31]. The later models use this basic model as a starting point while designing a new system. Once some kind of mathematical model for a structure of human hearing system is obtained, it can be appended to the core model. For instance, once the form of neural messages is found out, the signals in the model can be processed in such a way that they remind neural messages. On the other hand many of the suggested additions are based on pure heuristics deriving from psychoacoustics.

This model is implied by so called coincidence principle. The term derives from the fact that it (1) delays one input signal with respect to the other, and (2) gives as output the time delay for which the signals corresponded to each other best. That specific time value corresponds the path difference between signals. Mathematically this is modeled via calculation of correlation between two signals and evaluation of the time value for which the correlation function reached its maximum value:

\[ R(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} s_i(t)s_r(t + \tau)dt \]
\[ \hat{\tau} = \max_{\tau} R(\tau) \]  

(3.1)

Now this \( \hat{\tau} \) value is substituted into a priori agreed spatial location to obtain value for angle.

Colburn model

Jeffress’s early system models only the ITD detection of human hearing but it can be stated that it is able to conduct the lateralization of sound sources. Colburn
presented a model in references \[14\] and \[15\] which introduced no modifications to the core model of Jeffress but it appended more features that had been discovered in human auditory system.

If the Jeffress model was capable of modeling lateralization Colburn’s model is, at least in principle, able to model localization (\(\Rightarrow\) capable of lateralization). The most advanced of the models suggested by Colburn is presented in \[16\]. We stress only the new blocks compared to Jeffress model instead of full description. It is based on the summary written by Palomäki in \[46\].

One prominent feature in Colburn model is bandwise processing. Based on the knowledge that the function of cochlea in human auditory system is basically to perform initial frequency analysis, a filterbank is introduced as a modeling structure. Signals received from different bands are converted from sensory messages to neural messages (fluid movement in cochlea). Additionally, a structure producing neural firing patterns is introduced. This structure corresponds to the transition region between fluid in cochlea and neural terminals in inner ear. Auditory cortex was modeled with the similar structure to Jeffress model thus obeying the coincidence principle. Additionally auditory periphery was taken into account in this model. The function of this stage is to model how IID is built between ears. Yet, the information acquired by this subsystem was combined with the information produced by “auditory cortex” enabling a combined analysis of the lateral position of a sound source.

In \[14\] Colburn wished to build a model that could explain the conversion of sound waves to neural firing patterns, not just to build a system capable of localization. Instead, the localization ability was used more like a performance measure enabling the comparison between the model and human. The significance of Colburn’s model is the fact that it took a big step in modeling auditory system by introducing several new structures that seem to model the processing quite well.

**Lindemann model**

After Jeffress and Colburn had established the core structures, the binaural models have developed only “internally”. That is, the improvements have been achieved by modifying the basic structures. However, some of the modifications have introduced a model of some known feature of human auditory system.

Let us consider Figure 3.1 which illustrates the basic operation of the revised core. Lindemann’s model is presented in \[36\] and \[37\]. It not only introduces the utilization of IID but also tries to imitate the precedence effect (discussed on page \[21\] \[46\]). The model of the effect is implemented as “inhibition” messages which basically prevent the signal propagation at that instant of time as human auditory system would inhibit the observation. “Inhibit the observation” means that the propagation of neural impulses is prevented. Psychoacoustical experiments, in turn, suggest that the precedence effect is more about preventing the detection of a stimulus response in auditory cortex, in physical level. This claim is justified by the fact that small children do not seem to have this ability \[32\].

In spite of the modeling precedence effect as a low-level process, while psychoacoustical experiments suggest that the phenomenon is based on experience, it should not
be considered as a fully invalid approach. Organs or processes that are involved in the phenomenon are still not known exactly. Thus, one can not pick out one theory and decline all others [38].

**Gaik’s improvements to Lindemann model**

The performance of Lindemann model differed from the results obtained with real subjects. This occurred in certain cases alluded to as “natural combinations of interaural parameter differences” [25]. The concept embraces a hypothesis which claims that ITD and IID are dependent on each other, or once ITD is fixed, a certain range of values for IID is more likely than some other range. Lindemann model suggested, for instance, more spatially spread auditory event compared to the subjects’ response as a result of neglecting this concept [46]. This is due to the fact that Lindemann’s model does not account for auditory periphery modifications in a way as human auditory system presumably does. That is, the signals should be modified before the cross-correlation structure.

Gaik introduced a weighting function in parallel to inhibition mechanism that produced more psychoacoustically plausible results than leaving monaural analysis completely isolated as it is done in Colburn model (presented in [16]). The parameters of weighting function are adjusted according to DOA and frequency band that is in question.

**Conclusions on binaural models**

Whether the presented models are the best achievable with the current knowledge, is not known. At least the models developed so far are able to respond similar to human auditory system in certain situations. Thus, the models seem to be valid from physiological viewpoint. Lindemann’s model was developed to correspond to the psychoacoustical measurements achieved with subjects and Gaik's improvements strengthened this relation even further. Thus it is well argued to claim that the goals set to models are achieved at least partly since systems that have comparable performance to human auditory system can be built.

A lot of work is still to be done because no system has been published that could model all the known phenomena that human auditory system uses in localization.
Yet, they lack the capability of estimating all the coordinates simultaneously to give an unique estimate of the location of a sound source. Even best performing models have restrictions. For instance, signal content or environment are fixed. Often restrictions are not even recognized by the researchers developing these models. One can only question what constraints are unrevealed due to lack of test data or incapability to deduce them.

For instance, let us consider the modeling of the precedence effect. The precedence effect actually consists of several modes; normal operation mode (defined on page 21), build-up mode and breakdown mode just to mention a few. In the latest exhaustive research project on the precedence effect, conducted by Litovsky et al. in [38], it was stated that none of current model is able to describe all the phenomena which are already discovered in human hearing. The precedence effect has been suggested to result from low-level processing and high-level processing, or “bottom-up” processing and “top-down” processing, as they are often referred to. Psychoacoustically-oriented models and physiologically-oriented models have been proposed [38].

Finally, let us introduce an overview illustration of the binaural model according to Blauert. Figure 3.2 summarizes his views. Similar structures can be recognized while comparing this model to the previously presented models. What is worth noting in the figure, is that Blauert sees the models presented so far as preprocessors of data for a system that is able to classify data according to its attributes and the final deduction of the data is done at high-level.

In general, a system capable of classifying data and give a high-level interpretation as a final result is referred to as pattern recognition system. It connects the data with certain attributes to a specific location in feature space which, in turn, is divided into subspaces, classes, each of which corresponds to some high-level concept. For instance, data with attributes \( x = [r_0, \varphi_0, \delta_0]^T \) could connected to a class “right ahead quite far away”. In this case human brain is the pattern recognizer. Blauert’s model in Figure 3.2 is thus an example of “top-down” model.

![Figure 3.2. Core elements and basic operation of a binaural hearing model according to Blauert](image_url)
3.2 Binaural models applied to solve cocktail-party problem

The models of human auditory system capable of horizontal localization are referred to as binaural models. Thus, all the models presented in Section 3.1 may be put into the category of binaural models. A binaural model is just one constituent process in a system that tries to model the cocktail-party effect. However, it plays an important role in this type of a separation scheme and, thus, it is expected to perform in a proper manner.

Systems that employ binaural model to improve the intelligibility of the desired sound source are nowadays referred to as cocktail-party processors. To stress the difference between goals of methods for resolving cocktail-party problem we use the expression improving intelligibility instead of separation or segregation while discussing the solutions based on binaural models. Next, some well-known cocktail-party processors and their theory of operation is reviewed briefly.

Binaural deverberation system

One of the earliest model which can be arguable seen to belong under this topic was presented by Allen et al. in [1]. The technique applied in the paper was intended for dereverberation, that is, removing reverberation from signals and thus improve the intelligibility. Their intention was to build a system to improve the intelligibility of speech signals. This is a somewhat special case; noise type and signal type are fixed. The evaluation of the system was done by recording the speech signals in three reverberant rooms the characteristic reverberation times of which were 0, 1 and 2 seconds. To acquire the reference signal, the same signal was recorded in anechoic chamber in a separate recording session.

The system is illustrated in Figure 3.3. Basically, the idea is to emphasize the correlated components in channels x and y and attenuate the uncorrelated components.

They reported that signals suffering from serious reverberation improved “dramatically”. This was due to the fact that signal parts that did not correlate were removed

1. The argumentation on use of terms was discussed on page 27.
by “gain switching”. Processed signals were evaluated also by subjects\(^2\) who preferred the processed signal to the original.

**Binaural dereverberation – using binaural model as a “filter”**

Palm tried to apply Lindemann’s model (see page 34) on speech signals that were deteriorated by reverberation. Lindemann’s model was quite a novel by that time (presented in 1986, description can be found in [36] and [37]). Their hypothesis was that the ability to model the **precedence effect** (discussed on page 21) would help in removing the reverberation. As stated by Palm: “Our intent was to use the inhibition generated by the model to filter the reverberation from binaural recordings in typical office environments.” [45]. However, they found that the method was not successful in the sense of enhancing the intelligibility of signals. The subjects characterized processed signals as more noisy and more distorted than input signals.

The hypothesis behind this approach was somewhat similar to that of characterized by Figure 3.3. However, since the Lindemann model is intended to model ability to localize a sound source in horizontal plane as accurately as possible, it was not of interest to sustain or improve the signal quality in the processing. Several nonlinear operations are performed in the Lindemann model. Two major sources of additional noise are introduced while processing. (1) Firstly, the input signals are half-wave rectified. (2) Cross-correlation results in mixing of two input signals. The term mixing in this context refers to an operation the result of which is a frequency component at some frequency specified by of two signals that are involved in mixing process. The signal component resulting from mixing may not even appear in either of the originals. It is a well-known fact that human hearing is particularly sensitive to detect such changes which, in this case, are resulting in poorer intelligibility (see discussion concerning the sensitivity issues in Chapter 2).

**Binaural dereverberation – sub-band multi-channel processing approach**

The system is proposed by Yamada et al. in [65]. The fundamental idea of their system is completely different from the previously presented. They tried to cancel out the room reverberation by using the estimate of the room response. The estimate was acquired by using **adaptive filter approach**, that is, they filtered the recorded signals by a filter the coefficients of which were adjusted using the original stimulus as a reference. Actually, the processing is done bandwise, and each band has a dedicated filter.

The determined filters are applied to recorded signals, and the outputs of each band are summed which results in dereverberated signal. However, some constraints exists. (1) The system requires a proper reference signal (in [65] it was the original signal). (2) The filters have to be invertible. Since they are adaptive, the invertibility has to be monitored during the processing. If (2) was not satisfied, the band was not filtered. They reported that system performed well in their tests and the invertibility constraint was satisfied in most cases they studied.

\[\text{the authors point out that these evaluations were not a proper listening test}\]
Cocktail-party processor – Bodden model

So far the most wide consensus while speaking of binaural separation techniques has received the model presented by Bodden in [6]. The theory of operation is summarized in [5]. In fact, this model is also the first one, not only concentrated on reverberation removal but also a suggestion to model the cocktail-party effect.

The fundamental idea is to use a binaural model to determine the location of a sound source and basically amplify sound waves from that direction.

Let us describe the system briefly using Figure 3.4. The symbol of human head illustrates the fact that signals are driven through human auditory periphery modeled by HRTFs. “Preprocessing” is basically a filterbank which models the cochlea (bands of the filter in the bank are adjusted to correspond human hearing frequency resolution). The output of each band is passed to “binaural model” which is a coincidence detection system based on cross-correlation, similar to Lindemann’s core system. The second output from “preprocessing” is the signal stream which is the target of processing, not the target of analysis as the first stream. In order to improve the intelligibility, Wiener filterbank is used. Wiener filters are controlled in a way that certain spatial direction is emphasized. In Bodden’s processor the guidance of the filter is provided by “evaluation module” which adjusts the coefficients. The exhaustive description of its operation can be found in [5].

Wiener filter in this particular case is implemented as a time-dependent coefficient which multiplies each frequency component of a sub-band. The evaluation module decides whether, at a particular band the corresponding Wiener filter, should amplify or attenuate the signal. [5]

The effect of the system on a sound source at some spatial location, can be described with sensitivity patterns as in Figure 3.5. Note that the sensitivity diagrams are presented here for only four bands while the total amount of bands in the original system is 24. The diagrams are interpreted in such a way that semi-circles corresponds half the horizontal plane of Figure 1.2. The curves, in turn, denote sensitivity level boundaries. The sensitivity is bigger within the curves. Free field sensitivity is described by the set of semi-circles of different radii.

In the experiments with subjects, the improvement in sentence intelligibility was
3.3 Auditory scene analysis models applied to solve cocktail-party problem

In the models that try to solve cocktail-party problem using the knowledge of ASA the most outstanding feature they share is the use of only one channel signal. Effects of spatial attributes are in a basic ASA system completely ignored. The ASA approach assumes that the only preprocessing tool that human auditory system has enabling separation, or segregation, is frequency analysis of a signal. Despite the fact that the same information is included in the time domain signal and in the converted signal, the latter enables the interpretation of the signal. Using these two presentations of the same observation the desired components should be searched and synthesized resulting in the desired sound source. Table 3.1 presents a few additional concepts enabling us to understand the discussion below. Additionally, the reader is encouraged to review the concept of stream on page 28. Note that the definition of stream does not require that streams should be contiguous, instead, it can consist of segments, separate in time. Neither is a stream required to have continuous content. However, ASA based systems often assume that signals consist of continuous stream segments.

ASA based separation is often alluded to as computational auditory scene analysis (CASA). This is intuitive because CASA means using computational means to model human processing of sound mixture. The processing is assumed to begin with a somewhat thorough analysis of mixture. This means splitting the input mixture to its primitive constituents. For instance, amplitudes and phases associated to certain frequency range are determined over time. This process may be viewed as a formation of sets each of which has an infinite number of members. For instance, a set of frequencies, a set of amplitudes, a set of phases, and so forth could be formed to account for the input. The sets are infinite because, in theory, we have a continuous-time signal the attributes of which are also continuous to obtain an
### Table 3.1. Concepts related to streams as viewed by CASA

<table>
<thead>
<tr>
<th>Concept</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>sound fragment, or fragment</td>
<td>single component, contiguous component of a sound (or stream)</td>
<td>* a pure tone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>* any component of a played piano note</td>
</tr>
<tr>
<td></td>
<td></td>
<td>* noise burst</td>
</tr>
<tr>
<td>fundamental frequency ($f_0$) component</td>
<td>the fragment of a continuous stream with the smallest frequency</td>
<td>* $f_0$ of a tone is the frequency of the tone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>* the component with the smallest frequency of a played piano note</td>
</tr>
<tr>
<td>pitch of a sound</td>
<td>concept of psychoacoustics for fundamental frequency</td>
<td></td>
</tr>
</tbody>
</table>

Equivalent presentation compared to the original signal. Due to the extraction process of the primitive elements, the information concerning how an element in a set is connected to another element picked out from any other set, is in most cases available inherently. It is known to which element in the frequency set and to which element in the phase set an element picked out from the amplitude set is connected to. Referring back to Table 3.1 if this linkage is preserved over time it may be called a fragment.

In addition to fragments, the mappings to integrate them to form a sound are needed, or using ASA-terminology from page 28, to form a stream. This is the hard part in any CASA system; each of the fragments should be picked out and compared to all the other searching for correlation in some predefined dimensions. A high degree of correlation in any measured dimension suggests that the fragments arise from the same source. The degree of correlation between fragments is found using the principles presented in the introduction to ASA (see 2.3.3). Using this formulation the linking between the concept of separation and CASA is more or less justified.

Finally, let us present the basic structure of the CASA separation scheme. Three basic building blocks can be recognized in any CASA system: (1) analysis stage, (2) grouping stage and (3) synthesis stage. Implementation of the stages varies and working solutions can be achieved in various ways. A few solutions are presented below to enlighten the ideas of how to implement algorithms that are performing the operations of human auditory system from ASA viewpoint. There is ongoing debate within CASA community concerning the major principles of such systems. The different approaches are brought out in the discussion below.

**Parsons model – early model for speech separation**

This model was developed already in 1970s and it is presented by Parsons in [47]. His ultimate goal was to study whether it was possible to separate speech signals using their frequency distribution. It can considered as an early work that implements some of the basic principles of ASA enabling separation. This is somewhat astonishing because the principles are not announced until over ten years later.

It employs the same basic principles which can be recognized while reviewing the
operation of the models which are based on Bregman’s book, published in 1990. Block diagram is presented in Figure 3.6. Connections to the previous discussion can be made; the analysis stage corresponds to “Preparation” which covers A/D-conversion, time domain weighting and Fourier transformation. “Re-construction” equals to synthesis part conducting inverse Fourier transformation of the segments and attaching the segments into a continuous stream and D/A-conversion.

Let us consider this approach in more detail. The main assumption or establishment of viewpoint is made on signal content. The purpose is to utilize the well-known fact that in a speech signal vowels reign over consonants. This results in abundance of periodic information which is organized in a harmonic manner. The fundamental idea was to extract the harmonic components for each speech signal. The consonants were completely neglected but as Parsons puts it: “this restriction may not be a serious one” [47]. As the ultimate goal, they wish to find out whether it was possible to build a system that could (1) separate the two speeches automatically without the utilization of binaural information and (2) is this kind of processing meaningful; does it increase the intelligibility of speeches or not.

The brain of the system is “grouping stage”. It is formed by three blocks in the middle (Figure 3.6). The analysis stage provides output as “peak tables” the elements of which consists of amplitude, frequency and phase of each frequency component of Fourier transformation. The parameter values in the tables, at this stage, are held as coarse estimates. The parameter estimates suffer from interaction of the signals. “Peak separation” has to resolve the amount of interaction and revise the parameter values according to the estimated interaction and refine the coarse estimates. “Pitch extraction” can now use the refined peak tables to extract the pitch of the speech signal. “Pitch extraction” includes a process alluded to as peak assignment which basically predicts the frequency of a harmonic component associated with certain pitch value and searches for the nearest value in peak tables. Two new peak tables can be formed for each harmonic sound. Once the pitch extraction is performed, actual values for a pitch are available. The two main tasks that “Tracking” has are: (1) assign the corrected pitch values, and (2) predict the pitch values for next signal segment.

![Figure 3.6. Speech separation system by Parsons (adapted from 47)](image-url)
They found that the system was able to achieve their goals discussed above. Even if consonants were neglected, it was found that once the processed signals were played to the subjects they preferred the processed signals to the originals. It was found that the subjects experienced an illusion of consonant information in the separated signals.

CASA – data-driven architecture

Data-driven approach is, in a way, the traditional approach to build a CASA system. The theory of operation is put into one sentence by Ellis who is one of the pioneers in the area: “The sound is converted to its spectrum, cues are picked out, and representations of cues are grouped into an abstract description of initial input.”[20]. The basic structure is presented in Figure 3.7.

“Front-end” is performing the task of auditory periphery, that is, cochlear processing (see Figure 2.2). In Parsons’s early model the task was taken care of by DFT. More sophisticated solutions for this have been searched for as the research on the area has evolved. Constant-Q filterbank, gammatone filterbank, or transmission line model just to mention a few. Each of the conversions are aiming at modeling auditory periphery, one claiming to outperform the other. However, they share the common goal of modeling the behavior of auditory periphery on different frequency ranges. The filterbanks are often followed by additional processing to make outputs of filterbank to remind neural messages. For instance, constant-Q filterbank may be followed by amplitude envelope extractor. Some solutions may even include HRTF filters forming the cascade: “head & upper body”, “auditory periphery”, “filterbank” and “conversion to neural message”. However, even these most advanced solutions and plain DFT share the common goal of finding out the energy associated to certain frequency range.

Even if the fundamental task of “Front-end” is conversion of signal to more interpretable form, some architectures have also cue detectors included in this stage. Usually, the cue extraction is performed from the result of the transform. For instance, Brown [11] and Mellinger [41] use onset detectors in the output of filterbank and thus for a time domain signal. A new representation, a map, is formed in time-frequency plane out of which the likelihood of sound onset associated at particular time and at particular frequency can be numerically read.

Once signal has been transformed to its features (see Figure 3.7) we have not done basically anything but a conversion of number sequences into another sequences, especially in the case of DFT. “Object formation” stage is to conduct another conversion resulting in elements. Element is any kind of representation that contains...
energy only from one source. For instance, a good choice for an element deriving from a periodic sound is a sinusoid. The difference to signal features, is basically the fact that the duration information is appended\(^3\). Sophisticated CASA systems introduce objects such as sinusoidal tracks and synchrony strands\(^20\), for instance.

Figure 3.8 gives an example of object formation resulting in elements. The signal features are extracted at discrete time instants tagged with markers: □, x, *, ◦ and △. Using search algorithms associated with more or less heuristic measures elements described by contours can be formed.

After converting the data into a more abstract form we can begin to apply the rules presented in Section 2.3.3 to assign each of the extracted element to a set. The number of element sets, in an ideal case corresponds, of course, to the number of actual sound sources. However, as the “Grouping rules” consists of the algorithms derived from auditory perception (perception discussed on page 28, the perceptual grouping results in assigning the elements into streams. Stream, in turn, means perceived source. Thus, in an ideal case, the number of element sets is equal to the number of perceived sources.

CASA systems use different amounts of cues suggested by Bregman (discussed in Section 2.3.3). Very rarely all the known cues are used because they are not all that significant. The significance of the cues is discussed in 9 and in 44. For instance, Brown used in his system onset/offset, periodicity and modulation cues\(^11\). CASA systems differ from each other not only in the amount of cues they utilize but also in the way grouping is performed.

As an example of grouping let us briefly explain the grouping scheme that Ellis in 19 has used. The system consists of two stages. In the first stage elements, or sinusoidal tracks, are form large number of groups. The second stage is to refine and to assure the correctness of the first stage grouping. The first stage grouping is done using harmonicity, common onset, continuity (reconnecting tracks with short breaks into one element) and proximity (intended to group together the many short tracks arising from non-harmonic noise energy). Groups of elements that have common

\(^{3}\) In this example stationary source and DFT as “Object formation” are assumed.
onset are formed, groups of elements obeying harmonicity rule is formed and so forth. The first stage results in large number of groups. In the second stage the idea is to reduce the number of groups. Sets of elements that were grouped similarly in the harmonicity grouping process and in the onset grouping process are searched for and merged. Groups resulting from the continuity grouping and the proximity groupings are merged if they were similarly grouped similarly in the onset grouping.

After the grouping, the synthesis is performed using the inverse process. For instance, in Parsons model the synthesis was done basically via inverse DFT resulting in time domain separated signal. However, as Ellis points out, it depends on the application what should be done with separated groups of elements. For instance, the data may be already in a proper form, or at least, sufficiently preprocessed for a speech-recognition system.

CASA – prediction-driven architecture

Prediction-driven architecture to implement a CASA system is proposed using by Ellis in his dissertation [20]. In fact, Ellis is the issuer of the term prediction-driven CASA. The system that is used as an example is illustrated by Figure 3.9. Note that in order to understand the principles one should review the data-driven approach prior to studying the procedure below.

Let us review the operation of system by its core principles. The complete theory and experiments with the system can be found in [21]. Analysis of auditory environment is conducted obeying analysis-by-prediction-and-reconciliation principle. It means that the cues for the time-frame following the current frame are predicted based on current state. As the next time-frame is obtained from “Front-end”, the informa-

---

4. referred to as “contextual memory” by Ellis
tion, that is, the measured evidence is reconciled with the predicted information by modifying internal state.

Internal state gives arise to predictions. It is a world model which, in turn, is an abstract representation of auditory events (see definition on page 12 of external world. Core world model is world model particular to this system. It refers to basic elements from which every possible auditory event can be built. To be able to reconcile the information of internal state, prediction block converts the information to the same form as “Front-end” outputs.

The system aims at complete explanation of signal mixtures, or auditory environments. Rather than trying to divide mixtures into the desired part and the undesired part, the system must process all the data in order to make the prediction work properly.

For “Representation” the same subsystems as for data-driven approach can be basically applied. However, to obey the principle of complete explanation, the elements should be chosen in such a manner that also non-periodic components can be accounted for (otherwise the world model is incomplete). In Ellis’s system three different type of basic elements are used. Noise clouds model the unstructured parts of sound, wefts model continuous parts and transients model short bursts of noise. Each of these, of course, have their parametric presentation that is discussed in detail in [20].

The system employs so called competing hypotheses strategy. It refers to a mechanism that is used to resolve ambiguities. The system may not be able to determine “the identity of a sound”. That kind of a situation may occur, for instance, in the case of a tone and noise burst mixture. While the tone and a noise burst co-exist, there may arise an ambiguity of whether the sinusoidal actually exists underneath the noise burst or not (see page 47 for more detailed discussion). Thus two parallel explanations must be made and convey them as long as the choice becomes clearer or even obvious. The actual implementation is so called “blackboard system” that is particularly implemented for this kind of purposes. A detailed description can be found in [20].

Data-driven approach versus prediction-driven approach

Data-driven approach is able to utilize signal content. However, it is not able to take advantage of a higher level property; context. That is, the information of “what is going on in the mixture” is neglected. A prediction-driven architecture is built to consider also this aspect. This is again one step further in building a computational system that tries to utilize yet further features of human hearing to enable sound separation. Another outstanding improvement over the data-driven architecture is the ability to account for non-periodic components. In fact, the processing of non-periodic data is a prerequisite for a prediction-driven system to work. But more importantly than just being able to model non-periodic sounds, the system should be considered as a higher level model compared to the data-driven approach.
Data-driven approach builds the elements according to available raw data, that is, from the observation. This is to say that at every instant of time basically “undistorted evidence” of sound events must exist to obtain the correct amount of elements. Note that, making of the assumption can partly be avoided by processing the element space afterwards. In general, however, if the cues to be measured are masked by a strong interferer in the mixture, signal feature extraction suffers leading to failure in cue extraction, object formation, and finally resulting in more or less deteriorated output. Consider a contiguous tone masked by noise bursts. As a noise burst and the tone co-exist the data-driven approach may fail to build contiguous elements as described in Figure 3.8 resulting in too many elements, or even worse, not detecting some of the elements at all. In the previous case, the object formation of the system results in multiple elements which have the same frequency and which are separated with gaps the length of which is defined by the length of noise bursts. In general, the resulting output is not the same as what is the actual case, or more precisely, the resulting element space from the desired sound source viewpoint is different from the case without the disturbance. To reflect this particular example to earlier discussion the data-driven approach can be said not to fulfill principle of continuity (see Section 2.3.3 for grouping principles).

3.4 Other sound source separation schemes

First of all, the presentation of the following sound source separation schemes has to be justified since they have no connection or they do not even aim to separate sound sources as a human being. However, this does not change the situation in the sense that a separation system built according to the principles presented below still can be used similarly to the previously presented systems.

The systems presented in this section are introduced only briefly since they are not directly related to human hearing and the issues discussed in Chapter 2. Thus no background knowledge is available to the reader to understand the origin of the schemes. However, the systems are discussed in such a manner that the basic principles can be understood. The methods are brought out to show that there are alternatives for sound source separation that derive from completely different theory.

**Blind source separation scheme**

One of the most often presented alternative for binaural separation techniques and CASA techniques is blind source separation (BSS). The term refers to any method that is used to resolve the original signal from the processed data, or observed data. BSS methods do not perform any type information search as the previously presented methods. The mixture is processed without any segmentation (in CASA systems the processing begins with mixture analysis). To what kind of processing the original data has been exposed, is not known. The fact that has lead to the introduction of the term “blind”. By definition, in a BSS system the knowledge of exact processing is not even needed but on the other hand several things in BSS have to be hard-wired due to this. 

In Figure 3.10 the general form of processing is presented. That is, the model of path which sound waves propagate, and the processing to which they are exposed before they are received by sensors is presented. \( \{ o_i | i \in \mathbb{N} \} \) represents sound sources, \( \{ x_j | j \in \mathbb{N} \} \) refers to signals that are recorded by microphones and \( \{ \hat{o}_k | k \in \mathbb{N} \} \) represents the separation result. The arrows describe the paths that sound waves propagate. On the top of each arrow is the impulse response of each path to model the alteration to which sound waves are exposed to. Note that only a few possible propagation paths are presented for the sake of clarity. \( \mathbf{M} \) describes (1) a set of all possible propagation paths and (2) the alteration process related to each path. (1) \& (2) are modeled by impulse responses which are considered to be independent of each other and, in general, different for each path. Note that the total amount of various paths is actually infinite. \( \mathbf{W} \), in turn, describes as well (1) and (2) resulting in the same type of a set but now each element in \( \mathbf{W} \) describes a path from a sensor to a source. In other words, we have a process and its inverse process in the form of \( \mathbf{M} \) and \( \mathbf{W} \), respectively. Note that despite the fact that the same notation is used here for sources as for auditory objects (definition on page 12) in Figure 2.10 source \( o_i \) here refers to a physical source.

In BSS, the problem of separation is formulated as a task to find out \( \mathbf{W} \) as accurately as possible. Once \( \mathbf{W} \) is solved it can be applied to \( \{ x_j | j \in \mathbb{N} \} \) resulting in \( \{ \hat{o}_k | k \in \mathbb{N} \} \). The problem of retrieving the desired result is thus well defined when it comes to the questioning of what information is needed to solve the problem. Furthermore, the problem may be expressed using matrix algebra, that is, \( \mathbf{M} \) and \( \mathbf{W} \) are matrices. Using \( \mathbf{M} \), the matrix \( \mathbf{W} \) can be solved.

In literature, the matrices \( \mathbf{M} \) and \( \mathbf{W} \) are alluded to as mixing matrix and unmixing matrix, respectively. Practical methods that implement BSS separation paradigm do not, however, perform separation via matrix calculus, instead, multiple adaptive algorithms to obtain \( \mathbf{W} \) has been proposed. Infomax-algorithm by Bell and Sejnowski that was particularly tested for the cocktail-party problem is presented in [3].
theory concerning Infomax as well as other approaches is exhaustively presented by Saarelainen in [49]. In addition to Infomax, multichannel blind least-mean-square (BLMS) is discussed. BLMS could be applied for audio signal separation as well. The performance evaluation of the methods using real-world audio signals are presented in his thesis.

Separation model based on neural networks

Explaining the basics of neural networks is not trivial and thus not possible here. A complete description from the basics to advanced theories can be found in [29].

The fundamental idea of the separation scheme derives from the modeling of neurophysiological structures of human brain. That would imply basically a primitive level approach when it comes to the discussion whether a system based on neural networks is a low-level or a high-level approach. However, the separation model based on neural networks as well as neural networks in general, inherently model high-level abilities of human brain. For instance, they can be basically used for modeling perception arising from sensation rather than just giving responses to stimuli with no whatsoever deduction (see [28] for definitions of terms).

The system used as an example is intended to perform the following task: organize the input mixture consisting of two streams (definition on page [28]). The system is illustrated in Figure 3.11. In principle there is no restriction concerning the signal type. Also the amount of sources is not limited [63]. The system is rather an instance of a grouping logic than a separation scheme. However, it has all the basic blocks – analysis, grouping and synthesis. In this sense it fulfills the characteristics of a separation system. It is worth mentioning that the core of the model is used as a grouping engine in more recent structures that fall into category of CASA based sound separation. A CASA system utilizing this kind of grouping engine is presented by Wang in [64].

The system in the figure is based on neural oscillatory network. The core element of the system is an oscillator described by ellipses. The network is spanned by ellipses and the connections between them. The system is described by Wang in [63]. That publication is actually a summary of the previous work. Wang is one of the pioneers of this type of approach. His earlier work can be found in [60], [61] and [62].

From Figure 3.11 it can be seen that the input mixture is divided in time and in frequency to a discrete number of oscillators resulting in a matrix each row of which is associated with a certain frequency channel, and each column of which is associated with a certain time instant. The number of rows is specified by frequency resolution, and the number of columns is specified by time resolution, and of course, by the length of a sample. “Input end” is basically a filterbank to obtain the frequency axis discretization. Each row thus accounts for some frequency range. Each output of “Input end” is connected to a certain row in the matrix and thus the oscillators in this row represent oscillators of certain frequency range. Each column in the row differ from each other by a delay which increases from left to right meaning that the activation of oscillators begins from the left side of the matrix. Finally, there is
a mechanism illustrated by “Global inhibitor”. It is connected to each oscillator in the matrix. It is a special oscillator that is excited by each oscillator, and once it is activated, it inhibits the output of each oscillator. Global inhibition mechanism is partly controlling the network. Its theory of operation is too complicated to be explained here (see [63] for a detailed description).

There are two types of connections between oscillators in the network. The mutual connection of the oscillators \( j \) and \( i \) is fully defined by permanent weight \( T_{ij} \) and dynamic weight \( J_{ij} \). Both describing the strength by which two oscillators are connected to each other.

The system implements an ingenious logic to perform the task of stream formation. The logic is referred to as selective gating in [63]. It means that once any oscillator switches to active mode it triggers also the oscillators stimulated by the same stimulation pattern simultaneously preventing the stimulation of all the other oscillators. This in turn results in the synchronization within each pattern and desynchronization between the separate patterns.

The model is able to perform segregation based on multiplicity of cues. This is obtained by the fact that one layer to accounts for one cue which is “observing”. For instance, the system could consist of “stream formation layer”, “onset synchrony layer” and so forth [63].
Chapter 4

Mid-level representation

In Chapter 2, a generally accepted model concerning how sound waves are processed by human beings as they propagate from the source is presented. Sound waves produced by a source finally result in reporting a change in auditory environment. A detailed analysis of human hearing requires modeling techniques for stimuli and subprocesses of human hearing that are involved in the detection of these changes.

Signal representations corresponding to sound waves are usually referred to as low-level representations. Representations capable of describing perceptions (see page 28 for definition) are referred to as high-level representations. Developing signal processing systems related to modeling human hearing embraces not only recognizing the different stages in the processing but also modeling of signals as they propagate from one stage to another. Yet, more interpretable forms than low-level representations are desired to enable high-level interpretations of the mixture. Modeling techniques that are applied at this level are referred to as mid-level representations. That is, it is neither seen as a mixture of pressure waves, nor does it yet have context (see page 46 for definition). See Figure 2.2 in order to recognize the stage at which a signal can be modeled with a mid-level representation. A mid-level representation is capable of modeling signals in between “Auditory periphery” and “Auditory cortex”.

The need of a more convenient presentation of a signal than a time domain presentation was discussed in Chapter 3 while introducing CASA architectures for sound source separation. It was pointed out that such a representation is a fundamental part of any CASA separation system. That is, tools for building elements (e.g. sinusoidal tracks) are needed (see page 43). In general the set of elements extracted from a signal by utilizing the tools is referred to as a mid-level representation of a signal.

Localization systems also utilize mid-level representations. However, the role is more or less supplementary than a necessity. Still, by utilizing a mid-level representation, such algorithms may receive a significant improvement in performance (e.g. robustness, accuracy, computational efficiency).

Mid-level representation techniques are often alluded to as spectrum modeling. This concept embraces two things: (1) in practice a mid-level representation is strongly correlated to signal spectrum. (2) Modeling stresses the fact that it is not a necessity, but often favorable however, to have a mid-level representation that is fully
invertable conversion of a signal (e.g. exact spectrum). It was pointed out in [22] that any conversion made to a signal that produces approximately equal amount of information has failed as a mid-level representation by not discarding the unimportant information.

4.1 Desirable properties of mid-level representations

A more specific description or a definition of mid-level representation than the discussion above is hard to give. However, demands for a mid-level representation are easier to specify since it is known what exactly is desired. The fundamental requirements for a mid-level representation according Ellis and Rosenthal are discussed in [22]. The properties are summarized in Table 4.1.

<table>
<thead>
<tr>
<th>property</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. sound source separation</td>
<td>Ability to organize the mixture to sound events instead of a plain conversion</td>
</tr>
<tr>
<td>2. invertibility</td>
<td>A representation should carry at least as much information as it is needed to reproduce the original in the extent (or quality) as is desired</td>
</tr>
<tr>
<td>3. component reduction</td>
<td>Mixture may be viewed as an array of energy levels. A mid-level representation should decrease the total amount of elements in the array and, yet, increase the meaningfulness of an individual element in the array</td>
</tr>
<tr>
<td>4. abstract salience of attributes</td>
<td>The link of an extracted feature, and represented by a mid-level representation should reflect the perceptual attributes rather than an algorithmic attribute (e.g. loudness vs. sound intensity level)</td>
</tr>
<tr>
<td>5. physiological plausibility</td>
<td>No hypotheses that are inconsistent with physiological knowledge of human hearing should be employed</td>
</tr>
</tbody>
</table>

The list can be interpreted in such a way that if a mid-level representation satisfying the properties 1. - 5. was discovered, an ideal (multipurpose) mid-level representation would be available. Instead of fulfilling the demands in Table 4.1, most of the recently discovered, and in use mid-level representations do not fulfill even one of the properties fully. However, they perform satisfactorily for the purposes they are applied.

Ellis and Rosenthal classify mid-level representations using “a mid-level representation space”. The classification and how different mid-level representation are mapped onto that space is illustrated by Figure 4.1. Axis “bandwidth variability” describes the choice between fixed bandwidth (e.g. DFT) and variable bandwidth (e.g. Constant-Q transform). “Discreteness” describes how a mid-level representation divides the input into meaningful constituents, that is, number of different elements. “Dimensionality”; time and frequency axis are the “natural” dimensions of a mid-level representation of an audio signal. Advancing to the right on this axis increases dimensions, often just one more though. Note that the representations in the figure are just a few examples. Many others exist and they may even be combinations of some other mid-level representation.

A comparison between some mid-level representations can be found in [22], in which the idea of “Weft” is also presented. Note that according to Figure 4.1 Weft is the
most advanced mid-level representation compared to all others. Without describing Weft, it can be pointed out that it outperforms many other in use mid-level representations in view of Table 4.1 and Figure 4.1. However, a mid-level representation should be evaluated not only in the sense of how many of the properties are fulfilled but also in the sense to what extent each property is satisfied. Yet, the target application defines the validity of a mid-level representation.

### 4.2 Sinusoids + noise modeling

Sinusoidal modeling is presented by McAulay and Quatieri in [40] for speech coding purposes. Smith and Serra introduced practically the same kind of a model for music signals. Their system was published soon after the previously mentioned and it is presented in [53].

The fundamental idea is to present a signal with time-varying sinusoids each of which accounts for one harmonic component in a signal. The analysis is performed by dividing a time domain signal into sequential pieces (e.g. approximately 20 ms in length), estimating the periodic components of the sound by calculating the spectrum of each piece, or frame, and picking the most prominent peaks from the spectrum. For each periodic component, or sinusoid, amplitude, frequency and phase are estimated.

Note that up to this stage practically the same type of processing was used already in the separation system the introduction of which starts on page 41. However, the processing in sinusoidal modeling techniques continues by trying to form sinusoidal tracks. The calculation of the spectrum for each frame produces only three values per sinusoid. Thus, instead of hundreds of observations these three values suffice for presenting a component in a frame. Usually, the amount of components is relatively low. The tracks are formed by connecting sinusoids the frequencies of which are close to each other in adjacent frames (see Figure 3.8). As a result of the processing, the parameters of connected sinusoids can be interpolated between frames allowing smooth synthesis. Additionally, start and end time instant are known for each track. The synthesis may be implemented, for instance, using a bank of sinusoidal oscillators or calculating inverse DFT. The latter option is better if computational efficiency is of interest [35].
The signal split into sinusoidal part and noise part, or deterministic part and stochastic part, is first known to be suggested by Serra [51]. The idea of trying to extract as much as possible sinusoids and model the remaining part, or residual, by filtering white noise in a particular manner, has been used in many systems since Serra had introduced it. In the system by Virtanen [58] the deterministic part is obtained by searching the sinusoids from a signal in sinusoidal analysis. Using the estimated parameters, the deterministic part is synthesized in sinusoidal synthesis. Subtracting the deterministic part from the original, the stochastic part of a signal is obtained which is then the residual. Since the residual is stochastic and can be modeled using filtered white noise, it is often referred to as noise part of the signal. It must be noted that this is an ideal case. In practice, the estimation of all periodic components from a real-world signal is very difficult, and the residual may still contain components that would be presentable by deterministic means. Colloquially, some sinusoids still exist in the signal obtained in the subtraction.

In order to obtain a mid-level representation of the deterministic part, the spectral information is often of interest. It is used as a pre-analysis in forming elements, for instance, sinusoidal tracks. Also for the stochastic part the calculation of the spectrum is the next thing to do after dividing it into frames. However, the parameter extraction is not as “high-resolutional” compared to the deterministic part analysis. Two basic techniques are employed: time-varying filter related to the residual signal spectrum or estimating short-time energies in each predefined frequency band [58]. The parameters needed to form elements that represent stochastic entities are actually either the parameters of the time-varying filter or short-time energies, respectively.

4.3 Deterministic + stochastic model as a mid-level representation

Let us review how the properties presented in Table 4.1 are satisfied by sinusoids + noise techniques according to Virtanen in [58]. Criteria from 1. to 4. are satisfied at least in some extent. From the fifth feature, it is pointed out that sinusoids + noise model is poor by being rather physically than physiologically oriented modeling technique. It is not necessarily a shortcoming in the sense that data the model produces is not only in a mid-level form but also at so a generic stage that it can be further processed by higher level techniques to obtain, for instance, the same kind of context interpretation as would have been achieved with a more advanced mid-level representation. “The model produces oversimplified data, for which only a minimum amount of deduction has been done. If higher level information is desired the data can be easily analyzed using an upper level analysis which for example combines the sinusoids into separate sound sources.” [58]. In this respect it satisfies better than well the third property.

The first property is fulfilled once a higher level processing is fitted to use the data that the modeling produces. However, it must be pointed out that the noise part,
in general, is neglected as it is stochastic in nature, and thus a straightforward de-
duction concerning its origin is not possible. For the deterministic, or the sinusoidal
part, a more reliable deduction is possible. Several implementations concerning sep-
oration using sinusoids + noise have been published practically proving that the
property is satisfied in some extent.

The second and the third criterion are satisfied because the analysis by which the
mid-level representation is obtained produces parameters which, in turn, are used
in the synthesis. Usually, the number of parameters is significantly reduced.

4.4 Parameter extraction of sinusoidal modeling

The role of the parameter extraction is critical. How well the modeled signal corre-
sponds to the original signal is affected mainly by this stage. Next, a few algorithms
to extract the parametric representation are reviewed. This is more or less an in-
troduction to techniques. The reader is encouraged to check the references for more
details.

Parameter extraction can be further divided to finer stages. One possible segmenta-
tion can be found in [58]. (1) Peak detection, (2) peak interpolation, (3) parameter
estimation and (4) peak continuation can be recognized in sinusoids + noise mod-
eling systems. The segmentation arises from the fact that for each stage there are
clearly independent algorithms.

The first stage is compulsory and has the most significant role in general. Stage 2 is
used for better frequency resolution. Stage 4 covers an advanced set of algorithms
that are used for connecting sinusoidal tracks from frame-to-frame. Exhaustive stud-
ies of the stages can be found in [58] and in [35].

Sinusoidal track formation, which is the ultimate goal, (see Figure 3.8) is possible
without performing any operations in stage (2). Stage (4) is also optional in the
sense that its purpose is to perform inter-frame analysis. That is, it tries to connect
the sinusoidal tracks in the previous frame to the sinusoidal tracks estimated in
the current frame. Instead, a more simplified solution is to perform an independent
processing in each frame. However, it must be pointed out that signals for which
sinusoidal modeling is intended to (e.g. abundant periodic content), using the algo-
rithms developed for stages (2) and (4) a significant improvement has been achieved
in the applications.

Due to the important role of peak detection, two methods and their properties are
presented. Issues regarding parameter value refinition, that is, theories and imple-
mentations of stages (2) and (4) are not covered here.

4.4.1 Peak picking

Peak picking is the simplest method for extracting the deterministic information in
a signal. The name of the algorithm arises from the fact that it examines the signal
spectrum searching for local maxima. Once a maximum is found its spectral key
CHAPTER 4. MID-LEVEL REPRESENTATION

figures are recorded. This is the convenient property of the method; an impulse in
frequency domain corresponds a stable tone in time domain. Reading the frequency
domain values for the impulse (magnitude, location on frequency axis, and argument)
will result in the parametric representation of the tone.

Fourier series expansion is capable of exact representation of a periodic signal with
an infinite sum of sinusoids once Dirichlet’s conditions are met [30] p. 301]. Real-
world signals, however, are not periodic over an arbitrary time period, and thus they
are not presentable via Fourier series expansion. Instead, Fourier transformation is
a powerful tool for an arbitrary signal.

The usage of Fourier transformation embraces calculating integral over time range
of \((-\infty, \infty)\) for a continuous signal. In practice, the signal processing is done to a
signal which only approximates the original by a number sequence. Thus instead of
integration, a sum is calculated resulting in a continuous transformation of a discrete
sequence. The final result is known as the discrete-time Fourier transform (DFT) of
a signal. However, the practical solution instead of calculating DFT for a signal is
to use short-time Fourier transform (STFT) which introduces splitting a signal into
pieces and applying the formulae of the DFT on each piece separately.

For each signal piece, or frame, a complex number sequence is obtained. Applying
absolute value operator for the sequence will result in amplitude spectrum and ap-
plying argument operator will result in phase spectrum of the current frame. From
the amplitude spectrum it is easy to pick a desired number of local maxima. The
values of the local maxima of the amplitude spectrum correspond to amplitudes of
sinusoids. The index of the picked value multiplied with frequency resolution cor-
responds, in turn, the frequency of the sinusoid. Using the index, the phase of the
sinusoid is obtained from the phase spectrum.

4.4.2 Cross-correlation method

The cross-correlation method employs also STFT as a pre-stage to the actual detec-
tion algorithm. Rather than just detecting local maxima in amplitude spectrum, a
deeper analysis is made for the STFT sequence.

It was pointed out earlier that a sinusoid corresponds an impulse in frequency do-
main. Since DFT does the mapping to frequency domain, an impulse-like representa-
tion is obtained for a single sinusoid. Limiting the DFT into a finite time interval,
resulting transform for a sinusoid, is not a pure impulse. The impulse is smeared in
frequency domain, the shape of smearing depends on the utilized window function.

In the cross-correlation method the detection is based on calculating the STFT of
a signal and isolating a certain frequency range for examination. The isolated fre-
cuency range, that is, a part of the STFT of a signal, is then compared to the STFT
resulting from an ideal sinusoid. This is performed by calculating cross-correlation
between these two STFTs. The obtained correlation function is scaled in such a way
that it gets the values between 0 and 1. The scaled correlation function is sinusoidal
likeness measure.
Let us denote the spectrum of a sampled ideal sinusoid $\psi(k)$ by $\Psi(\omega_m)$. $k$ denotes sampled time axis and $\omega_m$ sampled frequency axis. Let us denote the spectrum of the analysis window, or frame, that is used to calculate $\Psi(\omega_m)$ by $H(\omega_m)$. $\Psi(\omega_m)$ can be interpreted as frequency-translated, scaled and phase-shifted version of $H(\omega_m)$ [58]. Natural signals are viewed thus as a sum of several $H(\omega_m)$ with proper scaling and phase shift associated to each frequency $\omega_m$ from Fourier analysis viewpoint. Let us denote STFT for an arbitrary signal $s(k)$ by $S(\omega_m)$. The correlation is calculated over a bandwidth $[−W, W]$. The resulting correlation function $r(\omega)$ can be expressed then by Equation (4.1).

$$r(\omega) = \sum_{m, |\omega - \omega_m| < W} H(\omega - \omega_m)S(\omega_m) \quad (4.1)$$

The equation produces non-zero values at locations in frequency axis if there exists energy associated at a certain discrete frequency $\omega$. The calculation of $r(\omega)$ means examining the frequency axis at $\omega$ and its environment constrained by $W$ which is chosen based on signal content and considering the requirements of the application.

$$|H_\Omega|^2 = \sum_{m, |\omega - \omega_m| < W} |H(\Omega - \omega_m)|^2$$

$$|S_\Omega|^2 = \sum_{m, |\omega - \omega_m| < W} |S(\Omega - \omega_m)|^2 \quad (4.2)$$

$|H|_\Omega^2$ and $|S|_\Omega^2$ in Equation (4.2) define certain norms for window and signal at frequency $\Omega$, respectively. Using the norms, absolute values of $r(\omega)$ at frequency $\Omega$ can be properly scaled.

$$\nu_\Omega = \frac{|r(\Omega)|}{|H_\Omega| |S_\Omega|} \quad (4.3)$$

In Equation (4.3), $\nu_\Omega$ is the desired likeness measure. The scaling ensures that $\nu_\Omega$ is in the interval of $[0, 1]$. 1 indicates that there is a strong sinusoidal component at frequency $\Omega$ and 0, in turn, indicates that there is no periodic content at frequency $\Omega$.

Some threshold between 0 and 1 has to be set to check the validity of a peak via the calculated likeness measures. A peak is valid if it satisfies Equation (4.4).

$$\nu_{\omega_m} > \alpha \quad m = 0, \ldots, M - 1 \quad (4.4)$$

As stated above, in addition to $\alpha$, another adjustable parameter of the cross-correlation method is bandwidth $W$. $W$ specifies the search range in the environment of some frequency $\omega$. The search range is $[\omega - W, \omega + W]$. The selection of both parameters depends strongly on signal content, thus, neither of them can be fixed as long as the content is unknown.
After selecting some plausible threshold \( \alpha \) and feeding a signal to the system results basically in the knowledge of the amount of periodic content. The answer has been found to the question: “how much of the signal can be presented with a sum of sinusoids?”. \( \nu_\Omega \) is calculated in each frame and compared to the threshold. The values that exceed the threshold refer to the same kind of spectrum that results from a sinusoid.

The peak picking method had the convenient property for “implementing” the stage 3, or parameter estimation. Also in the case of the cross-correlation method the stage is implemented almost as inherently. It can be done via already calculated \( r(\omega) \).

\[
A(\Omega) = \frac{|r(\Omega)|}{|H_\Omega|^2} \tag{4.5}
\]

\[
\phi(\Omega) = \text{Arg}\{r(\Omega)\} \tag{4.6}
\]

In Equation (4.5) \( A(\Omega) \) is the amplitude and in Equation (4.6) \( \phi(\Omega) \) the phase of the sinusoid at frequency \( \Omega \). The equations provide thus a convenient way to access the parameters without complicated calculation.

### 4.4.3 Peak picking and Cross-correlation – properties and comparison

Short comparison between the two presented methods is appropriate by reviewing their properties briefly.

**Demands and properties**

The fundamental task of any peak detection method is to find the “meaningful peaks” in order to achieve a proper representation of a signal. It is important to stress the issue because on the one hand it is desired to get rid of unimportant information to achieve the demands set for a mid-level representation, and on the other hand it is desired to achieve the goal of modeling. The goals are somewhat opposite. The former is ready to discard more information than the latter which tries to be as close as possible to the original signal.

“Meaningful peak” is difficult define, since the amplitude does not only specify “the meaningfulness” of a sinusoid. However, meaningful peaks may be thought of as sinusoidal components that are very characteristic to a signal. Unmeaningful peaks may be caused by noise-like components in a signal, or still, they may be actually presentable by sinusoids but due to their weakness, they can be neglected. Additionally, the noise-like components would require a considerable amount of sinusoids to achieve either of the goals discussed above.

The peak picking algorithm detects a pre-selected amount of peaks. No care is taken whether or not a signal in the current frame contains any information that is presentable by deterministic means. Many natural signals, for instance speech and music, are presentable largely with sinusoids. However, a speech signal in general contains also a lot of stochastic information but also pauses during which basically only
ground noise exists. If the peak picking algorithm is used for noise-like signals with the same parameter set as for periodic content, the processing usually results in poor modeling of the original signal. Colloquially, the amount of sinusoids is absolutely too low.

One improvement to the simplest method is to set an amplitude threshold and regard the peaks that exceed the threshold as meaningful peaks. It was pointed out in [58] that still problems remain. Simple amplitude rule cannot guarantee that a peak is not resulting from the stochastic part. Yet, in case of natural sounds, lower harmonics have usually significantly larger amplitude than upper harmonics, which are possibly ignored due to the simple amplitude rule.

A method which detects peaks while taking into account the above property of natural sounds is the cross-correlation method. Another powerful method that is also able to make use of this higher level of knowledge on signals is known as f-statistics presented by Thomson in [55]. The performance of the method including an exhaustive comparison to other methods is presented in [58]. In the cross-correlation method these spectral properties are taken care of by scaling the absolute value of the cross-correlation function with rough spectral shape (see Equation (4.3)). The effect is that the peaks due to noise may be more reliably avoided from detecting them as sinusoids since they are usually low in amplitude. Noise peak at $\Omega$ will result in smaller value of $\nu_\Omega$ compared to the sinusoidal component at $\Omega$. This is the case even if the noise peak had the same amplitude as an ideal sinusoid occurring at the same frequency. It is due to the fact that noise-like signals have energy over the whole bandwidth of $[\omega - W, \omega + W]$ while the energy of a sinusoid is concentrated only in the vicinity of $\omega$ resulting in smaller $|S_\Omega|$. The frequency range over which any method operates is basically limited only by sampling rate. More interesting property is frequency resolution. For the peak picking method frequency resolution, or the minimum separation of two ideal sinusoids, is defined by sampling rate and the length of the analysis window. The minimum distance between two frequency components is $f_s/K$ where $f_s$ is sampling rate and $K$ is the length of the analysis window. In practice, $K$ is the only parameter by which frequency resolution can be affected. Obviously, it should be selected as large as possible for the best frequency resolution. However, since $f_s$ is usually quite large and $K$ is limited by the fact that reasonable time resolution is also of interest (see page 62). Due to these facts, frequency resolution may be unacceptably poor.

However, better frequency resolution without losing in time resolution can be achieved by so called zero-padding. In zero-padding the analysis window is extended by adding samples the amplitude of which is zero to the end of true samples in each analysis window. Let us assume an analysis window length of 1000 samples and sampling rate of 48 kHz. The resulting frequency resolution is 48000 Hz / 1000 = 48 Hz. The analysis window is zero-padded by adding 3000 zeros after the first 1000 samples. With this trick the analysis window is 4000 samples and the resulting frequency resolution will be 48000 Hz / 4000 = 12 Hz. Note that the information is not increased but the method interpolates more frequency points.

Finally, it can be stated that due to the variable weighting of the frequency axis
the cross-correlation method outperforms the peak picking method if the reliability and validity are the arguments. However, in low noise conditions the peak picking method may perform good enough for some purposes. In [58] it was pointed out that since the peak picking method produces a fixed bitrate (produces the same amount of sinusoids for each frame) and relatively low computational load, it is an attractive method for audio coding purposes.

**Implementation issues**

Various amount of sinusoids per frame were tested in the initial simulations of the system proposed in this thesis. For instance, 25, 50 and 100 sinusoids were tested. After the initial evaluation it was concluded that raising the amount of sinusoids above 25 with the test material that is used in this work, no significant improvement in the modeling was gained (signals are listed in Appendix A).

At this point it is worth pointing out the problem related to the amount of sinusoids and the simplicity of the peak picking algorithm. Figure 4.2 illustrates the problem. Instead of detecting one sinusoid, which would be correct from the viewpoint of modeling, several near-to-each-other peaks are detected. This is due to the fact that the peak picking method finds a fixed amount of sinusoids in each frame. In the case of one sinusoid, this may result in perceivable artifacts after synthesis. With natural signals, like speech, for instance, the algorithm may neglect several strong peaks in the spectrum because all the available sinusoids have already been spent on a peak that could have been modeled with just one sinusoid.

The cross-correlation method introduces a likeness measure that enables a certain level interpretation of signal spectrum (Equation (4.3)). Using the likeness measure, and fixing the threshold \( \alpha \), enables the evaluation of each prominent spectral peak in the sense that should that particular peak be spared a sinusoid or not. For instance, consider the region after 1000 Hz in Figure 4.3. Right after 1000 Hz boundary there is a strong peak. Also \( \nu_{1000Hz} \) is near 1. However, the next peak at higher frequency the corresponding likeness measure value is only about 0.6. From this it may be deduced that the latter peak results from the previous peak or from background.
noise. Anyhow, it may be better to model the first peak with a sinusoid but neglect the latter.

To illustrate the performance using the same 1000 Hz sinusoid as with the peak picking method is presented in Figure 4.4. Apparently, the cross-correlation method fixes at least this particular problem. Formally, referring the issue back to the theory of modeling it can be stated that the cross-correlation method results in better modeling of a signal than the peak picking method.

### 4.5 Transient modeling

A transient can be vaguely defined as an abrupt very short burst of noise. This is to say that the energy related to transient rises “without warning” and falls quite fast too. This is the very fact that makes the stochastic part of the sinusoids + noise model to perform relatively poorly with transient-like sounds. The energy of “ordinary noise” varies little or, at least, very slowly compared to the energy of transients.
It must be pointed out that some transients are partly periodic, the energy of transients rising rapidly but dampening being periodic. In view of this property, basically the deterministic part of a sinusoids + noise modeling system could be used to present not only sinusoids but also transients. However, the properties of a transient prevent the usage in practice. It was pointed out in [58] that transients often have a very large bandwidth. From basic Fourier analysis theory it is known that any signal can be presented as a sum of sinusoids. The more complex the signal, the more sinusoids are required to obtain the equivalent of the original. The Fourier series expansion of a broadband transient results in a very large number of terms. Additionally, a problem called pre-echo occurs if sinusoidal modeling is applied to a transient [35]. Pre-echo is due to the fact that time resolution is poor for transients.

The practice introduces another constraint for the usage of deterministic part in addition to the theory discussed above. Discrete-time Fourier transform is used to obtain the parameters; amplitudes, frequencies and phases, for sinusoids. There is a trade-off between time resolution and frequency resolution. This is an instance of Heisenberg’s uncertainty principle. A good time resolution results in poor frequency resolution and vice versa. In order to obtain a good estimate for the deterministic part in sinusoidal analysis, often a relatively long analysis window is required. For transients, the analysis window of that length is absolutely too long. A proper length of the analysis window for modeling a transient, from practical viewpoint, is approximately the same as the length of the transient.

Due to the limitations in presenting broadband transients using sinusoids or noise, there is an apparent need for handling transients and transient-like sounds in a signal. The fundamental issue in transient modeling is the detection of transient regions, on the one hand in time domain and on the other hand in frequency domain. As transients occupy the majority of the total energy while they occur in a signal, it is sufficient for most purposes to evaluate the location only in time domain. During transient regions, the sinusoids + noise analysis of a system is practically turned off. Levine in [35] presents one implementation of a transient detector used as a part of an audio coding system. The technique employs a sinusoids + noise modeling technique to obtain the parametric representation for the deterministic part and the stochastic part. Transients are detected, and then they are presented with a non-parametric representation obtained via transform coding. The purpose of the transform coding is in that particular application data compression. However, if data compression is not of interest, a transient may be as well left untouched but the detection of transients and the labeling of transient regions still have crucial roles to prevent the usage of sinusoidal modeling on transients.

4.6 Transient detection

Transient detection methods have received very little interest compared to sinusoid detection methods. Thus not so many implementations to detect transients in signals are available. This is somewhat surprising since their presence or absence is perceived in listening tests, for instance, in the case of speech signals. However, in
some applications the lack of transients is not a serious shortcoming. For instance, computational efficiency and the delay resulting from processing may be more crucial.

### 4.6.1 Transient detection algorithm

The detection algorithm presented here is proposed by Levine in [35]. Levine pointed out the existence of another detection algorithm by Masri et al. which can be found in [39].

In the sinusoidal detection methods the fundamental issue is defining the concept of the meaningful peak. In the transient detection basically the same kind of question arises. It may be formulated: “Is the energy rise so abrupt in the current signal segment that it is plausible to consider it as a transient?” In the extreme case the whole signal can be considered as a sequence of transients. This is justified by the fact that a natural signal varies rapidly over any arbitrary time period excluding differentially small periods.

Some higher level knowledge has to be relied on while constructing a transient detector. The knowledge is that natural signals do not contain several transients over short sequential segments in time. Thus a mechanism limiting the transient occurrence rate can be built to improve the detection. However, such a transient detector may discard some more prominent transients and detect the one that may well be discarded by human hearing. Neglecting the important transient results in even worse phenomenon in a system that applies sinusoidal modeling. The strong transient region is processed using the parameter extraction for sinusoids (see [4.4]) which results not only in “wrong analysis” but also in a deteriorated signal if it was synthesized.

Another important issue is of course to find out the time instants at which transients begin. These time instants are referred to as onset time or attack time. The duration of a transient is more less a free parameter, the selection of which however, is not trivial.

Let us review the system by Levine ([35]) in which the problems discussed above are taken care of. The block diagram of the system is illustrated in Figure [4.5]. Two methods to obtain reliable accurate onset time estimates are employed. The first method is “a rising edge detector”. It searches for rapid changes in signal energy and labels them. The energy estimates are calculated by splitting the signal into pieces practically in the same way as in sinusoidal analysis which is presented previously in this chapter. However, in order to obtain better time resolution, much shorter analysis window is used. The detector itself is a simple predictor that compares the current frame energy to a weighted sum of the previous frames’ energies. If the energy is significantly higher than the average calculated using the previous frames, the current frame is labeled as a candidate transient frame.
The energy of a frame is calculated using Equation (4.7).

\[
E^s(n_{tr}) = \sum_{k=0}^{K_{tr}-1} h(k)[s(n_{tr}K_{tr} + k)]^2
\]

\[n_{tr} = 0, \ldots, N_{tr} - 1\] (4.7)

where \(n_{tr}\) refers to the current frame, \(k\) is discrete time variable, \(K_{tr}\) is analysis window length in samples, \(h(k)\) is Hamming window function, \(s(\cdot)\) is the original signal and \(N_{tr}\) is the amount of frames. The overlap percentage is fixed to 50%.

Once the processing begins energy values for each frame are stored. In order to detect rapid energy rises in a frame, some reference value has to be established. Mean energy of past frames is used as the reference. However, it is not plausible to use all the past frames but just a few. \(P\) frames, just preceding the current frame, are taken into account. The reference for frame \(n_{tr}\) may be then expressed by Equation (4.8).

\[
E^{ref}(n_{tr}) = \frac{1}{P} \sum_{p=1}^{P} E^s(n_{tr} - p)
\] (4.8)

where \(P\) is the amount of past frames used in comparison.

Finally, in each frame the comparison in Equation (4.9) is performed.

\[
E^s(n_{tr}) > \epsilon E^{ref}(n_{tr})
\] (4.9)

where \(\epsilon\) is a parameter allowing the adjustment of the threshold. If Equation (4.9) holds, the current frame is marked as a potential transient frame.

The second method requires sinusoidal model with synthesis capability, since it uses also [the residual signal] (see page 54 for definition). The detection is based on observing the performance of sinusoidal modeling in the current frame. If sinusoidal modeling performs poorly, it is probable that there is a transient in the current frame.

The residual energy is calculated using Equation (4.10) and the original signal energy using Equation (4.7).

\[
E^r(n_{tr}) = \sum_{k=0}^{K_{tr}-1} h(k)[\hat{s}(n_{tr}K_{tr} + k) - s(n_{tr}K_{tr} + k)]^2
\]

\[n_{tr} = 0, \ldots, N_{tr} - 1\] (4.10)

where \(\hat{s}(\cdot)\) refers to the estimated and synthesized deterministic part of the signal.

For each frame, the ratio described by Equation (4.11) is calculated.

\[
\rho(n_{tr}) = \frac{E^r(n_{tr})}{E^s(n_{tr})}
\] (4.11)
Equation (4.11) produces values on the interval of [0, 1]. Values near zero indicate that the signal had such a content that much of it could be presented with a sum of sinusoids. In other words, it is not likely that the current frame has a transient. On the other hand values near one suggest that a significant part of the signal is not presentable by a rational amount of sinusoids referring to a transient content in the current frame.

A scalar threshold is needed to classify the frame between transient and non-transient frame. The comparison specified by Equation (4.12) is performed.

\[ \rho(n_{tr}) > \eta \]  \hspace{1cm} (4.12)

where \( \eta \) is the threshold. Theoretically it is limited to range [0, 1]. If Equation (4.12) holds, the current frame is marked as a potential transient frame. Actually, Equation (4.11) may produce values exceeding unity due to the fact that perfect synthesis is unachievable.

The resulting information of the transient detection methods may be viewed as two sets which carry the information concerning the transient content in a signal. In order to obtain reliability in the detection, both methods must indicate a transient within a frame before it is considered valid. Thus the set of valid transient frames is the intersection of these sets. A transient is valid if and only if Equation (4.9) holds and Equation (4.12) holds. This process is illustrated by “thresholding” in Figure 4.5.

As a result, a presentation concerning the transient content of the signal has been obtained.

After transients have been located on time axis, so called transient regions are formed. Levine fixed 66 ms limit to the duration of a transient region [35]. In addition to forming transient regions, 50 ms before and 150 ms after the time instant are labeled as non-transient regions. As a consequence of the labeling, a transient are not detected more than once. If the labeling was not used, the same transient could be detected as the next frame is processed. A transient could also be detected “too late” without this operation. Additionally, it reduces the total amount of transients to five per second.
Chapter 5

Description of the system

This chapter covers the description of the separation system that was implemented to evaluate the validity of the hypotheses discussed in Chapter 1. Despite the fact that the separation is the ultimate goal, the localization system has a crucial role in the separation system. Thus, a lot of effort was put into implementation and evaluation of the localization system.

At first, the signal model is briefly presented to justify the solutions presented for the localization system and the separation system. System overview in 5.2 can be used as a outline to this chapter.

5.1 Signal model

Any signal produced by musical instruments or by a physical system may be viewed to consist of two parts; the deterministic part and the stochastic part, which can be modeled as a sum of sinusoids and as filtered white noise, respectively\(^{52}\). The signal model is defined by Equation (5.1)

\[
s(t) = \sum_{m_s=0}^{M_s-1} a^{(m_s)}(t) \cos \left[ \theta^{(m_s)}(t) \right] + r(t) \tag{5.1}
\]

where \(a^{(m_s)}(t)\) is amplitude and \(\theta^{(m_s)}(t)\) instantaneous phase of \(m_s\)th sinusoid. \(r(t)\) embraces the stochastic part at time instant \(t\).

It must be pointed out that \(r(t)\) contains not only that part of a signal what is in general modeled with filtered white noise but it often contains a lot of periodic information. That is, the first term in Equation (5.1) is used to account for only the most prominent periodic components in a signal. Thus a reasonable upper limit for \(m_s\) can be fixed. Consider a sinusoid modeled using the equation. It can be modeled with \(M_s = 1\). Still, \(r(t) \neq 0\) since perfect estimation of the deterministic part is impossible. Thus, \(r(t)\) also accounts for the errors made in the estimation.

In this particular system sinusoids are assumed to be locally stable, that is, no changes are allowed to occur within a time interval fixed by analysis window. This
is of course not true in a general case but on the other hand it is not totally unreasonable assumption since the time interval is usually small. In fact, deterministic + stochastic models in general assume that sinusoids do not exhibit rapid changes. However, slow variation of amplitudes and frequencies are allowed over analysis window \[ 52 \].

Keeping the total complexity of the system low enough was one of the goals while designing the system. Therefore, certain stages in the deterministic + stochastic modeling stage were neglected. For instance, the peak interpolation stage in a sinusoidal modeling system (see page \[ 55 \]) is not used since it is used for refining the parameter estimates.

### 5.2 System overview

The system which was implemented to evaluate the hypotheses and the applicability of algorithms for the localization and for the separation with the given measurement setup is illustrated in Figure 5.1. The setup, in turn, is discussed in Chapter 6. Figure 5.1 illustrates the system built in the project related to the thesis from the separation viewpoint. However, it is worth pointing out that DOA estimation has a significant role in the overall performance since the location of a sound source is used as a primary cue for separation.

The illustration of the system is fitted into the general separation scheme discussed in Section 3.3. The three major stages are implemented – “Analysis stage”, “Grouping stage” and “Synthesis stage”. Next, each stage is discussed in detail. However, due to the fact that the analysis stage has the most important role to obtain desired results in the processing, the major emphasis is on that stage.

### 5.3 Direction-of-arrival estimation

The preliminary requirements for the DOA system in this thesis with a particular measurement setup are the following: (1) operate reliably on data recorded in real-world environments, (2) handle arbitrary signal contents, (3) make no assumptions concerning environment. The setup as well as the environments and the signal types
are described in detail in Chapter 6. It is worth pointing out that DOA systems in general do not perform either reliably or accurately with two-sensor setups. Instead, advanced signal processing techniques have to be utilized.

A DOA system that largely fulfills the requirements is introduced by Liu et al. in [23]. The basic structure is illustrated in Figure 5.2. The implementation in this thesis contains certain simplifications and modifications to make it more usable considering the goals. However, the basic structure is exactly the same. Note that the system is strongly related to the models in Chapter 3. That is, it utilizes the same principles as the early models of human hearing.

Before the details, let us describe the processing roughly by citing Liu et al. “Two input signals are transformed to frequency domain and analyzed for coincidences along left/right-channel delay-line pairs. The coincidence information is enhanced by a nonlinear operation followed by a temporal integration. The azimuths of the sound sources are estimated by integrating the coincidence locations across the broadband of frequencies in speech signals.” [23].

### 5.3.1 Frequency analysis

The signals recorded by the measurement setup covers the processing up to stage in which left and right channel are referred to as $s_{ln}(k)$ and $s_{rn}(k)$. $k$ is the discrete time axis. Frequency analysis is a frame-based operation. The subscripts $ln$ and $rn$ refer to left/right signal in frame $n$. As a result, Fourier transforms $S_{ln}(m)$ and $S_{rn}(m)$ are obtained, where $m = 0, \ldots, M - 1$. $m$ is the discrete frequency axis the total amount of frequency points is then $M$.

First, a signal is split in pieces, or frames, the total amount of which is defined by the frame length $K$, the overlap percentage and, of course, the duration of a signal. In order to avoid the Gibbs phenomenon (detailed discussion of the phenomenon in [30]), sampled Hamming window is used to shape each frame before the actual frequency analysis.
After windowing, STFT can be performed. The combined windowing and transform can be then expressed in the form of Equation (5.2).

$$S_{(l,r)n}(m) = \sum_{m=0}^{M-1} h(k)s_{(l,r)n}(k)\exp\left(-\frac{j2\pi nk}{M}\right)$$  \hspace{1cm} (5.2)

where $h(k)$ is Hamming window function.

The frame length $K$ was desired to correspond approximately to 20 ms in continuous time due to the properties of test signals. Using 48 kHz results in analysis window length $K = 1024$. The overlap percentage was fixed to 75. In order to achieve a plausible frequency resolution $M = 8192$ was used which results in frequency resolution about 6 Hz. Note that $M = 8192$ may be exaggeration since a good frequency resolution is not the primary interest. However, this frequency resolution was used to enable the comparison of the results presented by Liu et al. in [23].

### 5.3.2 Coincidence detection

The ultimate goal of the coincidence detection is to find the spatial location of a sound source. Therefore, it has the most important role in the localization system. The theory regarding the coincidence principle is presented in Section 3.1. Two signals are fed to the delay line and the coincidence detection finds out that particular location in the delay line where the difference between the signals is at a minimum. In the theoretical model, the delay is found based on calculating cross-correlation $R(\tau)$ between left and right channel (see Equation (3.1)). The estimate of the time delay is obtained by seeking the maximum of $R(\tau)$.

In the system presented by Liu et al. in [23], the task of Equation (3.1) is done in another way. An outstanding difference to the theoretical model is that the processing is done in frequency domain. In Figure 5.2 the frequency domain notation (capital letters) is used for the two input sequences. The minimum difference between two input signals is found by feeding the coincidence detection system by $S_{lm}(m)$ and $S_{rn}(m)$ which are STFTs of respective discrete time signals $s_{lm}(k)$ and $s_{rn}(k)$ (see Figure 5.2). The signals are time delayed with each value $\tau_i$ : $i = 1, \ldots, I$. The resulting STFTs of the delayed signals are presented in Equation (5.3). The calculation of STFTs can be limited to the range of $[0, M/2 - 1]$ due to the fact that the interval $[M/2, M]$ consists of complex conjugates of the previous range.

$$S_{lm}^i(m) = S_{lm}(m)\exp(-j2\pi f_m\tau_i)$$  \hspace{1cm} (5.3a)

$$S_{rn}^i(m) = S_{rn}(m)\exp(-j2\pi f_m\tau_{I-i+1})$$  

$$m = 0, \ldots, M/2 - 1$$  \hspace{1cm} (5.3b)

$$i = 1, \ldots, I$$

where $S_{lm}^i(m)$ is the STFT of the left channel delayed with the time delay $\tau_i$ and $S_{rn}^i(m)$ is the STFT of the right channel delayed with the time delay $\tau_{I-i+1}$. Note

---

1. It was pointed out in [23] that the used method is an analogous operation to the operational principle of biological systems which is presented by Feng and Shofner in [24].
especially that in Equation \( (5.3) \) \( m \) is the index of frequency axis and \( f_m \) is frequency in Hertz which is obtained using the relation presented in Equation \( (5.4) \).

\[
f_m = \frac{m}{M} f_s
\]
\[
m = 0, \ldots, M - 1
\]

\( (5.4) \)

where \( f_s \) is sampling rate in Hertz.

\( I \) variants of STFTs are produced by Equation \( (5.3) \) for each channel. Next, each variant is compared between left and right channel. For each time delay \( \tau_i : i = 1, \ldots, I \), a similarity measure is formed. This is described by Equation \( (5.5) \).

\[
\Delta S^i_n(m) = \left| S^i_{ln}(m) - S^i_{rn}(m) \right|
\]
\[
m = 0, \ldots, M/2 - 1
\]
\[
i = 1, \ldots, I
\]

\( (5.5) \)

where \( \Delta S^i_n(m) \) is the similarity measure which enables the actual comparison. Thus, it has the role of \( R(\tau) \) in the theoretical model.

Finally, the output of this stage is found by picking the minimum of the sequence \( \Delta S^i_n(m) : i = 1, \ldots, I \) and recording its place in the delay line, that is, the index \( i \). The delay index estimate is obtained by evaluating \( i_p(m) = \arg \min[\Delta S^i_n(m)] \). In an ideal case \( \Delta S^i_n(m) \) equals to zero with that time delay \( \tau_i \) which corresponds to the true spatial location. However, in practice only near-to-zero values are possible (see Figure \( 5.3 \)). The output of this stage consists of sequence of estimates for the correct sector, that is, estimates evaluated in each discrete frequency point \( m \).

The processing is easy to understand by thinking of input signals as sums of sinusoids. Using this interpretation, \( M/2 \) sinusoids are fed at a time to the coincidence detection. This corresponds to fixing \( m \) in the equations above and then examining the value at each time delay \( \tau_i \).
The coincidence detection produces at a maximum $M/2$ candidates for the time delay. However, in the case of one broadband source at a certain spatial location, each $i_n(m) : m = 0, \ldots, M/2 - 1$ is the same in an ideal case.

The parameter that can be adjusted in the coincidence detection is sector count $I$. Various values for $I$ were tested. $I = 361$ was a common choice offering reasonable resolution with reasonable computational load.

5.3.3 2D coincidence patterns

This stage has two main purposes: (1) provide a representation that describes the coincidences in sector/frequency -plane and (2) enhance the expression of the coincidence location for each frequency. Figure 5.4 illustrates a 2D coincidence pattern for one frame.

The idea behind introducing the mechanism that provides the second purpose derives from human hearing. The modeling of this feature in localization systems is presented, for instance, in Colburn model introduced in Section 3.1. This stage produces the same kind of neural firing patterns as in human hearing system.

The actual forming of 2D coincidence patterns is simple. Kronecker delta is applied to the similarity measure $\Delta S^i_n(m)$. Equation (5.6) is the definition of Kronecker delta function.

$$\delta(a - b) = \begin{cases} 1 & , a - b = 0 \\ 0 & , a - b \neq 0 \end{cases}$$

where $a$ and $b$ are arbitrary scalars. Using the notation of the localization system the conversion to neural firing patterns is achieved by inserting $a = i$ and $b = i_n(m)$. Thus applying $\delta(i - i_n(m))$ to the neural coincidence pattern will result in a sequence of zeros but unity when $i = i_n(m)$ which corresponds to the sector estimate produced by the coincidence detection. The desired representation as in Figure 5.4 is achieved by applying Kronecker delta function in each frequency band $m : m = 0, \ldots, M/2 - 1$, and collecting the resulting sequences into a matrix the rows of which correspond to frequency points $m$ and the columns of which are sectors $i$. Figure 5.4 is thus a visual view of the matrix in frame $n$. Note that in Figure 5.4 $m$ is converted to Hertz using Equation (5.4) to offer an example of how the coincidence detection varies along real-world frequency axis with natural signals.

5.3.4 Integration over time

Time domain integration has been utilized in very few localization systems based on binaural models (Section 3.1). In this system the past response of the system is utilized while evaluating the current response. 2D coincidence patterns (Figure 5.4) are multiplied by a factor and integrated over time. This processing stage improves the robustness against spurious responses occurring especially in natural environments and multi-source cases. Both the cases represent actually the same situation that results in a problem known as the interaction of sources. See Section 2.2.3 for theoretical discussion on this subject.
The coincidence patterns from the previous stage are integrated between sequential frames. The first part of this stage is described by Equation \((5.7)\).

\[
P'_n(i, m) = \sum_{n'=1}^{n} \beta^{n-n'} \delta(i - i_{n'}(m))
\]

\[
m = 0, \ldots, M/2 - 1
\]

\[
i = 1, \ldots, I
\]

where \(P'_n(i, m)\) refers to the coincidence pattern similar to Figure 5.4 but it represents several frames instead of one. Note that \(n\) is the current frame and \(n' \in [1, n)\) refers to all the frames before \(n\). \(\beta\) is a scalar coefficient that affects the amount by which the earlier frames affect to the integrated coincidence pattern. It is limited to the interval of \((0, 1)\). The current frame is taken as such in the sum \((\beta^{n-n} = 1)\) whereas the previous frames are de-emphasized exponentially.

Secondly, the resulting patterns \(P'_n(i, m)\) are refined to eliminate spurious responses with a simple thresholding. Equation \((5.8)\) illustrates this.

\[
P_n(i, m) = \begin{cases} 
P'_n(i, m) & P'_n(i, m) \geq \gamma \\ 0 & P'_n(i, m) < \gamma \end{cases}
\]

where \(\gamma\) is a scalar threshold. \(\gamma \geq 0\) are valid. It is chosen according to signal and environment. It has the effect of removing the coincidences that arise from phantom sources.

In this stage the parameters that can be varied are \(\beta\) and \(\gamma\). Choosing \(\beta\) near to unity results in emphasizing the past coincidence patterns. This makes the system robust against source dynamics; onsets, offsets and sound source movement whereas small values guarantee that a spurious response at some time instant has not a big effect over time. Liu et al. in [23] evaluated the system using \(\beta \in [0.85, 0.99]\). In the simulations done in this thesis it was fixed to 0.98. \(\gamma\) improves the system reliability against interaction between sources. A large \(\gamma\) is able to remove the spurious effects.
that last longer. Liu et al. fixed $\gamma = 1$. This choice proved to serve the purposes of this thesis also. $\gamma = 1$ proved to be capable of removing the coincidences resulting from the interaction of sources in one time frame in the simulations done by Liu et al. in [23].

5.3.5 Integration over frequency

The purpose of frequency integration is the same as integration over time: reliability. The frequency integration is performed by summing the coincidence patterns similar to Figure 5.4. However, the input to this stage is provided by “Integration over time” meaning the patterns represent a longer time interval than a single frame.

$$H_n(i) = \sum_{m=0}^{M-1} P_n(i, m)$$  \hspace{1cm} (5.9)

where $H_n(i)$ describes the reliability of coincidence in frame $n$ at particular sector $i$. $P_n(i, m)$ is defined by Equation (5.7) and Equation (5.8).

Ideally, in the sequence $H_n(i)$ all other values are zero except those sectors that correspond to the true location of a sound source. In practice, it is common that $\forall i |H_n(i)| > 0$. Still, the detection of the correct sectors is possible due to the fact that the peaks resulting from true sources are considerably greater than the peaks that result from the interaction of sources.

5.3.6 Peak classification

“Integration over time” and “Integration over frequency” together are able to attenuate the peaks in the 2D coincidence patterns that are not originating from any true sound source. Still, in the case of several sound sources or equally well one source at some reverberant environment, the phantom source phenomenon may cause a problem to the system.

As a solution to this problem Liu et al. in [23] propose a simple classification system, the basic idea of which was adapted to the system presented in this thesis. A simple clustering of the peaks in $H_n(i)$ is made. Two classes are established; the class of valid peaks and the class of phantom peaks.

Since $H_n(i)$ contains a lot of rapid variation over sector axis $i$, it is smoothened with a low-pass filter. Local minima of the smoothened $H_n(i)$ are then searched and their locations on $i$-axis are recorded. After that, all the local maxima in the original $H_n(i)$ are tested using the local minima. If a local maximum occurs between two local minima it is considered as a valid peak.

5.3.7 Mapping to horizontal angles

The final estimate of the horizontal angle of a sound source is achieved by a conversion of sectors $i$ to angles using a simple lookup table. The lookup table derives
from geometrical facts and the knowledge on velocity of sound waves in air. Note that since the measurements were done using a simple two-microphone setup this simple mapping is possible (see Figure 6.2 in Chapter 6). However, if for instance a dummy head or head-and-torso-simulator was used, the different behavior of low frequencies and high frequencies of sound waves as they meet obstacles should be considered (A head-and-torso-simulator is illustrated in Figure 2.3). This issue is discussed in detail in Section 2.1.2.

The relation between the horizontal angle $\varphi_i$ and the discrete sector $i$ is defined by Equation (5.10).

$$\varphi_i = \frac{\pi}{2} - \frac{i - 1}{I - 1} \pi$$

$$i = 1, \ldots, I$$

### 5.4 Sinusoids + transients modeling

The processing in this stage consists of forming a mid-level representation of a signal. The theoretical discussion of mid-level representations can be found in Chapter 4. Especially, Section 4.4 and Section 4.5 provide rather a thorough description of the implementation of this stage.

#### 5.4.1 Sinusoids modeling

Deterministic modeling can be performed basically in parallel to the location estimation (see Figure 5.1). However, in the grouping stage an estimate of the sound source location has to be available which is why “delay” blocks are in the figure before the modeling. The modeling techniques applied in this stage are presented in Chapter 4. In this thesis, stages (1) peak detection and (3) parameter estimation of the general parameter extraction model were implemented to cover deterministic modeling (see Section 4.4). Two methods for the peak detection stage were evaluated, that is, “peak picking” and “cross-correlation”. The methods are explained in Section 4.4.1 and Section 4.4.2 respectively.

Any method that is employed in sinusoidal modeling produces $a_{(m_s)}^{(l,r)}(n)$, $\omega_{(m_s)}^{(l,r)}(n)$ and $\theta_{(m_s)}^{(l,r)}(n)$ which are the amplitude, the frequency and the phase of a sinusoid $m_s = 0, \ldots, M_s - 1$. $M_s$ is the maximum amount of sinusoids per frame $n$. The processing is made independently for left channel ($l$) and right channel ($r$).

In the initial evaluations it was discovered that the performance of the peak picking method was insufficient for the purposes in this project. Some of the shortcomings are discussed and illustrated in the previous chapter (Figure 4.2). Briefly, the modeling capabilities of the method proved to be insufficient. The cross-correlation method, instead, performed in the desired manner.

Both methods share three parameters which contribute the signal modeling; the length of the analysis window $K$, the length of DFT $M$ and the maximum amount of
sinusoids per frame $M_s$. For speech signals, for instance, the parameter set $K = 1024$, $M = 8192$ and $M_s = 25$ is a plausible choice.

The cross-correlation method introduces several additional parameters. The most important are $\alpha$ and $W$. The effect of each is discussed in the previous chapter. Values for these parameters are selected according to signal content. $\alpha$ is selected from range $(0, 1)$, and a plausible choice for $W$ is in the range of $[10, 100]$ in the case of natural signals.

5.4.2 Transients

The transient detection utilizes the idea of tracking energy changes. The complete theoretical discussion on the method can be found in Section 4.6.1. The basic idea, however, can be understood from Figure 4.5.

The transients are searched from the residual part of a signal. The residual in this context refers to a signal from which the deterministic part is removed, that is, transients and other stochastic-in-nature information is included. This part of a signal corresponds to $r(t)$ in Equation (5.1).

The residual is obtained by synthesizing the sinusoids and subtracting them from the original ($r(t) = s(t) - \tilde{s}(t)$, where $s(t)$ and $\tilde{s}(t)$ represent the original signal and the synthesized sinusoids, respectively). Note that Figure 5.1 this “extra synthesis” is included in the block “deterministic + stochastic modeling”. The residual processing consists of the detection of transient locations in a signal, and the formation of transient regions. The detection of transients is equally important in the residual processing as the detection of sinusoids in the sinusoidal modeling. Due to this, two methods are employed to achieve the reliable onset estimate for a transient. Even if all the transients are detected and their locations are found, a lot of information remains in the “final residual signal” which is a signal from which all the detected transients are removed. This part can be modeled as properly filtered white noise as stated earlier. Due to the separation method, the final residual is not processed in this system since no directional information can be extracted reliably. On the other hand, the spatial origin of a transient can be deduced using its onset time. The theory concerning transient processing techniques as well as stochastic modeling techniques in general can be found in Chapter 4.

The reference system of this stage is introduced by Levine [35]. Speech signals were used also in his study so largely the same system parameters can be used. The length of the analysis window $K_{tr}$ is 512, and sequential windows are overlapping by 50%. Transients are assumed to last 64 ms which differs from 66 ms that is used in the reference system due to the fact that the sampling rates are 48 kHz and 44.1 kHz, respectively. 64 ms results in labeling 12 sequential frames as transient frames. Constraining the maximum transient rate is done similarly to Levine’s system. The frames corresponding time regions 50 ms before and 150 ms after the frame in which the transient is initially detected, are labeled as non-transient regions. Other transients detected within this region are considered as invalid.

The transient detection algorithm outputs the transient frame indices for each channel.
An important issue concerning the transient regions is worth pointing out. A transient detected in the frame $n_{tr}$ results in labeling one frame less before the transient frame than after the frame. However, this is justified by the fact that in general a transient-like sound contains more energy after its onset than before it. In general, the amplitude of a transient rises rapidly to its maximum but dampens relatively slowly.

### 5.5 Grouping stage

The basic idea of the grouping stage in this system is very simple (see Figure 5.1). It receives the parameter stream modeling periodic content and the information of frames that are detected as transient regions in a signal. Additionally, the reference information, that is, the horizontal DOA estimate is fed to this stage enabling the grouping.

#### 5.5.1 Grouping of sinusoidal components

In the frame $n$, the deterministic part of the signal is modeled as a sum of sinusoids (see Equation (5.1)). Three parameters are produced by the sinusoidal modeling for each component in one frame. Thus the parametric presentation may be expressed with three sets per channel $a_{(l,r)}(n)$, $\omega_{(l,r)}(n)$ and $\theta_{(l,r)}(n)$ which represent a set of amplitudes, a set of frequencies and a set of phases, respectively.

In the grouping, it is assumed that the desired signal is the same in the left channel and in the right channel. The desired signal in one channel is only delayed compared to the signal received by the other channel. Using the DOA estimate which is obtained for the desired source, the delay can be estimated. On the other hand, using the phase information carried by the sinusoids, for each sinusoid pair the corresponding delay is easily estimated. Those sinusoid pairs for which the delay is near to the delay calculated using the DOA estimate are considered as arising from the desired sound source.

The reference information needed to group the sinusoids between the desired part and the undesired part is provided by the DOA estimation system. A DOA estimate has to be converted to a form that enables the phase comparison. The conversion is made by calculating first the time it takes for sound pressure waves to propagate between two sensors, and yet, converting seconds to samples. Figure 5.5 illustrates these aspects. Time differences are obtained using Equation (5.11).

$$\Delta t_{\text{ref}} = f_s \frac{D}{c} \sin \varphi \quad (5.11)$$

where $f_s$ is sampling rate, $D$ is the distance between the sensors, $c$ is propagation velocity of sound wave fronts and $\varphi$ is the DOA estimate.

Two things in Equation (5.11) are worth pointing out. (1) Plane wave propagation is assumed which is common also in the systems that are dedicated to DOA estimation. The equation may be derived for instance using Figure 6.2 (2) $\varphi$ is constant implying
that sources are not allowed to move. For moving sources the $\varphi$ should be updated at frame rate. Additionally, since the DOA information has to be available in the grouping, additional delay units should be introduced into the system.

The grouping of sinusoidal components is based on a phase constraint, which is calculated using the time difference (Equation (5.11)). Since the behavior of a sinusoidal function is exactly known, a frequency component in the left channel is selected, the same component in the right channel is examined by checking how much the phase of the component in the left channel has changed compared to the right channel. The reference for the maximum allowed phase change is provided by the DOA system. Using the sets defined above $\omega_{\{l,r\}}(n)$ and $\theta_{\{l,r\}}(n)$ are needed in the grouping. $\omega_{\{l,r\}}(n)$ is utilized while searching for the equivalent components in each channel, and $\theta_{\{l,r\}}(n)$ is utilized in the DOA estimation.

The phase comparison makes sense only if the components have exactly the same frequency. At first, a sinusoid with the smallest frequency in the left channel is selected for the reference. Next the same frequency is selected from the right channel enabling the actual phase comparison. However, since the processing in the analysis stage is independent for each channel, an element that represents the same component in $\omega_l(n)$ may differ from the element in $\omega_r(n)$. Thus a small deviation has to be allowed despite the fact that phase comparison makes sense in theory only between two sinusoids that have exactly the same frequency.

Thus firstly, sinusoid pairs are formed which fulfill the frequency deviation criterion: $|\omega_l^{(m_s)}(n) - \omega_r^{(m_s)}(n)| < 2\pi f_{\text{dev}}$, where $f_{\text{dev}}$ is the chosen deviation in Hertz. Secondly, the phase difference is calculated as $\Delta \theta_{\{l,r\}}(n) = \theta_l^{(m_s)}(n) - \theta_r^{(m_s)}(n)$, where $m_s\{l, r\}$ is the index of a sinusoid in each channel, $m_s$ refers then to a sinusoid that passed the deviation test.

The choice of $f_{\text{dev}}$ is not obvious. Signals in each channel are processed independently basically starting from the sensors. Thus already in the analysis stage the signals may differ from each other quite drastically, at least, from the modeling viewpoint. Several values for $f_{\text{dev}}$ were tested. However, 10 Hz maximum deviation seemed to be a plausible choice in general. It may be stated that it is rather a large value. However, it was concluded to use such a high value for $f_{\text{dev}}$ since with some smaller
values a lot of components did not pass the deviation test resulting in losing a large amount of components. Also occasional checks made on $\omega^{(m_s)}(n)$ proved that this is a plausible choice.

$\Delta\theta^{(m_s)}(n)$ has to be converted to time difference enabling the grouping. Equation (5.12) describes the conversion.

$$\Delta t^{(m_s)}(n) = f_s \frac{\Delta \theta^{(m_s)}(n)}{\omega^{(m_s)}(n)}$$  \hspace{1cm} (5.12)

where $f_s$ is sampling rate. Note that for $\omega^{(m_s)}(n)$, either $\omega^{(m_s)}_l(n)$ or $\omega^{(m_s)}_r(n)$ can be used since only the frequencies which fulfill the condition discussed above are used.

The final decision between the parameter sets that represent the desired part and the undesired part is presented by Equation (5.13).

$$M_D(n) = \left\{ m_s \left| \Delta t^{(m_s)}(n) - \Delta t_{ref} < \Delta t_{dev}^{(m_s)} \right. \right\}$$  \hspace{1cm} (5.13)

where $M_D(n)$ represents the set of elements which consists of admissible frequency component indices and $\Delta t_{dev}^{(m_s)}$ is a parameter which allows some tolerance to the angle estimate. $\Delta t_{dev}^{(m_s)}$ is obtained applying Equation (5.12) and setting the desired tolerance to $\varphi_{dev}$. Several tolerance values were tried but $\varphi_{dev} = 10^\circ$ was chosen because it was the most suitable to cover all the cases. In general, the more noisy and the more broadband is the signal, the more tolerance is needed. However, one value for the parameter is plausible because it enables the comparison between different environments to some extent.

### 5.5.2 Grouping of transient regions

The transients do not have similar parametric representation as sinusoids by which the DOA estimate of a frequency component is easily obtained. However, the spatial origin of a transient may be extracted using a broadband DOA system to estimate the time difference for each transient region.

Consider a transient region that has been found valid. Let us denote the transient region related to a transient frame by $tr_{\{l,r\}}(g) : g \in \mathbb{Z}_+$. Note again that both channels are taken into account. Before any deduction of a transient region can be made concerning its spatial origin, synthesized transients have to be available. This means synthesizing all the detected transients. However, it may be considered as a minor drawback since the occurrence rate of transients is not so high, and yet, the transient synthesis in this system does not result in increase of computational load.

First, transient regions $tr_{\{l,r\}}(g)$ are utilized to find out the absolute locations of transients on time axis from the residual $r_{\{l,r\}}(k)$ (corresponds $r(t)$ in Equation (5.1)}
in discrete-time). Then, for each region \( tr_{\{l,r\}}(g) \) the estimate of DOA \( \varphi_{tr_{\{l,r\}}}(g) \) is obtained by feeding each transient region to the DOA estimation subsystem. Note that the duration of transient regions is sufficient in the sense that DOA system is able to produce plausible estimates. Finally, the grouping between the desired transients and the undesired transients is made based on Equation (5.14).

\[
T_D = \left\{ tr(g) \mid \left| \varphi_{tr_{\{l,r\}}}(g) - \varphi_{\text{ref}} \right| < \varphi_{\text{dev}} \right\}
\]  (5.14)

where \( T_D \) represents the set transients that seem to arise from the same spatial origin, \( \varphi_{\text{ref}} \) is the estimated horizontal angle and \( \varphi_{\text{dev}} \) is the maximum allowed deviation from \( \varphi_{\text{ref}} \).

### 5.6 Synthesis stage

#### 5.6.1 Overlap-add synthesis of sinusoidal components

The sinusoidal synthesis employs overlap-add principle. In general, sinusoidal synthesis is quite straightforward since it is nothing but inserting the estimated parameters; amplitude, frequency and phase to Equation (5.1). However, while considering the final resulting signal the transients have to be taken into account. More precisely, their location information on time axis in each channel has to be available. Using this information, the sinusoidal modeling is turned off and the transient modeling is turned on as the detected transients occur.

The information needed to turn off the sinusoidal modeling is obtained using the transient region information in \( T_D \). Let us denote by \( N_D \) the set of frames that do not overlap on the transient regions, \( k_{(g)}^{(r)} \) is a time range on discrete time axis related to the transient region \( tr(g) \) and \( T_D \) is the set of transients that fulfill the constraint in Equation (5.14).

At this stage all the needed information is available to conduct the actual synthesis. A synthesized frame is presented by Equation (5.15).

\[
\tilde{s}_{\{l,r\}}(n)(k) = \begin{cases} 
\sum_{m_s \in M_D(n)} a_{\{l,r\}}^{(m_s)}(n) \cos \left[ \omega_{\{l,r\}}^{(m_s)}(n) k + \theta_{\{l,r\}}^{(m_s)}(n) \right], & \forall n \in N_D \\
0, & \forall n \notin N_D \\
n = 0, \ldots, N - 1
\end{cases}
\]  (5.15)

where \( a_{\{l,r\}}(n) \), \( \omega_{\{l,r\}}(n) \) and \( \theta_{\{l,r\}}(n) \) are the amplitude set, the frequency set and the phase set for each channel.

The sequential frames \( \tilde{s}_{\{l,r\}}(n) \) are built to a contiguous discrete-time signal using the overlap-add synthesis principle. Let us assume that the frame \( n \) is about to be reconstructed and the frames \( 0, \ldots, n - 1 \) are already reconstructed. The frame \( n \) is formed by windowing \( \tilde{s}_{\{l,r\}}(n)(k) \) with a smoothing window function and linearly
adding it partly on the top of the previous frame \( n - 1 \) and on the top of the next frame \( n + 1 \). The smoothing window is sampled Hanning window. The overlap coefficient \( \zeta \) is fixed to 0.5. Summing sequential Hanning windows with this overlap, results in constant function the value of which is 1.

### 5.6.2 Transient synthesis

In Levine’s system, which used used more or less as a reference study, audio compression is the primary goal. In the system presented in this thesis there are, however, no intentions to compress the data which the analysis stage and the grouping stage produce. Levine used a transform coding for transients which would require decoding in this stage\(^{35}\). In this system that pre-stage is not needed. Thus the transient synthesis consists of simply utilizing the locations of admissible transient regions and copying each transient region to its correct position.

Using \( tr_{l,r}(g) \in T_D \), the correct indices for the desired transient regions are found. Finally, a transient signal is reconstructed by basically forming a zero-amplitude signal equal in length compared to the original input signals, and adding each transient region in each channel to the zero-amplitude signal.

### 5.6.3 Combining sinusoidal part to transient part

The synthesis stage is terminated by utilizing synthesized sinusoids and transient regions, and linearly adding them to a single stream (see Figure 5.1). Note especially that switching off the sinusoidal stream while transients occur and vice versa takes care of the fact that the deterministic modeling and the stochastic modeling are not used simultaneously. The simultaneous modeling results in even more serious deterioration of the synthesized signal than, for instance, applying sinusoidal modeling on a transient.

Equation (5.16) characterizes the resulting signal.

\[
\hat{s}_{\{l,r\}}(n) = \begin{cases} 
\hat{s}_{\{l,r\}}(n) & n \in N_D \\
\hat{r}_{\{l,r\}}(n) & n \notin N_D 
\end{cases}
\]  

(5.16)

where \( \hat{s}_{\{l,r\}}(n) \) is the signal produced by the sinusoidal synthesis, \( \hat{r}_{\{l,r\}}(n) \) is the signal that contains the detected transient regions. \( N_D \) is explained in Section 5.6.1.
Chapter 6

Audio measurements

One of the goals of the thesis was to build a system capable to operate with real-world data. Instead of generating a mixture from pure undistorted signals using a computer, it was desired that the signal goes through the whole chain altering the sound waves as they propagate from source (see discussion in Chapter 2). In most of the systems that are published so far, the performance evaluation is done with artificial signals. In the case of a tone, basically the only factor that causes an error to the signal compared a sinusoid generated by an ideal oscillator is finite sampling rate. Angles corresponding to a certain direction-of-arrival are easily generated by adjusting the delays between the channels of a multi-channel signal. Additionally, it causes no trouble to generate a great variety of combinations.

Using the artificial data, several factors, mostly signal-distorting, issued by everyday auditory environments are neglected. However, the use of generated stimuli is well-argued especially in the preliminary evaluation of an algorithm to cancel out the factors that can cause misconclusions concerning, for instance, the parameter evaluation of an algorithm.

To evaluate the real-world performance of a system some real-world data has to be available. In the end, it is often the case that either a direction-of-arrival system or a sound source separation system is built to perform a predefined real-world task. Some applications that utilize the theory of sound source separation already exist for hearing-impaired persons.

In this thesis it was desired to evaluate the selected localization algorithm and the selected separation algorithm with a vast variety of material. That is, several types of signals were used to experiment the performance of the algorithms. The signals were recorded using high-quality equipment. However, it was in our interests to use affordable-price equipment. This enables also the initial evaluation of the algorithms in the sense that can consumer level equipment be used with this type of a separation system. 135 different mixtures were measured which equals to 25 minutes and 15 seconds of audio data for the evaluation of the algorithms.
6.1 Signal content

The signal combinations including the sound source placement information in each environment is collected to Appendix A. The signals included natural sounds such as speech, music and some generated signals. No special constraints on signals were fixed. It was one of our interests to test the system with practically any kind of real-world data. The operation of some DOA and separation systems is restricted to signals the spectral content of which is fixed in advance. Additionally, certain consonant and vowel content in the case of speech is often expected. Also special signal-and-masker combinations are targeted by some separation systems. Arbitrary combinations result usually in poor performance.

The signals were extracted from audio compact discs (music samples) and generated with computer (tones, noise signals). Additionally, some signals such as speech were extracted from Audio Research Group databases. The acquired signal samples were cut to 8 seconds (see Section 6.6). This was considered as sufficient for the performance evaluation of each algorithm to see how they perform with a certain signal type. The level of environmental sounds in each environment was practically constant during the recording of each mixture. It is well-argued to use relatively long samples because it enables us to evaluate the convergence speed of the DOA algorithm. The signal set was collected to a compact disc. Different mixtures were generated by using two similar compact discs.

An outstanding feature in the selected signal mixtures is the great variety of speech mixtures. This is due to the fact that sound separation methods often address the problem of speech separation rather than separation of an arbitrary signal. Thus, a great variety of speech signals was also in our interests since it enables the comparison to other systems to some extent. In addition to that, speech signals are maybe the best selection while evaluating the usefulness of an algorithm from the application viewpoint.

From the DOA estimation viewpoint the selected signal mixture set could have been much smaller. Instead, the variety of the different source placement could have been more abundant. For DOA systems the evaluation can be done using basically only two types of signals; one representing narrowband sources (e.g. a tone) and another representing wideband sources (e.g. white noise).

Due to the facts alluded above, and taking into account the available resources, the signal set was selected by discussing the experts in audio signal processing.

6.2 Environments

The measurements took place in three different environments. These were an anechoic chamber, a classroom and a workstation class. Due to limited resources not only the signal content had to be limited but also constraining the environments to three was a necessity. Initially, it was in our interests to include some outdoor environments. However, it became soon apparent that with given resources, with the fixed signal set and with the fixed source location plan it was not possible to include
more than three environments. From the candidate environments it was concluded to stay with indoor setups as they already are challenging cases for DOA estimation systems and sound source separation systems. Outdoor cases would have resulted in not only even more demanding task from the viewpoint of the algorithms but also from practical viewpoint. The measurements in an outdoor environment would have raised several difficulties in the measurement procedure (e.g. adjusting recording signal levels in a windy environment).

Anechoic chamber, a zero-background-noise environment, was used in order to evaluate the performance of the algorithms in an “ideal environment”. The anechoic chamber of Institute of Measurement and Information Technology at TUT was used. It must be pointed out that anechoic chamber is maybe not a very common environment. However, it is still more realistic to evaluate the algorithms with the data recorded in such an environment than using the generated data. It may be regarded as an example of a very quiet natural environment.

Classroom, a reverberant environment, offered an example of a reverberant environment. The concrete walls, wooden tables, floor of concrete and practically no sound-absorbing material except chair paddings and two human beings who conducted the measurements were present. The measurement personnel characterized the space as disturbing environment due to the reverberation.

Workstation class, a reverberant noisy environment, was as an auditory environment quite similar to the classroom (concrete walls, wooden tables, sound-reflecting floor and ceiling), but offered the most challenging environment. 14 computer workstations acted as extra noise sources producing “office noise”. However, as the workstation class was more densely furnished than the classroom which, in fact, resulted in a lesser reverberance. The measurement personnel perceived not so prominent annoying features in the workstation class as in the classroom.

6.3 Measurement equipment

The measurements was conducted using the equipment specified in Table 6.1. The two-sensor setup is illustrated by Figure 6.1. As it was pointed out in the beginning of the chapter, a decent quality of the measured data was of interest. Decent quality refers to the fact that extra distortion in path from the compact disc to the sound-waves produced by speakers and, on the other hand, along path from sound waves to DAT-tape should be so small that it would have no effect on the performance of either the DOA estimation algorithm or the sound source separation algorithm.

Excluding the signal sources, that is the CD-players, the equipment can be regarded as professional-level hardware. This CD-player model represents a consumer-level product. However, nowadays they are of good quality satisfying our requirements. Additionally, as this player model is portable it provided a flexible solution to conduct the measurements in such large extent.
Table 6.1. Hardware used in the measurements.

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sony Walkman D-EJ625 CD-player</td>
<td>2</td>
</tr>
<tr>
<td>Genelec Biamp 1019A active speaker</td>
<td>2</td>
</tr>
<tr>
<td>AKG C460B microphone</td>
<td>2</td>
</tr>
<tr>
<td>Portable microphone 2-channel amplifier for AKG C460B</td>
<td>1</td>
</tr>
<tr>
<td>Sony TCD-D10 Pro DAT</td>
<td>1</td>
</tr>
<tr>
<td>Denon DTR-2000 DAT</td>
<td>1</td>
</tr>
<tr>
<td>Brüel &amp; Kjaer 2232 sound-level meter</td>
<td>1</td>
</tr>
<tr>
<td>SGI Octane workstation</td>
<td>1</td>
</tr>
</tbody>
</table>

The speakers which were used are a high-quality product. They are commonly used as studio monitors by music industry professionals. Furthermore, they contain an internal amplifier enabling flexible adjustment of signal levels in each measurement session.

The microphones are as well commonly used studio hardware. The characteristic pattern of C460B is omnidirectional, that is, no direction is preferred. As the microphones are so-called condenser microphones, that is, they are not operational without a power supply. A portable amplifier working with AA-size batteries was used.

The microphones were attached to a microphone stand with a custom-made bar attached on top of the stand. The bar, enabled the adjustment of the inter-microphone distance (see Figure 6.1). However, 10 cm distance was used in all the measurements.

The microphones were attached to the DAT-recorder via the portable amplifier. Sony TCD-D10 Pro is a portable professional-level DAT-recorder. TCD-D10 Pro has the important feature of input signal level adjustment which enabled us to guarantee
good signal-to-noise ratio by maximizing the input signal in each recording session. Certain signal-to-interferer ratios were desired while recording the mixtures. This was achieved using a sound-level meter to guarantee the desired ratios. The absolute values are not of importance, however, they are available in Appendix A. A high-quality sound-level meter by Brüel & Kjaer with a 7-segment display was used enabling a person to read the exact values easily.

Denon DTR-2000 was used to transfer the recorded audio signals from raw data on tapes to files on a hard disk. The samples were transferred by playing each sample with DTR-2000 and recording the sample with the Silicon Graphics Octane workstation.

In addition to the above hardware, a set of various cables was needed. This, of course, included the power cables for the devices but also the signal cables. In order to attach the portable CD-players to the speakers, high-quality coaxial cables with a 3.5 mm jack in one end and an XLR-connector at the other were used. The microphones were attached to the amplifier with the same type of cable but with XLRs at both ends. Exactly the same type of cable was used to connect the amplifier to TCD-D10 Pro. DTR-2000 was connected to the workstation via digital interface and with an optical cable.

6.4 The placement of the sound sources

The setup which was used in each environment is illustrated by Figure 6.2. The placement of sound sources in each environment was, firstly, defined by the fact that the performance of the selected DOA method was evaluated not only in one-source case but also in two-source case. It was desired to use the same angles in both cases thus also the placement in the one-source case is restricted by the placement in the two-source case. Secondly, for the selected separation method, it is an operational precondition that in two-source case the sources are spatially apart. That is, the sources must be placed with different angles ϕ.

At first, a new coordinate system is issued which is used from now on unless it is explicitly overruled. As the development of the DOA algorithm was launched, a coordinate system which is different from the commonly used system was specified. The traditional coordinate system is presented in Figure 1.2 on page 10. Let us define the new coordinate system using Figure 6.2. The angles of the microphone-oriented coordinate system in the figure are referred to with a mo-subscript to avoid confusion. What has changed to the original is that ϕ = 90° corresponds to −90° and ϕ = 270° corresponds to 90°. The critical values are presented in the figure. Colloquially, the left hand side of the horizontal plane is the “negative half” and the right hand side is the “positive half”.

Relatively easy source placement from the viewpoint of the algorithms was used. It is no use to start with difficult configurations, and place the sources near to each other. One source was placed at ϕ_mo ∈ [−90°, 0°] while the other was placed at ϕ_mo ∈ (0°, 90°] (see Figure 6.2). Sufficiently large spacing specified by ϕ_mo was fixed
to avoid source merging into one source from the viewpoint of the algorithm. For instance, one source placed at \( \varphi_{mo} = -5^\circ \) and the other at \( \varphi_{mo} = 5^\circ \), merging may occur resulting in only one DOA estimate instead of two. Note that the phenomenon has also been discovered in human hearing. The issue is discussed in more detailed on page 21.

Angles \( \varphi_{mo} = -30^\circ \) and \( \varphi_{mo} = -15^\circ \) were used in the anechoic environment and in the classroom environment for source 1. In the office environment, only the latter setup, that is, source 1 at \( -15^\circ \) was measured. Source 1 is regarded as the “primary source”, or the source of interest. Source 2, or the “noise source”, was always at \( \varphi_{mo} = 45^\circ \) once it was active.

In order to evaluate the accuracy of the DOA algorithm, the location of a source has to be determined as accurately as possible. Careful placement of the speakers and the microphone stand in each environment has thus a crucial role. It became soon apparent while determining the actual locations for the sources and the sensors that, in practice, this is not so an easy task as one may think at first. The equipment that was used for the placement, basically consisted of an ordinary three-meter rollable measuring tape and a calculator.

Using basic trigonometry, Figure 6.2 and the given value for \( \varphi_{mo} \), the locations could be determined. The trigonometric formulae which was used is \( \tan \varphi_{mo} = a_1/b \).

The placement of sources and sensors was difficult due to simple tools. 90° angle was difficult to guarantee while determining the correct places (see Figure 6.2). Additionally, \( b \) was fixed to 2.5 m while planning the measurements. The constraint was not wanted to be relaxed. However, this convention caused serious problems in the classroom and in the workstation class, that is, the dimensions of the rooms and the furniture prevented partly the desired placement (e.g. the sensors and the sources should be as far as possible from walls).

To overcome the difficulty concerning the 90° angle, the measurement personnel used the means offered by the environment such as floor tiles. Additionally, once the sources were placed using \( a_1 \) and \( b \), the distance \( c \) was measured, and compared to the value obtained using the equation \( c = a_1/\sin \varphi_{mo} \).

A few words are worth pointing out here related to the measurements. These issues should be remembered while studying the results. It cannot be denied that errors...
occurred in the placement ranging, for instance, from fine errors due to the uncali-
brated tools to some factors resulting in coarse errors. For instance, human being is
not able to conduct such tasks precisely and consistently, in general. Additionally,
assuming that the place of a source is correctly measured using for instance a laser
to determine the positions and a robot for the placement, still, a variety of errors
would remain. For instance, it may be questioned at what point exactly a wavefront
produced by a source has its true origin. However, even the coarse errors made by
the measurement personnel can be considered as minor from the viewpoint of this
study. Furthermore, once the source placement is done the errors are systematic, or
the bias in some variable is the same within a session because neither the micro-
phone stand nor the speakers were moved. Due to these facts the more specific error
analysis is neglected.

6.5 Measurement procedure

In each measurement session the sensors were placed first. The convention concerning
the distance $b$ caused certain problems in the classroom and in the workstation
class (Figure 6.2). It was hard to place the equipment due to the small rooms and
the furniture. Additionally, the measurement personnel encountered problems while
adjusting the sound levels. This was due to the fact that the speakers and the stand
had to be placed near concrete walls. Sound waves reflect from such structures almost
perfectly. In the worst case, the session had to be restarted because of the clipping
with some mixtures. The clipping was not first detected because the reflections
seemed to be more serious in one frequency range while in some other range, no
major rising in signal level was detected. However, the problem was overcome by
trying out various placements and adjusting the amplifier gains.

Once the places for the speakers and the microphones were marked, the devices were
placed at their positions. Because only the horizontal angle $\varphi_{mo}$ was of interest, no
specifications concerning the height of either the speakers or the microphone stand
were fixed. Because the height of the sources or the height of the sensors from floor
was not important the furniture in the rooms was utilized to place the equipment.
For instance, the tables in the premises were utilized to the place the loudspeakers.
In each environment it was checked visually that the microphones and the middle
points of the speakers were approximately equal in height. Note that it was one of the
preconditions for the system not be affected by the fact that actually there may be
a several-degree variation in the elevation angle between the different environments,
and even between the different session in the same environment. The latter is due
to the fact that in the anechoic chamber and the classroom source 1 was at two
different angles. Thus there may be a minor change in the elevation angle between
the sessions.

Cables were connected, all the devices were powered up, and the recording device
was set to monitor its input. The final tuning of the setup was done with a specific
calibrator to guarantee 10 cm difference between the microphones ($D$ in Figure 6.2).
The actual recording session began by adjusting appropriate signal levels. This included finding out appropriate gains for the amplifiers to be able to achieve the desired signal level differences, that is, 0 dB and 10 dB for each mixture within a session (see Appendix A). Simultaneously, possible problems were tried to observe that may occur during the actual recording (e.g. clipping and resonances).

To achieve the desired levels, the signal level meter was placed between the microphones. The device was set to monitor the sound level, and each mixture was played. The sound level was monitored by the measurement personnel. The gains of the CD-players were adjusted until the level difference in the two-source cases was the desired for each mixture. Once the correct level was found the gains of the CD-players were written down.

The recording for a session was made batch-like. It means that that nothing but the gains of the CD-players were tuned between each mixture. This type of procedure was used to minimize the errors within a session. The same operation took place in each session and in each environment.

Note that 0 dB and 10 dB are only approximative values. The gains for the CD-players were obtained by playing one source at a time and reading the sound level value, and after that the other source was adjusted according to this value. Thus the source interaction was neglected. Moreover, the variation of signal level is quite big for instance in the case of speech signals. Thus the person who monitored the meter estimated some sort of average of the absolute sound level. However, the actual difference is not important. It was sufficient that in the 0 dB case the sources were approximately equal in level, and in the 10 dB case that source 1 was stronger.

### 6.6 Converting the data from DAT-tapes to files

Signals from DAT-tapes were converted to files using Denon DTR-2000 and the workstation. More specifically, this was done by simply regarding DTR-2000 as the source and Octane as the recording device. The ADAT-interface of the workstation was used. Each recorded mixture was transferred to its own file. 16 bits per sample and sampling rate 48 kHz was used.

A certain problem has been observed in the previous projects with this equipment. The recording software seems ignore the data which results in zero-amplitude signal while the problem occurs. However, this is easy to detect by listening. However, there is another problem. The link between DTR-2000 and the workstation may be broken partly (e.g. bad cable or jacks), or it is an internal flaw in the workstation. Anyhow, the latter problem results in losing a few samples once in a while. Surprisingly, even if the latter problem occurs the resulting file is readable, which makes the detection of the data loss hard. One would expect a corrupted file while samples are lost.

To cope with the latter problem, the exact size of the audio file was calculated as the length of the mixture, bit resolution and sampling rate were available, and compared to the amount of raw data in each file. After the conversion process, it was concluded that no data was missing according to either “listening test” or “calculation test”.
Chapter 7

Results

The results are presented in two separate parts. The first consists of the estimation of DOA algorithm in various cases. The second part is reserved for a brief discussion of the performance of the separation scheme employed in this thesis.

In each environment fundamentally different configurations are evaluated. Case A refers to a configuration in which one sound source is present. The undesired part in a signal mixture received by the sensors thus consists only of the background noise specific to each environment. It should be kept in mind that in addition to the background noise, reflecting surfaces (e.g. walls, floor, ceiling, tables) introduce several weak, but not insignificant, noise sources in rooms in general. Including these interferers, in case B there is a second source acting as a primary interferer. The sound level of the primary interferer is 10 dB weaker measured at receiving end of the configuration compared to the source of interest. Case C corresponds best to the classic cocktail-party situation: the source of interest and the primary interferer are equal in loudness (see page 23 for the description). Additionally, an artificial situation was generated for some signal configurations in which sources are artificially placed at the desired angles. This case is utilized for debugging of the algorithms and for adjusting the parameters. For more detailed description of the different cases the reader is encouraged to consult Chapter 6 and Appendix A.

7.1 The performance DOA estimation algorithm

Successful estimation of DOA is a fundamental prerequisite to expect plausible performance from the overall system. The DOA estimation subsystem should be able to provide a valid estimate concerning the horizontal angle of the desired sound source. DOA estimates need relatively frequent update. The optimal update rate is the same as the frame rate of the overall system (see Figure 5.1 for the overview of the system).

In general, the main criterion by which DOA estimation systems are characterized is the accuracy of the DOA estimates. Some applications stress as an additional performance measure the speed at which an algorithm converges towards the correct DOA estimate. As stated above, in this work the primary interest is the consistency of DOA estimates, that is, no big deviation among the DOA estimates for a sample
are allowed. In the case of moving sound source, the estimates before and after the current time instant should experience graceful, or smooth, variation. For instance, a 5° or even bigger deviations to the actual horizontal angle are acceptable. Since it is assumed that sources are spatially more distant. Instead, the amount of outliers should be minimized.

At first, the general performance of the DOA estimation method is presented by summarizing the discovered issues in the easiest case, that is, a single sound source in the anechoic chamber. In the following sections, the difficulty level is increased by moving into a new environment. For case B and case C the aspect is changed by the fact that instead of describing the accuracy of the DOA method, the view is in the robustness. In case B and in case C only the results for the speech signals are presented since they are of the primary interest.

The results concerning case A as well as the other cases certain issues should be kept in mind. The results are presented as if the true horizontal angles were known. However, this is not the case since no matter how carefully the measurements were conducted there is no guarantee that a sound source is placed exactly, for instance, at horizontal angle $-15^\circ$ (see Figure 6.2). The formal estimation of the error related to this issue is neglected. However, due to the careful measurement procedure the error may be roughly estimated to be no more than $1^\circ - 2^\circ$. The description of the measurements can be found in Chapter 6. Especially the facts related to this issue can be found in Section 6.4.

### 7.1.1 Accuracy of DOA method in the case of one sound source (case A)

**General performance**

The results of the simulations are presented in the following tables for one source case. Each sound source was played in each of the environments with no other noise sources excluding those that were naturally present in each environment. In the anechoic chamber there were absolute silence which in this case means that the background noise level was below 34 dB(A). This is the minimum value for the sound-level meter (Brüel & Kjaer 2232). In the classroom environment the background noise consisted mainly of air-conditioning. Also quite strong resonances occurred caused by the furniture in the premises. The resonances resulted from the fact that relatively high sound levels were used. However, quite big attenuation was required to get rid of the resonances. Due to this, the strong levels were used and the occurrence of the resonances was accepted to gain sufficient desired-signal levels against interferers’ levels. The presence of the resonances, however, changes to some extent the original configuration. The resonance may be viewed as extra interferers in this auditory environment. In the office environment, the background noise consisted of air-conditioning and computer workstations. In this environment the same resonance problem was noticed. It is worth mentioning at this point that it was considered whether there is a need to repeat the complete recording set in some other premises to avoid these problems. However, this suggestion was postponed since the recordings in the anechoic chamber provide the information to make judgments for the need of further measurements.
In order to enable the evaluation of Liu’s algorithm\(^1\), the performance of a reference method is presented. It is based on calculating the cross-correlation between the left channel and the right channel (Equation (3.1)). It should be reminded that various scaling schemes concerning the calculation of cross-correlation function can be found in the literature. The selection of the scaling scheme has a crucial role in some applications. However, different schemes are not compared or discussed any further here since this method is used just to provide some sort of a reference to the primary method.

It was discovered during the simulations that practically either of the methods can be tuned by adjusting the parameters to operate in such a manner that the correct angle is found. However, a method should perform well with a fixed parameter set. Thus, some parameters, not arbitrary though, are selected and the performance with this parameter set is discussed and illustrated. The aspect is in stressing the problems that the algorithms have.

The selected parameters for the cross-correlation method and Liu’s method are summarized in Table 7.1. The window length refers to the time window over which the calculation of a method is performed. Since Liu’s method is designed to utilize several frames’ data, the segment length is used for adjusting to what extent the past time is taken into account. The discussion concerning the significance of other parameters can be found in Chapter 5.

Velocity of sound was assumed to be 343 m/s which is a valid estimate at 20°C \(^2\).

\[\text{Table 7.1. The parameters of the DOA estimation methods}\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Cross-correlation</th>
<th>Liu’s method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment length</td>
<td>106 ms</td>
<td>106 ms</td>
</tr>
<tr>
<td>Cross-correlation scaling</td>
<td>no scaling</td>
<td>Window length</td>
</tr>
<tr>
<td>Segment length</td>
<td></td>
<td>106 ms</td>
</tr>
<tr>
<td>Number of sectors ((I))</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>Temporal integration coefficient ((\beta))</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Phantom peak parameter ((\gamma))</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

In Figure 7.1 and Figure 7.2 the performance of the DOA algorithms is illustrated for a set of two-second excerpts of the recorded signals. The statistics are listed in Appendix B. Note that some of the samples appearing in Table B.1 are left out from the figures for the sake of clarity. From the figures it can be stated that the simple cross-correlation is able to provide reasonable estimates for all samples except for the transient sequence at time instant around 1.25 s. In Table B.1 this sample has a relatively large standard deviation \((\approx 13^\circ)\) compared to the other samples. The performance of Liu’s method is virtually the same as the performance of the cross-correlation method excluding the 100 Hz sinusoid. It fails completely by suggesting

\(^1\) In fact, the algorithm in this thesis is not exactly the same as Liu et al. have presented in \cite{23} but much simplified version of that. “Liu’s method” is a reference to the system presented in Chapter 5.
angle $0^\circ$ at every time instant, resulting thus in zero standard deviation in Table B.1. However, since Liu’s method is capable of locating multiple sources, it was set to produce more than one estimate also for the one-source case. It was observed that the correct source was detected as the second strongest source in the case of the 100 Hz sinusoid. Figure 7.3 and Figure 7.4 present the results obtained for the classroom environment. The statistical measures related to this environment are collected to Table B.2. In this environment the performance change is seen as drastic rise in the standard deviations. In general, it can be stated that the cross-correlation is able to estimate the DOA successfully. Liu’s method, instead, performs rather poorly. It fails again completely with the 100 Hz sinusoid. Also in the case of the transient sequence, the outcome is unacceptably poor. In general, rise in the standard deviations is obvious also within the samples that are located correctly almost at every time instant.

The office proved to be the hardest environment as expected. Figure 7.5 and Figure 7.6 reflect the challenges set by this space. Again, the performance of the cross-correlation does not drop as rapidly as with Liu’s method. The transient sequence and the pop music seem to cause major difficulties. Also the failure of Liu’s method with the 100 Hz sinusoid is worth noting.

**Speech signals in case A**

The significance of various parameters provided by Liu’s algorithm was not studied thoroughly. However, it was presumed that the segment length would have a great significance to the performance. Thus this parameter was searched for the best value. Moreover, it affects the dynamics of the overall system. The dynamics here refers to the fact that the shorter the segment length, the more rapid the response. Let us conclude this section by presenting the statistical measures calculated using the optimal segment length for the speech signals that are the most important for our purposes.

Figure 7.7 and Figure 7.8 account for the localization ability in the anechoic chamber. Surprisingly, the optimal segment length differs despite the fact that signal
type is the same. This is seen as more frequent updates of the DOA estimate in Figure 7.7 compared to Figure 7.8. However, no conclusive statements can be presented based on one pair of signals. In Table 7.2 the optimal segment lengths including the statistical measures related to the figures are presented.

The opposite behavior compared to the anechoic chamber is noticed while evaluating the results obtained using the classroom recordings. The results are presented in Figure 7.9 and Figure 7.10. In this case, the female speaker was located best using a shorter time period (see Table 7.2).

Finally, Figure 7.11, Figure 7.12 and Table 7.2 account for the office. In this case, 21 ms is sufficient for the male speech but 85 ms segment resulted in the best performance for the female speaker.

At least dozens of recordings using different speech signals in each environment are required to enable some sort of deduction concerning any parameter (see Table 7.1). However, using the statistical measures also with this amount of data seems to provide at least a plausible value for the segment length. $\gamma$ and $\beta$ were selected using the same values as in the reference system by Liu et al. [23].
7.1.2 DOA estimation in the case of strong interferer (case B and case C)

In the previous section, it was pointed out that DOA estimation based on calculating cross-correlation between the left channel signal and the right channel signal is able
Table 7.2. The performance of Liu’s method with the optimal segment length in the case of speech signals. The nominal direction-of-arrival is $-15^\circ$.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Sample</th>
<th>Segment length [ms]</th>
<th>$E(\hat{\varphi} - \varphi)$ $[^\circ]$</th>
<th>$\sqrt{\text{var}(\hat{\varphi})}$ $[^\circ]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anechoic</td>
<td>speech (male)</td>
<td>21</td>
<td>0.48</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>speech (female)</td>
<td>107</td>
<td>0.19</td>
<td>0.70</td>
</tr>
<tr>
<td>Classroom</td>
<td>speech (male)</td>
<td>107</td>
<td>-0.03</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>speech (female)</td>
<td>43</td>
<td>0.08</td>
<td>3.37</td>
</tr>
<tr>
<td>Office</td>
<td>speech (male)</td>
<td>21</td>
<td>-0.49</td>
<td>4.33</td>
</tr>
<tr>
<td></td>
<td>speech (female)</td>
<td>85</td>
<td>0.17</td>
<td>3.08</td>
</tr>
</tbody>
</table>

to provide valid estimates. It is able to outperform the much more sophisticated method with several signal types. Moreover, the cross-correlation method is capable of rapid response. However, in the case of several strong sound sources it is not able to operate in the desired manner. Liu’s method, in turn, is designed particularly for multi-source cases. More specifically, it is designed particularly for multi-talker speech signals.

The discussion in this section is restricted to DOA estimation using Liu’s method and the results are presented only for the speech signals (see Appendix A for the complete list of the two-source mixtures). However, let us point out that several two-source combinations reflected a well known problem. In several cases, the existence of phantom sources is observed which are resulting from summing localization (see page 21 for the description). This is a serious problem with some two-source combinations in case B, not to mention case C. Particularly, a broadband interferer such as “white noise” and “babble” resulted in failing to localize the desired sound source, for instance, the male speaker (source 1: speech (male) 64 dB, source 2: white noise 54 dB in Table A). In the anechoic chamber, despite the strong broadband interferer “babble”, for instance, the correct DOA estimate for male or female speaker is found once the segment length is chosen properly (basically sufficiently long).

Figure 7.13, Figure 7.14 and Figure 7.15 present the performance for two-simultaneous-talkers configuration in the anechoic room, in the classroom and in the office, respectively (case B). The male speaker is placed at $\varphi = -15^\circ$ and the female speaker at $\varphi = 45^\circ$ (see Appendix A Table A.2). The male speaker is approximately 10 dB louder than the female speaker.

Figure 7.13 and Figure 7.15 illustrates well the problems that begin to arise as the difficulty level in auditory conditions is increased. Speech signals contain a lot of silent periods. Thus, if the other speech stream is “silent” while the other speaker is uttering, naturally, the stronger of the sources is localized. This is the reason for the large variation between the strongest and the second strongest source (female speaker at $\varphi = 45^\circ$).

In Figure 7.13 and Figure 7.15 some erroneous estimates (not in the vicinity of either
$\varphi = -15^\circ$ or $\varphi = 45^\circ$) may result from phantom sources the existence of which is highly probable in the office environment (for instance at $t \approx 0$ s).

As it was pointed out in the previous section, any conclusive statements with this amount of recordings concerning the parameters are not appropriate. Table 7.3 presents the segment lengths which resulted in the estimates in the figures related to cases B and C. Note that for the anechoic environment 21 ms is basically sufficient in case B but the signal type has to be kept in mind in this case. However, considering the overall system the fluctuation in the DOA estimate, illustrated in Figure 7.13 is not acceptable. Instead, some other parameter configuration should be used.

Figure 7.14 illustrates the system’s operation in the classroom. The segment length is 341 ms is used which seem to produce valid estimates for the source of interest at $\varphi = -15^\circ$.

The trend, however, seems to be towards longer segments as the auditory conditions are made more challenging. Compare the segment lengths that are used in case A (Section 7.1.1). It is worth pointing out that increasing the complexity in auditory conditions while increasing the segment length resulted also in better results in the system by Liu et al. [23].

<table>
<thead>
<tr>
<th>Case</th>
<th>Environment</th>
<th>Segment length [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Anechoic</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Classroom</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td>Office</td>
<td>234</td>
</tr>
<tr>
<td>C</td>
<td>Anechoic</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Classroom</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Office</td>
<td>21</td>
</tr>
</tbody>
</table>

Figure 7.16, Figure 7.17 and Figure 7.18 illustrate the case in which the sound level of two talkers is equal at the receiving end. Since both sources are of interest, two-source tracking is enabled in the system. It means that in each segment two dominating sources are searched for instead of one. Two-source tracking can be utilized also in case B to get rid of the problems discussed above. Yet, it is worth noting that, for instance in the office, the segment length 21 ms is sufficient to obtain good results when two-source-tracking is enabled. A shorter segment length is of course better since the response is more rapid.

It is worth pointing out that, subjectively, 10 dB in sound level (see the measurement point in Chapter 6) is surprisingly small meaning case B and case C are virtually the same case for human auditory system.

### 7.2 Quality of separated signals

In this section, a brief discussion concerning the quality of the separated signals is presented. The performance evaluation of signal processing systems is usually made via presenting quality measures such as signal-to-noise ratio (SNR), mean-square error (MSE) and correlation between some reference signal and the processed signals. However, presenting this kind of quality measures is not possible here due to the fact that no proper reference signal is available (required by MSE evaluation and
cross-correlation measure). Although the original signal (stimulus) is available, it can not be used due to the fact that the transfer function, modeling the alterations to which signal is exposed on its path from the loudspeakers to the microphones would be neglected. In fact, the signals that include the effect of the transfer function are disposable for a reference signal (case A signals). However, also these signals can not be used as the reference signals since there is no guarantee that phase information is the same in different recordings. Time alignment of the signals without synchronization is also problematic. Yet, there is no proper estimate of noise power to enable the estimation of SNRs.

MSE alone is a poor measure of the perceptual quality of signals. Figure 7.19 illustrates this issue. It was obtained while searching parameters for the sinusoidal modeling with the following goal: search for the best $\nu$ and $W$ which result in the smallest MSE. That is, the arguments by which the minimum of the function in the figure is reached. The parameters suggested by the MSE were set and the mixture was fed to the separation system. It was soon discovered that subjective quality estimation resulted in more or less the opposite to this performance measure. For instance, choosing parameters which result in small values for MSE can result in significantly worse quality on a subjective scale. The same sort of tendency as for
MSE was experienced using SNR as a quality measure.

Due to the facts discussed here, no numerical evaluation is presented concerning the quality. However, it can be stated that the quality compared to the original signals is affected most by the following factors. Each natural environment introduce
some background noise to the recorded signals. Including the background noise, the transfer function may have a significant effect on the signals in certain cases (e.g., reverberant rooms). The separation system itself is another major factor affecting the quality. The quality compared to an original signal is degraded, firstly, due to the fact that the original signal is modeled which inherently results in lower quality. It is highly probable that in the separation subsystem several signal components are classified erroneously. Apparently, the quality can be affected in this system by tuning the modeling subsystem and the grouping subsystem. Thus, the sinusoidal modeling parameters were exhaustively searched for the best possible result. Finally, more or less the general trend of parameter behavior was found for speech signals.

**Generated signals**

Generated signals refers to the fact that a sound source is artificially placed at a certain spatial location by delaying the left channel signal compared to the right channel signal or vice versa. The generated signals were used, firstly, to obtain initial parameter estimates for the speech signals. Additionally, they provide some kind of an upper limit to the audio quality which is possible to achieve with this system. The same samples were used to produce the generated mixtures that were used in the measurements. Compared to the recorded signals, the generated mixtures lack the effect of room response which was observed to be significant. Even the mixtures that are recorded in the anechoic chamber, are altered. Minor differences are caused by A/D conversions and other data processing that is needed to provide the mixtures in the appropriate format to the system (see Chapter 6 for the discussion related to this issue).

The audio quality of the system is affected by a few major factors. The factors are discussed above. If the quality degradation introduced by the signal modeling is ignored, the resulting signal quality is affected by the performance of the grouping stage (see Figure 5.1). The grouping is performed based on DOA estimate. Thus, the performance of the DOA subsystem is directly affecting the quality. It was discovered while simulating the subsystems that each performed satisfactorily excluding some special cases. For instance, some tones caused problems to the DOA subsystem. However, any broadband stimuli resulted in exact DOA estimates.

Also the modeling subsystem operated in the desired manner. This was tested by comparing the original signal to the modeled signal. Modeling artifacts were tried to detect. It turned out that already setting an upper limit of 20 sinusoids, for instance, resulted in sufficient quality with the speech signals. Moreover, not all sinusoids are needed in every time frame.

The effect of the detected transients was studied by comparing the speech signals modeled using only the sinusoidal modeling to the signal which was modeled using not only sinusoids but also the detected transients. In the case of speech signals, virtually all the detected transients are actually consonants in the original signal. The existence of the consonants in the separated signals improves the quality. 20 sinusoids, for instance, is absolutely too little to model a transient, or a consonant, using the sinusoidal modeling.
The overall performance, that is, both subsystems operating and the grouping stage enabled, was tested with a few signal contents. Narrowband and broadband stimuli were generated to correspond to similar situations to case A and to case C. In the latter case, the mixtures were generated by first scaling both signals to the interval of \([-1, 1]\) in order to estimate the case C configuration (see Appendix A). Of course, the actual effect of this scaling is quite far from the case C since even the quantity that is used to control the difference is not the same. The mixture that was generated using the male speech sample and the female speech samples produced a good separated result no matter which of the talkers was selected as the desired source. For the female speech signal the quality is however a bit better. Probably, the modeling of the female speech sample is more successful than the male speech sample.

Recorded signals

Each subsystem operated in the desired manner also using the recorded signals. Even in the hardest acoustic conditions, that is the office environment, both subsystems performed well. However, there are a few problematic cases for the DOA subsystem. These aspects are illustrated and discussed in Section 7.1. Next, the overall performance in the case of the recorded signals is presented by describing the performance, and by stressing the noteworthy issues that were encountered.

In case A basically all the signal types were separated in such a manner that the background noise in the environments reduced compared to the original recorded signal. However, some artifacts caused by the sinusoidal modeling may be more disturbing with some samples than the background noise in the original signal. Particularly, if the parameters of the sinusoidal modeling (e.g. window length, the amount of sinusoids, \(W\) and \(\alpha\)) are not tuned to the type of signal under processing the artifacts are disturbing. The extreme cases of this type of signals are of course the noise signals and the noise-like signals which are not even modeled plausibly at all with a reasonable amount of sinusoids.

Presumably, the quality of the resulting signal is the best in the anechoic chamber. However, the existence of artifacts caused by modeling can be easily observed. In the classroom and in the office, the effect of the room response is prominent. It seems that the artifacts caused by the modeling get even “emphasized” in these environments. This is probable due to the fact that in these rooms the reverberation is significant. As a consequence of the reverberation and the modeling, many people may prefer the original signal to the modeled-separated signal.

In cases B and C the observations concerning the quality of the separated signals are somewhat overlapping. Despite the 10 dB level difference in case B for the desired sound source, no significant improvement was observed in the quality of the separated signals using case C signals. As it was pointed out earlier, the 10 dB difference in sound level at the receiving end of the configuration, is not so big subjectively. In addition to the phenomena observed in case A, “leaking” of the undesired sources to the separated desired source occurs. This leaking is particularly disturbing since it is random in nature. For instance, if male speech is separated from male + female mixture, complete vowels belonging to the female speech signal can be observed while
listening the separated male speech. This is probably results from the reflections in
the environments.

In fact, the leaking problem was studied exhaustively to track its origin. At first, the
operation of each subsystem was verified independently. As stated, each subsystem
as such was observed to operate in the desired manner in most cases including
the speech signals in all the environments. The testing of the DOA subsystem was
made as described in Section 7.1. The signal modeling unit was tested, firstly, with
the one-sound-source configuration. The grouping stage was disabled. The stimulus
was modeled with the same parameters. The quality of two resulting signals was
compared, and it can be stated that the quality is virtually the same excluding the
effect of ambience noise the influence of which can be observed in the former case.

However, once the grouping was enabled the quality drop was obvious. It was dis-
covered that the phase information of the received-modeled signal is corrupted. The
corruption of phase information was discovered by probing the grouping stage while
the recorded signals were fed to the system. The phase information of the signal
components that arise from the desired source did not correspond to the spatial
angle at which the source was actually located. The phenomenon was observed with
several signal types.

The corruption of phase information leads to false grouping, that is, the desired
components are grouped to the undesired part and vice versa. Based on the tests and
the probing of the phase information, it is well-argued to state that room response is
the major factor to the phenomenon. It should be pointed out that the phenomenon
occurred also in the anechoic chamber in which the effect of the room response is
relatively small. In the classroom and in the office the problem is significantly bigger.
Due to the existence of the phenomenon in the anechoic chamber, there may be other
factors that are contributing to the phase information of the recorded signals. Even
though these factors can not be presented here, they all are located in path between
the loudspeakers and the transformation process from DAT-tapes to wav-files.

The significance of the transient processing subsystem was conducted also in this
case (see Figure 5.1 for overview of the system). Using the generated signals, the
quality of the speech signals was improved. The transient detection operated also
with the recorded signals resulting in better intelligibility of the separated speech
signals over the case where only the sinusoidal modeling was used.

In this section a brief evaluation on the quality of the separated sounds was made.
However, a proper subjective evaluation requires listening tests with persons that
are not affiliated to the project. For instance, it was observed that human hearing
seems to interpolate missing constituents in the modeled-separated signals. Thus, the
quality estimate of a person who searches for the best parameters for the separation
system is not the same as the quality estimate of a person that not involved in the
project. Thus, the reader is encouraged to contact the author for demonstration
signals for personal judgments. A few examples can be found in a demonstration
page at
Chapter 8

Conclusions

In this thesis the goal was to develop a system that is capable of sound source separation using two microphones. The separation is based on (1) searching the spatial location of the sound source, (2) model the signal mixture using sinusoids + transients -model and (3) use the location information for grouping the modeled signal between the desired part (source of interest) and the undesired part (background noise, other strong sources). Ambitiously, the original intention was to build a system that is capable of operating in real-world environments.

The primary hypothesis is that spatial location of a sound source is one of the major cues for sound source separation. This hypothesis arises from the knowledge of human auditory system and from psychoacoustics. To evaluate the validity of the proposed hypothesis exhaustive literature review was made. The review was directed to two separate research areas – DOA estimation methods and sound source separation methods.

Human has a tremendous DOA estimation ability. It is excellent even though human DOA system has only two sensors collecting the auditory data. Even more amazingly, a human being is able to direct the attention to a certain sound source. Despite the presence of multiple sound sources, human can separate the sound source in a manner that all the other sources are basically ignored.

The term cocktail-party effect is used allude to the ability of sound source separation while human auditory system is concerned. The cocktail-party effect is easy to prove to exist basically with no scientific studies. The effect works also in rather hard auditory conditions. One is able to have a conversation in a space which is full of people, each having a conversation of their own.

Since the human auditory system performs so tremendously in such auditory conditions, it is well argued to state that the methods for the purposes of this project were to be found in the research areas that model human hearing system. Loosely speaking, the system that is needed to meet the goals of this project is a cocktail-party processor.

This type of systems have been developed but most of them work only if certain conditions are met. In this project, it was desired that neither assumptions of the signal type nor assumptions of auditory conditions are made. This is to say, the
goal was to develop a system that would operate equally well in anechoic chamber conditions and in a common room with basically any signal content.

Test signals that correspond to real-world separation problems were recorded to study the performance of the proposed system. The recordings took place in three environments: anechoic chamber, classroom and office. In each of these environments, the measurements consisted of three types of recordings which differed from each other basically in the amount of interfering sources. Various signal types were recorded to evaluate the influence of this aspect to the performance of the DOA subsystem and the sound source separation system. In addition to recorded data, some signals emulating certain spatial location and cocktail-party situation were generated (e.g. summing male speech signal and female speech signal). Despite the system was desired to be independent of signal type, speech signals finally were of the primary interest.

The performance of the individual subsystems is satisfactory and meets the set goals. The performance of the overall system is also good when it was evaluated using the generated signals. For instance, male speaker mixed with spatially apart female speaker are separated producing audio quality that basically suffers only from signal modeling. The performance of each subsystem with real-world recordings is in most cases satisfactory. The DOA of a sound source is estimated with the desired precision and robustness. Also the sinusoids + transients modeling produces the target audio quality. However, particularly in the classroom and in the office, especially, two-sound source configurations resulted in significantly degraded audio quality compared to the quality achieved with the generated mixtures. Excluding a few exceptions, also in the problematic cases the DOA subsystem and the modeling subsystem, as such, are working properly. While performing the separation, the quality drop is in some cases drastic due to the fact that signal components are grouped erroneously.

It is plausible to suggest that the room response is a major factor affecting the quality of the separated signals. In fact, this issue was exhaustively studied and discussed in more detail while presenting the results.

Finally, some ideas for further development of the system are presented. Firstly, the most important step is to identify the factors that cause the corruption of the phase information since it is the major factor affecting the quality of separated signals. As stated above, the DOA subsystem and the sound source separation subsystem already operate in the desired manner. One of the original goals was not making any assumption of signal type. However, this goal had to be given up to some extent. While performing the simulations, it was discovered that the adjustment of the parameters of each subsystem is required if the signal type changes drastically. For instance, a pure sinusoid requires usually different parameter set than a natural broadband signal like speech and music.

It is well-argued to suggest an additional stage to the system, that is, a unit capable of room response compensation. The actual implementation of this unit is, however, not a trivial task. Moreover, it can only be guessed what is the contribution of such a compensation unit to the resulting signals. However, it is highly probable that at least the problem with phase information is partly helped.
One may question what is the effect of the chosen signal modeling technique in relation to the primary signal type, that is, speech. It can be stated that using only the sinusoidal modeling, the achieved audio quality can be considered as sufficient for most purposes. Yet, the maximum amount of sinusoids per frame can be as low as 20, for instance. It must be pointed out that using the sinusoids + transients modeling may result in poorer subjective quality due to occasional transient bursts among smoothly behaving signal that is obtained by synthesizing sinusoids only. However, as stated in this thesis, the transients improve the intelligibility of speech signals.

Further improvement to the quality of the separated signals can be achieved by adding constraints in the grouping stage of the system. The constraints can be basically any of the cues affecting sound source separation. Initial evaluation using fundamental frequency as an additional cue was made. However, the estimation and the tracking of the fundamental frequency of a signal using the subsystem that was tested is not very robust. Enabling the grouping based on the fundamental frequency in addition to the grouping based on direction-of-arrival, resulted usually in lower amount of admissible sinusoidal components compared to the case in which only DOA was used as a grouping cue. This, in turn, resulted in poorer subjective quality obviously due to drop in the amount of admissible components. However, further development of the fundamental frequency estimation subsystem may significantly improve the audio quality. However, adding this cue to the system directs the overall system to be more dependent of signal type, that is, a continuous content is assumed.
References


[8] A. S. Bregman. When will we hear separate events in sequence of sounds? Presentation on ASA connected to human hearing.


REFERENCES

Appendix A

Measured signal combinations

Table A.1. Anechoic chamber, session 1, source 1 at $-30^\circ$ and source 2 at $+45^\circ$

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Level [dB]</th>
<th>Source 2</th>
<th>Level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>pink noise</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>sine 100 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>sine 2000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>brass music</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>pop music</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>transient (natural) sequence</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
</tbody>
</table>

SNR 0 dB

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Level [dB]</th>
<th>Source 2</th>
<th>Level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech (male)</td>
<td>57</td>
<td>white noise</td>
<td>57</td>
</tr>
<tr>
<td>speech (female)</td>
<td>57</td>
<td>white noise</td>
<td>57</td>
</tr>
<tr>
<td>speech (male)</td>
<td>52</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (female)</td>
<td>52</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>52</td>
<td>speech (female)</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>53</td>
<td>babble (dozens of people)</td>
<td>53</td>
</tr>
<tr>
<td>speech (female)</td>
<td>52</td>
<td>babble (dozens of people)</td>
<td>51</td>
</tr>
</tbody>
</table>

SNR 10 dB

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Level [dB]</th>
<th>Source 2</th>
<th>Level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech (male)</td>
<td>57</td>
<td>white noise</td>
<td>46</td>
</tr>
<tr>
<td>speech (female)</td>
<td>57</td>
<td>white noise</td>
<td>46</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (female)</td>
<td>63</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>speech (female)</td>
<td>51</td>
</tr>
<tr>
<td>speech (male)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>53</td>
</tr>
<tr>
<td>speech (female)</td>
<td>63</td>
<td>babble (dozens of people)</td>
<td>53</td>
</tr>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>music (brass music)</td>
<td>56</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>music (pop music)</td>
<td>51</td>
</tr>
</tbody>
</table>
Table A.2. Anechoic chamber, session 2, source 1 at $-15^\circ$ and source 2 at $+45^\circ$

<table>
<thead>
<tr>
<th>source 1</th>
<th>level [dB]</th>
<th>source 2</th>
<th>level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>pink noise</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>sinc 100 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>sinc 1000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>sinc 2000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>brass music</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>pop music</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>transient (natural) sequence</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-</td>
<td>ambient noise</td>
<td>&lt; 34</td>
</tr>
<tr>
<td>SNR 0 dB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech (male)</td>
<td>57</td>
<td>white noise</td>
<td>57</td>
</tr>
<tr>
<td>speech (female)</td>
<td>56</td>
<td>white noise</td>
<td>56</td>
</tr>
<tr>
<td>speech (male)</td>
<td>55</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (female)</td>
<td>52</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>54</td>
<td>speech (female)</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>53</td>
<td>babble (dozens of people)</td>
<td>53</td>
</tr>
<tr>
<td>speech (female)</td>
<td>52</td>
<td>babble (dozens of people)</td>
<td>51</td>
</tr>
<tr>
<td>SNR 10 dB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech (male)</td>
<td>57</td>
<td>white noise</td>
<td>46</td>
</tr>
<tr>
<td>speech (female)</td>
<td>56</td>
<td>white noise</td>
<td>46</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (female)</td>
<td>63</td>
<td>pink noise</td>
<td>52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>speech (female)</td>
<td>50</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>babble (dozens of people)</td>
<td>51</td>
</tr>
<tr>
<td>speech (female)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>51</td>
</tr>
<tr>
<td>speech (male)</td>
<td>66</td>
<td>music (brass music)</td>
<td>56</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>music (pop music)</td>
<td>51</td>
</tr>
</tbody>
</table>
## Table A.3. Classroom, session 1, source 1 at $-30^\circ$ and source 2 at $+45^\circ$

<table>
<thead>
<tr>
<th>source 1</th>
<th>level [dB]</th>
<th>source 2</th>
<th>level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>pink noise</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 100 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 1000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 2000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>brass music</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>pop music</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>transient (natural) sequence</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SNR 0 dB</th>
<th></th>
<th>SNR 10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>white noise</td>
<td>67</td>
</tr>
<tr>
<td>speech (female)</td>
<td>68</td>
<td>white noise</td>
<td>67</td>
</tr>
<tr>
<td>speech (male)</td>
<td>61</td>
<td>pink noise</td>
<td>60</td>
</tr>
<tr>
<td>speech (female)</td>
<td>61</td>
<td>pink noise</td>
<td>60</td>
</tr>
<tr>
<td>speech (male)</td>
<td>61</td>
<td>speech (female)</td>
<td>62</td>
</tr>
<tr>
<td>speech (male)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>63</td>
</tr>
<tr>
<td>speech (female)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SNR 10 dB</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>white noise</td>
<td>57</td>
</tr>
<tr>
<td>speech (female)</td>
<td>67</td>
<td>white noise</td>
<td>54</td>
</tr>
<tr>
<td>speech (male)</td>
<td>64</td>
<td>pink noise</td>
<td>54</td>
</tr>
<tr>
<td>speech (female)</td>
<td>64</td>
<td>pink noise</td>
<td>54</td>
</tr>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>speech (female)</td>
<td>56</td>
</tr>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>babble (dozens of people)</td>
<td>57</td>
</tr>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>music (brass music)</td>
<td>57</td>
</tr>
<tr>
<td>speech (male)</td>
<td>67</td>
<td>music (pop music)</td>
<td>57</td>
</tr>
</tbody>
</table>
Table A.4. Classroom, session 2, source 1 at $-15^\circ$ and source 2 at $+45^\circ$

<table>
<thead>
<tr>
<th>source 1</th>
<th>level [dB]</th>
<th>source 2</th>
<th>level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>pink noise</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 100 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 1000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>sinc 2000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>brass music</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>pop music</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>transient (natural) sequence</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-</td>
<td>ambient noise</td>
<td>37</td>
</tr>
</tbody>
</table>

SNR 0 dB

| speech (male)       | 64         | white noise          | 64         |
| speech (female)     | 64         | white noise          | 64         |
| speech (male)       | 59         | pink noise           | 59         |
| speech (female)     | 59         | speech (female)      | 60         |
| speech (male)       | 64         | babble (dozens of people) | 63       |
| speech (female)     | 64         | babble (dozens of people) | 63       |

SNR 10 dB

| speech (male)       | 64         | white noise          | 53         |
| speech (female)     | 64         | white noise          | 53         |
| speech (male)       | 59         | pink noise           | 50         |
| speech (female)     | 60         | pink noise           | 50         |
| speech (male)       | 64         | speech (female)      | 54         |
| speech (male)       | 64         | babble (dozens of people) | 54       |
| speech (male)       | 64         | music (brass music)  | 54         |
| speech (male)       | 64         | music (pop music)    | 54         |
Table A.5. Office, session 1, source 1 at $-15^\circ$ and source 2 at $+45^\circ$

<table>
<thead>
<tr>
<th>source 1</th>
<th>level [dB]</th>
<th>source 2</th>
<th>level [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>pink noise</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>sinc 100 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>sinc 1000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>sinc 2000 Hz</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>brass music</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>pop music</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>transient (natural) sequence</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-</td>
<td>ambient noise</td>
<td>48</td>
</tr>
<tr>
<td>SNR 0 dB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>white noise</td>
<td>70</td>
</tr>
<tr>
<td>speech (female)</td>
<td>70</td>
<td>white noise</td>
<td>70</td>
</tr>
<tr>
<td>speech (male)</td>
<td>64</td>
<td>pink noise</td>
<td>63</td>
</tr>
<tr>
<td>speech (female)</td>
<td>63</td>
<td>pink noise</td>
<td>63</td>
</tr>
<tr>
<td>speech (male)</td>
<td>62</td>
<td>speech (female)</td>
<td>62</td>
</tr>
<tr>
<td>speech (male)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>64</td>
</tr>
<tr>
<td>speech (female)</td>
<td>64</td>
<td>babble (dozens of people)</td>
<td>64</td>
</tr>
<tr>
<td>SNR 10 dB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>white noise</td>
<td>60</td>
</tr>
<tr>
<td>speech (female)</td>
<td>70</td>
<td>white noise</td>
<td>60</td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>pink noise</td>
<td>59</td>
</tr>
<tr>
<td>speech (female)</td>
<td>70</td>
<td>pink noise</td>
<td>59</td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>speech (female)</td>
<td>60</td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>babble (dozens of people)</td>
<td>60</td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>music (brass music)</td>
<td>59</td>
</tr>
<tr>
<td>speech (male)</td>
<td>70</td>
<td>music (pop music)</td>
<td>60</td>
</tr>
</tbody>
</table>
Appendix B

Simulations

Table B.1. Anechoic environment. Nominal direction-of-arrival is $-15^\circ$

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tdoff $\hat{\phi} - \phi$ [$^\circ$]</th>
<th>$\sqrt{\text{var}(\hat{\phi})}$ [$^\circ$]</th>
<th>$\hat{\phi} - \phi$ [$^\circ$]</th>
<th>$\sqrt{\text{var}(\hat{\phi})}$ [$^\circ$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-1.50</td>
<td>1.82e-15</td>
</tr>
<tr>
<td>pink noise</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-1.44</td>
<td>0.16</td>
</tr>
<tr>
<td>sine 100 Hz</td>
<td>-1.71</td>
<td>2.051</td>
<td>15.00</td>
<td>0.00</td>
</tr>
<tr>
<td>sine 1000 Hz</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-1.66</td>
<td>0.14</td>
</tr>
<tr>
<td>sine 2000 Hz</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-3.30</td>
<td>0.78</td>
</tr>
<tr>
<td>pop music</td>
<td>-0.01</td>
<td>2.12</td>
<td>-0.13</td>
<td>0.23</td>
</tr>
<tr>
<td>brass music</td>
<td>1.40</td>
<td>1.95</td>
<td>0.11</td>
<td>0.43</td>
</tr>
<tr>
<td>transient sequence</td>
<td>1.73</td>
<td>13.37</td>
<td>-1.36</td>
<td>5.30</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-0.97</td>
<td>3.39</td>
<td>0.63</td>
<td>0.83</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-2.65</td>
<td>3.81</td>
<td>0.19</td>
<td>0.59</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>-0.24</td>
<td>2.53</td>
<td>0.00</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table B.2. Classroom environment. Nominal direction-of-arrival is $-15^\circ$

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tdoff $\hat{\phi} - \phi$ [$^\circ$]</th>
<th>$\sqrt{\text{var}(\hat{\phi})}$ [$^\circ$]</th>
<th>$\hat{\phi} - \phi$ [$^\circ$]</th>
<th>$\sqrt{\text{var}(\hat{\phi})}$ [$^\circ$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>white noise</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>pink noise</td>
<td>-1.65</td>
<td>3.66e-15</td>
<td>-1.44</td>
<td>0.16</td>
</tr>
<tr>
<td>sine 100 Hz</td>
<td>3.97</td>
<td>2.02</td>
<td>14.97</td>
<td>0.12</td>
</tr>
<tr>
<td>sine 1000 Hz</td>
<td>1.40</td>
<td>1.95</td>
<td>12.63</td>
<td>0.30</td>
</tr>
<tr>
<td>sine 2000 Hz</td>
<td>2.34</td>
<td>1.00</td>
<td>3.58</td>
<td>0.91</td>
</tr>
<tr>
<td>pop music</td>
<td>-2.47</td>
<td>6.01</td>
<td>0.69</td>
<td>3.38</td>
</tr>
<tr>
<td>brass music</td>
<td>-6.64</td>
<td>6.73</td>
<td>-2.05</td>
<td>8.78</td>
</tr>
<tr>
<td>transient sequence</td>
<td>0.71</td>
<td>9.37</td>
<td>8.30</td>
<td>16.99</td>
</tr>
<tr>
<td>speech (male)</td>
<td>-2.62</td>
<td>9.01</td>
<td>3.27</td>
<td>6.36</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-1.38</td>
<td>13.78</td>
<td>2.55</td>
<td>6.81</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>0.91</td>
<td>4.14</td>
<td>-0.25</td>
<td>1.62</td>
</tr>
</tbody>
</table>
### Table B.3. Office environment. Nominal direction-of-arrival is −15°

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tdoff</th>
<th>Method</th>
<th>Liu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\mathbb{E}(\hat{\phi}) - \phi)</td>
<td>(\sqrt{\text{var}(\hat{\phi})})</td>
<td>(\mathbb{E}(\hat{\phi}) - \phi)</td>
</tr>
<tr>
<td>white noise</td>
<td>2.58</td>
<td>5.48e-15</td>
<td>1.69</td>
</tr>
<tr>
<td>pink noise</td>
<td>2.58</td>
<td>5.48e-15</td>
<td>1.75</td>
</tr>
<tr>
<td>sine 100 Hz</td>
<td>1.17</td>
<td>2.06</td>
<td>15.50</td>
</tr>
<tr>
<td>sine 1000 Hz</td>
<td>-1.65</td>
<td>3.65e-15</td>
<td>-1.69</td>
</tr>
<tr>
<td>sine 2000 Hz</td>
<td>-1.65</td>
<td>3.65e-15</td>
<td>-2.00</td>
</tr>
<tr>
<td>pop music</td>
<td>-5.93</td>
<td>15.15</td>
<td>8.58</td>
</tr>
<tr>
<td>brass music</td>
<td>4.19</td>
<td>3.25</td>
<td>7.33</td>
</tr>
<tr>
<td>transient sequence</td>
<td>8.57</td>
<td>5.53</td>
<td>8.52</td>
</tr>
<tr>
<td>speech (male)</td>
<td>3.23</td>
<td>5.23</td>
<td>4.00</td>
</tr>
<tr>
<td>speech (female)</td>
<td>-2.04</td>
<td>7.19</td>
<td>6.13</td>
</tr>
<tr>
<td>babble (dozens of people)</td>
<td>4.20</td>
<td>2.53</td>
<td>4.25</td>
</tr>
</tbody>
</table>