FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Computational Auditory Scene Analysis (CASA) for Separating Monophonic Music
Student:	Nikita Mortuzaiev
Supervisor:	Ing. Mgr. Ladislava Smítková Janků, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2022/2023

Instructions

1. Study and describe the basics of sound recognition and auditory scene analysis in humans.
2. Study and describe the existing algorithms for computational auditory scene analysis and their applications for sound recognition and separation.
3. Prepare a dataset that consists of recorded monophonic piano sounds.
4. Using the prepared dataset, try to apply CASA algorithms to separate tones from the background.
5. Evaluate the proposed CASA system in connection with a simple classifier. Evaluate the performance in comparison to non-CASA classification.
6. Perform experiments and evaluate them.

Electronically approved by Ing. Karel Klouda, Ph.D. on 1 February 2022 in Prague.

Bachelor's thesis

COMPUTATIONAL AUDITORY SCENE ANALYSIS (CASA) FOR SEPARATING MONOPHONIC MUSIC

Nikita Mortuzaiev

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Ing. Mgr. Ladislava Smítková Janků, Ph.D.
April 28, 2022

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Mortuzaiev Nikita. *Computational Auditory Scene Analysis (CASA) for Separating Monophonic Music.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2022.

Contents

List of Figures

Chtěl bych poděkovat především sit amet, consectetuer adipiscing elit. Curabitur sagittis hendrerit ante. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on April 28, 2022 .....

Abstract

The problem of source separation has become a reasonable challenge for the machine listening models. Often, the recorded sounds don't come from a single source, and thus it is needed to firstly preprocess them by splitting into separate streams. Humans, however, developed an astounding ability to analyze the auditory scene and focus their hearing on target sounds, while trying not to be distracted by the background ones. Looking at how easily the auditory system manages to do this, researchers have been trying to implement computational models that simulate this ability, and thus the first studies of computational auditory scene analysis have found their place in the world. This thesis is meant to be an introduction to the field of CASA. It investigates the theory behind auditory modeling and existing approaches to it, provides a practical example of a CASA system for separating music from noise, and experimentally evaluates it on a set of piano recordings.

Keywords computational auditory scene analysis, auditory modeling, source separation, cocktail party problem, music processing

Abstrakt

Problém oddělení zdrojů se stal rozumnou výzvou pro modely strojového slyšení. Nahrané zvuky často nepocházejí z jediného zdroje, a proto je nutné je nejprve předzpracovat rozdělením do samostatných proudů. Nicméně, lidé si vyvinuli vynikající schopnost analyzovat sluchovou scénu a zaměřit svůj sluch na cílové zvuky, přičemž neztrácet pozornost kvůli těm na pozadí. Vzhledem k tomu, jak snadno to sluchové ústrojí zvládá, se výzkumníci stále pokoušejí implementovat výpočetní modely, které tuto schopnost simulují, a tak první studie výpočetní analýzy sluchové scény našly své místo ve světě. Tato práce by měla být úvodem do oblasti CASA. Zkoumá teorii za sluchovým modelováním a existující přístupy k němu, poskytuje praktický příklad systému CASA pro oddělení hudby od šumu a experimentálně jej vyhodnocuje na sadě klavírních nahrávek.

Klíčová slova výpočetní analýza sluchové scény, sluchové modelování, oddělení zdrojů, cocktail party problem, zpracování hudby

Acronyms

ACF	Autocorrelation Function
ANSI	American National Standards Institute
ASA	Auditory Scene Analysis / American Standards Association
CASA	Computational Auditory Scene Analysis
CCF	Cross-Correlation Function
CCCF	Cross-Channel Correlation Function
ERB	Equivalent Rectangular Bandwidth
FIR	Finite Impulse Response
IBM	Ideal Binary Mask
IIR	Infinite Impulse Response
JND	Just-Noticeable Difference
SACF	Summary Autocorrelation Function
SNR	Signal-to-Noise Ratio
T-F	Time-Frequency

Chapter 1

Introduction

Imagine a party. You can hear a variety of sounds: music in the background, conversations between people, noises of somebody coughing, maybe even a dog barking outside... These sounds merge into a single stream that approaches your ears by vibrations in the air and then goes through different physical, biological and psychoacoustical processes to finally come in a form of electrical impulses to the brain. Despite all these sounds from different sources are mixed on the way to your ears, the brain can segregate one (or several) of them. You can focus your hearing on these "target" sounds and separate them from the complex mixture, leaving other sounds in the background. This phenomenon has been described as a "cocktail party effect", and the process of integrating separate sounds into meaningful streams, or "auditory objects" – auditory scene analysis, or ASA.

In machine perception — specifically in machine hearing — a related concept is referred to as Computational ASA (CASA) and is tightly connected to the fields of sound recognition and digital signal processing. CASA systems indeed aim to separate sounds from mixtures, but they differ from BSS (blind source separation) systems in that they try to do this in a way a human ear does. Being based on and trying to combine works from different fields of science, CASA systems can bring new solutions and insights to the complex problem of signal separation.

The main objective for this thesis is to describe the principles and goals of CASA, existing applications and approaches. Another objective is to practically apply the theoretical knowledge by implementing a simple CASA system to separate monophonic music from noise and experimenting with it. But before all of this, since this thesis is made for an IT-oriented audience, it is needed to make a brief introduction to the underlying physics and biology.

The thesis is split into four large chapters. Chapter 2 is made to provide theoretical background for computational auditory scene analysis and will gather related knowledge from different fields of science: physics, biology and math. Physical background is needed for some basic understanding of what sound actually is. Moreover, since the implemented system from the practical part aims to segregate music from noise, special attention there will be given to harmonic sounds and pitch perception. Next, biological background theory will introduce the reader to the mechanisms in the human ear that CASA systems try to mimic. There, auditory scene analysis according to Bregman will be described as well. The following section dedicated to the related math will make an introduction to the field of digital signal processing, including a conversation about digital filters and filterbanks. And finally, background theory for computational auditory scene analysis will follow, giving a description of its main principles, goals and applications, along with an overview of major works in the field.

Then, to make a step closer to the actual implementation, a chapter dedicated to the methodology will follow. Chapter 3 will firstly provide an overview of the mathematical concepts used there, and then will give a detailed description of the architecture of a typical CASA system.

Next, chapter 4 will describe the system implemented for this thesis. There, a conversation about the selected algorithms and methods will take place, along with their input parameters. Each stage of the architecture described previously will be given some special attention.

The final chapter (chapter 5) will gather information about the experiments made to test the implemented system, along with the observed results.

Chapter 2
Theoretical Part

2.1 Physical Background

Before starting to ponder the structures of the human ear, it is necessary to understand the basics of how sounds work in the real world. It is safe to say that many people don't ask this question – they just make sounds or react to them, unconsciously knowing the outcomes. Human mind has already developed a deep understanding of which sounds are produced under different circumstances – you can easily say what to expect when somebody scratches a blackboard or rings a bell. Some could say that sounds are just "pressure waves that propagate through the air", but in reality, there is a lot of interesting and complex things beyond this definition to pay attention to. This section will introduce the reader to the underlying physics of sound and some interesting related concepts. A special focus will be made on describing harmonic sounds, which are essential to understand to be able to work with music and pitch.

2.1.1 What a Sound Is

The definition of sound above, saying that it is just vibrations in the air, is hard to be called incorrect from the scientific point of view. Of course, there are improvements to be made: for example, that sound can propagate not only through the air, but through any medium that has inert mass and is "elastic", or stiff, meaning that it will respond to forces applied to it. Making those corrections, it is also important to note that the definition above relates to sound as a physical phenomenon, but there is another definition that people use mostly in psychology and physiology, saying that the sound is a perception in the brain, or auditory sensation of the concept described above, or "an object of hearing". It is possible to argue about the question of "What Is Sound?" for a long time, as people still haven't come to a single definition and tend to mix the concepts [1], but in this thesis, the term "sound" will be used primarily in the first, physical sense, unless specified differently.

For better understanding of how physical sound works, keep in mind the mass and elasticity of the air mentioned above. Overall, mass and elasticity (not only of the air, but of any medium) play a very important role in the related studies: mass-spring systems are a highly discussed topic, along with the type of oscillations they tend to have. Any object that can produce sounds may be considered a mass-spring system: a bell, a guitar string, or even air or water, which can be thought of as many small masses connected by invisible springs... This knowledge is quite staggering – in most cases, it is hard to imagine such system, because there could be no obvious mass nor elasticity. Consider an example for explaining resonant cavities: why a can of soda

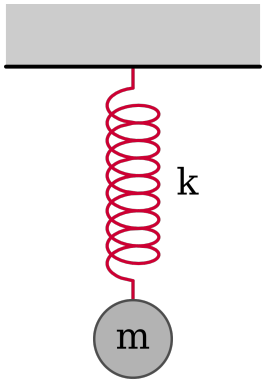


Figure 2.1 A simple vertical mass-spring system. Taken from `https://commons.wikimedia.org/`

makes that clicking sound when it is being opened? The air is the answer. When you open the can, some parts of the air near its top act as a mass, and other parts near the bottom as a spring. The pressure in the can drops, and the "spring" at the bottom tries to suck the "mass" back in, producing the expected sound [2].

Now, if you imagine the simplest of such systems, like the one on figure 2.1, you can notice that when a particular force is applied to it, it tends to oscillate in a sinusoidal manner (due to some famous laws of physics, which will not be further discussed here). In fact, this can be applied to all mass-spring systems: they naturally "want" to vibrate in a sinusoidal fashion with a preferred frequency, called resonance frequency. Sinusoidal vibrations will be given more attention in chapter 2.3.1.

When someone talks about applying forces to objects, they can probably say that an impulse is delivered. In classical mechanics, impulse is a widely used concept, but for the purposes of this thesis, it is rather important to note how objects respond to impulses. When an impulse is delivered, the object starts to vibrate at all possible frequencies, but having in mind an understanding of resonance frequencies, it is safe to say that not all applied frequencies sound the same in the end. Thus, some tones in the resulting sound tend to be louder, and others, if not completely silent, highly attenuated. This frequency selectivity is based on the object's properties: the material of which it is made, its form, mass, ... In signal processing, the notion of impulse response is widely used and will be referenced once again when discussing digital filters in chapter 2.3.2. The above-mentioned "chosen" frequencies will be described a bit more in the next section.

Another essential topic to mention here is why sounds fade in time. This is again connected to the concept of mass-spring systems and the amplitudes of their vibrations. Usually, the greater these amplitudes are, the louder the resulting sound is, so if the amplitudes didn't become smaller, we would live in constant unbearable noise. In brief, the fading is caused by the resistance of the medium, in which the sound propagates, and the manner of this propagating. It also depends on the material of which the sound source is made. If you imagine air as it was described above — as many masses connected by invisible springs — the mechanics of the propagation becomes clear: the sound source pushes the closest mass near it, which due to elasticity pushes its neighbors and returns to its starting location. Then its neighbors, in turn, push their neighbors and return, and so on, until these vibrations come to your ears. The air masses must be pushed again and again for the sound to spread, so it tends to lose its strength along the way, and the further from its source it travels, the smaller the amplitudes of the vibrations become.

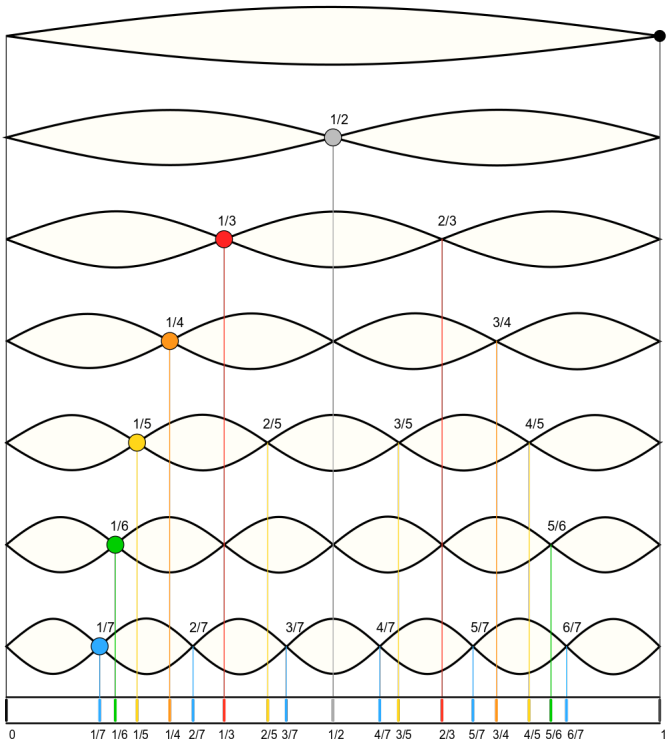


Figure 2.2 First seven harmonics of a sound wave (or first seven modes of vibration of a string). Taken from `https://commons.wikimedia.org/`

2.1.2 Harmonicity and Pitch

The conversation about how harmonics (or overtones) appear was already started in the previous section. In simple words, not all frequencies of the vibrations caused by delivering an impulse to an object keep their amplitudes for long. The ones that benefit the most from this phenomenon are harmonics, which are the periodic waves with frequencies that are positive integer multiplications of a specific frequency called fundamental (figure 2.2). For example, if the fundamental frequency is 200 Hz, the corresponding harmonics are 400 Hz, 600 Hz, 800 Hz, 1 kHz and so on. Each harmonic can be labeled with a number – the fundamental frequency one is also called the 1st, so the wave with frequency of 1 kHz from the example above would be the fifth. However, the scientific notation for harmonics might be confusing – some authors refer to the fundamental frequency as f_0 (and the fifth harmonic would be f_4 in that case), others as f_1 (and f_5 respectively). In this thesis, fundamental frequency will be notated as f_0.

Another explanation of how harmonics work might be found in [2]. When you pluck a guitar string, it doesn't vibrate only as a whole. The same string might be thought of as two halves, or three thirds, or even one hundred one hundreds, and that each part of it vibrates separately. So, when the string is plucked, all its harmonics are excited, and the resulting sound is not a pure tone, but a complex one. This behavior is often called "modes of vibration" and might be observed not only in strings, but also, for example, in sheets of metal.

The most interesting property of harmonics is that they are all periodic at their fundamental frequency. If you sum up any number of adjacent harmonics of a wave, the period of the resulting wave would be equal to the period of the fundamental. This property plays an important role in the perception of pitch and is often used for its estimation in machine hearing systems.

Now, what is pitch exactly? To start using this term in the thesis, it is important to provide a clear definition, but in fact, there is none that is considered a standard. Two most widely used ones were given in [3]. The first one was provided by the American Standards Association (ASA) in 1960 with a reference to music – they defined pitch as *"that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale"* ([3], p. 1). The second one was given by the American National Standards Institute (ANSI) in 1994 without a reference to music, saying that *"Pitch [is] that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends primarily on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus"* ([3], p. 1). For this thesis, it is enough to consider pitch as the auditory sensation mentioned in both definitions that can be ordered on a scale.

It is important to note that pitch is not a physical property of sound, but perceptual. When someone says "low pitch" or "high pitch", it is not certainly clear where this "low" or "high" is – low pitch for some people might be high for others. In the related studies of pitch perception in psychophysics, a term "just noticeable difference" (JND) is used, and there are references that humans can distinguish about 1 400 points on the pitch scale.

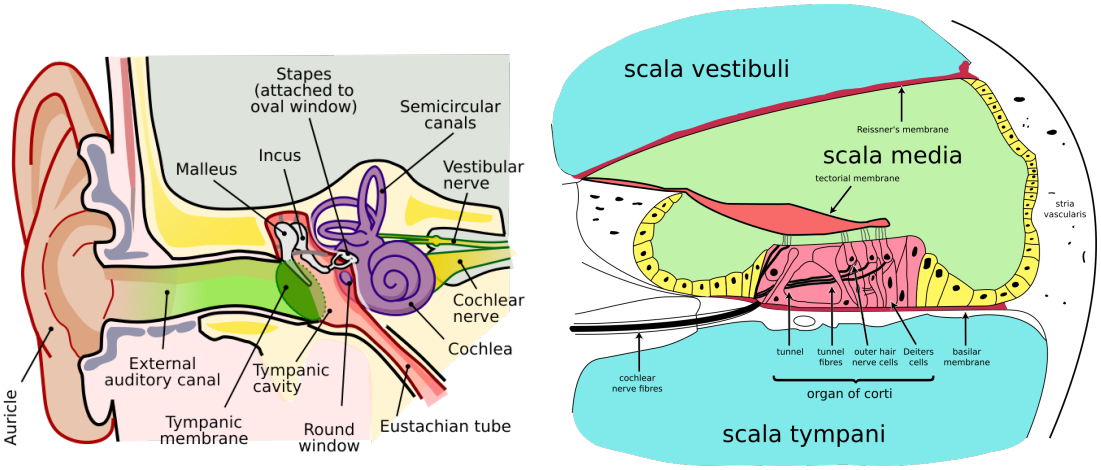
Pitch is often associated with fundamental frequency, though it is not fully equivalent to it. Experiments have shown that for some short periodic sounds the perception of pitch might not appear at all, though it was clear that they had a fundamental frequency. On the other hand, there are reports saying that sounds with a missing fundamental (the ones made only from higher harmonics) could evoke the perception of pitch associated with the missing frequency, thus giving the illusion of what was not present in reality [2]. Either way, pitch is a major attribute used while describing tones in Western music, as well as is loudness, duration and timbre. Pitch also plays an important role in auditory grouping, given the fact that same sound sources tend to produce sounds that are close in pitch. Auditory grouping in humans will be given more attention in chapter 2.2.3.

2.2 Biological Background

Sounds are for sure one of the most important sources of information in our everyday life. By listening to them, one can describe what is happening around, understand how to react to occurring situations, or even tell if a danger is approaching, and it is time to take action. It is hard to imagine human sensation without hearing, but as easy as this may sound (no pun intended), the biology behind it is quite complicated. This section will introduce the reader to how sound as a mechanical phenomenon is converted to sound as perception and provide a basic overview of the structures in the human ear, along with the mechanical and neurobiological processes happening inside of them.

2.2.1 Outer and Middle Ear

At the beginning, sound approaches the ear by vibrations in the air (or any other elastic medium) and enters the outer ear, which consists of the visible part (called the auricle, or the pinna) and the ear canal. The auricle is a thin plate of elastic cartilage, covered with integument, and connected to the surrounding parts by ligaments and muscles; and to the beginning of the ear canal by fibrous tissue. The ear canal is a tube leading from the bottom of the auricle to the middle ear, separated from it by the eardrum (or tympanic membrane). The main purpose of the ear canal is to focus the sound energy gathered by the auricle on the eardrum. It also amplifies



(a) **(b)**

Figure 2.3 **(a)** Anatomy of the human ear. The ossicles of the middle ear are shown in white. The inner ear is shown in purple. **(b)** Cross-section of the cochlea showing the organ of Corti and three chambers filled with cochlear fluids. Both pictures were taken from `https://commons.wikimedia.org/`

frequencies between 3 kHz and 12 kHz [4].

Being gathered on the eardrum, the mechanical vibrations propagate through the middle ear. Three bones (called the ossicles) are located inside of it. The malleus (also called the hammer) is connected to the eardrum and transfers the vibrations from it to the incus (the anvil). These vibrations are chaotic, but the malleus is connected to the eardrum in a linear manner, helping the ear to respond more linearly and smoothly. The incus, in turn, connects to the stapes (the stirrup). The footplate of the stapes introduces pressure waves in the inner ear, which starts with the oval window of the cochlea. The structures of the middle ear can be seen on figure 2.3a.

It may sound redundant to have additional structures in the ear which propagate the vibrations even further, when they could travel just one centimeter more in a way like before, in the ear canal, but in reality, the pressure of these mechanical vibrations is too small to cause the waves of the same velocity in the cochlear fluids. The ossicles help to amplify the pressure of these vibrations. They are positioned to form a lever, and, because the oval window is about 14 times smaller than the eardrum, the pressure gain becomes quite significant in the end – at least 18.1 times [4].

To regulate the middle ear and protect it from damage due to very loud sounds, two muscles are located inside of it: the stapedius muscle and the tensor tympani muscle. These muscles are controlled by unconscious reflexes and hold the ossicles when the vibrations become too intense. To provide ventilation and drainage of the middle ear and to equalize pressures in this isolated environment, the middle ear is connected to the back of the throat by the eustachian tube [2].

2.2.2 Inner Ear

The inner ear starts with the above-mentioned oval window, which is connected to the stapes of the middle ear. The oval window is a part of the cochlea — a structure of the inner ear dedicated to hearing. Along with the cochlea, the inner ear also contains the vestibular system, which is responsible for the sense of balance and spatial orientation and uses the same kinds of fluids and cells as the cochlea does. The vestibular system will not be covered in this thesis, but the fluids

and cells will be described in more detail later in the section.

The cochlea itself is a spiral-shaped cavity made of bony tissue, which makes about 2.75 turns around its axis and is about 3 cm long. The core component of it is the basilar membrane, which runs along almost its entire length and separates two of the three chambers of the cochlea filled with different fluids [2]: the tympanic duct (scala tympani) filled with perilymph, and the cochlear duct (scala media) filled with endolymph. The third chamber, the vestibular duct (scala vestibuli), is separated from the cochlear duct by the Reissner's membrane and is filled with perilymph (figure 2.3b). When the footplate of the stapes of the middle ear introduces movements to the cochlear fluids, the basilar membrane is affected too, and the endolymph in the cochlear duct moves along.

The most interesting property of the basilar membrane is that its stiffness and width is different throughout its length – the membrane is narrow and stiff at the basal end of the cochlea, and wide and floppy at the apical end. And here sound waves have two possible routes to take while propagating through the basilar membrane: a shorter path, which includes going through the stiffer parts of it, or a longer path, which means travelling along the membrane until it becomes easier to pass through, but pushing more fluid on the way. In fact, high-frequency waves tend to choose the shorter path, and low-frequency waves – the longer one. The distribution of frequencies passing through the basilar membrane is not linear, but close to logarithmic. In machine hearing systems, equivalent rectangular bandwidth (ERB) scale is usually used (see chapter 3.1 for a definition).

Thus, the basilar membrane moves in different places depending on the frequencies of the vibrations. The organ of Corti, which sits on top of it and runs along its entire length, contains displacement cells able to detect movements of the fluid nearby and excite the nearby neurons to send electrical impulses. Such cells are packed with a bunch of stereocilia (hair) that stick out of its top, and thus are called hair cells. These cells can be of two types: inner hair cells that are located closer to the center of the cochlea, and outer hair cells that sit closer to its outer side. Inner hair cells are less numerous than outer hair cells and form a single row along the organ of Corti, while outer hair cells usually form three rows [2].

Now, it is important to mention that the endolymph in the cochlear duct contains high amounts of positively charged ions (primarily potassium and calcium). When it moves in response to the sound pressure, the stereocilia of the inner hair cells are deflected, and tiny ion channels open in them. This allows the charged ions from the endolymph to enter the stereocilia. The cell becomes depolarized, and a receptor potential is produced. This results in releasing the neurotransmitters at the basal end of the cell and then triggering action potentials in the nerve nearby. In this way, inner hair cells detect movements around them and convert mechanical sound waves to electrical nerve signals.

Outer hair cells, in turn, serve as amplifiers of the quiet sounds. Their receptor potentials are converted to cell body movements, thus increasing the sound pressure [5].

2.2.3 Auditory Scene Analysis

To close up the section, it was decided to make an introduction to auditory scene analysis according to Bregman [6]. His book named *"Auditory Scene Analysis: The Perceptual Organization of Sound"* (1990) made a big influence on further researches, as it attempted to bring together the theoretical knowledge in the field that did not have any clear base to build on. Bregman's book is now widely recognized as this base, so it is necessary to list at least the primary concepts of ASA described there. This section could have been put to either of the chapters in the theoretical part

of the thesis, because it is connected to every field being discussed, but it resides in the biological part, because most of the addressed experiments were testing human auditory perception and are highly connected to the related studies in Gestalt psychology.

To start off, Bregman brings to the world a new term related to auditory perception. If you recall the two definitions of sound from chapter 2.1.1, you may remember that there were two of them: one related to sound as a physical phenomenon, and another related to perception in the brain. Bregman introduced a term "auditory stream", or "auditory object", to address the second definition. He made an analogy with vision and how humans tend to group separate surfaces of the same object on their eye retina to see the object as a whole and referred to "auditory streams" as to the same kind of objects, but for audition. He said that the term "sound" is not really well suitable in this case, because for example a melody in a recording of music consists of different sounds (notes), but people often percept this melody as a whole and group the sounds into something greater in their perception. Bregman's definition of auditory streams became very popular, so it will be used throughout this thesis too.

Bregman defines ASA as the process of separating these auditory streams from mixtures and refers to it as to a two-stage process. The first stage (segmentation) is said to be splitting the auditory input into so-called "segments", just as the visible object is split into surfaces in the human eye. The second stage is grouping and refers to integrating the segments together based on the grouping cues. With references to experiments from his lab he describes two possible approaches to grouping and searching for cues: simultaneous (which is also called vertical, or spectral) and sequential (or horizontal). While simultaneous grouping takes into account the segments that appear at the same time, but relate to different frequencies (are spread in space), sequential grouping works with segments that share the frequency component, but are located at different points in time. As an example of a cue for simultaneous grouping one could take common onset and offset, because it is usual for sounds (or different frequency components of the same sound) from the same source to start at the same time. For sequential grouping, probably the most common cue is pitch. If two sounds have pitches associated with fundamental frequencies that are close to each other, the sounds will likely be grouped to the same stream. When the frequencies are further away, the sounds will most certainly appear in different streams.

For a demonstration of sequential grouping, consider a galloping sound that consists of two alternating tones: "A-B-A". The pattern repeats endlessly with a speed set by the researcher. The fundamental frequencies of the tones A and B can also be varied beforehand. Experimentally [2], it was discovered that when the fundamental frequencies are close to each other and/or the speed of repeating the pattern is small, then there is only one resulting stream in the perception of the sequence. On the other hand, when the frequencies are further away and/or the sequence is being repeated at a faster rate, the subjects report that they can hear two streams: one consisting of repeating A-sounds ("A---A-A---A"), and another of repeating B-sounds ("--B----B--"). The subjects also reported that they can only focus their hearing only on one of the two streams, and the other one was heard in the background. Interestingly, when the repeating rate and the difference in frequencies were set to some specific values in between, the subjects reported that the perception of two streams was alternating with the perception of one stream every 15-20 seconds. This phenomenon was experimentally verified during the research for this thesis with the help of the website made for [2].

In his book [6], Bregman described a similar experiment, but the repeating pattern was more complex. There were six tones: three lower ones (1, 2 and 3) and three higher ones (4, 5 and 6), and they were repeating in a pattern like this: "1-4-2-5-3-6". Bregman asked the subjects to report the order of the heard tones, and at faster rates of repeating they were failing to do this for both groups at the same time. When they were focusing the hearing on the lower tones, the

higher ones were heard in the background, and thus it was difficult to correctly determine the order.

As an example for simultaneous grouping, Bregman makes another experiment. This time, the pattern consists of two alternating sounds: a pure tone (A) and a complex tone that consists of two pure tones (B and C). Again, the pattern repeats endlessly, and both the speed of repetition and fundamental frequencies of the tones might be changed. As a result, Bregman reports that the tone B was *"an object of rivalry"* ([6], p.654): in cases when A was close to it in frequency, they were grouped together in a simultaneous manner, but when they were further away from each other, the tone B was rather grouped simultaneously with C, thus creating a richer tone BC.

Bregman's theory is highly related to the studies in Gestalt psychology. He was drawing parallels between vision and hearing, and found a lot of similarities between them that were supported by the Gestalt laws of grouping. He described the concepts of "belongingness" and "exclusive allocation", and the principles of similarity, proximity and closure. Also, he was questioning whether scene analysis is an innate process, or the one acquired by learning.

Finally, to make a parallel with computer modeling, Bregman referred to the notion of heuristics. In his words, heuristics are *"the procedures that are not guaranteed to solve the problem, but are likely to lead to a good solution"* ([6], p. 32). He believed that outputs from multiple heuristics should be used at the same time to find the good solution, and that there are similar mechanisms in human perception. For example, when there is evidence about common onsets and offsets of two frequency components of a sound, and it was supported by the fact that these components were different harmonics of the same fundamental frequency, the probability of being incorrect after guessing that these components should be grouped in the same stream becomes very close to zero. Here, two heuristics contributed to the decision of whether to group the sounds into a single stream.

2.3 Mathematical Background

Next, to address the practical part of the thesis, it is needed to give some attention to the underlying math. Considering that the system described in chapter 4 extensively uses techniques and concepts from the field of digital signal processing, it was decided to make a brief introduction to the basics of it for a reader that might be confused by the variety of new terms. Thus, the present section will firstly describe how sounds are represented computationally, and then will provide a few examples of the most common ones. Next, a section dedicated to digital filters and filterbanks will follow, while other specific mathematical concepts used in the implementation part will be addressed later in chapter 3.1.

2.3.1 The Basics of Digital Signal Processing

As it was described in chapter 2.1.1, physical sounds in real world spread in the environment in a form of pressure waves. These waves are continuous, thus to be able to work with sounds via computers it is usually useful to convert them to some kind of a discrete representation. The notion of a discrete, or discrete-time, signal is used in these cases and is defined as a time series sampled at equally-spaced points on the time axis, or as a function of discrete time (for example, $x(n)$) [7]. Digital signals are, in simple words, encoded representations of the discrete-time signals. The digital signal's sampling frequency f_s is defined as a number of samples observed during a unit of time. Sometimes, digital signals are represented as vectors [8]:

$$\mathbf{x} = [x(0), x(1), \ldots, x(N-1)]^T \qquad \mathbf{x} \in \mathbb{R}^N \tag{2.1}$$

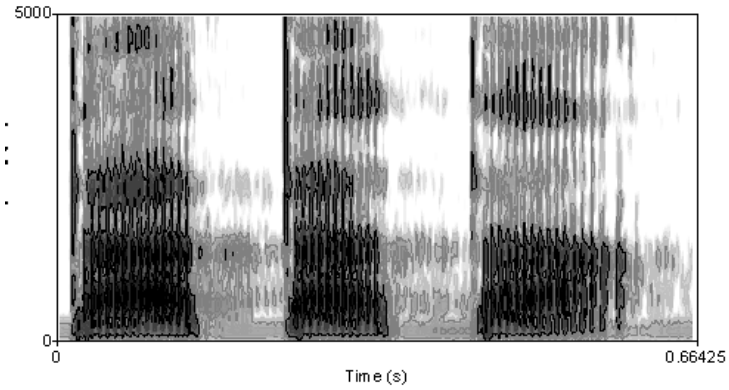


Figure 2.4 A spectrogram of a male voice saying "ta-ta-ta". Time is shown on the horizontal axis, and frequencies are shown on vertical. The color intensity increases with density. Three separate syllables are clearly visible. Taken from `https://commons.wikimedia.org/`

where N is the overall number of samples.

If you recall the conversation about resonant frequencies and the sinusoidal manner of vibrations from chapter 2.1.1, you may remember that when an impulse is delivered to an object, the object responds to it by entering into vibrations. It starts to vibrate at all possible frequencies, but not all of them survive. An important mathematical instrument for frequency analysis is Fourier transform (along with Fourier series). Fourier theorem states that any periodic function might be represented as a sum of sines and cosines, so technically, any waveform of a sound (including the ones for sound waves) might be decomposed and represented in such a way. This decomposition may also be given by an amplitude spectrum and a phase spectrum [2].

The most basic example of such frequency decomposition is for a pure tone (the first row on figure 2.5). Pure tones are impossible to find in nature, or even perfectly produce with a speaker. Pure tones produced computationally sound flat and unnatural, but they are the basic building blocks of other sounds. The waveform of a pure tone is a sine wave and is defined as a function of time t:

$$y(t) = a\sin(2\pi ft + \varphi) \tag{2.2}$$

where f is the frequency, a is the amplitude and φ is the phase. The frequency decomposition of a pure tone will contain only one peak at frequency f.

Another example of a common sound would be a click. Clicks are instant modulations in amplitude of a sound waveform, or waves that instantly go up and down at certain points in time. The most interesting fact about a click is its frequency decomposition being an infinite set of sine waves. More about clicks can be found, for example, in [2].

The last important sound that will be mentioned here is white noise. The waveform of a white noise is completely random, so its frequency decomposition is random too. White noise is used in the thesis for experiments with the resulting CASA system and its ability to separate music.

Continuing the conversation about frequency decomposition, it is necessary to note that it is impossible to extract any information about the time component from the output of the Fourier transform. Thus, it is useful to firstly split the wave into separate intervals on the time axis

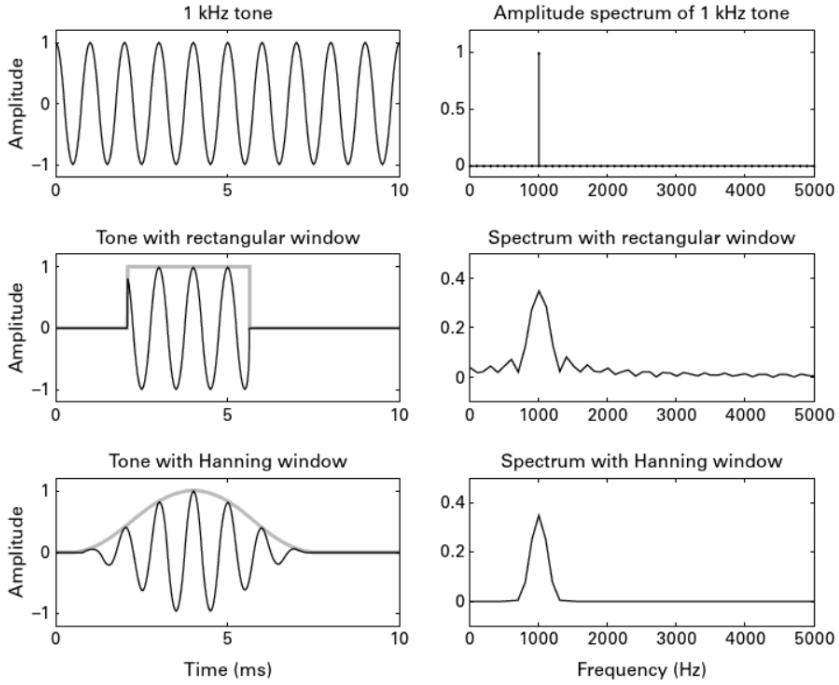


Figure 2.5 The problem of discontinuities that arises when rectangular windows are used. The first row depicts a 1 kHz pure tone (a sine wave) and its amplitude spectrum. The second row demonstrates the amplitude spectrum of the same tone masked by a rectangular window. The third row shows the spectrum of the same tone masked by a Hanning window. The window functions are shown in gray. Taken from [2].

(which are often called windows), and only then compute their frequency decompositions. The resulting time-frequency representation of the sound wave is called spectrogram. An example of a spectrogram is shown on figure 2.4.

However, there is a known problem that emerges when windowing is used with the Fourier transform. When the window is rectangular (the wave is cut off vertically from both sides) and is not aligned with the period of the signal, the onset and offset of the wave become abrupt. These sudden changes in amplitude result in the necessity of adding countless additional sine waves to the frequency decomposition, and it begins to contain chaotic information, which makes the valuable parts of the spectrum less precise. An example of this behavior was given in [2] and is shown on the second row of figure 2.5.

A viable solution of this discontinuity problem comes with attempts to smooth the abrupt ends of the masked wave. Windows that have some kind of ramping on both sides are used in this case. The ramping helps to smoothly turn the sound on and off and reduce the ”spectral splatter” [2]. An example of such window (Hanning window) is shown on the third row of figure 2.5.

It is also worth noting that when the masking window is short, the resulting amplitude spectrum becomes wider, and vice versa: when the precise frequency representation is needed, the time window must be wide enough. This property is called time-frequency trade-off and can be observed in spectrograms: the spectrograms with high frequency resolution usually have low time resolution, and the ones with high time resolution have low frequency resolution.

2.3.2 Filters and Filterbanks

In chapter 2.1.1, when there was a conversation about the objects' impulse responses, some attention was given to the selectivity of frequencies. It was said that frequencies that don't align with the object's resonance frequency are attenuated, or filtered out. Thus, the mentioned object might be thought of as a mechanical linear filter. The linearity comes from the fact that the force applied to the object is proportional to the amplitude of the output signal (in simple words, the harder you pluck the guitar string, the louder the resulting sound will be). The linear filter's impulse response is a function of time that depicts how it responds to a simple external impulse, and its frequency response is a function that shows how much different frequencies are affected after the filtering.

The filters used in the implementation part of the thesis are digital, meaning that they operate on discrete-time or digital signals by performing different mathematical operations. Their counterpart is analog filters operating on continuous-time (also called analog) signals [7]. Linear filters might be of two types: infinite impulse response (IIR) or finite impulse response (FIR). Impulse response of an IIR filter does not become equal to zero after a certain point in time, but continues infinitely, whereas impulse response of an FIR filter is given only for a certain time interval. FIR filters are usually non-recursive and less efficient, while IIR filters are recursive and computationally better.

The technique that is used for filtering of signals is called convolution. For digital signals, it is defined as follows [2]:

$$(f * g)(n) = \sum_{m=0}^{N-1} f(m)g(n-m) = \sum_{m=0}^{N-1} f(n-m)g(m) \tag{2.3}$$

where $f(n)$ is the input signal, $g(n)$ is the filter impulse response, m is the delay, or lag, and N is the overall number of samples. Convolution is commutative, thus the functions for the input signal and the filter impulse response may be swapped.

The last term for this section is a filterbank. Basically, a filterbank is a collection of filters with different properties. In the implementation part, a filterbank of gammatone filters is used to simulate the basilar membrane of the human inner ear.

2.4 Computational ASA

Now, having described the underlying concepts from different fields of science in previous sections, it is time to finally focus on computational auditory scene analysis. CASA is said to be the study that groups practical, programmable solutions for auditory scene analysis problems, or the study of ASA by computational means. CASA systems are used primarily for source separation, meaning that they are machine listening systems that aim to separate "target" sounds from mixtures, just like people do when try to focus on a specific sound and not to be distracted by others. In that CASA systems differ from systems for blind signal separation – they try to mimic (at least to some extent) the mechanisms inside the human ear, which were discussed in chapter 2.2. In this chapter, main principles of CASA systems will be described, along with a typical architecture, goals and applications. In the second part, major works that use computational auditory scene analysis for source separation will be reviewed and compared.

2.4.1 Principles, Goals and Applications

Having the definition of CASA above, to be able to limit the requirements to the models it is necessary to describe the principles of CASA and common concepts across different systems. As the most major one, one could pick the restriction of number of microphones used in the input. Being based on the mechanisms of the human auditory system, CASA models only use recordings from one or two microphones (to simulate one or two ears), thus being split to monaural and binaural. Monaural models are researched better, but can't be used for extracting features based on the location of the sound, which is possible to some extent in binaural models, when time differences between the two recordings might be used.

To discuss the goals of CASA, it is useful to refer to the goals of ASA. According to Bregman [6], the primary goal of auditory scene analysis is to produce separate streams from the auditory input. Here, the term "stream" refers to a representation of a distinct sound source in the acoustic environment (see chapter 2.2.3), but, for example in [9], the authors also use it when talking about these representations in computer memory.

For CASA, Wang and colleagues [10] proposed that the goal should be to find an ideal binary mask (IBM) for the time-frequency (T-F) representation of the input. If the input is split into T-F units, where time is on horizontal axis and frequency on vertical, the IBM is a binary matrix that has ones in places where the target sound is stronger, and zeroes elsewhere for background units. Ideal binary masks (and time-frequency masks overall) will be given some more attention in chapter 3.2.3.

The research of CASA systems and their applications in science [11] has been quite diverse recently. Some of the models are inspired by various biological experiments [12][13], while others are trying to address the cocktail party problem in natural environments [14]. Some models try to explicitly simulate perceptual data, but others may refer to perception only very slightly. The expected output for the system implemented in this thesis is to find an IBM to be able to mask noisy background in monophonic piano music.

Aside from pure scientific interest, CASA systems find useful applications in everyday life [9]. Some of them are listed below.

Speech recognition Apparently, speech recognition is the most popular field, where CASA systems have been used. Many speech recognition systems have performance losses in acoustic environments, where multiple sources of sound are present. The development is often put in contrast with computer vision systems that basically fulfill the same purpose, but for another human sense.

Automatic music transcription A complex problem on its own (even human experts can come up with different solutions) becomes more complicated when multiple musical instruments are involved and need to be transcribed separately. CASA can certainly bring new insights to the field.

Hearing prostheses Modern hearing aids made for people suffering from hearing loss don't separate speech in noisy environments, amplifying the noisy background too. CASA could address this problem to filter the noise out at least to some extent.

Audio information retrieval Recordings on the Internet usually contain mixtures of sounds from different sources, thus it is necessary to separate them to be able to search efficiently.

2.4.2 Major Works

Chapter 3
Methodology

3.1 Mathematical Concepts Used in the Thesis

This section will provide more specific information about the mathematical concepts used in the implementation part. They are listed below.

Equivalent rectangular bandwidth Equivalent rectangular bandwidth, or ERB, is a measure used for computing bandwidths of the filters for human auditory system. It was defined by Moore and Glasberg in 1983 as [15][16]:

$$ERB(f) = 6.23f^2 + 93.39f + 28.52 \tag{3.1}$$

In 1990, the authors published another approximation (linear) [17][9]:

$$ERB(f) = 24.7\,(4.37f + 1) \tag{3.2}$$

where f is frequency in kHz.

ERB-rate scale In psychoacoustics, ERB-rate scale is used to uniformly distribute the filter center frequencies based on their ERB bandwidths. This scale is similar to the critical-band scale of the human auditory system. In 1983, Moore and Glasberg defined it as follows [15]:

$$E(f) = 11.17 \ln \left| \frac{f + 0.312}{f + 14.675} \right| + 43.0 \tag{3.3}$$

Using the latest approximation of ERB by Moore and Glasberg (1990), ERB-rate scale function is approximated as [17][9]:

$$E(f) = 21.4 \log_{10}(0.00437f + 1) \tag{3.4}$$

Gammatone filter A gammatone filter is a linear filter, whose impulse response is a product of a sinusoidal tone and gamma function [9]:

$$g_{f_c}(t) = at^{L-1}\, e^{-2\pi t b(f_c)}\, \cos(2\pi f_c t + \varphi)\, u(t) \tag{3.5}$$

where a is the filter amplitude, L is its order (number of iterations of filtering), f_c is its center frequency, φ is the phase, $u(t)$ is the unit-step function ($u(t) = 1$ for $t \geq 0$, and 0 otherwise), and $b(f_c)$ is the function that determines the bandwidth for a given center frequency [9]:

$$b(f) = 1.019\, ERB(f) \tag{3.6}$$

Gammatone frequency response is defined as follows [16]:

$$G(f) = \left[1 + \frac{j\,(f - f_c)}{b(f_c)}\right]^{-L} + \left[1 + \frac{j\,(f + f_c)}{b(f_c)}\right]^{-L} \qquad (-\infty < f < \infty) \tag{3.7}$$

However, when modeling human auditory system, the second term from the definition above can be ignored for sufficiently large $\frac{f_c}{b(f_c)}$ [9][16]:

$$G(f) \approx \left[1 + \frac{j\,(f - f_c)}{b(f_c)}\right]^{-L} \qquad (0 < f < \infty) \tag{3.8}$$

Autocorrelation Autocorrelation, or autocorrelation function (ACF), is a function that is used to find periodicities and other cues in the input signal. It is defined as the correlation of the signal with its shifted copy. In this thesis the simulated auditory nerve responses will be used to compute it [9], and the result will be normalized to receive a set of Pearson's coefficients:

$$ACF(n, c, \tau) = \frac{\sum_{k=0}^{K-1} a(n-k, c)\, a(n-k-\tau, c)}{\sqrt{\sum_{\tau} a^2(n-k, c)}\sqrt{\sum_{\tau} a^2(n-k-\tau, c)}}\, h(k) \tag{3.9}$$

where $a(n, c)$ represents the simulated auditory nerve response for frequency channel c and discrete time n, τ is the time lag, K is the length of the sampling window, and $h(k)$ is the window function (usually Hanning, exponential or rectangular).

Summary autocorrelation Summary autocorrelation function, or SACF, is defined as [9][18]:

$$SACF(n, \tau) = \sum_{c} ACF(n, c, \tau) \tag{3.10}$$

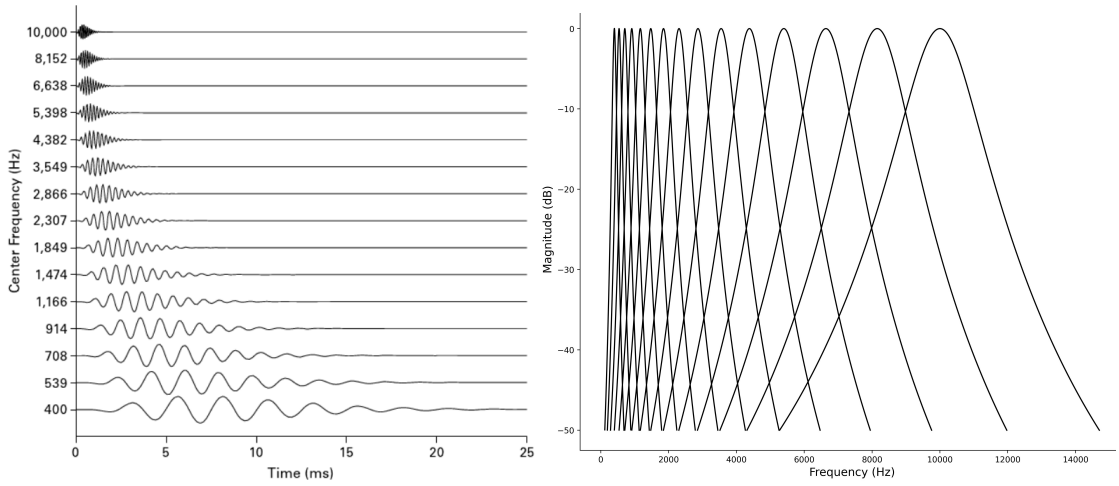
Cross-channel correlation Cross-channel correlation is a correlation between signals from different frequency channels. In this thesis it is defined for each two neighboring channels (c and $c+1$) in each time window as normalized correlation [9][18]:

$$CCCF(n, c) = \frac{\sum_{k=0}^{K-1} a(n-k, c)\, a(n-k, c+1)}{\sqrt{\sum_{k=0}^{K-1} a^2(n-k, c)}\sqrt{\sum_{k=0}^{K-1} a^2(n-k, c+1)}} h(k) \tag{3.11}$$

where $a(n, c)$ is the simulated auditory nerve response for frequency channel c and discrete time n, K is the length of the sampling window, and $h(k)$ is the window function.

3.2 Architecture of a CASA System

Next, to successfully address a concrete implemented system, it is necessary to understand its architecture. The architecture described in the following subsections is based on [9], [18] and [19], though it is impossible to say that it is used in all systems – in different sources the authors use different approaches and methods, and thus different structures of the models.



(a) **(b)**

Figure 3.1 **(a)** Impulse responses of 15 gammatone filters in a gammatone filterbank. The filters' center frequencies are equally spaced between 400 Hz and 10 kHz on the ERB-rate scale and are shown on the vertical axis. Taken from [2]. **(b)** Frequency responses of the same filters. The vertical axis depicts the changes in amplitudes in decibels. Note that the filters with higher center frequencies have wider bandwidths.

3.2.1 Peripheral Analysis

Usually, a model for computational auditory scene analysis begins with the peripheral analysis of the input sound. Here, first preparations of the input sound for further processing take place. The expected result of this stage includes a time-frequency representation of the input sound – a set of so-called T-F units. Since the models try to mimic the human cochlea — and a lot of scientific attention was given to researching it — the outcome of this stage is almost always a cochleagram.

Cochleagrams are in many cases produced by a filterbank of gammatone filters (see chapter 3.1 for a definition). Gammatone filters were picked as the most precise ones to simulate points on the basilar membrane of the inner ear. The number of filters in the filterbank might be chosen by the researcher, but the most frequent choice is 128. The center frequencies of the filters are uniformly distributed on the ERB-rate scale (see chapter 3.1 for a definition), as it was developed to be similar with the critical-band scale of the human auditory system. The filters' impulse and frequency responses are shown on figure 3.1, however it is worth noting that the filters on the figure are normalized, and in practice frequencies higher than 2 kHz are often amplified [9].

If you look at a cochleagram and a spectrogram of the same sound at the same time, you probably won't notice many differences. The cochleagram similarly shows the densities of frequencies at different points in time, though in this case it may be thought of as a set of frequency channels, where each channel is the output of a certain gammatone filter from the filterbank. The gammatone filters are meant to change the signal so that the frequencies near the center frequency are kept, and the ones that are further away are attenuated (see figure 3.1b). At this stage, the windowing techniques described in chapter 2.3.1 are also used for long signals.

3.2.2 Feature Extraction

After the cochleagram is computed, a classic CASA system involves computing a correlogram. In this thesis, correlograms will be put to the feature extraction stage, however some sources discuss them along with the cochleagrams in the peripheral analysis stage [19].

According to [9], the term correlogram was introduced by Slaney and Lyon in 1990 [20] as *"an animated picture of the sound that shows frequency content along the vertical axis and time structure along the horizontal axis"* (in this thesis the correlogram is a three-dimensional array). The authors used autocorrelation function (see chapter 3.1 for a definition) on each frequency channel to compute it and described that if the sound is periodic, then the ACF will have peaks in places corresponding to the lags that are equal to the period of repetition. This property of the autocorrelation function has been extensively used in signal processing to estimate pitch and the corresponding fundamental frequencies.

So, the fundamental frequency is estimated using the correlogram and the representations of the higher harmonics in it. If you recall the conversation about harmonics from chapter 2.1.2, you may remember that harmonics are periodic at their fundamental frequency, so if the fundamental frequency is, for example, 100 Hz (the period is 10 ms), the second harmonic is 200 Hz (5 ms), and thus repeats every 10 ms as well. The correlogram depicts this property in different frequency channels, and when the autocorrelations are summed up, the resulting summary ACF (see chapter 3.1 for a definition) will have peaks in places corresponding to the period of the fundamental frequency.

Of course, correlogram might be used not only for the f_0-estimation. In the implementation part, for example, cross-channel correlation is computed from it too. Cross-channel correlation was defined in chapter 3.1 according to [18] and may be used to find similarities between units in adjacent frequency channels.

Some authors also work with cross-correlograms at this stage. According to [9], cross-correlogram is based on the simulated auditory nerve responses from the left and right ears (thus it may be used in binaural systems) and is defined as follows:

$$CCF(n, c, \tau) = \sum_{k=0}^{K-1} a_L(n-k, c) a_R(n-k-\tau, c) h(k) \tag{3.12}$$

where $a_L(n, c)$ and $a_R(n, c)$ is the above-mentioned simulated auditory nerve response from the left and right ears respectively, τ is the lag, $h(k)$ is the window function, and K is the size of the sampling window.

3.2.3 Mid-Level Representation and Scene Organization

Next, after the feature extraction stage, follows some work related to segmentation and grouping, which according to Bregman are the main two steps of auditory scene analysis (see chapter 2.2.3). For this stage, different authors come up with different names and implementations, but the most common approach is to split it into two sub-stages as in [9] and [19]: mid-level representation and scene organization.

Mid-level representation stage includes the first step of Bregman's ASA — segmentation. Here, the T-F units computed during the peripheral analysis stage are segregated into segments with the help of the extracted features. These segments, or mid-level representations, are meant to split the cochleagram into separate continuous parts and then are grouped to form auditory

streams.

Scene organization stage, as it was mentioned, serves to group the mid-level representations into meaningful auditory objects, or streams (see chapter 2.2.3) that come from individual sound sources. Here, for example, separate harmonics are grouped together to form the richer musical sound, or the melody played by the pianist's right hand is grouped with the accompaniment played by the left hand. As a part of this stage, various techniques from the field of artificial intelligence might be used to achieve better results.

A notable outcome of the whole segmentation-and-grouping stage is a time-frequency mask of the input cochleagram (or any other chosen T-F representation). The main idea behind masking is to emphasize the T-F units corresponding to the target sound and attenuate those from the background. The mask might have either binary or real values from the $[0, 1]$ interval. If the former method is chosen, the task of searching the T-F mask might be interpreted as a binary classification problem. In case of the latter representation, the values from the mask might be interpreted differently: for example, as the signal-to-noise ratios (SNR) that show the ratios or differences between the target sound energy and overall energy in the T-F unit, or as probabilities that the T-F unit belongs to the target sound [9].

As it was mentioned earlier, Wang and colleagues [10] proposed ideal binary mask (IBM) to be the primary goal of CASA. The IBM is defined as follows [9][18]:

$$IBM(n, c) = \begin{cases} 1 & \text{if } SNR(n, c) \geq LC \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where $SNR(n, c)$ is the signal-to-noise ratio, or the difference between the target sound energy and the interference energy in the T-F unit, and LC is the local criterion chosen as a threshold for the SNR function (most commonly 0 dB).

3.2.4 Resynthesis

Finally, some systems include the resynthesis stage to convert the masked T-F representation back to the sound waveform. This is useful for evaluation of the resulting CASA system in cases when listening experiments are held, or signal-to-noise ratios need to be compared before and after the processing [9]. The technique for resynthesis was proposed by Weintraub in his PhD thesis in 1985 [21]: the outputs from each gammatone filter in the filterbank were reversed in time to remove the phase shifts, then passed backwards through each filter and reversed back in time again.

Chapter 4
Implementation

The implementation chapter will provide specific information about the CASA system implemented for the thesis. It will give more attention to the algorithms used in different parts of the architecture and their input parameters. The main goal of the system was to separate monophonic piano music from the background noise by finding an ideal binary mask for the cochleagram. The follow-up experiments will be described later in chapter 5.

It should also be noted, that the implemented system is rather simple, thus it should not be expected to observe source separation of too high quality. The implemented model serves only as an example of a CASA system and is not aiming to separate the sources using the most modern and sophisticated algorithms and approaches.

The system is implemented in Python with the help of *numpy* [22], *scipy* [23], *statsmodels* [24], *matplotlib* [25] and *brian2hears* [26] packages. *skimage* [27] and *brian2* [28] were used as supporting ones. For evaluation, a simple multi-class classifier was implemented with the help of the *scikit-learn* [29]. The main Python scripts contain functions for different parts of the resulting architecture, and then an example of their usage along with the experiments overview is given in the supporting Jupyter notebooks (see the structure of the attached medium for more information).

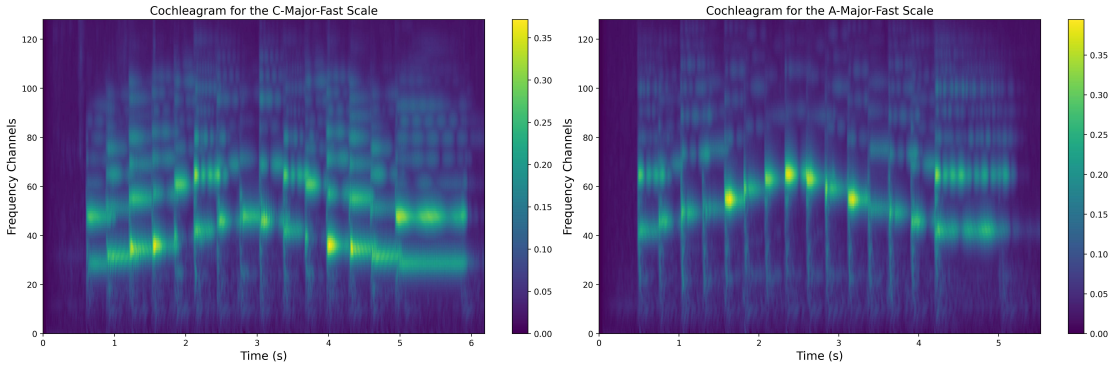
4.1 Cochleagram

The cochleagram described in chapter 3.2.1 was implemented with the help of *brian2hears* [26] package. At the beginning, an array of center frequencies was computed using the ERB-rate scale defined in chapter 3.1. In the main example notebook, there were 128 center frequencies uniformly distributed on it between the values of the lowest and the highest fundamental frequencies on the standard piano keyboard (containing 88 keys) – from 27.5 Hz (note A_0) to 4.186 kHz (note C_8).

As a next step, a gammatone filterbank was used to split the input into 128 corresponding frequency channels. The filters were implemented as cascades of four IIR filters of order 2 (which corresponds to single gammatone filters of order 4 [26]). The approximate impulse response was similar to the one defined in equation 3.5:

$$g_{f_c}(t) = t^3\, e^{-2\pi b ERB(f_c)t}\, \cos(2\pi f_c t) \tag{4.1}$$

where $b = 1.019$ and the ERB function was the same as in equation 3.2 for frequency in Hertz. Finally, a unit-step function was applied to the output along with a cubic root function that



(a) **(b)**

Figure 4.1 Cochleagrams for C-major **(a)** and A-major **(b)** scales. The ascending and descending note progressions and the note harmonics are clearly visible in both cases, as well as the difference in fundamental frequency between the tones (all notes from A-major scale are higher).

helped to better emphasize low amplitudes in the cochleagram. The resulting cochleagrams for C-major and A-major scales are shown on figure 4.1.

After the cochleagram was computed, is was needed to split the output into windows for further processing. A rectangular window of size 20 ms was used as a default with an overlap of 10 ms.

4.2 Correlogram and Other Features

For the feature extraction stage described in chapter 3.2.2, a correlogram was implemented using the autocorrelation function implemented in the *statsmodels* package. The provided implementation could compute the ACF similarly as defined in equation 3.9, however its another variant that uses Fast Fourier Transform was used for higher efficiency. The default number of lags for the autocorrelation function was chosen to be equal to the number of samples in the sampling window, i. e. 20 ms times the samplerate of the input sound (48 kHz).

Thus, the resulting correlogram was a three-dimensional array of floats in $[-1, 1]$ range. The first dimension was time frames, the second was frequency channels and the third was lags for the autocorrelation function.

Next, a summary autocorrelation function was computed to help with extracting the fundamental frequencies for separate time frames. The formula was exactly the same as in equation 3.10. Also, for demonstration purposes, cross-channel correlation was computed as defined in equation 3.11.

Finally, the SACF function was used to estimate the fundamental frequencies of the signals in each time frame. For this step, the ”dominant” lags were firstly found, meaning equally spaced lags with the highest sum of SACF values, and then the fundamental frequency was estimated for the current time frame using the distance between two adjacent lags. The default number of ”dominant” lags was initially chosen to be 5. The resulting correlogram for time frame 150 is shown on figure 4.2 along with all extracted features.

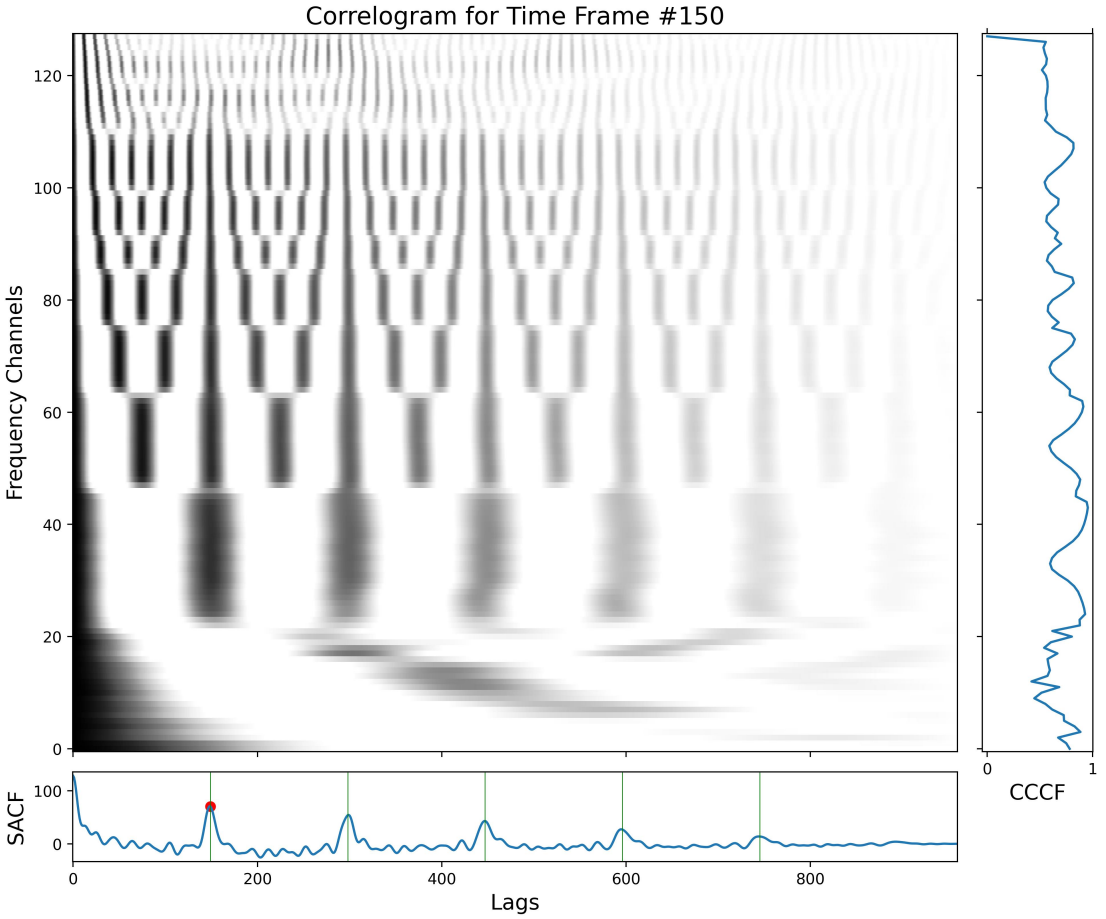


Figure 4.2 An example of a correlogram for C-major scale for time frame 150. Note that distinct frequency channels have similar repeating patterns corresponding to the repetitions observed in harmonics. The pattern starts around the frequency channel 40 and repeats twice as fast around channel 60, then three times as fast around channel 70, and so on. Cross-channel correlation is shown on the left panel, and the summary autocorrelation is shown on the bottom panel. Note the peaks on the plot for SACF that emerge when all harmonics become "synchronized" (emphasized with green lines). The lag corresponding to the fundamental frequency is marked as a red dot.

4.3 Masking

The segmentation-and-grouping stage involved the task of estimating the ideal binary mask for the cochleagram computed in the first part. This step was done by combining two matrices described below.

The first matrix playing a big role in the resulting IBM was an "energy mask". This mask helped to emphasize the regions of the cochleagram that contained high sound energy by comparing the mean value of the samples in a T-F unit with a threshold. As a result, silent regions of the cochleagram — like the ones that usually appear at the beginning of a recording — were addressed and considered as ones not associated with the target sounds (the resulting binary energy mask contained zeroes for such T-F units). The default value for the threshold was chosen to be 0.05.

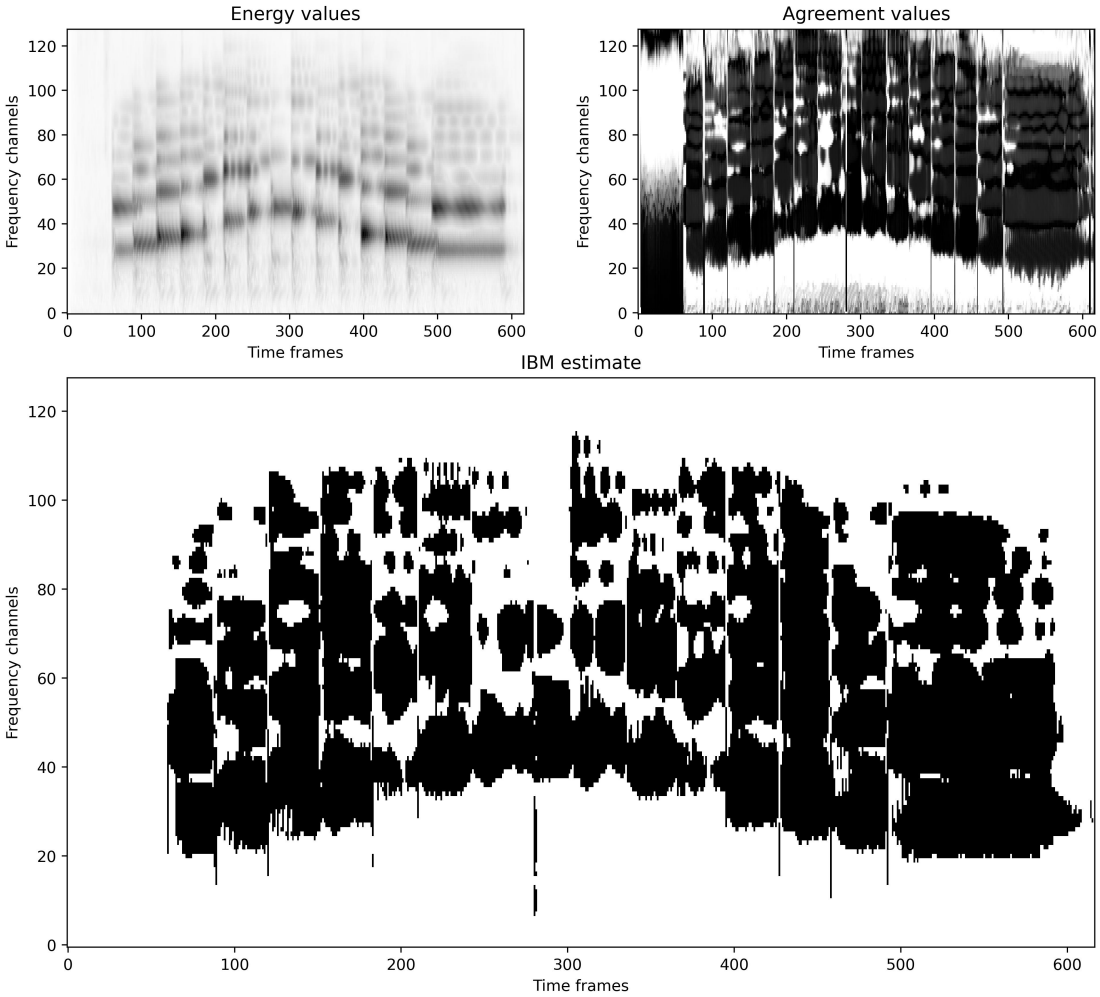


Figure 4.3 An example of an ideal binary mask for C-major scale. The mask was estimated from combining two other masks computed by comparing sound energy and agreement values with thresholds.

The second matrix was an ”agreement values mask”. This one was extracted from the correlogram by taking values at the estimated ”dominant” lags. These measures technically showed which frequency channels contained the harmonics of the fundamental frequency, and which did not. The values were also normalized by the maximum value of the autocorrelation for the T-F unit (which was equal to 1 in most cases). To make a binary decision about the measures of agreement, a threshold was used, and it was set to 0.7 by default.

Finally, an ideal binary mask was estimated by combining the two above-mentioned masks using the logical ”and” operation. In the end, the values in the resulting IBM were set to 1 for T-F units that had mean sound energy higher than the threshold and were in agreement with the estimated fundamental frequency. An example of energy and agreement matrices and the corresponding IBM for C-major scale is shown on figure 4.3.

The next step was to apply the estimated mask to a cochleagram. This was done by rebuilding the cochleagram from its windowed representation and multiplying the samples in the windows by the values from the mask. The result for C-major scale is shown on figure 4.4.

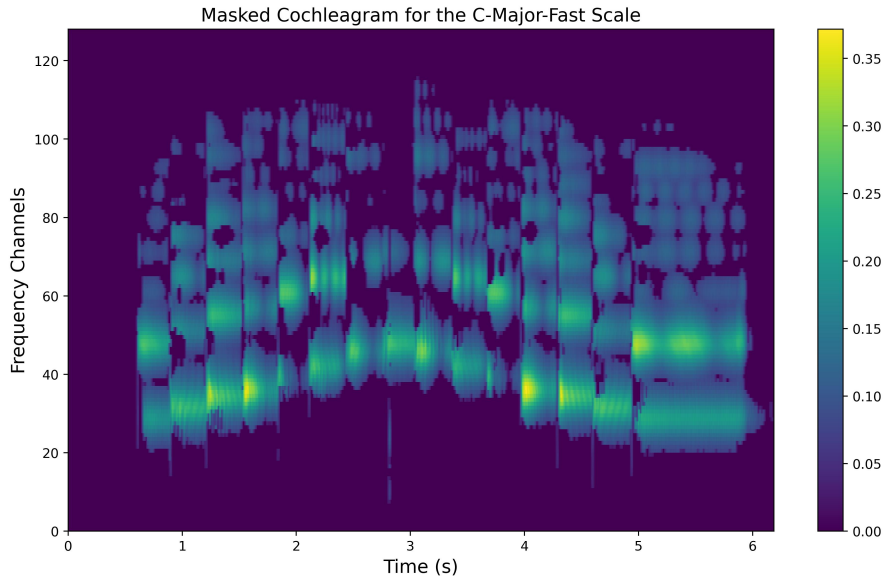


Figure 4.4 The cochleagram from figure 4.1a masked by the IBM shown on figure 4.3. The main energy regions corresponding to distinct notes were kept, while the background sounds were filtered out. The ascending and descending note progressions as well as higher harmonics of the notes are also clearly visible after the masking.

As it can be seen from the approach chosen for estimating the IBM, Bregman's references to ASA as to a two-stage process were not addressed. The segmentation-and-grouping stage was implemented by working with all T-F units at once, without firstly segregating them into segments and then grouping the segments together based on the grouping cues. As a part of future work it is planned to give this stage more scientific attention and hopefully receive better outcomes.

Chapter 5

Experiments and Results

5.1 Dataset Overview

The set of input data for testing of the implemented system consisted of 34 piano recordings of various scales and intervals. The files were converted to WAV file format to address the limitations of the used libraries. Below is a more detailed overview:

- A-major scale played fast and slow in ascending and descending order, starting from A_3
- A-minor scale played fast and slow in ascending and descending order in three modes: natural, harmonic and melodic, starting from A_3
- C-major scale played fast and slow in ascending and descending order, starting from C_4
- C-major scale, where each note is repeated 3 times, in ascending order in four variants: starting from C_3, C_4, C_5 and C_6
- D-major scale played fast and slow in ascending and descending order, starting from D_4
- E-major scale played fast and slow in ascending and descending order, starting from E_4
- E-minor scale played fast and slow in ascending and descending order in three modes: natural, harmonic and melodic, starting from E_4
- F-major scale played fast and slow in ascending and descending order, starting from F_4
- G-major scale played fast and slow in ascending and descending order, starting from G_4
- H-major scale played fast and slow in ascending and descending order, starting from H_4
- Perfect melodic fourths, where the lower notes are in range starting from C_3 and ending at C_6, in ascending order
- Perfect melodic octaves, where the lower notes are in range starting from C_3 and ending at F_5, in ascending order
- All semitones (or all keys one after another), starting from C_4 and ending at H_5, in ascending order
- All semitones (or all keys one after another), starting from C_2 and ending at H_3, in ascending order

5.2 Experiments with White Noise

The first set of experiments involved experiments with white noise levels. The white noise level was an input parameter of the CASA system, and was used before the peripheral analysis stage as an amplitude of white noise added to the target signal (this noise was considered the background). The example set of values was as follows: {0; 0.005; 0.01; 0.02; 0.04}, and the results for it are shown on figure 5.1.

On the figure, the first row depicts the outcomes of processing a clean target signal (A-major scale played fast), and each next row shows the results for the added white noise levels in ascending order. When comparing the cochleagrams, the rising amounts of background interference can be clearly seen (the dark-blue color in the first cochleagram becomes brighter with each iteration). From the resulting masking, it can be observed that the model began to lose the quality of source separation when the noise level was around 0.02, which is clearly a major amount of randomness in the input, and can be well heard by a human listener. Given the simplicity of the algorithm for the segmentation-and-grouping stage, the outcomes can be considered a success.

As an improvement of the resulting source separation, the masks computed for clean target signals were used on cochleagrams for signals with noise. This method was inspired by the fact that human brain tends to remember sounds heard before (the word ”sound” is used here in a more psychological sense), and then use this knowledge to analyze a complex auditory scene more easily. This memory-based approach, however, would not have been possible, if target sounds had not been known beforehand.

5.3 Experiments with Other Backgrounds

Besides the white noise, other sounds were used in the background too. Among those were various rattling, clanging, ringing or grinding sounds, clatter of kitchen utensils, sounds of rustling paper or a plastic candy wrapper, ticks of a clock, clicks of a computer mouse, and so on. Mainly, these sounds were not harmonic, and thus did not evoke a perception of pitch. All of them may be found on the attached medium.

For the demonstration of these experiments, the noise level parameter was chosen dynamically to better adapt to varying input sound levels and nature of the sounds. Otherwise, the experiments were similar to the ones for white noise levels, and are shown on figure 5.2. On the figure, each row represents results of processing of G-major scale played fast with different background sounds: the first and the second are rustlings of paper and a plastic candy wrapper respectively, the third one is clatter of kitchen utensils, the fourth is a sound of a bending metal plate, and the last one is a sound of a ticking clock.

The most interesting results can be seen in the fourth and fifth rows. For the bending metal plate case, the background interference can be well seen on the cochleagram in lower frequency channels. This type of background may be considered harmonic (at least to some extent), and thus some kind of polyphony was created, where the metal plate sounds are heard simultaneously with the piano sounds. Being built to separate only monophonic inputs, the system was not able to process this mixture well enough.

The fifth row, in turn, is interesting due to the noise level argument for the clock ticks in the background. In this case, the amplitude was multiplied by a factor of 15.0 (!), however the outcomes don’t seem too much affected. This may be explained by the nature of the clock ticking sounds – they are not harmonic and don’t have pitch. All the mixtures for these experiments for

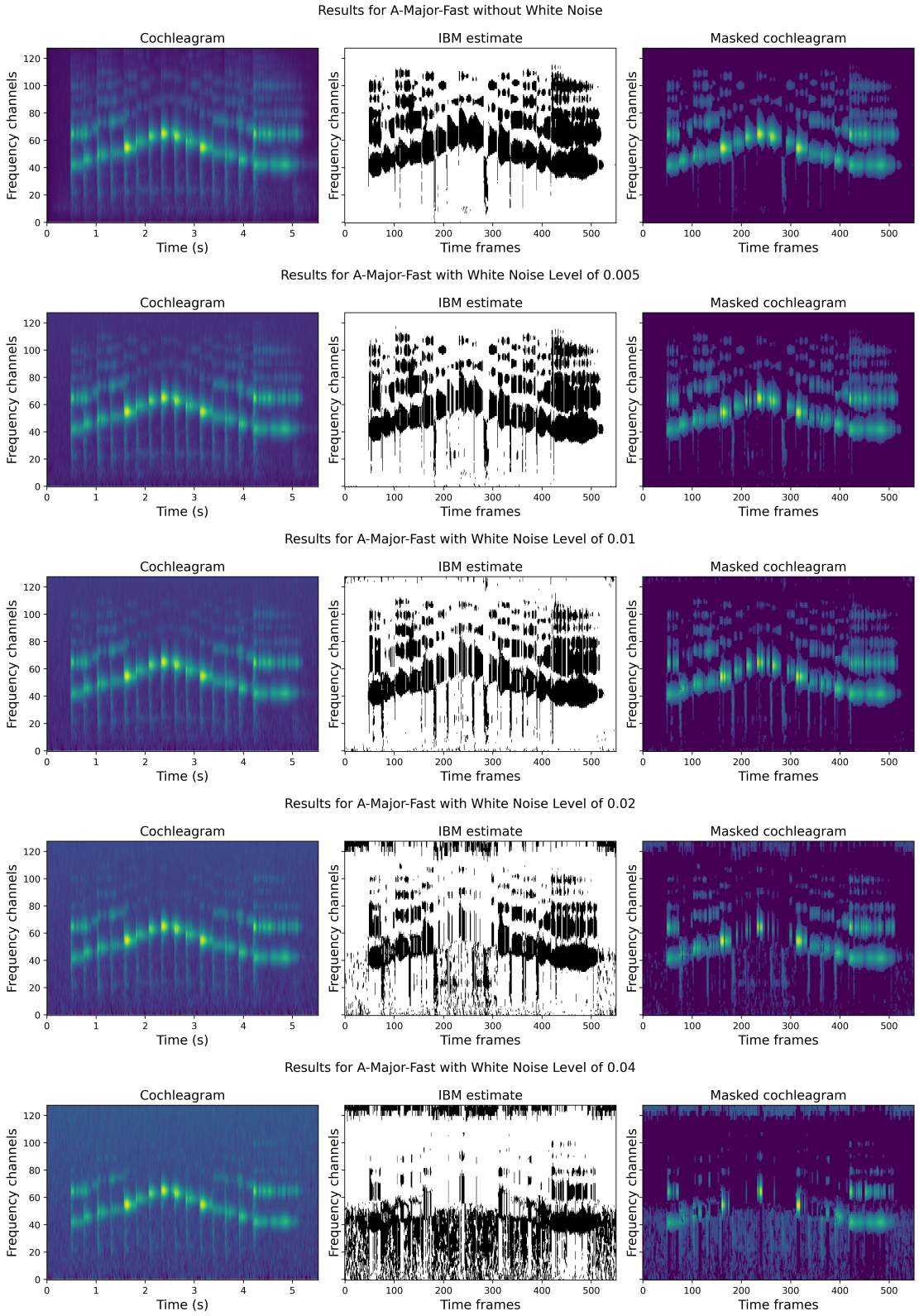


Figure 5.1 Results of the experiments with white noise levels for A-major scale. In each row, the first image is a cochleagram of the input sound, the second is the estimated IBM, and the third is the masked cochleagram.

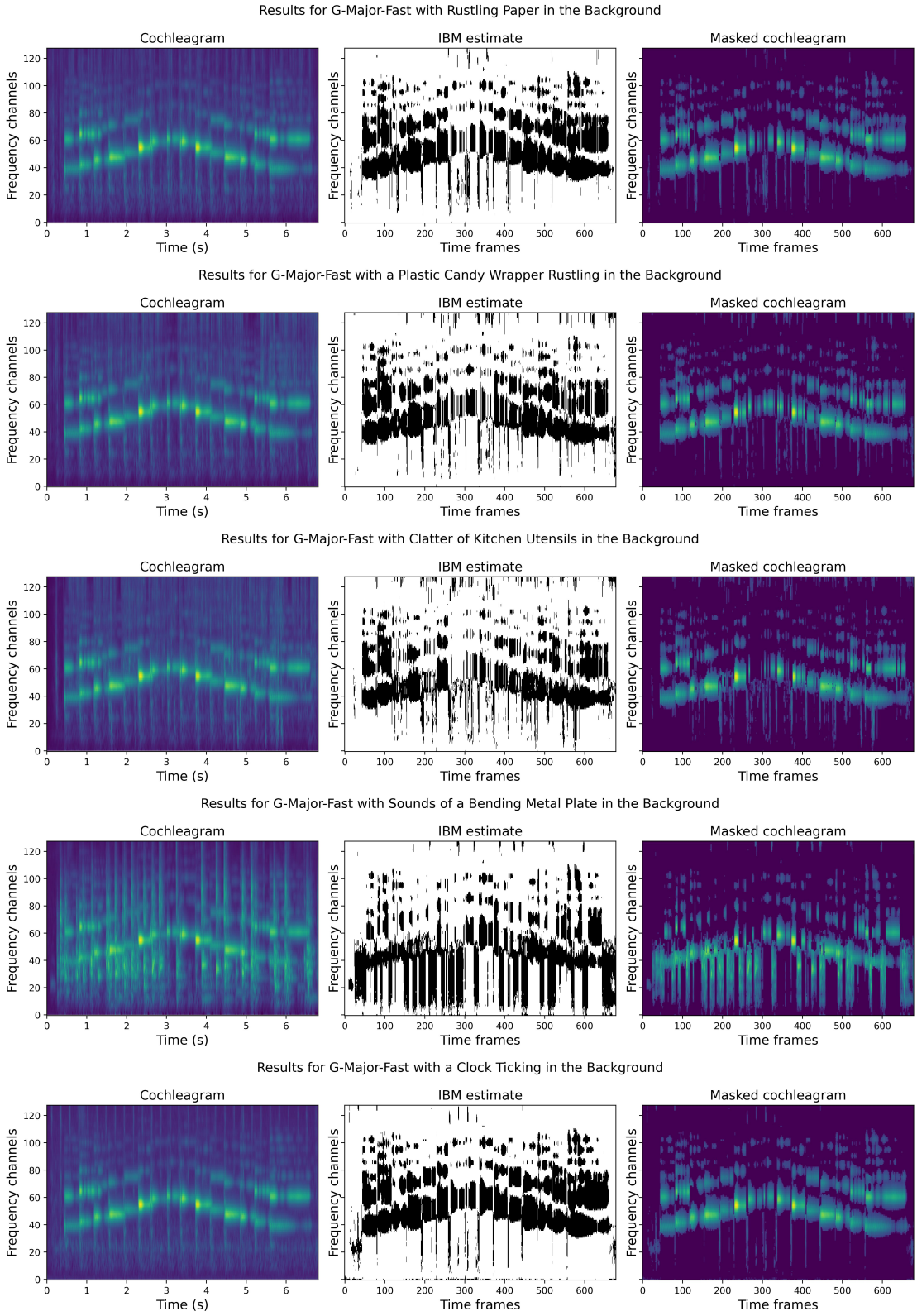


Figure 5.2 Results of the experiments with other background sounds for G-major scale. In each row, the first image is a cochleagram of the input sound, the second is the estimated IBM, and the third is the masked cochleagram.

different noise level values may be found on attached medium.

5.4 Experiments with a Simple Classifier

The next set of experiments was meant to test the implemented CASA system in connection with a simple classifier. The main thought behind it was to practically prove that CASA processing brings non-zero improvements to the model's accuracy scores, when the predictions are made for noisy data. For this purpose, the decision tree classifier implementation was chosen from the *scikit-learn* [29] package, because decision trees are easier to understand and don't require additional data preparations.

The model was trained on two experimental sets of data computed from the datasets described in the chapter 5.1. Both sets consisted of masked cochleagrams, among which were those for clean input signals, signals with random amounts of added white noise, and signals for sounds with other artificially added backgrounds. The labels for the samples were numerical representations of their corresponding input sounds. The numbers of samples in each group were also chosen experimentally, being one sample per each clean input signal, two samples per an input signal with random white noise levels, and two samples per an input signal with other random backgrounds with random amplitudes in the $[0; 1]$ range. The argumentation behind this ratio is simple – the model was going to make predictions primarily for sounds with noisy backgrounds, thus it was trained on high amounts of these, but not left without an ability to recognize the clean ones.

Before the training, 40 % of the samples were reserved for model validation, because the hyperparameters for the model were chosen in a more simple manner (iteratively). The implementation, of course, saved space for improvement, as well as it allowed to manually override the hyperparameters.

Finally, using the best hyperparameters, the model was trained and tested. The tests were using sets of 100 samples produced from random input sounds with random backgrounds of random amplitudes. The results are more or less expected: in the case when the samples were produced from the unmasked cochleagrams, the accuracy scores were around 10–20 %, and 70–80 % for the masked ones.

5.5 Other Experiments

As a possibility for deeper research, the implemented system allows to alter other input parameters of its algorithms. For example, different numbers of filters in the gammatone filterbank, their minimum and maximum center frequencies, or parameters of sampling windows might be tested at the peripheral analysis stage; number of autocorrelation lags or harmonics being searched in the summary autocorrelation might be varied at the feature extraction stage; or different numerical thresholds might be chosen for the segmentation-and-grouping stage. The attached notebook with experiments provides examples of experimenting with the number of searched harmonics. There, some more limitations of the implemented system might be seen.

Chapter 6

Conclusion

The main goal of this thesis was to describe computational auditory scene analysis. For this task, theoretical background was provided in chapter 2, which hopefully helped the reader to dive into the topic and get used to the main concepts and ideas. There, sounds were firstly discussed from the physical point of view, and particular attention was given to harmonic sounds and the perception of pitch. Then, biological theory followed, including an overview of the structures in the human ear and an introduction to Bregman's ASA. Next followed a section dedicated to the basics of digital signal processing, along with a brief conversation about digital filters and filterbanks. Finally, chapter 2 also described the principles, goals and applications of CASA, and reviewed major works in the field.

Chapter 3 gathered and described some related mathematical concepts, such as ERB-rate scale, gammatone filter and autocorrelation function. It also included an overview of the architecture of a typical CASA system, and its main stages: peripheral analysis, feature extraction, mid-level representation, scene organization and resynthesis. There, the problem of finding an ideal binary mask for the cochleagram was referred to as the main objective of CASA.

Chapter 4 addressed the next objective, and as a result, a simple CASA system was implemented to separate monophonic piano music from background noise. The main focus in this chapter was given to describing the used algorithms and their input parameters. In the end, the system appeared to give appropriate results, although the techniques used at some stages were quite primitive.

Chapter 5 was the final one and provided an overview of the experiments for the implemented system. The dataset for them contained a variety of piano recordings, among which were different scales and intervals. They were processed either separately, or in mixtures with white noise or other backgrounds. A memory-based approach was tested as well, giving considerably better results if compared with straight-forward attempts for source separation.

Overall, the objectives set for the thesis were successfully achieved. For the author, this thesis became a big inspiration to continue studying digital signal processing, psychoacoustics and music. The final section will be dedicated to an overview of what can be done next.

6.1 Future Work

The field of computational auditory scene analysis is relatively new and, of course, requires more scientific attention. In the author's opinion, the main reason why it is not yet extensively researched and did not catch every scientist's eye is that it requires deep knowledge in several fields that are not usually taught in parallel. This thesis provided a decent proof of this: to begin talking about CASA it was appropriate to introduce the reader into the underlying physics and biology, as well as to digital signal processing, which is often given separate university courses. Thus, the possibilities for improvements come from different fields of science, but could bring valuable solutions to all of them at once.

The first improvement the author sees for the implemented system is the one for the segmentation-and-grouping stage. In practice, different authors approach it differently and bring solutions that could have not many things in common, but almost all of them refer to the Bregman's definition of ASA. Thus, for a CASA system it is not very natural to work with all time-frequency units at once and compute the resulting mask by exploiting their common features. More sophisticated techniques are usually employed in this case, including, for example, machine learning algorithms for the grouping stage.

Another limitation was that the output masks were binary, and thus background sound energy observed in unmasked T-F units was not attenuated at all. A possible improvement in this case would be to work with real-valued masks that were briefly discussed in chapter 3.2.3.

The next task may be to remove the word "monophonic" from the name of the thesis. This will include research of the algorithms for multiple f_0 estimation and solving the related problems, one of the main ones being, for example, how to estimate the fundamental frequency of a note that is the same as one of the harmonics of the other note. A separate chapter in [9] is dedicated for this task and may be taken as a starting point.

Of course, the above-mentioned polyphony may be applied differently. Some may refer to it in terms of a single musical instrument, when two or more notes are played simultaneously, while others may think of it as played by multiple instruments at once. Each of these cases brings new challenges to the computational models for the cocktail party problem, but they are certainly worth the work to be done.

Another improvement might address binaural recordings. In fact, binaural sounds enable the possibility of applying sound localization techniques, which employ the notions of interaural time and intensity differences. The approaches to sound localization might involve computing a cross-correlogram, which was given some attention in the thesis. Overall, the topic is also quite interesting and challenging, and may bring new solutions to the feature extraction stage. The chapter dedicated to binaural sound localization from [9] might be studied for further inspiration.

Finally, if one returns back to the basic fact that CASA systems try to simulate the human ear, further improvements can be made in the "amount" of this simulation. This thesis was mimicking the activity of the basilar membrane of the inner ear by using a cochleagram, however many authors improve the outcomes by involving models that simulate neural activity of the hair cells. A notable example of such model might be found under the name "Meddis hair cell".

Bibliography

1. PASNAU, Robert. What Is Sound? *The Philosophical quarterly*. 1999, vol. 49, no. 196, pp. 309–324. ISSN 0031-8094.
2. SCHNUPP, Jan; NELKEN, Israel; KING, Andrew. *Auditory Neuroscience: Making Sense of Sound*. Cambridge, Mass: MIT Press, 2011. ISBN 026228975X, ISBN 9780262518024, ISBN 9780262289757, ISBN 0262518023, ISBN 026211318X, ISBN 9780262113182.
3. PLACK, Christopher J.; OXENHAM, Andrew J.; FAY, Richard R.; POPPER, Arthur N. *Pitch: Neural Coding and Perception*. New York, NY: Springer New York, 2005. ISBN 0387234721, ISBN 9780387234724.
4. STANDRING, S.; BORLEY, N.R. *Gray's Anatomy: The Anatomical Basis of Clinical Practice*. Churchill Livingstone/Elsevier, 2008. ClinicalKey 2012. ISBN 9780443066849.
5. HUDSPETH, A.J. Making an Effort to Listen: Mechanical Amplification in the Ear. *Neuron*. 2008, vol. 59, pp. 530–45. Available from DOI: 10.1016/j.neuron.2008.07.012.
6. BREGMAN, Albert S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford Books, 1990. A Bradford book. ISBN 9780262022972, ISBN 9780262521956, ISBN 0262022974, ISBN 0262521954.
7. SHENOI, B. A. *Introduction to Digital Signal Processing and Filter Design*. 1. Aufl. Chichester: Wiley-Interscience, 2005. ISBN 0471464821, ISBN 9780471464822, ISBN 9780471656388, ISBN 0471656380.
8. ABOOD, Samir I. *Digital Signal Processing: A Primer with MATLAB*. 1st ed. London; New York; Boca Raton: CRC Press/Taylor & Francis Group, 2020. ISBN 9781000765571, ISBN 1000765571, ISBN 9780367444938, ISBN 0367444933.
9. WANG, DeLiang; BROWN, Guy J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006. ISBN 9780471741091, ISBN 0471741094.
10. WANG, DeLiang. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. In: Boston, MA: Springer US, 2005, pp. 181–197. Speech Separation by Humans and Machines. ISBN 1402080018, ISBN 9781402080012.
11. SZABÓ, Beáta T.; DENHAM, Susan L.; WINKLER, István. Computational Models of Auditory Scene Analysis: A Review. *Frontiers in Neuroscience*. 2016. ISSN 1662-4548.
12. WANG, Deliang; CHANG, Peter S. An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*. 2008, vol. 2, pp. 7–19.
13. BOES, Michiel; DE COENSEL, Bert; OLDONI, Damiano; BOTTELDOOREN, Dick. A biologically inspired model adding binaural aspects to soundscape analysis. In: Institute of Noise Control Engineering Japan, 2011.

14. ELHILALI, Mounya; SHAMMA, Shihab A. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*. 2008, vol. 124 6, pp. 3751–71.
15. MOORE, Brian C. J.; GLASBERG, Brian R. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*. 1983, vol. 74, no. 3, pp. 750–753. ISSN 0001-4966. Available from DOI: 10.1121/1.389861.
16. HOLDSWORTH, John; NIMMO-SMITH, Ian; PATTERSON, Roy; RICE, Peter. *Implementing a GammaTone Filter Bank*. 1988. Tech. rep. Cambridge Electronic Design; MRC Applied Psychology Unit.
17. MOORE, Brian C.J.; GLASBERG, Brian R. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*. 1990, vol. 47, no. 1, pp. 103–138. ISSN 0378-5955. Available from DOI: 10.1016/0378-5955(90)90170-T.
18. WANG, DeLiang; NARAYANAN, Arun. Computational Auditory Scene Analysis and Automatic Speech Recognition. In: VIRTANEN, Tuomas; RAJ, Bhiksha; SINGH, Rita (eds.). *Techniques for Noise Robustness in Automatic Speech Recognition*. 1. Aufl.;1; New York: Wiley, 2012, chap. 16, pp. 433–462. ISBN 9781119970880, ISBN 1119970881, ISBN 9781118392669, ISBN 1118392663.
19. JASTI, Venkata. *Computational Auditory Scene Analysis (CASA) Technique Inspired by Human Auditory System (HAS) - A Review*. 2020. Available from DOI: 10.13140/RG.2.2.18406.45125.
20. SLANEY, M.; LYON, R. F. A perceptual pitch detector. In: IEEE, 1990, vol. 1, pp. 357–360. ISSN 1520-6149.
21. WEINTRAUB, Mitchel. *A Theory and Computational Model of Auditory Monaural Sound Separation*. 1985. PhD thesis. Department of Electrical Engineering and the Commitee on Graduate Studies, Stanford University.
22. HARRIS, Charles R.; MILLMAN, K. Jarrod; VAN DER WALT, Stéfan J. et al. Array programming with NumPy. *Nature*. 2020, vol. 585, no. 7825, pp. 357–362. Available from DOI: 10.1038/s41586-020-2649-2.
23. VIRTANEN, Pauli; GOMMERS, Ralf; OLIPHANT, Travis E. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 2020, vol. 17, pp. 261–272. Available from DOI: 10.1038/s41592-019-0686-2.
24. SEABOLD, Skipper; PERKTOLD, Josef. statsmodels: Econometric and statistical modeling with python. In: *9th Python in Science Conference*. 2010.
25. HUNTER, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 2007, vol. 9, no. 3, pp. 90–95. Available from DOI: 10.1109/MCSE.2007.55.
26. STIMBERG, Marcel; GOODMAN, Dan FM. *Brian 2 hears*. 2019. Available also from: https://brian2hears.readthedocs.io/en/stable/.
27. VAN DER WALT, Stefan; SCHÖNBERGER, Johannes L.; NUNEZ-IGLESIAS, Juan et al. scikit-image: image processing in Python. *PeerJ*. 2014, vol. 2, e453.
28. STIMBERG, Marcel; BRETTE, Romain; GOODMAN, Dan FM. Brian 2, an Intuitive and Efficient Neural Simulator. *eLife*. 2019, vol. 8, no. e47314. Available from DOI: 10.7554/eLife.47314.
29. PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011, vol. 12, pp. 2825–2830.