UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The investigation of 3D space, frequency, and rate contributions to solve the cocktail-party problem using signal detection theory

Permalink

https://escholarship.org/uc/item/5p04s5h9

Author

Afghah, Tahereh

Publication Date

2020

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

The investigation of 3D space, frequency, and rate contributions to solve the cocktail-party problem using signal detection theory

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Music

by

Tahereh Afghah

Committee in charge:

Professor Miller Puckette, Chair
Professor Sarah Creel
Professor Sarah Hankins
Professor Molly Henry
Professor Tamara Smyth
Professor Shahrokh Yadegari

2020

Copyright

Tahereh Afghah, 2020

All rights reserved.

The Dissertation of Tahereh Afghah is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

3D Three-Dimensional

CD Coefficient of Determination

FAR False Alarms Rate

FC$_D$ Difference between the average Carrier Frequencies of target and masker

FC$_M$ average Carrier Frequencies of Masker

FC$_T$ average Carrier Frequencies of Target

FMOD$_D$ Difference between the Modulating Frequency of target and masker

FMOD$_M$ Modulating Frequency of Masker

FMOD$_T$ Modulating Frequency of Target

HR Hit Rate

HRTF Head Related Transfer Function

ILD Interaural Level Differences

ITD Interaural Time Difference

QC Quadratic Coefficient

RT Reaction Time

SDT Signal Detection Theory

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Prof. Miller Puckette, I would like to thank you first and foremost for everything that you have done for me. You were always benevolent, compassionate, and above all inspiring. Your ever so cheerful tone never failed to motivate me throughout my Ph.D.

Prof. Molly Henry, words do little justice in acknowledging your role in bringing this thesis to fruition. Thanks for being so kind and thanks for all your help and support. Your guidance and supervision in the short time frame was of immense help.

Prof. Tamara Smyth, thanks for your support and giving me the opportunity to be your assistant. I have learned a lot from you.

Prof. Shahrokh Yadegari, thanks for your support and very helpful advice throughout my study. Your invaluable input always stood me in good stead.

Prof. Sarah Hankins, thanks ever so for showing me commercial applications and opening fresh prospects for my study.

Prof. Sarah Creel, thanks for your very helpful comments and profound insight which served well in my analyses.

Prof. Peter Otto, simply thanks for changing my life. Thanks for trusting in me and giving me the opportunity to grow as a researcher.

Elliot Patros, thanks for four years of productive collaboration. I enjoyed every second of that and learned a lot from you.

And I would like to thank my parents and siblings for being there for me throughout.

VITA

2007-2012 Bachelor of Engineering, Electrical Engineering, Dr.Shariaty Technical College

2013-2015 Master of Science, Media Technology-Audio Engineering, Technical University of Ilmenau,Germany

2019-2020 Visiting Scholar, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

2016-2020 Doctor of Philosophy, Music, University of California San Diego

PUBLICATIONS

Miller Puckette, Elliot Patros, Tahereh Afghah."Acoustic beamform-ing technique for the Pressure-Matching Beamforming Method.",Publication Date:20.06.2019, Pub. No.: WO/2019/118521 International Application No.:PCT/US2018/065047

Elliot Patros, Tahereh Afghah, and Peter Otto."Reproduction and analysis of near and far stereophonic sound fields with compact uniform linear arrays", The Journal of the Acoustical Society of America146(4):3058-3058, October 2019

Afghah, Tahereh, Elliot Patros, and Miller Puckette. "The physical evaluation of the efficiency of an enhanced pressure-matching beamforming method using eigen decomposition pseudoinverse mathematical approach.", The 2019 Audio Engineering Society International Conference On Immersive And Interactive Audio, York, UK, March 2019

Afghah, Tahereh, Elliot Patros, and Miller Puckette. "A Pseudoinverse Technique for the Pressure-Matching Beamforming Method." Audio Engineering Society Convention 145. Audio Engineering Society, 2018.

Afghah, Tahereh, Andrew Allen, Peter Otto, and Aravindan Joseph Benjamin. "The Evaluation of the Effect of Sound Directionality in Horizontal Plane on the Human Auditory Distance Perception in a Large Reverberant Room." In Audio Engineering Society Convention 142. Audio Engineering Society, 2017.

"The evaluation of auditory distance perception after the continuous change of distance", Master's thesis, University of Ilmenau, Germany, 2015

ABSTRACT OF THE DISSERTATION

The investigation of 3D space, frequency, and rate contributions to solve the cocktail-party problem using signal detection theory

by

Tahereh Afghah

Doctor of Philosophy in Music

University of California San Diego, 2020

Professor Miller Puckette, Chair

The cocktail party effect describes the human ability to detect a specific sound of interest in a noisy environment. Scientists have evaluated different influential parameters affecting auditory stream segregation. Previous studies have focused on individual cues while the interaction between them appears to be poorly studied. This thesis investigates the interactions between the temporal and spatial cues, and the spectral and spatial cues in a cocktail party scenario. Two streams were presented at the same time from $\pm 7.5^\circ$ or $\pm 45^\circ$ or $\pm 90^\circ$. One stream (the target) was to be attended, and the other stream (the

masker) was to be ignored. The stimuli were frequency-modulated narrowband noise. The modulation-rates were randomly chosen between 0.7 and 3.3 Hz (prosodic fluctuations in natural speech) so that the interaction of modulation-rate difference and stream segregation could be examined. Transient gaps were embedded in both streams and the task was to report only the gaps in the target. If the subject could detect gaps in the target (not masker), the stream segregation was accurate. Performance was evaluated as a function of source separation using signal detection theory. The overall performance was significantly poorer at $\pm 7.5^\circ$ than at $\pm 45^\circ$ and $\pm 90^\circ$. The task was evaluated as a function of the difference between the carrier frequency of target and masker that was chosen randomly. The results showed that the effect of the frequency difference on performance was significant at all the angles and it was significantly stronger at $\pm 7.5^\circ$ than at $\pm 45^\circ$. The task was analyzed as a function of the modulation-rate difference between the streams and showed it did not play the role of a segregation cue at any angle. The reaction time at $\pm 7.5^\circ$ was significantly slower than that at $\pm 45^\circ$ and $\pm 90^\circ$. In summary, consistent with previous work on stream segregation, spatial cues influenced performance. When they were not strong enough cues, making a decision took a longer time. Frequency was a stronger cue for stream segregation when spatial cues were less clear. Regardless of the location in space, no interaction between the rate and space was found.

1 Introduction

1.1 Overview

Humans have the ability to focus on a desired, specific sound amid a noisy environment. The sound field in such an environment could constitute a mixture of speech and music sources. The "Cocktail Party Effect" refers to the ability to recognize a stimulus, especially a specific speech stimulus originating from a crowd. Determining how influential a given parameter is in solving the cocktail party problem or, in auditory scene analysis in general, is the initial step in clarifying how the human auditory system evaluates the surrounding acoustic environment [8–13].

Preliminary research on the cocktail party problem and auditory scene analysis was conducted by Colin E, Cherry [14] and later by Albert S. Bregman [15]. Their experiments to define the parameters and variables that contribute to human auditory scene analysis and to solving the cocktail party problem laid the foundation for similar studies that would follow.

These parameters include but are not limited to frequency-domain cues, time-domain cues, spatial cues, speech perception (exclusive) cues, intensity profile, visual cues, and rhythm (reviewed in chapter 2). Among all of these parameters, the focus of this research is on identifying the interaction between the sound source position in space (in the respective horizontal plane in this study), rate, and frequency content in a real-world (still simulated) cocktail party scenario.

The initial motivation for conducting this research was to create a novel and generic experimental design that would allow future researchers to quantify speech perception in a cocktail party scenario while overcoming the potential limitations and challenges

faced in similar studies conducted to date that are explained in this chapter.

The experiment was designed taking to the account the combination of the following bases; the choice of stimulus simulating the speech, the physical layout of the experiment emulating the real cocktail party scenario, and the assigned engagement or experimental task. These three bases together, allowed to use the Signal Detection Theory (SDT) [16–18] to evaluate the performance of the subjects in segregating two auditory streams.

1.1.1 Stimulus

In this study, the stimulus was a stereo sound made of narrowband (bandwidth of 200 Hz) frequency-modulated noise. One channel contained the desired sound (target) and another channel contained the undesired sound (masker). The stimulus was designed to satisfy the following criteria:

1.1.1.1 Prosodic rate

The prosody of speech reflects the emotion of the speaker and hence has a significant impact on how one relates to what was spoken. Several parameters together create the speech prosody including the fluctuation in the modulation-rate and rhythm of speech, the changes in the pitch and frequency content of speech, the variation of the loudness of the voice, and alternation of the duration of pronouncing a segment of a sentence [19–24].

The rates of the prosodic variations of speech in time, approximately correspond to a range of 1 - 4 Hz [21, 25–28]. Therefore, in this study, in order to simulate different rates of the prosodic fluctuation, modulating frequencies were set at random and ranged between 0.7 to 3.3 Hz. The modulating frequency range of the target and the masker

was in the range of 1.5 - 2.5 Hz (centered around 2 Hz; the center of prosodic rate) and 0.7 - 3.3 Hz, respectively. A larger frequency range of the masker allows for having the possibility of comparing the effect of the modulating frequency difference between the target and masker in different trials. This includes the cases where the target and masker modulating frequencies overlap or are very different.

1.1.1.2 Linguistic background and familiarity effect

In most speech perception studies, the stimulus presented was a phrase in a specific language, mostly the subjects' native language and accent. This implies that the results of the experiments could only be expanded to the possible behavior of a larger sample of subjects speaking the same language [29–31]. One motivation to conduct this study was to reproduce human speech rhythm independent of the specific feature of an individual language. This allowed the researcher to evaluate the same concept as that with spoken language on people of different linguistic backgrounds. This can potentially be used to expand the claim to a larger group of people of different nationalities.

In some of the pitch or speech rhythm perception studies, the perception was evaluated using a familiar voice or a popular melody. This means that the results possibly did not reflect the absolute influence of rhythm or pitch, but the rhythm or pitch perception of a familiar sound that could involve long-term memory as an additional cue [32–34].

This is not always the case in a real crowded place like a party scene with a lot of superimposed sounds that could be unfamiliar and heard for the very first time. Therefore, the task in these studies was simplified for the subjects. Furthermore, the results of these investigations could be possibly different if another familiar/unfamiliar sound was played. In this research, the stimulus did not have any familiar content and it

was never heard before. Therefore, it did not involve long-term memory in the perception task.

If the target sound is played initially and the subject is asked to remember that sound and continuously track that overtime after the masker sound is added with some delay, the short-term memory could potentially play the role of a cue which could affect the segregation ability [35]. In this test, the target and masker sounds were presented at the exact same time with no relative delay to avoid this effect.

1.1.1.3 Temporal continuity

In the case of using a phrase or a sentence of a language as a long stimulus, there will be several temporal onsets and offsets in time in the stimulus. These onsets and offsets are also used in the studies where the general concept of regularity with a sequences of Sine tones or musical rhythm perception with musical notes were evaluated [25, 36–40]. The onsets and offsets provide additional scene analysis cues (section 2.2.6). In this study, the stimulus was temporally continuous and did not have intermediate amplitude onset or offset to exclude their extra impact on the auditory perception, making the rhythm perception evaluation, a more exclusive analysis.

1.1.1.4 Two regular sounds

If in a multiple sound source scenario of a rhythm perception study, a rhythmic stimulus is compared to a non-rhythmic or an irregular stimulus, the subject attempted to differentiate regularity versus irregularity of the sounds, accordingly [33, 40]. These results do not necessarily justify real-world complex scenarios where there could be more than one rhythmic sound in the scene. Case in point, if two conversations (speech stimulus) in a party are heard at once and one has to be ignored, two rhythmic sounds are being

compared. Separating two or more regular and rhythmic sounds could prove to be a more complicated task.

1.1.2 Physical layout

In a handful of studies where rhythm and regularity perception, selective attention, or pitch perception in a cocktail party scenario were studied, the interaction of these concepts with the physical location of the sound sources relative to the listeners was not taken to account. In these studies, the stimulus presentation method was mostly limited to using headphones [19, 29, 33, 35, 37, 38, 40–44].

Consequently, the experiments excluded the effect of sound origin position. Therefore, their results could not possibly offer a generic explanation for the real effect in a Three-Dimensional (3D) space. Unlike this simulated scenario, in a real cocktail party scenario, multiple sounds are presented, simultaneously, from different directions. Therefore, in this case, the stream segregation task demands more attention and involvement. To increase the reality of the situation to an actual cocktail party scenario, at least two sound sources (the desired and competing one) that are separated in space are needed.

1.1.3 Task

In some cocktail party experiments, the traditional method of directly asking the subject to report whether one or two auditory streams were perceived or to report if one stream sounds different from another, was applied. However, an indirect test design can be used to assign subject to a more engaging and challenging task. If the task is completed successfully, the researcher can conclude that the participant was able to segregate the two streams. Directly asking the subjects as explained earlier, appears

effortless to them as they are not performing a task but are reporting the number of streams they perceive. Furthermore, observing a difference in the reaction time of the subjects in stream segregation is only possible when they are assigned to an indirect and challenging task [36, 38, 45–48].

In this research, to render audible differences perceivable between the two streams played through distinct speakers, transient gaps were included in both streams. If a subject correctly reported the gap of the target sound, he/she segregated the two simultaneous auditory streams. If a subject mistakenly reported the gap of the masker sound, he/she integrated the two simultaneous auditory streams.

This task is similar to the real cocktail party scenario task which requires selective attention to extract one sound at the presence of the noise. In this experiment, the task had to be done at the appropriate time to be accepted which was within a 1.5 second from the time the gap was presented.

1.2 Research questions and hypothesis

The following three research questions were asked; 1. Does the physical location of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation? 2. Does the frequency content difference of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation? 3. Does the modulation-rate difference of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation?

To answer all the questions, SDT was applied to the subject-based performance of this study. The comparisons of the performances were made with respect to the hit rates, false alarms, and sensitivity index. These parameters were derived by applying the SDT model. This was to define the degree to which the participants were successful at segregating the auditory streams in different directions.

To answer the first question, the results of the model were calculated as a function of the physical location of the sound sources. To answer the second question, the results of the model were calculated as a function of the difference between the frequency content of the target and masker sounds. To answer the third question, the results of the model were calculated as a function of the difference between the modulation-rate of the target and masker sounds.

The null hypothesis of this research was that the physical location of the desired and competing sound sources, their frequency content difference and their modulation-rate difference do not influence the gap-detection and auditory stream segregation ability. The alternative hypothesis is that the physical location of the two sound sources, their frequency content difference and their modulation-rate difference significantly influence the gap-detection and auditory stream segregation ability.

This research was an initial step to prove a general effect of interactions of these variables in the full 3D space. Therefore, only a limited number of sound source locations on the horizontal plane were initially selected. As humans are more precise in the localization of sound in the frontal plane, the two by two comparison of the following angles were assumed to be initially enough to reflect the existence of an effect (if any); speakers located at $\pm 7.5^\circ$ vs. $\pm 45^\circ$, $\pm 7.5^\circ$ vs. $\pm 90^\circ$, and $\pm 45^\circ$ vs. $\pm 90^\circ$ (where 0° is considered to be in front of the subject in the direction of the Sagittal plane).

Lastly, the subjects' reaction time of making a decision to report the gaps in the target (hits) and the gaps in the masker (false alarms) was evaluated as a function of speaker locations.

1.3 Application areas

To design an artificial intelligence system that analyses the auditory scene, at first, the acoustical cues applied by the human auditory system to analyze a scene must be defined. This would enable the system to automatically analyze an auditory scene similarly as the human auditory system. To be specific, the ultimate goal of designing a hearing aid device is to create a real-time computational system that performs the sound analysis as the human auditory system in a similar manner in terms of detecting a sound and keeping track of the stimulus over time. To do so, it is essential to reduce the overall noise which in this case would be any content that is not desired and enhance the perceived quality of the sound of interest to make it more distinguishable [9, 34, 49, 50].

One of the key points in creating such an algorithm is to improve the accuracy of estimating the sound direction of arrival in the horizontal plane. Two sounds with the same temporal and spectral content arising from two different directions relative to the head carry distinctive information about the surrounding environment. The effect of

this separation in space must be always considered in recreating the performance of the human auditory system [51–53].

Another advantage of determining the effects of the cues that are needed to be synthesized by a computational system to solve the cocktail party problem for speech is to improve the effectiveness of the automatic speech recognition algorithms. This is to be able to precisely distract the speech content from the environmental noise. It has several uses including the applications in security systems, voice-user interface to control interactive systems and speech to text assistance technology for disabled people [54–57].

3D audio applications including spatial sound design for 3D video games and movies, artistic spatial sound installations, and spatial music composition demand full awareness of space. It is vital to ensure that the sound synthesized is heard in terms of pitch and rhythm as it was intended to be perceived in the sweet spot of the speaker's setup. Therefore, if the spatial location of a sound source affects the perception of these two fundamental parameters, the perception difference that is caused by sound source position has to be taken into account and compensated to create a similar sound at different locations of the space.

1.4 Outline of the dissertation

Chapter 2, Literature Review: This chapter is started by defining the common terms in this research area, followed by reviewing effective parameters that contribute to the cocktail party effect and auditory scene analysis.

Chapter 3, Methodology: This chapter describes the experimental design of the psychoacoustic listening test. This includes the process of stimulus generation, test participants, the different phases of the training procedure and pre-test evaluations. This chapter concludes with an explanation of the structure of the main listening test.

Chapter 4, Results: This chapter, at first, overviews the overall outcome of the segregation task analysis as a function of speaker location by applying a SDT model. This is followed by analyzing the effects of frequency content difference between the target and masker sound on the stream segregation task. Next, the effects of the modulation-rate difference between the target and masker sound on gap-detection are provided. Finally, the results of the reaction time analysis as a function of speaker separation are discussed.

Chapter 5, Follow-up Experiments: Two follow-up experiments that were designed to compare and discuss their outcome with that of the first experiment are explained in this chapter.

Chapter 6, Discussion: A discussion of the results acquired from this study is provided in this chapter. The results are compared with the previous works in this field. Nextly, the experimental design limitations are explained.

Chapter 7, Conclusion: The outcome of the study is summarised in this chapter. Some recommendations for future research ideas are provided here.

2 Literature Review

2.1 Terminology

2.1.1 Feature, Event, Object, Stream

When sound reaches the ears, different components of it are extracted by the auditory system. This process gives us the ability to distinguish distinct sound features, sound events, auditory objects, and auditory streams [58, 59]. In the following, the definition of these fundamental terms is reviewed.

A "feature" of a sound is referred to as the characteristics of a specific frequency component at a certain time. A "sound event" relates to a set of frequencies generated over a while. Sound feature and sound event are physical parameters of a stimulus which can be defined regardless of how they might be perceived by the auditory system. A sound event formation occurs based on the formation of a collection of sound features at both low-level and high-level analysis [58, 60, 61].

The human auditory system can form a set of sound events and create an auditory "object". The perception of an auditory scene starts with auditory object selection which is choosing the sound source of interest among many. When a sound source is chosen, the temporal and spectral components of it are analyzed by the auditory system which provides the definition of the perceived boundaries of a single object [58, 61–63].

However, the perceived temporal and spectral sensory information might have been originally generated by different objects but perceived as one single "stream" of sound. The auditory system continuously segregates or integrates these components to detect different sound streams [64].

2.1.2 Temporal grouping vs. Spectral grouping

Temporal grouping or "Sequential Integration" refers to the ability to detect the temporal content of a composite sound stream over time. This includes integrating the temporal components of a single sound stream and separating temporal components of this stream from other existing streams on the scene [9, 49, 61].

Spectral grouping or "Simultaneous Integration" refers to the capability of analyzing the frequency components of the auditory scene, combining the frequency components of a single target source, and differentiating that from the spectral components of other present sources [9, 49].

2.1.3 Bottom-up processing vs. Top-down processing

Bottom-up Processing or Data-driven processing applies to the initial and mostly subconscious analysis of a sound event which is usually performed with the help of the basic auditory cues. Top-down Processing or Schema-driven processing starts after the basic analysis was applied to the incoming streams. This higher level of analysis demands more attention and concentration on the sounds reaching the ears. For instance, the familiarity of the sound source, prior training and memory become effective auditory cues during the Top-down processing stage [8, 49, 64–67].

2.2 Effective parameters that contributes to the cocktail party effect

For decades, many studies discovered the parameters which were found to be influential on human scene analysis and the ability to solve the cocktail party problem.

In the following section, these factors are briefly reviewed.

2.2.1 Frequency domain cues

2.2.1.1 Fundamental frequency and harmonicity

The fundamental frequency and its harmonics are among the strongest cues which give the auditory system the ability to segregate or integrate auditory streams. If the harmonics of a fundamental frequency are correctly detected, this fundamental frequency and its harmonics are grouped together and separated from other sets of frequencies and corresponding harmonics which results in successfully solving the cocktail party problem [8, 9, 49, 58, 62, 68–70].

2.2.1.2 Spectrum separation and spectrum continuity

If there is no overlap between the frequency range of the detected object and the spectrum of other produced sounds, the auditory object recognition would be an easier task in comparison with the case that multiple sound origins share frequency components. After the frequency profile of a part of a stream is detected, it is more likely that the upcoming similar frequency components would be categorized as they have been generated by the same already detected sound source [8, 13, 68].

2.2.1.3 Spectrum onset and offset synchrony

If the start and end moment of the presence of frequency components of a desired and a distractor sound are different, they would be more likely to be determined as separate streams by the auditory system. Furthermore, the overtones which have the

same attack time are more likely to be integrated into one stream [49, 58, 69].

2.2.1.4 Timbre

When it comes to differentiating the sounds of different musical instruments, timbre perception turns to be among the fundamental cues to separate the sources. Once the timbre of an instrument is detected, over time, the sound with the same timbre is more likely to be assumed to come from the same instrument on stage or in a recorded piece of music [58, 68].

2.2.1.5 Frequency masking

If the frequency of the distractor and the desired sound fall into the same critical band, the amplitude of the desired sound must be higher than the unwanted one to be audible to reach the threshold of hearing at that particular frequency, otherwise, it could be masked by the competing sound [11, 71].

2.2.2 Time domain cues

2.2.2.1 Temporal continuity

If a sound is temporally continuous (without interruption), it is more likely to be perceived as a single stream after the sound source is detected for the first time. However, if an auditory scene is made of the several distinct streams from different sources, segregating the short events arriving from a single source from another short-time events is a more challenging task [8, 72].

2.2.2.2 Forward masking vs. Backward masking

The audibility of a target sound is totally affected by the features of other sounds coming from other sources, simultaneously. The masking effect does not only apply when the masker and the target sound are existent at the same time but also before and after the time that the masker is presented. Forward masking is a rarely detectable auditory phenomenon that indicates how the target sound is masked by the masker even before the masker is presented in a period of a maximum of 50 ms. Backward masking is a more common and distinguishable effect that shows how the masker could still affect the perception of the target sound after it is stopped. In this case, in the first 15-20 ms after masker removal, the making effect is noticeable and after that, it gets less significant [67, 73].

2.2.3 Binaural masking level difference

A signal that is masked at the presence of noise can become audible if the phase of one of them is inverted in one ear while the masker and the target are both projected to both ears. In this case, after the phase inversion, the inaudible signal (maskee) reaches the perception threshold of hearing. The amount of change in the intensity level of the audible maskee and inaudible maskee is referred to as binaural masking level difference [66, 73–75].

2.2.4 Spatial cues

The monaural and binaural sound localization cues including Head-Related Transfer Function (HRTF), Interaural Level Difference (ILD), Interaural Time Difference (ITD), reverberation, Pinna factor, etc together enable the auditory system to detect the position

of a sound source in an environment. Once the auditory system localizes the positions of the sound sources on a scene, the streams coming from different sound source locations are more likely to be segregated. Hence, over time, the target sound is released from being masked again by other sound sources in space. This is referred to as spatial release from masking. The spatial location maintenance (continuity) can be an effective perceptual cue to remember the extracted auditory content emanating from a specific source and differentiate that from the other sources on the scene [8, 49, 62, 76–83].

2.2.5 Intensity profile

Human uses the intensity of masker and maskee as a cue for segregating them. The sound events with lower intensity are more likely to be grouped and differentiated from the sound events in the scene that have a higher intensity level [8, 58, 68].

2.2.6 Amplitude modulation

Continuous synchronized changes in the amplitude of sound events suggest that these events with the similar amplitude fluctuation pattern are made by the same sound source. [49, 58, 62, 77].

2.2.7 Visual cues

The correlation between the perceived streams and the visual cues could be a strong cue to integrate or differentiate the sound sources. This could be more effective when it comes to speech perception. The body gestures, lip-syncing, body movements, etc could have a profound impact on solving the cocktail party problem for speech [8, 84].

2.2.8 Familiarity and predictability of sound

Sound source separation would be an easier task with the help of memory if the sound of interest is familiar to the listener and the sound source characteristics have remained the same over time [33, 63].

2.2.9 Speech recognition exclusive cues

In case the target source is speech, more cues are available to the listener including the speaker's accent, lip movements, the familiarity with the native language, sentences stress, speaker's gender, familiarity with the voice of known people, etc [48, 62, 72].

2.2.10 Regularity and rhythm

The human auditory system is able to detect repetition in some of the sound features. This includes fluctuation in sound intensity, temporal pattern, frequency component, and so on [59, 62, 85, 86]. Some studies showed that the temporal regularity can act as a secondary cue after the auditory object is formed by applying the bottom-up cues to keep track of the already recognized source over time [38, 64].

3 Methodology

3.1 Overview

The first part of this chapter focuses on the preparation of the listening experiment. This includes the stimulus generation procedure and the method of generating and applying the gaps. Next, the experimental design is explained. This part covers the description of the equipment used in the test, the test environment, and the information regarding the subjects that participated in the test. The last part illustrates each phase of the procedure of the listening experiment which contained the subject's preparation, training procedure, gap length threshold measurement, and finally the main listening test operation.

3.2 Stimulus generation

3.2.1 Narrowband noise generation

The stimulus was a two channels sound, one channel contained the target sound and the other channel contained the masker sound. Both target and masker sounds were frequency-modulated narrowband noise with the modulation depth of 25%. The target and masker differed in terms of frequency content, modulation-rate and the location which they were played from.

For a single stimulus, the modulation frequency of the target was a random value in a range of 1.5 - 2.5 Hz. The target noise frequency components fall into the 9th critical band and were in a range of 900 - 1100 Hz (i.e. a narrowband noise with a bandwidth of 200 Hz). This range was chosen as the equal-loudness contour centered around 1 kHz

is almost flat. Owing to this, no loudness adjustments were necessary when comparing the frequency range of target and masker if range mentioned was used. The narrowband noise was generated by adding thirty pure sinusoids of random frequencies (30 carrier frequencies) in the aforementioned range. The cumulative amplitude resulting from such a summation was divided by the number of sinusoids to prevent amplitude distortions and clipping.

For a single stimulus, the modulation frequency of masker was a random value in a range of 0.7 - 3.3 Hz. The masker was also a narrowband frequency-modulated noise with a bandwidth of 200 Hz, but the noise frequency components (30 carrier frequencies) varied from trial to trial in range of 550 - 1450 Hz. The target and masker could randomly fall into the same or different critical bands. Table 1 illustrates the target and masker frequency content of 5 trials.

Table 1: The frequency content of target and masker of 5 trials.

Target frequency range (Hz)	Target bandwidth (Hz)	Target critical band	Masker frequency range (Hz)	Masker bandwidth (Hz)	Masker critical band
900 - 1100	200	9	550 - 750	200	6 & 7
900 - 1100	200	9	650 - 850	200	7 & 8
900 - 1100	200	9	700 - 900	200	7 & 8
900 - 1100	200	9	1000 - 1200	200	9 & 10
900 - 1100	200	9	1250 - 1450	200	9 & 10

The starting phase of each stimulus was randomly chosen of the following values: 0, $\pi/4$, $\pi/2$, $3\pi/4$, π, $5\pi/4$, $3\pi/2$, or $7\pi/4$. This was to avoid any undesired effect of the initial phase on the stimulus perception (in case the initial phase plays the role of a cue). If the target sound had always the same initial phase, the subject could possibly recognize that and rely on it as a cue for detecting the target sound.

At the beginning and end of each signal, a 10 ms cosine window was applied to avoid acoustic artifacts from the instantaneous onset or offset of the stimulus.

3.2.2 Gap generation

The gap was applied as an amplitude window to a stimulus. It was made of two 10 ms sine ramps with a zero amplitude sustain part in the middle. Figure 1 shows the full window that makes a 45 ms gap (the average gap measured in this experiment) and figure 2 shows the result of applying this window to the amplitude of a white noise.

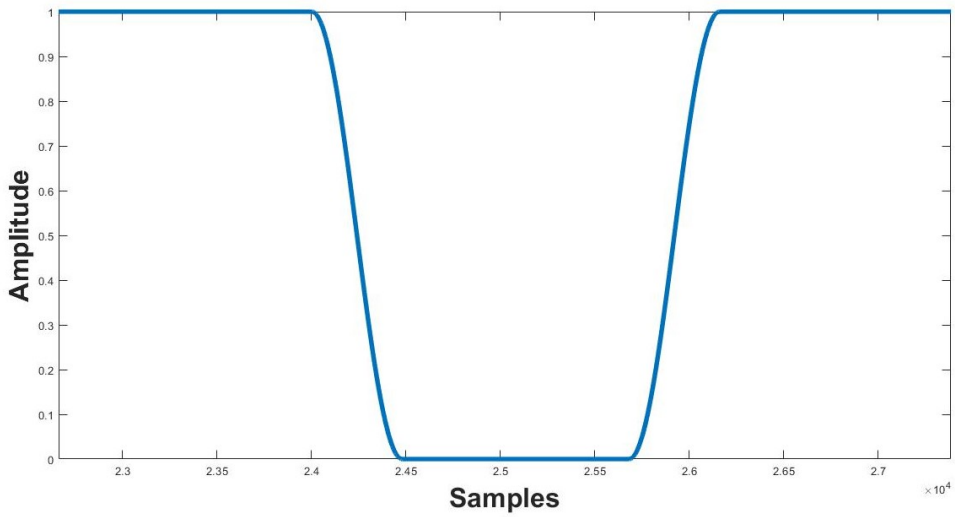

Figure 1: A sample window that applies a 45 ms gap to the stimulus.

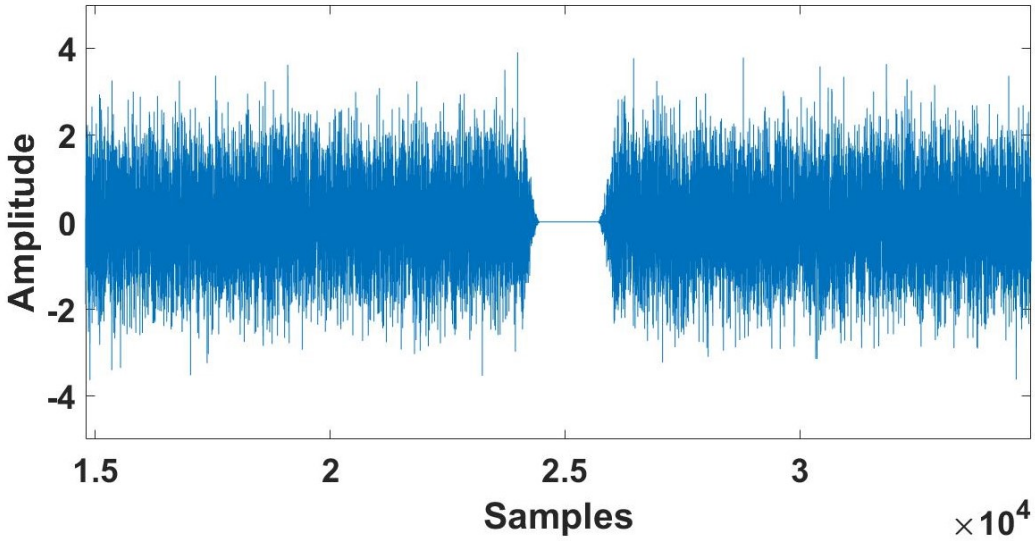

Figure 2: The change in the amplitude of a white noise after the gap was applied.

3.2.3 Gap location

To ensure that the number of gaps within a trial is unpredictable, different coupling patterns of gap numbers in target and masker in a single trial were chosen as shown in table 2. The total number of gaps in a 10 sec stimulus could be 2,3, or 4. The total number of gaps remained the same for all the subjects.

Table 2: The possible number of gaps in two channels of one stimulus.

Occurrence percentage in 180 trials	Nr. of gaps in target per trial	Nr. of gaps in masker per trial
30 %	1	1
20%	2	1
20%	1	2
10%	2	2
10%	3	1
10%	1	3

The time duration of a single trial was 10 seconds. A minimum of 1.5 seconds time difference was considered between all the existing gaps of a two-channel sound to give the subject at least 1.5 seconds to react to the previously presented gap. No gap was located at the start of the stimulus (within the first 1.5 seconds) to provide the subject sufficient time to become oriented to the task at hand. The last 1.5 seconds of the stimulus was gap-free owing to the subject not having enough time to report it before the stimulus terminates.

In the total of 180 stimuli of the main test, 540 gaps were located; 270 gaps in the target sounds and 270 gaps in the masker sounds. In total, 90 gaps were embedded in the trials of each speaker pair ($\pm 7.5°$, $\pm 45°$, $\pm 90°$).

The gaps locations were defined in such a way that for an individual subject, the generated gaps were uniformly spread out over all the trials in terms of time and

the instantaneous phase where the gap was applied. This is to ensure that the subject's performance was evaluated over all the possible gap locations (in time) and the instantaneous phase. Figure 3 shows the distribution of all 270 gaps of target sounds and 270 gaps of masker sounds of 180 stimuli over the 10 seconds time period generated for a single subject. The gaps were scattered uniformly at the entire time period of 1.5 - 8.5 seconds. Figure 4 shows the distribution of the instantaneous phases of the signal where the gaps were generated for a given subject. Similar to the distribution in time, the gaps balanced over the full possible phase range $(-\pi : \pi)$.

In summary, the following parameters changed within trials made for a subject: the assignment of the target and masker to the left and right channel that was equally often, the noise frequency content and modulating frequency of target and masker, gaps location in time and phase, the trial initial phase, and number of gaps per trial. The following parameters remained constant between the subjects: the total number of trials (180), the total number of trials per speaker pair (60), the total number of gaps per speaker pair (90), the total number of gaps in target (270) and in masker (270).

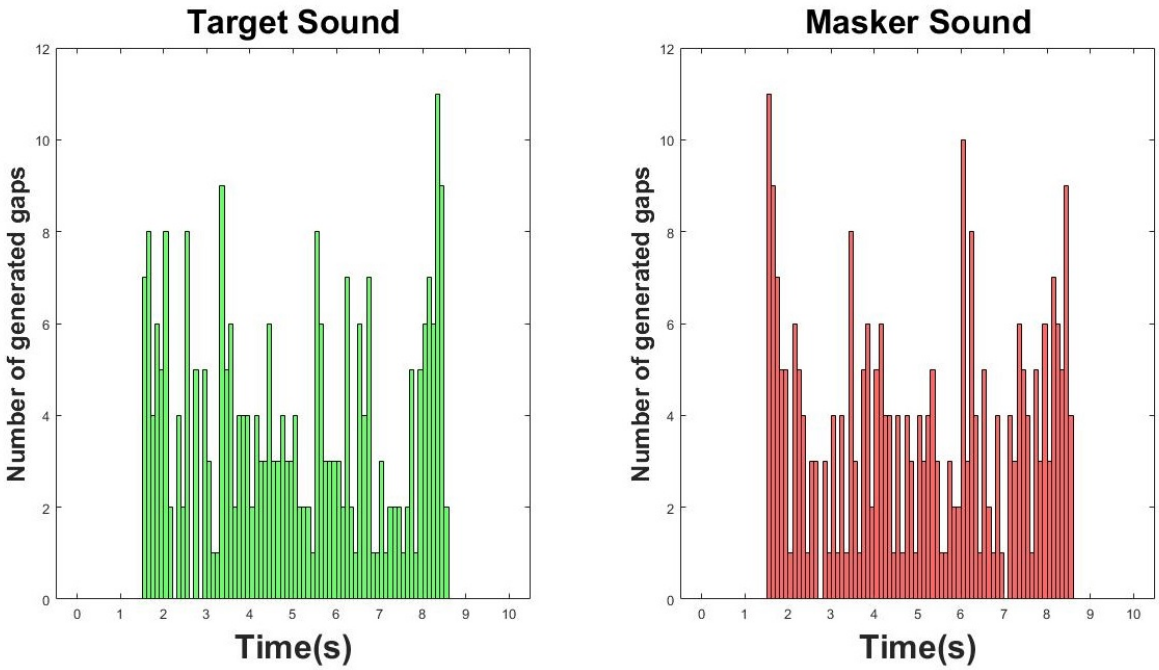

Figure 3: The distribution of all 270 gaps of target and 270 gaps of masker of 180 stimuli generated for a single subject over 10 seconds time period.

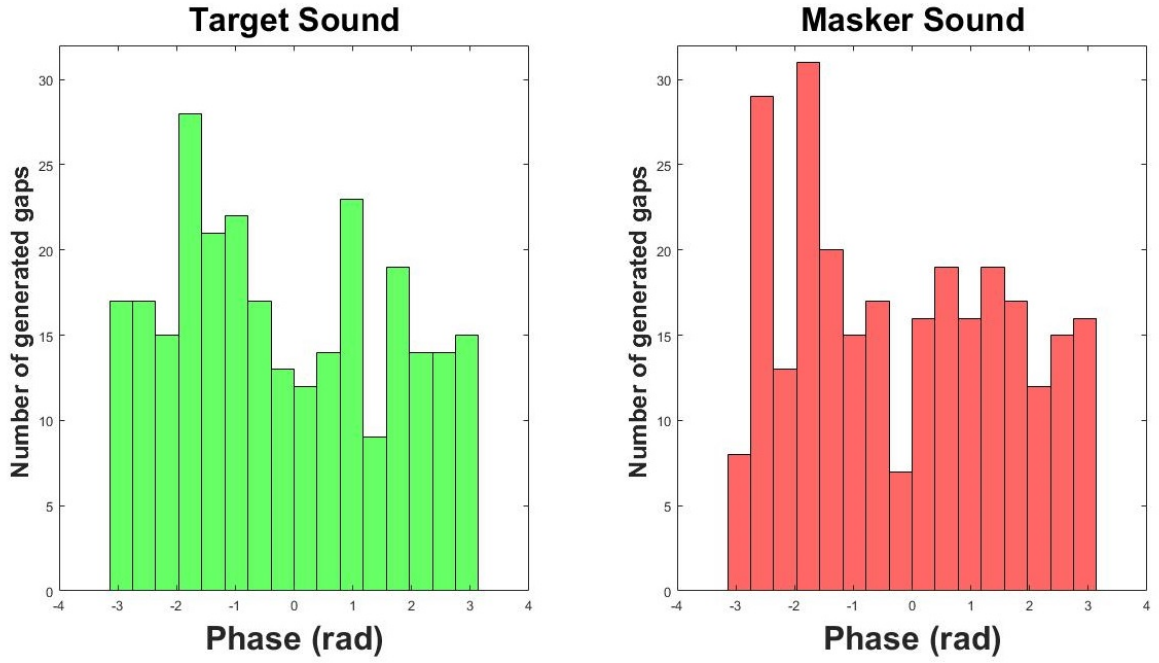

Figure 4: The distribution of all 270 gaps of target and 270 gaps of masker of 180 stimuli for a single subject over the instantaneous phases where the gaps were generated.

3.3 Experiment design

3.3.1 Physical layout

Figure 5 shows six hypothetical sound sources in a cocktail party scenario and figure 6 shows the corresponding setup implemented to simulate such a scene. A speaker array made of six loudspeakers was designed. All the speakers had the same distance from the subject (0.9 m).

Figure 7 demonstrates a scheme of the physical layout of the experimental design as the visual cue of the test. The subject was trained to map the visual cue with the implemented setup during the training procedure.

As the visual cue given to the subjects contained the information about the location of the loudspeaker which played the sound in each trial, the sound source localization task was not demanded and as such, was excluded from the subject's task. This could potentially make a fair situation in which people with a higher and lower localization ability would still have the same preconditions for doing this test.

At different trials, the target and masker were played through the following speaker pairs. Each case was used for 30 out of 180 trials of the experiment.

Target: +7.5°, Masker:-7.5°

Target: -7.5°, Masker: +7.5°

Target: +45°, Masker: -45°

Target: -45°, Masker: +45°

Target: +90°, Masker: -90°

Target: -90°, Masker: +90°

Figure 5: Six hypothetical sound sources in a party!

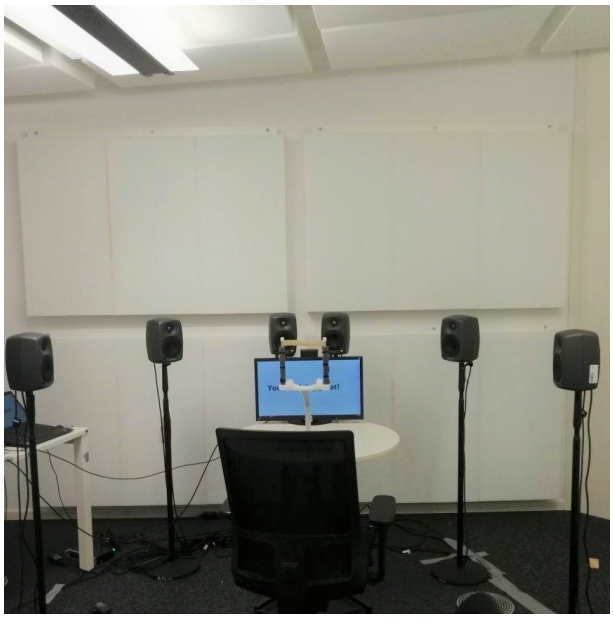

Figure 6: The implementation of the scenario in the lab.

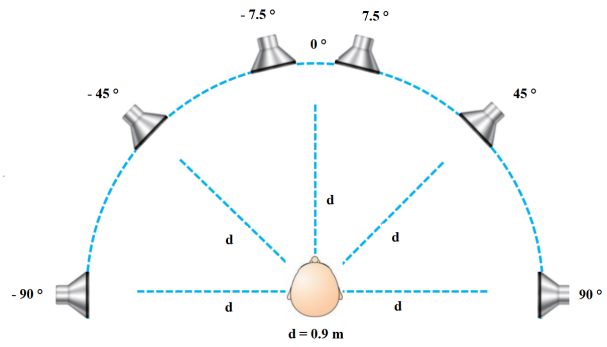

Figure 7: The visual cue illustrating the physical layout of the experiment.

3.3.2 Equipment

To keep the participant's head fixed and avoid any head movement during the study, a medical chin rest device was used (figure 8). The height of the device was adjustable from 29.5 cm to a height of 54 cm and was set to the right height from the table based on the height of the subject.

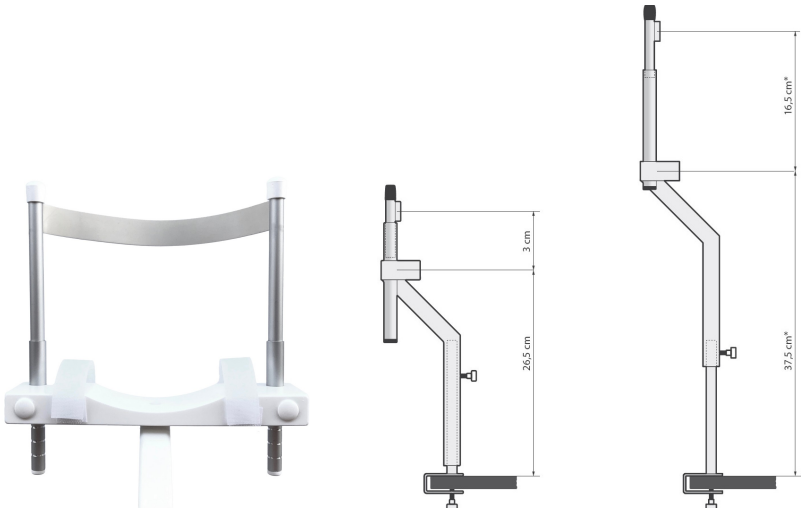

Figure 8: "Hasomed" Medical Chin Rest, adapted from [2].

Figure 9: A demonstration of the chin rest device usage.

The reference point of the speaker (according to [3]) was positioned at a height of 160 cm from the floor, measured by a laser distance meter. To calibrate the loudness of the speakers, a microphone was located at the hypothetical midpoint of the subject's head attached to a sound pressure level meter to measure the loudness of the incident signal. The used equipment is shown in figure 10 and the calibration setup is shown in figure 11.

Figure 10: Genelec 8040 Studio Monitor with the reference point marked in Red according to [3], adapted from [4], SPL Meter NTi Audio, Model:XL2 and M4261 microphone [5], Fluke 424D Laser Distance Meter [6].

Figure 11: The speaker loudness calibration setup.

3.3.3 Playback environment

The experiment was conducted in the acoustic lab at the Max Planck Institute for Empirical Aesthetics, Frankfurt. The reverberation time (RT_{60}) of the room is 0.16 s at 1000 Hz. Two sound-absorbing panels were located behind the subject to avoid any reflections from the back of the room. Between the panels, a camera was located to monitor the subject from the control room where the researcher ran the test. The video was only streamed and no data was recorded. A Talkback system was located below the camera and the subject was allowed to communicate with the researcher when needed.

Figure 12: The relative position of the subject, camera and the Talkback system.

3.3.4 Participants

In total, 33 participants were subjected to the listening test. The results of the first 7 tests were considered the pilot outcome. Out of the 26 main subjects, the results of four of the subjects were removed from the final dataset as their answers were found to be unreliable (detailed reasoning explained in Chapter 4). The outcome of this study is reported based on the data gathered from 22 subjects with an equal number of males and females. Their ages were in the range of : 19 - 34 (μ =26, σ =3.89).

The minimum required sample size to have a test power of 0.8 with a medium to large effect size was calculated according to the ANOVA power analysis as there were three groups (angles) to compare. The results showed that to have an effect size of 0.75, a minimum of 21 subjects is needed. Therefore, evaluating the data from 22 subjects was still sufficient to have 80% test power with a large effect size.

The recruitment of research participants was done through subjects database who previously signed up to participate in the studies of their own volition. 18 out of 22 final subjects were registered through this system. They have been reimbursed 28 EUR for the two hours tests. The other four subjects were the institute staff. The participants were asked to sign the consent form in the language of their choice; English or German. The general information sheet was also provided in both languages.

3.4 Experimental procedure

3.4.1 Subject's positioning

At first, the subject was asked to remain seated on the chair and the researcher adjusted the height of the chair and the chin rest. The height of the ears from the floor

was adjusted in such a way that they were at the same height as the speaker's reference point (160 cm from the floor) to be able to waive any spatial effect in the vertical plane as much as possible.

3.4.2 Introduction

The subject was asked to watch a 7 minutes introductory video. The goal here was to familiarize the subject in steps with the surrounding space, six different directions of arrival, stimulus type, and the concept of a "silence" moment. The content of the introduction is explained below:

Phase 1: To instruct the subject on how the relative position of him/her and the speakers are demonstrated on the scheme, the following sentences and picture displayed on the screen.

Text on screen: *"As you see, there are 6 loudspeakers around you. This is how we show them on screen."*

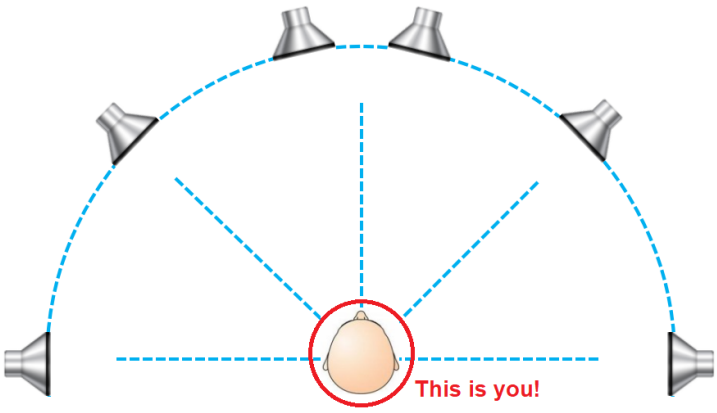

Figure 13: The scheme was used to clarify the relative position of the speakers and the subject.

Phase 2: Four seconds of white noise was played through the speakers in the following order: +90°, +45°, +7.5°, -7.5°, -45°, -90°. This was to familiarize the subjects

with different sound source locations with a known sound. While the sound was played, the speaker corresponding was highlighted in Green. Figure 14 shows two samples of these images at +90° and -45°.

Text on screen: *"Now, we play some noise from each loudspeaker. The speaker marked in Green is playing the sound."*

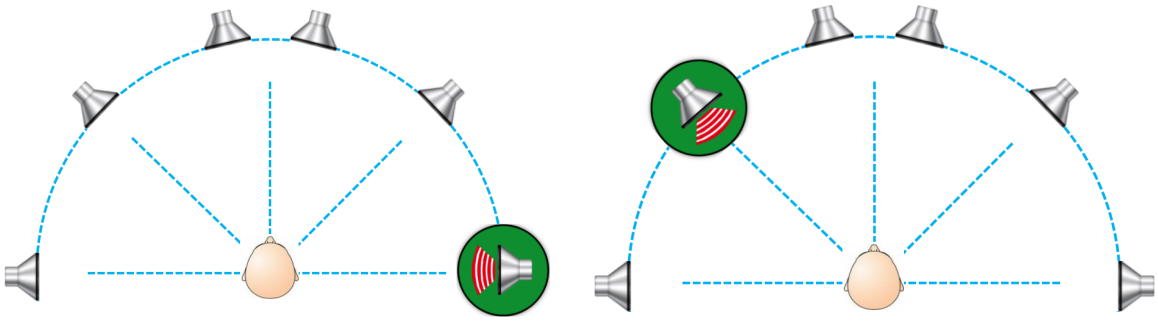

Figure 14: The examples of the figures shown on the screen while the sound was being played from one speaker, left:+90°, right:-45°.

Phase 3: Four seconds of white noise was played through the speakers in the following order: -90°, -45°, -7.5°, +7.5°, +45°, +90°

Text on screen: *"Second round! This time in the opposite direction!"*

Phase 4: A four seconds long stimulus without any gap was played through the speakers in the following order: +90°, +45°, +7.5°, -7.5°, -45°, -90°. This was the first time that the subject heard the stimulus of this experiment.

Text on screen: *"Now listen to a different type of sound!"*

Phase 5: Four seconds long stimulus without any gap were played through the speakers in the following random order: +45°, -45°, -7.5°, -90°, +90°, +7.5°

Text on screen: *"Now, from here and there!"*

Phase 6: After familiarization with space and stimulus type, from now on, the

sound was being played with a speaker pair. The icons designed for the pair (taregt and masker) were shown on the screen as figure 15. The speaker marked in Green showed the source of the target sound and the subject was asked to only pay attention to that and ignore the one without the Green circle which contained the masker sound.

Text on screen: *"Now, we play sound from two speakers, simultaneously, shown as below. From now on, only focus on the sound coming from the speaker marked in Green."*

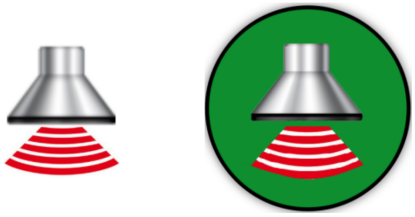

Figure 15: The icons designed for the speaker which played the masker sound (left) and the target sound (right).

Phase 7: Four seconds long stereo stimuli (target and masker) with no gap were generated. The playback order was as the following: ±45° (target: +45°), ±90° (target: -90°), ±7.5° (target: +7.5°), ±45° (target: -45°), ±90° (target: +90°), ±7.5° (target: -7.5°). Figure 16 previews two examples of these combinations.

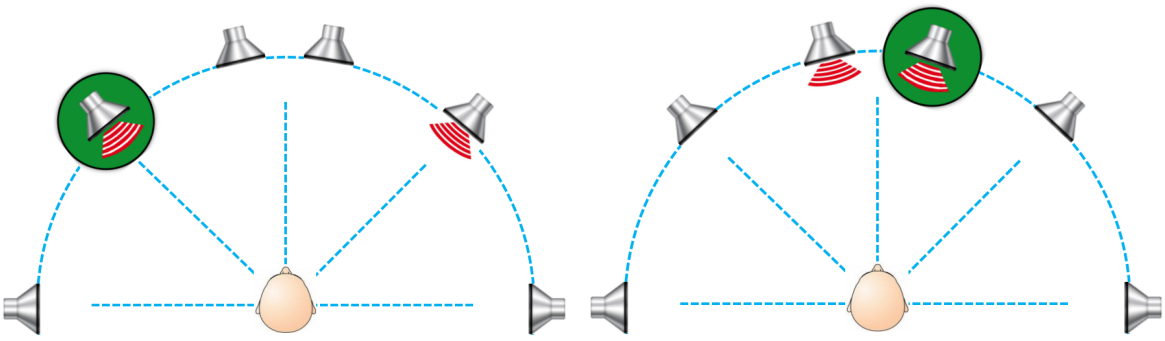

Figure 16: The examples of the figures shown on the screen while the sound was being played from two speakers, left:±45° (target: -45°), right:±7.5° (target: +7.5°).

Phase 8: In this step, the concept of a gap or silence was explained:

Text on screen: *"The "Silence" is the short moment in time that the sound is interrupted and you notice a discontinuity in the sound. It sounds like a short tiny pause or a distinct break or a gap in time. Your task is to detect that gap or "Silence" moment. In the following part, you will hear some examples."*

Phase 9: In this phase, in 11 short examples, gaps with different lengths were presented from different directions.

Example 1: No gap, at -7.5°.

Text on screen: *"Here is a sound with no gap in it similar to what you heard so far!"*

Example 2: 900 ms gap at -7.5°.

Text on screen: *"Here is the same sound with a large gap in it!"*

Example 3: 500 `ms` gap at -7.5°.

Text on screen: *"Here is the same sound with a shorter gap in it!"*

Example 4: 250 `ms` gap at -7.5°.

Text on screen: *"Here is the same sound with an even shorter gap in it!"*

Example 5: 250 ms gap at +90°.

Text on screen: *"Now, listen to the same sound coming from another direction!"*

Example 6: 120 ms gap at -45°.

Text on screen: *"Yet, a shorter gap and from a different direction!"*

From now on, two channels sounds (having both the target and masker sounds)

generated and there was only one gap in the target sound. To further emphasize on the idea of playing two sounds, simultaneously, two icons were presented again.

Text on screen: *"Now, we play sound from two speakers simultaneously, shown as below. From now on, only focus on "silence" moment in the sound of Green speaker."*

Example 7: 500 ms gap at ±7.5° (target: +7.5°).

Example 8: 300 ms gap at ±90° (target: -90°).

Text on screen: *"A shorter gap from a different direction!"*

Example 9: 250 ms gap at ±7.5° (target: -7.5°).

Text on screen: *"Yet, a shorter gap from a different direction"*

Example 10: 120 ms gap at ±45° (target: +45°).

Text on screen: *"Yet, a shorter gap from a different direction"*

Example 11: 90 ms gap at ±90° (target: +90°).

Text on screen: *"Yet, a shorter gap from a different direction"*

At this moment, the introduction phase ended.

Therefore, the shortest gap length presented in the introduction step was 90 ms. This value was larger than any measured minimum gap length threshold for the seven pilot subjects. The following gaps length thresholds were measured for the pilot subjects; 45, 52, 30, 65, 28, 85, 51 ms. In the end, considering the final 22 main subjects, as the largest measured gap for a subject was 82 ms, stopping the training at 90 ms found to be the right choice.

Finally, the subject was asked to define the gap in their own words to ensure

that he/she understood the concept. Moreover, the subject was inquired as to his/her ability to hear the gaps when it was presented. According to what they reported, all the subjects clearly understood the topic after the introductory session was presented.

3.4.3 Training

To evaluate the capability of the subject in performing the task, an initial test with 12 trials with large gaps in them was conducted during the training phase. To collect the data a response device shown in figure 17 was used. The device was connected to the main computer with a USB-3 cable (3ms latency) and the data was collected with the MATLAB Psychophysics Toolbox extensions functions [87].

The subject was required to push the Green "Silence" button as quickly as possible when he/she heard the gap or push the Red "Nope" button when the sound stopped with no gap detected. The Yellow "I'm back" button, was pushed by the subject when he/she got back to the room from a short break during the main test. By pushing this button, the audio was continued. The rest of the buttons were deactivated and not used at any point.

Figure 17: The response device used in the experiment, Cedrus RB-740.

Text on screen: *"Now, we play a couple of sounds for you. If you hear the silence, right away push the "Silence" button! If you didn't hear it, after the sound stops, push the "Nope" button"*

The sounds were four seconds long and there was only one gap in the target sound and no gap in the masker. Participants passed the test when they correctly responded to 9 out of 12 trials. The gap length was similar for all the subjects. The gap duration was large in all the stimuli to simplify the task. The gap duration of the 12 trials was as the following; 400 ms, 600 ms, 300 ms, 350 ms, 130 ms, 200 ms, 150 ms, 300 ms, 560 ms, 225 ms, 445 ms, 190 ms. Each speaker was used twice in a random order. This process took approximately 4 minutes.

All of the subjects passed this test successfully showing that the step by step instructions was clear and sufficient for them. Figure 18 demonstrates the distribution of the number of times the subjects did not detect the gap at this step.

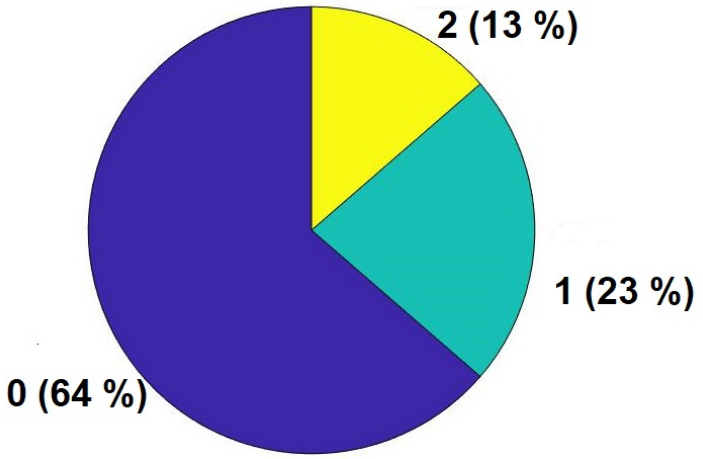

Figure 18: The percentage of the number of wrong answers (0,1,2) out of 12 trials over all the subjects.

3.4.4 Custom adaptive tracking procedure

This step aimed to measure the minimum required gap length for each subject to be able to do the stream segregation task. The participant was assigned to a similar task as the training test, but here, the length of the gaps varied from trial to trial based on the subject's performance in the previous trial. The trials were randomly played through different speakers.

Text on screen: *"Now, you know what we mean by a gap! In the following examples, if you hear the silence, push the "Silence" button as soon as possible. If you didn't hear it, push the "Nope" button after the sound stops. The silence can happen at any time while the sound is being played."*

Two sequences of trials were considered here based on the gap length values and the trials of these two sets were randomly played. The descending sequence started with a stimulus with a 150 ms gap in it and followed the logic stated: If the subject was able to detect the gap correctly (pushed the "Silence" button at the right time), the gap length

was reduced in the next trial to 100 ms. By any chance, if the subject did not push the "Silence" button, the next trial also was made with a 150 ms gap.

If the gap length reached a value smaller than 100 ms and larger than 20 ms, the duration adjustment was done by adding or subtracting either 4, 5 or 6 ms (randomly chosen). This is because according to the pilot study results, the gap duration of 7 subjects was in a range of 28 to 85. Therefore, small changes in gap duration in the range of 20 - 100 ms were needed to be able to estimate the gap duration with the highest resolution possible. If it reached the range of 20 - 10 ms, the changes were done by either 8, 9 or 10 ms (randomly chosen). If it reached 10 ms, the next trial was repeated with a 10 ms gap.

The second sequence was initiated with a 5 ms gap and the length was increased trial by trial if the subject answered the questions correctly. The amount of increment to the gap length followed the same logic as that of the descending sequence. The logic for the opposite direction was as follows: The gap duration change was either 4, 5 or 6 ms if the gap was in the range: 20 - 100 ms and was either 8, 9 or 10 ms changes after the trials with less than a 20 ms gap.

The subject was not aware of the logic behind the test and as the trials were being played randomly and with a random amount of change, it was not possible for them to predict the change in the next trial.

After each trial, the status of each sequence was checked. If the status changed from increasing the duration in one trial of the sequence to decreasing the duration in the next trial or vice versa, a status changed occurred. After the status of a sequence changed 12 times (12 reversals), the process of making new trials for this sequence was stopped. When both sequences met this requirement, the test stopped altogether. If the latter requirement was not met, the test was stopped by default after 34 trials for each sequence. This process took a maximum of 10 minutes. Figure 19 shows the results of

this test for two subjects.

The median of the measured gap duration in the last four trials of both sequences was calculated. This value was considered as the minimum individual required gap duration to be able to the segregation task when there is only one gap in a short trial. The range of calculated gap duration for the 22 subjects was: 20-82 ms (μ = 45.36 ms σ = 17.76).

Considering the results of the pilot subjects and the fact that in the actual test the existence of multiple gaps in a 10 seconds long target and masker sounds made the task more difficult than the task of this section, the gap duration had to be increased with the same ratio for all the subject to the level that the task would be doable for them (according to the individually calculated sensitivity index, explained in chapter 4). After repeating the test with different levels of increment, the value of 2.5 found to be suitable to multiply with the original gap duration calculated for all the subjects.

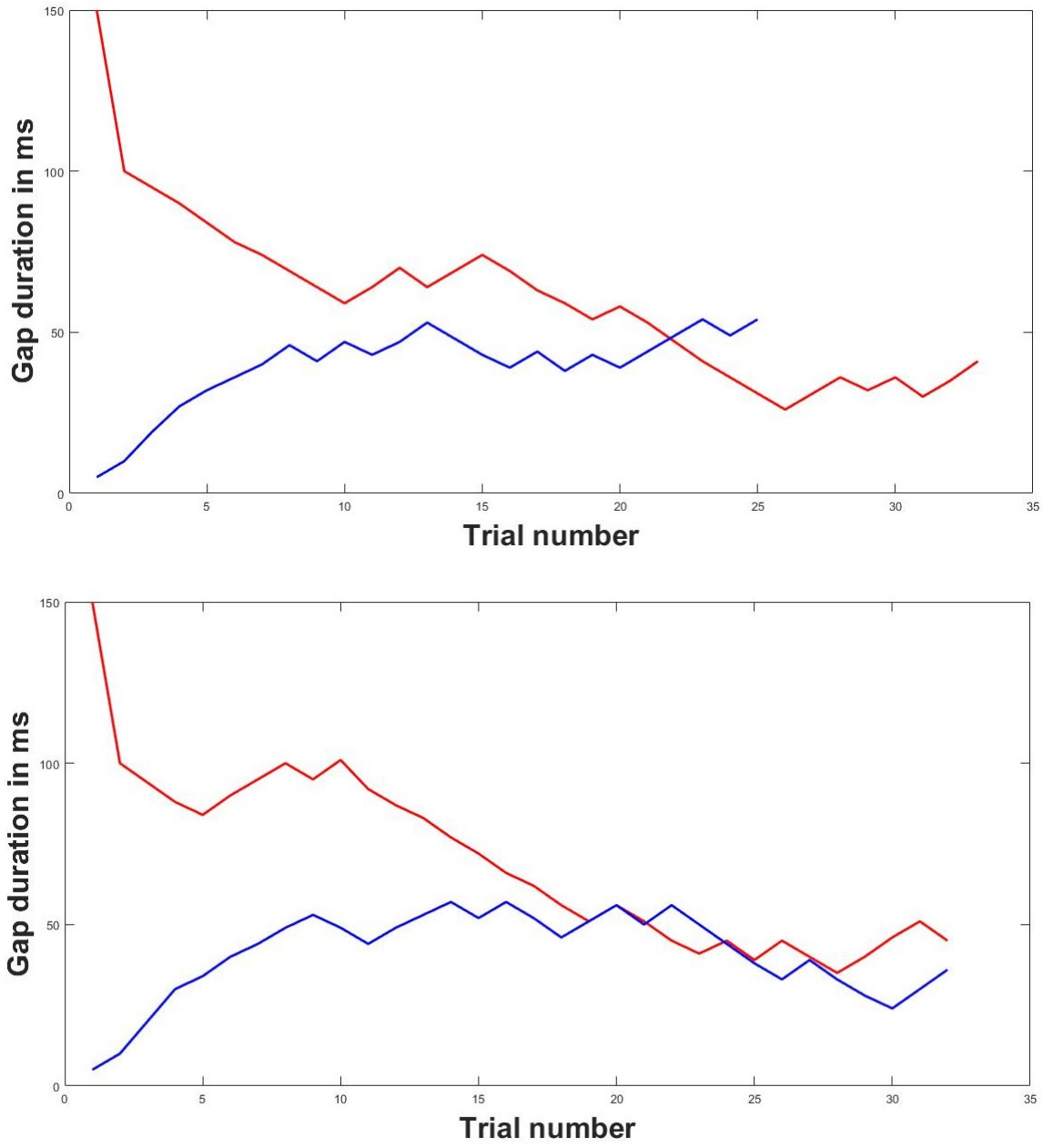

Figure 19: The change in gap duration values trial by trial for two subjects, Blue: ascending sequence, Red: descending sequence.

3.4.5 Stimulus generation

At this stage, for an individual subject, all the stimuli were made based on the calculated gap length for him/her and the stimulus generation procedure explained earlier.

3.4.6 Main test

The instruction of the main task was given verbally and written on the screen as well. The visual cue showed up 1 second before the audio signal started to provide the subject enough time to prepare to react to each stimulus.

Text on screen: *"Let's start the main task! From now on, you don't need to use the "Nope" button. We play sound from two speakers as before, but there are gaps in the sounds coming from both speakers and the sounds are longer in time. Only focus on the gaps coming from the speaker marked in Green. At any time you hear a gap coming from the Green speaker, push the "Silence" button as soon as possible."*

The stimuli generated for each speaker set were randomly played and the random order differed for subjects. After each sound, there was a 3 seconds pause. After a set of 10 trials, a pause of 10 seconds was provided. The subject was advised to move his/her head to avoid any discomfort caused by keeping the head fixed on the chin rest device for the time needed. After every 45 trials (total of 3 times in the whole experiment), a 5 minutes optional break was given. Playing the trials took 30 minutes and considering all the pauses and breaks, this process lasted up to 1 hour depending on the subject's choice to take the optional breaks or not.

4 Results

4.1 Overview

The results of this experiment are represented in four steps;

1. The results of the gap-detection and segregation task performance as a function of speaker separation.

2. The results of the gap-detection and segregation task performance as a function of the difference between the frequency content (carrier frequency) of the target and masker.

3. The results of the gap-detection and segregation task performance as a function of the difference between the modulation-rate (modulating frequency) of the target and masker.

4. The results of the reaction time to detect a gap in the stimulus as a function of speaker separation.

4.2 Analysis of the number of reported gaps

As the first step of the data analysis procedure, an overall assessment of the gap-detection performance of the subjects was made. To do so, the number of times a subject pushed the "Silence" button meaning he/she reported that I detected a gap was counted. The reported gaps could fall into three categories; 1) the gaps that were reported within 1.5 seconds of the gaps embedded in target, 2) the gaps that were reported within 1.5 seconds of the gaps embedded in masker 3) the gaps that were reported anytime other than the acceptable time window of the gaps embedded in the target or masker. In other words, in this case, the button was pushed while there was no gap either in the target or masker.

Next, the number of times that a gap fell in to the third category over all the

180 trials was counted for each subject. The average of this overall value was 8.72 with a standard deviation of 8.34. Figure 20 demonstrates the distribution of this value over all the subjects. As a result of this analysis, the data collected from four subjects (Red bars in figure 20) were removed from the final dataset of this experiment. This was because, in comparison with other subjects, this variable was unusually high for them; 72, 170, 40, and 67 times during the experiment. The explanation of the potential causes of this behavior is beyond the scope of this study.

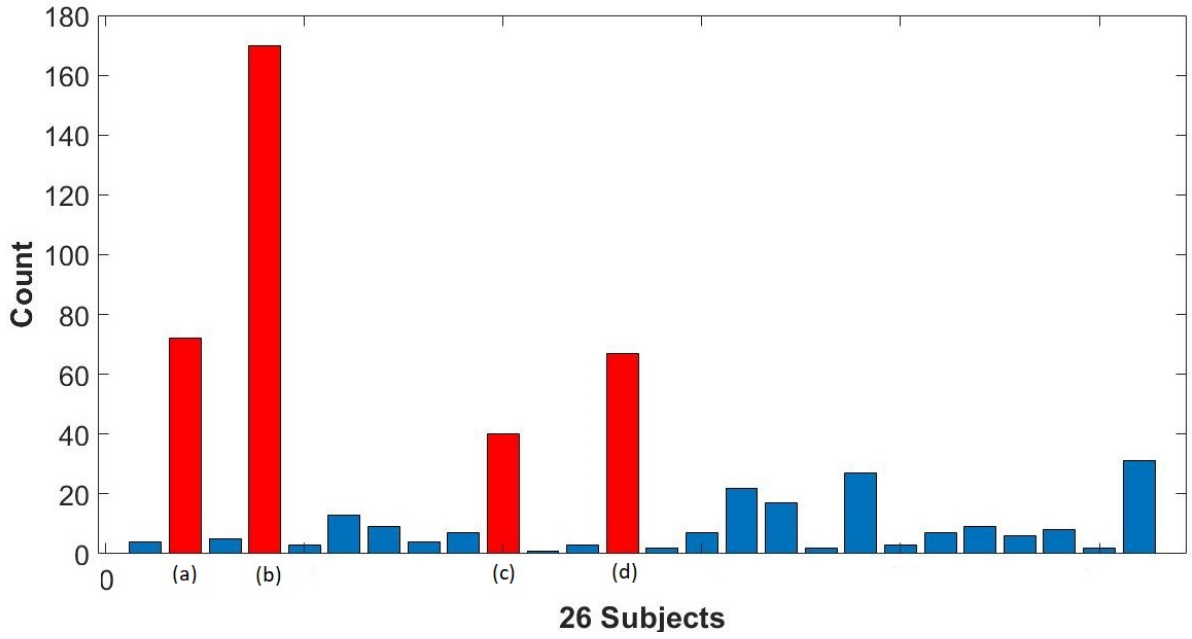

Figure 20: The number of gaps reported by each subject which were neither in target nor in masker over the entire experiment. The results of the subjects a,b,c, and d were removed from the experimenter's dataset.

Figure 21 shows the number of gaps in each category reported by subject b. A similar rate of the reported gaps in the first and third categories can be seen in this figure.

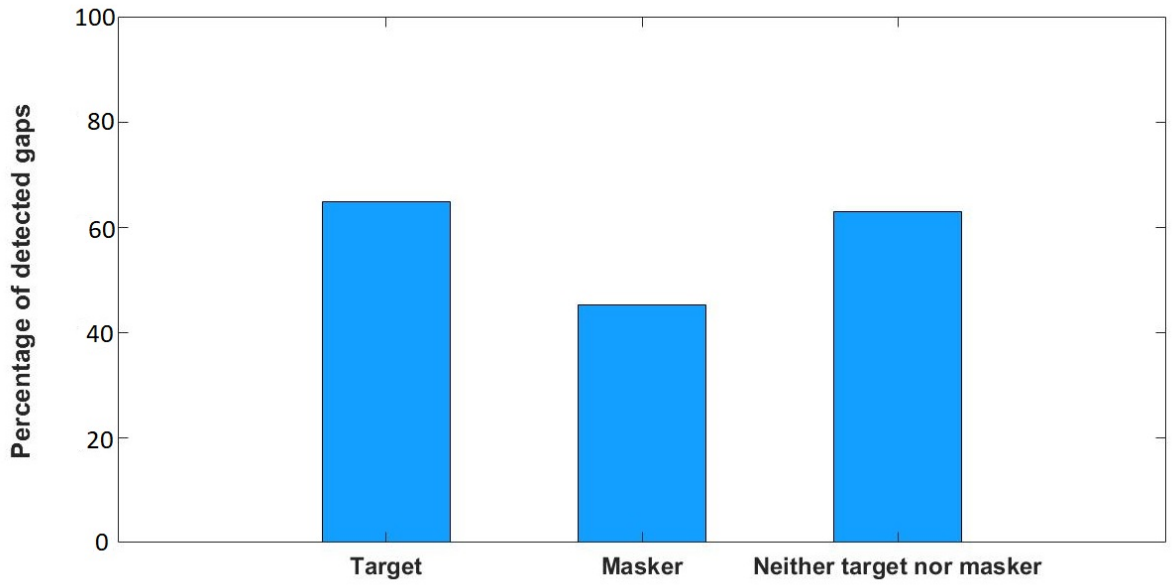

Figure 21: The comparison of number of reported gaps in the target, masker, neither in the target nor in the masker by subject b.

4.3 Analysis of the segregation task performance as a function of speaker separation

4.3.1 Signal detection theory

To evaluate the human perception of a stimulus at the presence of noise, Signal Detection Theory (SDT) was applied. According to this theory, four potential responses can be reported by a human observer. By evaluating these responses, the examiner can asses the participant's personal strategy in the decision-making [1] [17] [18]. Table 3 demonstrates these responses at the presence or absence of a stimulus.

Table 3: The four possible responses to a stimulus categorized according to Signal Detection Theory, adapted from [1].

	Decision: Participant's response	
Reality	Yes	No
Signal Present	Hit	Miss
Signal Absent	False Alarm	Correct Rejection

4.3.2 Analysis of hits

In this study, a hit is defined as the embedded gap in the target sound that is reported in a time window of 1.5 seconds by a subject. For each subject, the number of hits per angle (±7.5°, ±45°, ±90°) was calculated. Table 4 summarizes the initial analysis of hits over all the subjects.

Note that, at each angle, the results of both speakers at the right and left of the head are combined. For instance, the results of +7.5° and -7.5° are merged and presented as the results of ±7.5°.

Table 4: The initial analysis of hits over all the subjects.

Speaker pair	μ (%)	σ (%)	Min (%)	Max (%)
±7.5°	42	11	23	70
±45°	62	13	34	84
±90°	63	12	42	79

To represent the maximum difference between the individual performances, the distribution of hits over each angle for two subjects is compared in figure 22. The bar plot on the left illustrates the results of the subject who had the highest overall hits and the right plot shows the results of the subject who had the poorest performance in terms of overall hits.

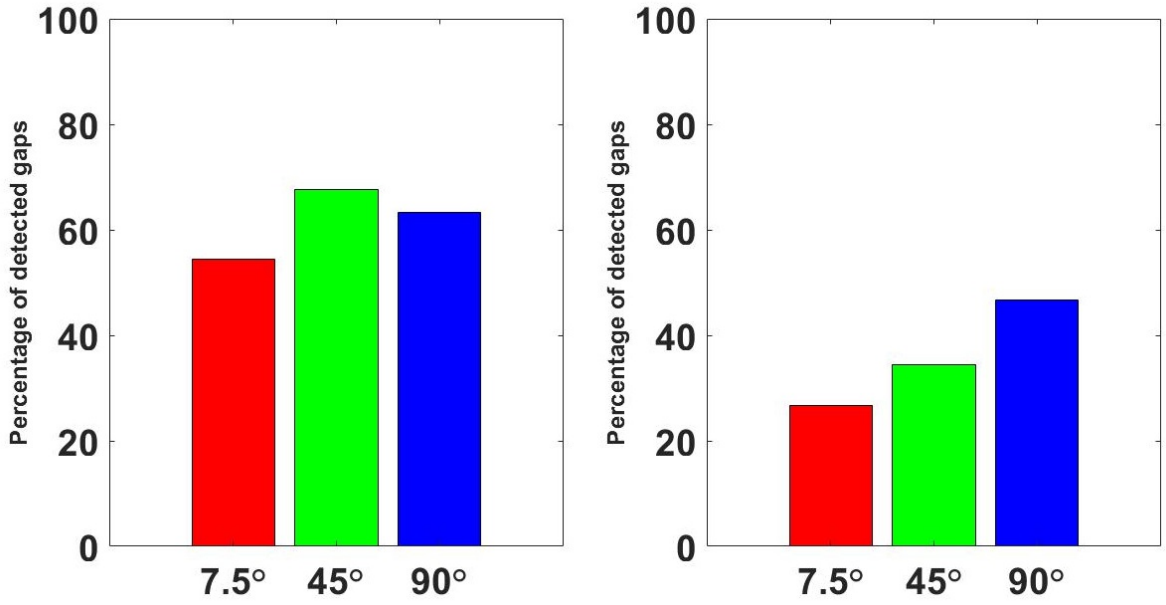

Figure 22: The comparison of the number of hits per angle for the subject who in total had the highest number of hits (left) and the subject who in total had the lowest number of hits (right).

Figure 23 illustrates the distribution of hits as a percentage evaluated for all of the subjects at the respective angles. The median number of hits over all the subjects at a given angle was: 39 % at ±7.5°, 65 % at ±45°, and 63 % at ±90°. The maximum percentage of the gaps detected at ±7.5° was 69%. This maximum value was close to the median observed for the other two angles. The interquartile range observed at ±7.5° falls outside (no overlap) that at ±45° and ±90°. The similarity of the median and the interquartile range of the hits distribution at ±45° and ±90° is also apparent. All the observations made here intimate that the performance of the subjects in the stream segregation task at ±45° and ±90° was comparable. Among the angles discussed, the gap-detection task at ±7.5° was relatively poor.

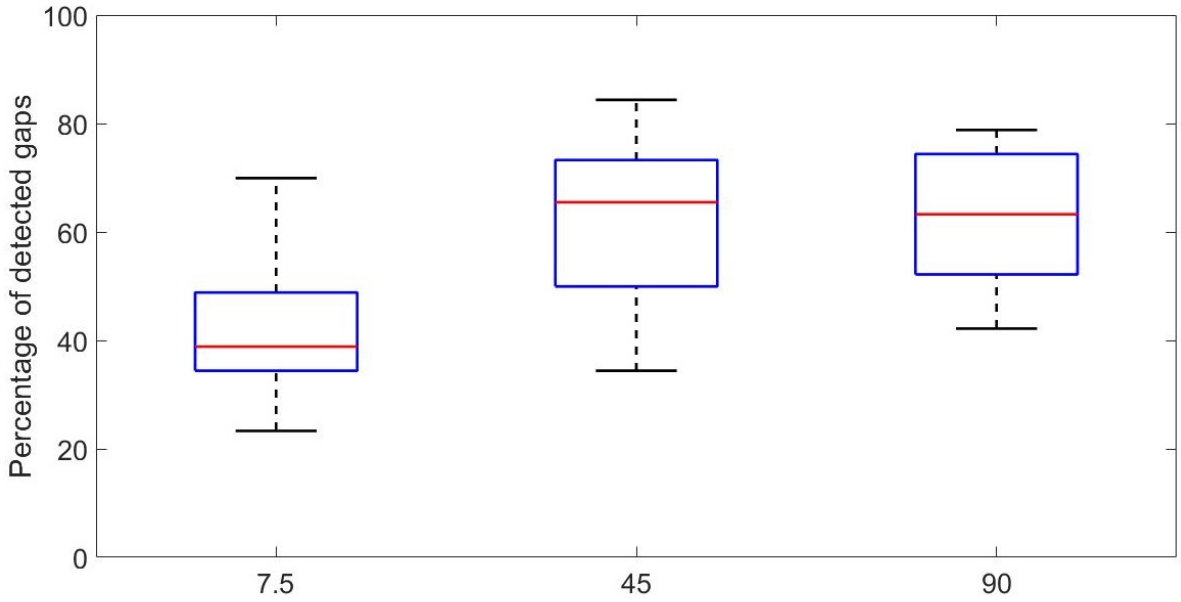

Figure 23: The distribution of the number of hits over all the subjects at each angle.

To define the type of the required statistical data analysis approach, the normality of the above distributions was evaluated by applying the Kolmogorov-Smirnov test to the data from each angle. The results proved that none of the distributions were normal. Therefore, the Kruskal-Wallis test that is a non-parametric statistical analysis method was applied. If the p-value of this test is lower than the significance level, it shows that there was a significant difference between the medians of the hits distributions of the given angles, but it does not specify that the data from which of the pairs was significantly different. The H statistics of the test is calculated as the following;

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{1}$$

Where k is the number of the groups, n_i is the sample size of the group i, R_i is the sum of the calculated ranks of the group i and N is the total sample size which includes the values of all the groups. To evaluate the significance of the results, H is

compared with the chi-squared critical value with the degree of freedom of $k-1$. If H is larger than χ^2_{k-1}, it is concluded that there is a significant difference in this dataset [88–90].

Throughout the data analysis of this study with three groups (data collected at $\pm 7.5°$, $\pm 45°$, and $\pm 90°$) the above variables have the following constant values; $k = 3, n_1 = n_2 = n_3 = 22$ (the number of subjects), $N = 66$, degree of freedom=2. According to [91], in this case, $\chi^2_2 = 5.991$.

The statistics of the Kruskal-Wallis analysis that was applied to the hits distribution of the three angles was $H = 25.24$ with a p-value< 0.00001 which is highly significant. However, as it was mentioned before, this test does not reflect a pairwise p-value. Therefore, as the second step, the Wilcoxon signed-rank test that is a non-parametric test was applied to the pairs to identify the hits distribution of which of the pairs was significantly different. The z-score of the test can be calculated using the test statistics value (W) for the sample size of n as shown in equation 2. The corresponding p-value is accordingly calculated based on the z-score [88, 90, 92–94].

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{2}$$

Table 5 summarises the p-values resulted from applying Kruskal-Wallis and Wilcoxon signed-rank tests on the hits distribution of the speaker pairs.

Table 5: The results of the statistical data analysis applied to the hits distributions at different angles.

Directions	Statistical Test	P-Value
($\pm 7.5°$ & $\pm 45°$ & $\pm 90°$)	Kruskal-Wallis	<0.00001
$\pm 7.5°$ vs. $\pm 45°$	Wilcoxon signed-rank	0.00003
$\pm 7.5°$ vs. $\pm 90°$	Wilcoxon signed-rank	0.00003
$\pm 45°$ vs. $\pm 90°$	Wilcoxon signed-rank	0.7142

The results showed that the overall number of hits at ±7.5° was significantly lower than the number of hits at ±45° and ±90°. This means that the gap-detection and the segregation task were significantly less accurate at this angle in compression with the other two. It was concluded that in terms of the number of hits, the stream segregation ability was significantly affected by speaker separation at ±7.5° vs. ±45° and ±7.5° vs. ±90°. There was no significant difference between the hit distribution at ±45° and ±90°. It means there was no significant difference between the difficulty of performing the task at these two angles.

4.3.3 Analysis of false alarms

False Alarm in this study is defined as the embedded gap in the masker sound that is reported in a time window of 1.5 seconds after the gap was presented. Table 6 illustrates the initial analysis of false alarms over all the subjects.

Table 6: The initial false alarm analysis over all the subjects.

Speaker pair	μ (%)	σ (%)	Min (%)	Max (%)
$\pm 7.5°$	11	11	0	30
$\pm 45°$	9	7	0	22
$\pm 90°$	8	6	0	25

Figure 24 illustrates the difference between the performance of the subject who had the highest overall false alarms (left) and the subject who had the lowest overall false alarms (right).

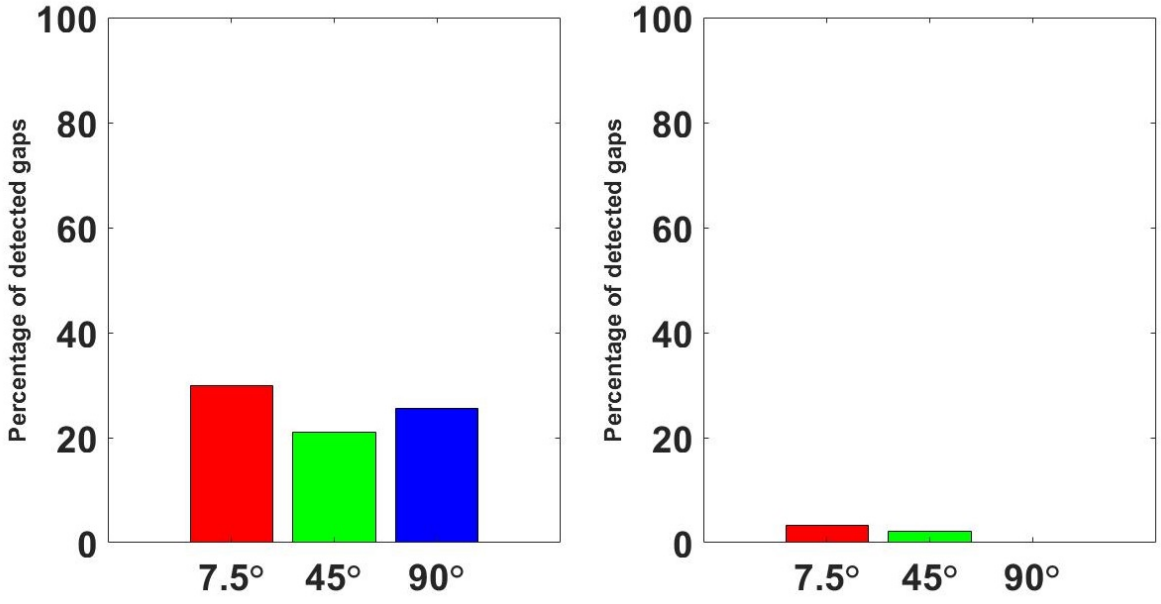

Figure 24: The false alarm comparison between the subjects who had the highest overall false alarms (left) and the subjects who had the lowest overall false alarms (right).

Figure 25 illustrates the distribution of false alarms over all the subjects at each angle. For the three angles shown in the figure, at least one participant was responding with no false alarm. The medians observed here were as follows : 11 % at ±7.5°, 10 % at ±45°, and 8 % at ±90°.

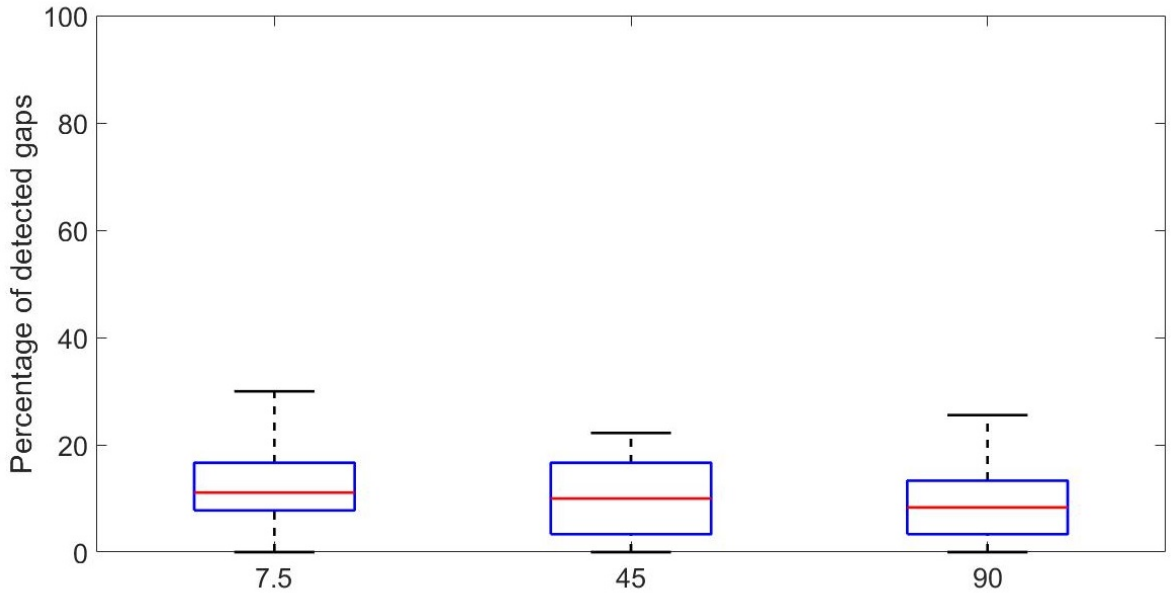

Figure 25: Distribution of false alarms over all the subjects at each angle.

The results of the Kolmogorov-Smirnov normality test demonstrates that none of these datasets were from a normal distribution. Similar to the hits analysis, as false alarms distribution was not normal and there were three angles to compare, the Kruskal-Wallis test was applied. The resulting p-value was 0.5259 with $H = 1.29$ which showed there was no significant difference between the distribution of false alarms of these three angles. This means that in terms of false alarms, speaker separation did not affect the segregation task performance.

4.3.4 Analysis of sensitivity index

The sensitivity index of SDT is calculated as

$$\acute{d} = \mathbf{Z}(HitRates) - \mathbf{Z}(FalseAlarmRates) \tag{3}$$

Where **Z** is the Inverse Cumulative Distribution Function of normal distribution [16–18]. According to this equation, the sensitivity index is advantageous as it takes into account both parameters that reflect a successful task performance; a high number of hits and a low number of false alarms.

Here, the sensitivity index is used as a single value that can reflect the overall performance of the subjects in segregating the target and masker sounds in such a way that a higher sensitivity index represents a relatively more success in performing the sound source separation task. This parameter was calculated for the individual subjects. The overall evaluation of the sensitivity index is given in table 7.

Table 7: The sensitivity index statistics over all the subjects.

Speaker pair	μ	σ	Min	Max
$\pm 7.5^\circ$	1.09	0.55	0.48	2.25
$\pm 45^\circ$	1.77	0.69	0.79	3.19
$\pm 90^\circ$	1.84	0.71	0.86	3.30

The difference between the performance of the subject who had the highest overall sensitivity index and the subject who had the lowest sensitivity index is shown in figure 26.

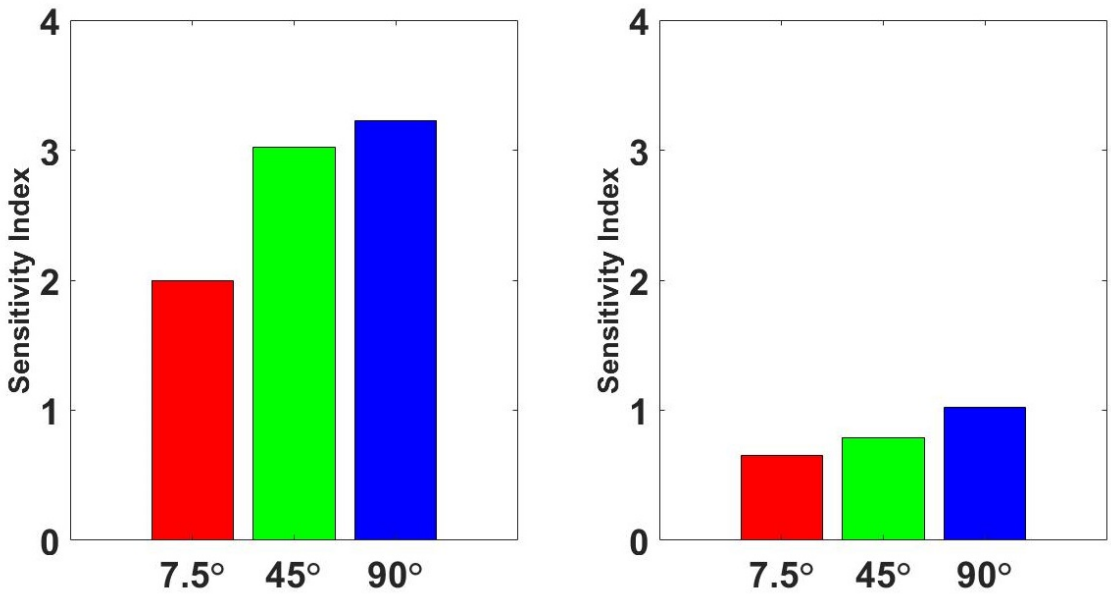

Figure 26: The sensitivity index estimated for the subject who had the highest overall sensitivity index (left) and the subject who had the lowest overall sensitivity index (right).

Figure 27 demonstrates the distribution of the sensitivity index value at each angle. The interquartile range of the distribution at $\pm 90^\circ$ was 0.71. The third quartile was 2.07. Therefore, the upper fence of the boxplot was 3.13 $((0.71 \times 1.5) + 2.07)$. As the highest sensitivity indices at this angle were above 3.13 (3.16, 3.23, 3.30), they were shown as outliers.

To select the appropriate statistical data analysis approach, the Kolmogorov-Smirnov test was applied. It demonstrated that these distributions were not normal at any angle, therefore, the Kruskal-Wallis test was applied to the sensitivity index values of all the subjects. The resulting p-value was 0.00029 with $H = 16.24$ which showed there was a significant difference. Similar to the approach used to analyze the hits, here, in order to compare the results in pairs, the Wilcoxon signed-rank test was applied. Table 8 summarises the results of the statistical data analysis applied to the sensitivity index.

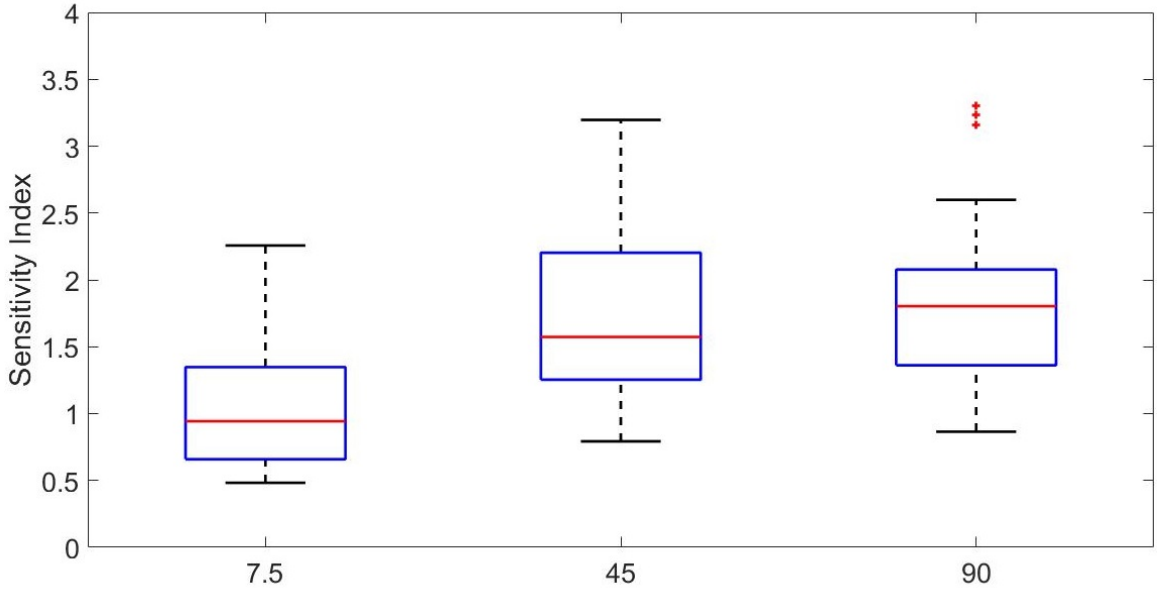

Figure 27: Distribution of the sensitivity index evaluated for all the subjects at each angle.

Table 8: The results of the statistical data analysis applied to the sensitivity index.

Direction(s)	Statistical Test	P-Value
($\pm 7.5°$ & $\pm 45°$ & $\pm 90°$)	Kruskal-Wallis	0.00029
$\pm 7.5°$ vs. $\pm 45°$	Wilcoxon signed-rank	0.00004
$\pm 7.5°$ vs. $\pm 90°$	Wilcoxon signed-rank	0.00004
$\pm 45°$ vs. $\pm 90°$	Wilcoxon signed-rank	0.3896

The results of $\pm 7.5°$ vs. $\pm 45°$ and $\pm 7.5°$ vs. $\pm 90°$ were significant, meaning the overall performance of the subjects which here is measured with the sensitivity index was significantly poorer at $\pm 7.5°$ in comparison with $\pm 45°$ and $\pm 90°$. This shows that the segregation task at $\pm 7.5°$ is the most difficult among the three angles.

The results of $\pm 45°$ vs. $\pm 90°$ show that there was no significant difference between the performance of the subjects at these two angles and their different amount of angular distance in space did not cause a statistically significant difference on the gap-detection ability.

4.4 Analysis of the segregation task performance as a function of the difference in frequency content

In the following, the gap-detection and segregation task performance as a function of the difference between the frequency content of the target and masker is analyzed. The carrier frequencies of the sinusoids which were summed to create the target sound (narrowband noise) were averaged and represented as FC_T. The carrier frequencies of the sinusoids which were summed to create the masker sound (narrowband noise) were averaged and represented as FC_M. To calculate the difference between the frequency content of the target and masker, the difference between their average carrier frequencies was considered. This difference is represented as FC_D (as shown in equation 4). This was to evaluate how FC_D influenced the auditory segregation task at each angle and to compare this effect at different angles.

$$FC_D = FC_T - FC_M \tag{4}$$

As it was previously mentioned, the target frequency range was 900-1100 Hz. The highest possible masker frequency range was 1250-1450 Hz. The lowest possible masker frequency range was 550-750 Hz. Hence, considering the means of the above ranges, FC_D was in the range of -350 (650 - 1000) Hz to +350 (1350 - 1000) Hz. The FC_D range (700 Hz) was divided into 10 bins, each with the frequency range of 70 Hz. The mid-frequency of the bins were as follows; -315, -245, -175, -105, -35, 35, 105, 175, 245, 315 Hz.

The number of hits per bin was divided by the number of the generated gaps per that bin and represented as Hit Rate (HR). The distribution of HR over FC_D suggested a non-linear relationship between the two. Therefore, a quadratic polynomial regression

was used to model the distributions of all the angles. Equation 5 shows the relationship between the observed data that was modeled ($\mathbf{Y}$) and the quadratic polynomial fit ($\mathbf{X}$).

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{X}^2 + \epsilon \tag{5}$$

Where ϵ represents a random error and $\beta_0, \beta_1, \beta_2$ are the regression coefficients. The coefficients are calculated based on the number of data-points (i) that are used to create the model as the following [95–97];

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_i & x_i^2 \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \end{bmatrix} \tag{6}$$

In this study, for each distribution, 10 data points (10 bins) were used to find the coefficients ($i = 10$). After estimating the coefficients, further data analysis was done based on the comparison of β_2; Quadratic Coefficient (QC) of different distributions at each angle.

Figure 28 shows a hypothetical HR over FC_D distribution over the 10 bins. The fit polynomial is shown as a black curve. In this example, as FC_D increases, the HR increases. This means, at large FC_D, it was more likely that the subject successfully segregates the two auditory streams. On the other hand, when FC_D was low (middle of the range), HR decreases. This implies that the segregation task was poorly performed when FC_T and FC_M were more similar.

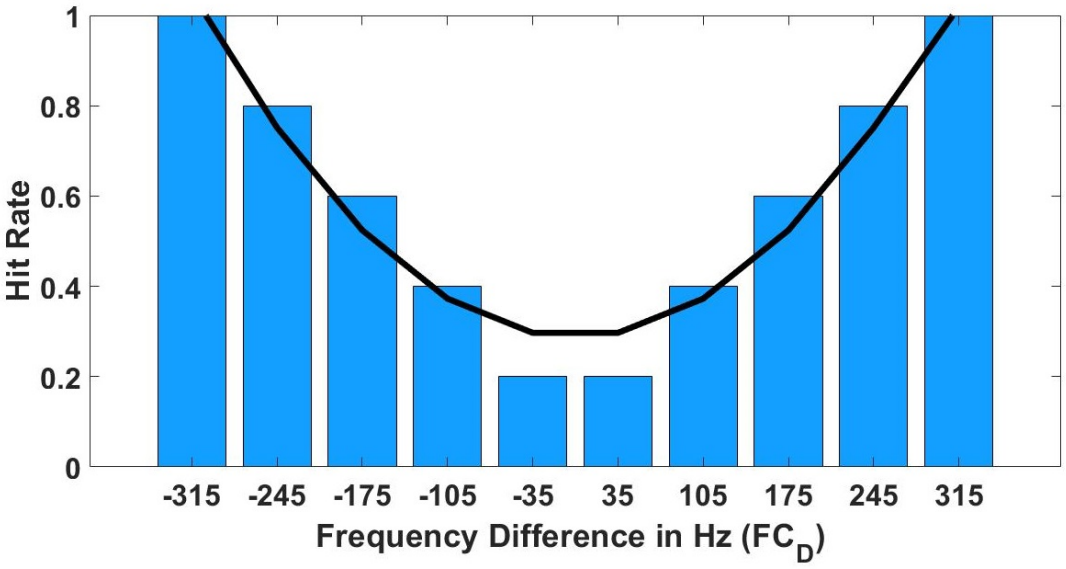

Figure 28: A hypothetical scenario showing the relation between HR and FC_D over 10 bins. The black curve shows the fit polynomial.

Figure 29 illustrates the QC of the polynomial fit calculated for each subject at each angle.

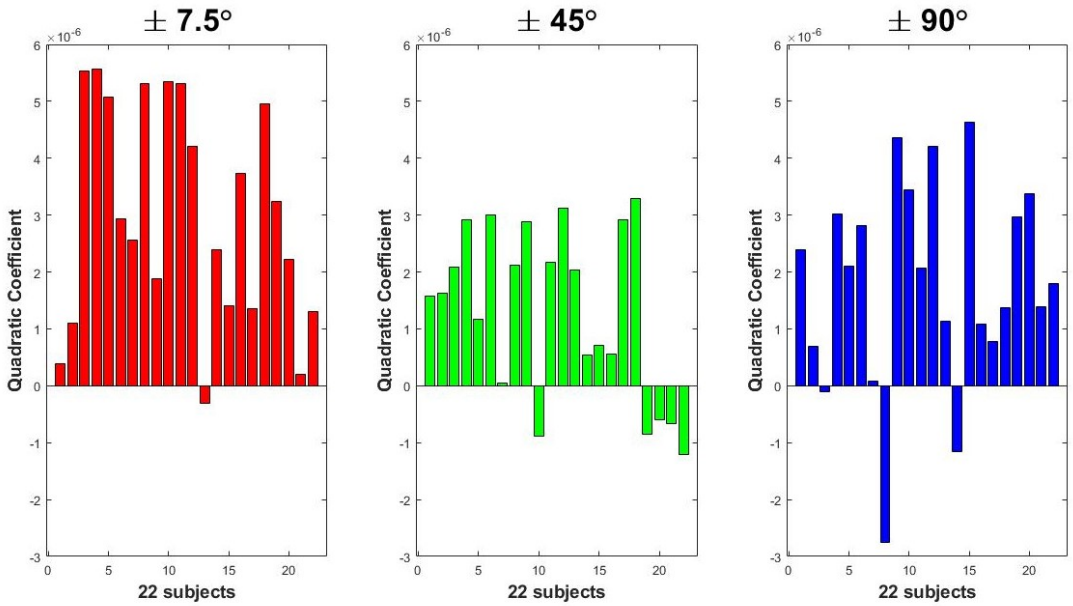

Figure 29: The QC of the polynomial of degree two fit to the distribution of HR over FC_D for individual subjects.

To evaluate the validity of fitting a quadratic polynomial to data to model it, the Coefficient of Determination (CD) was calculated as a goodness of fit evaluation approach. CD (represented as R^2) is a measure that reflects the degree to which the applied statistical model fits the actual collected data as shown in equation 7.

$$R^2 = 1 - \frac{\sum_{n=1}^{i}(y_n - x_n)^2}{\sum_{n=1}^{i}(y_n - \bar{y})^2} \tag{7}$$

Where i is the number of data points, y is the observed data, $\bar{y}$ is the mean of the obsereved data, and x represents the corresponding estimated model [98–101].

CD calculated over the average performance of the subjects was as the followings; $R^2 = 0.83$ at $\pm 7.5°$, $R^2 = 0.67$ at $\pm 45°$, $R^2 = 0.77$ at $\pm 90°$. According to this assessment, it was concluded that this method would be a suitable data modeling approach. The hypothetical data shown in figure 28 had a CD of 0.94.

In the following, the results of this analysis on the data from two subjects are described in more detail. Figure 30, top: The results of this subject represents a very similar HR distribution to the hypothetical data at $\pm 7.5°$ with $R^2 = 0.88$ and at $\pm 90°$ with $R^2 = 0.70$. Figure 30, bottom: The analysis of the results of this subject showed the highest positive QC at $\pm 7.5°$ (this can also be seen in figure 29). This means that the difference between the performance of this subject where FC_D was small (bin 5 and bin 6) and where FC_D was large (bin 1 and bin 10) was the largest among the subjects.

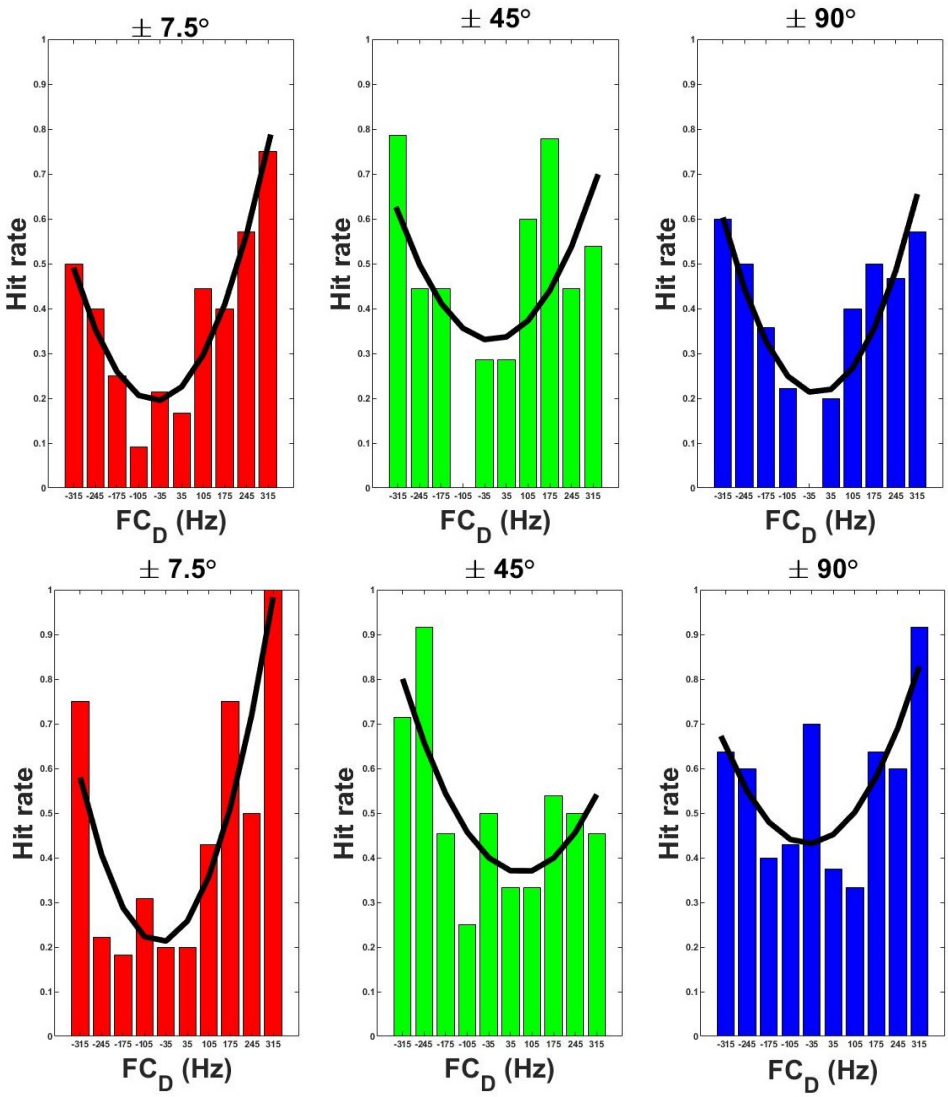

Figure 30: The distribution of HR over the FC_D range for two subject.

Figure 31-top demonstrates the average of the HR of all the subjects and the corresponding errorbar of each bin. A quadratic polynomial is fit to this mean distribution as well. The similarity of the distribution at $\pm 45^\circ$ and $\pm 90^\circ$ over all the subjects and high curvature of the polynomial at $\pm 7.5^\circ$ can be seen.

Among all the trials, where FC_D was a small value (in the range of -70 to 70 Hz) at $\pm 7.5^\circ$, on average, the subjects had the poorest performance of stream segregation as they were able to detect 26% of the gaps, correctly. This value increased to 55% at $\pm 45^\circ$ and to 50% at $\pm 90^\circ$.

The number of false alarms per bin was divided by the number of the generated gaps per that bin and represented as False Alarm Rate (FAR). Figure 31-middle shows the mean FAR over all the subjects at 10 bins.

Figure 31-bottom shows the mean sensitivity index over all the subjects. Similar to HR, the sensitivity index had a larger value where the frequency difference between the target and masker was higher. The highest difference between the sensitivity index at different bins can be seen at $\pm 7.5^\circ$.

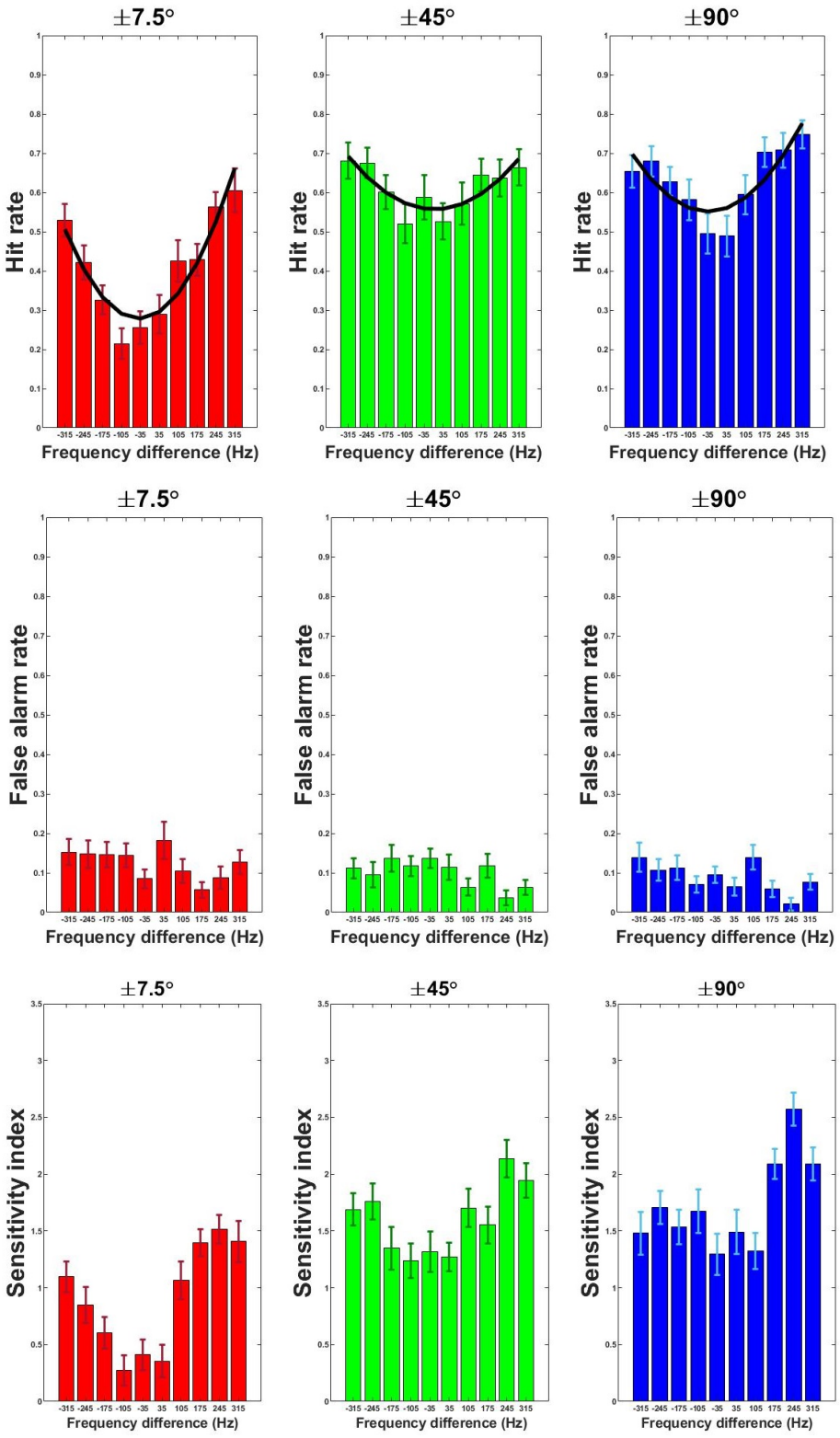

Figure 31: The average HR (top), the average FAR (middle), and the average sensitivity index (bottom) relative to FC_D over all the subjects at each angle.

4.4.1 Analysis of quadratic coefficients

To compare the effect of FC_D over all the subjects at different angles, statistical data analysis was applied to QC of all the distributions. Figure 32 shows the distribution of QC over all the subjects. The Kolmogorov-Smirnov test showed that none of the following distributions were normal.

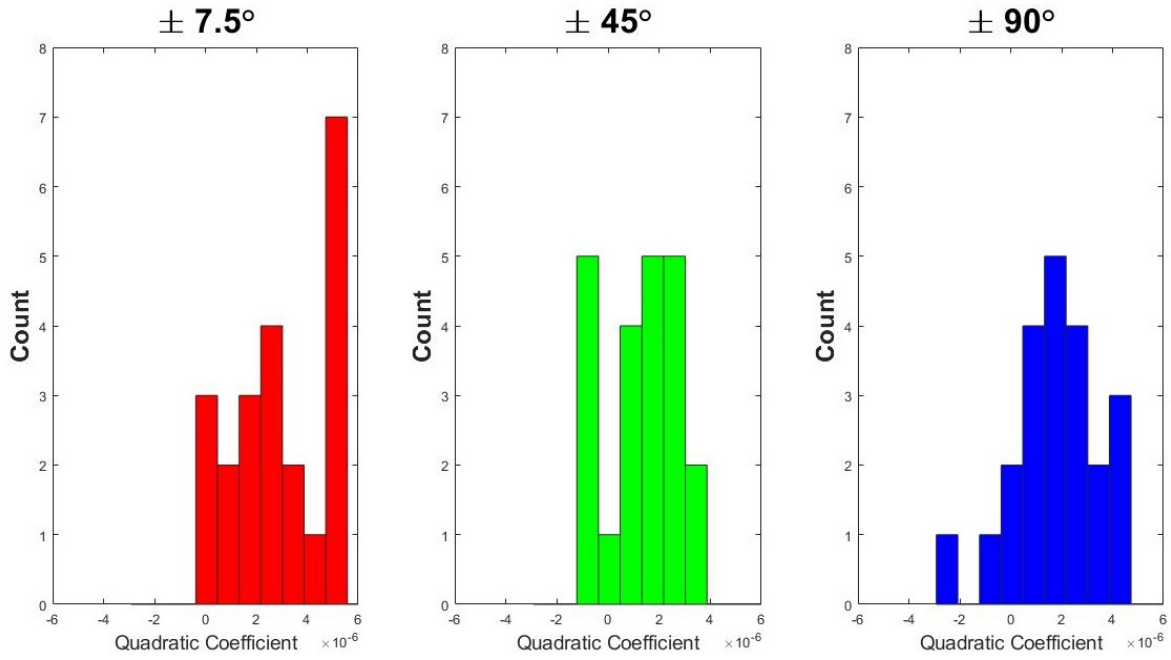

Figure 32: The distribution of quadratic coefficients over all the subjects at each angle.

Step 1: To determine whether, regardless of the sound direction of arrival, FC_D is a stream segregation cue or not, at first, a dataset was made of all the QC calculated for the three angles. The Wilcoxon signed-rank test was applied to this overall dataset. The result was highly significant, $p < 0.00001$ which showed FC_D generally played the role of a segregation cue.

Step 2: To determine the effect of FC_D on stream segregation cue at a single angle (whether HR has changed over FC_D range or not), the Wilcoxon signed-rank test

was applied to QC corresponding to that particular angle. All the p-values of this analysis showed a significant difference; $p < 0.0001$ at $\pm 7.5°$, $p = 0.0027$ at $\pm 45°$, $p = 0.0007$ at $\pm 90°$. This showed that at each angle, FC_D played the role of a segregation cue and the segregation ability was significantly affected by FC_D changes.

Step 3: To determine if this effect was significantly different among the angles, the Kruskal-Wallis test was applied to the overall dataset of QC. The results of this analysis was $p = 0.0221$ with $H = 7.63$. It showed that there was a significant difference where comparing QC at different angles.

Step 4: To clarify which speakers have a significant difference, the Wilcoxon signed-rank test was applied on the distribution of QC of the three pairs. The following p-values were calculated as the result of this analysis; $p = 0.0027$ at $\pm 7.5°$ vs. $\pm 45°$, $p = 0.0619$ at $\pm 7.5°$ vs. $\pm 90°$, and $p = 0.3896$ at $\pm 45°$ vs. $\pm 90°$. Table 9 summarises the results of quadratic coefficients statistical data analysis.

Table 9: The results of the Quadratic Coefficients analysis.

Direction(s)	Statistical Test	P-Value
($\pm 7.5°$ & $\pm 45°$ & $\pm 90°$)	One-sample Wilcoxon signed-rank	<0.0001
$\pm 7.5°$	One-sample Wilcoxon signed-rank	<0.0001
$\pm 45°$	One-sample Wilcoxon signed-rank	0.0027
$\pm 90°$	One-sample Wilcoxon signed-rank	0.0007
($\pm 7.5°$ & $\pm 45°$ & $\pm 90°$)	Kruskal-Wallis	0.0221
$\pm 7.5°$ vs. $\pm 45°$	Two-samples Wilcoxon signed-rank	0.0027
$\pm 7.5°$ vs. $\pm 90°$	Two-samples Wilcoxon signed-rank	0.0619
$\pm 45°$ vs. $\pm 90°$	Two-samples Wilcoxon signed-rank	0.3896

This outcome showed that FC_D significantly affected the segregation ability of the subjects at $\pm 7.5°$ in comparison with $\pm 45°$. The fact that the effect of FC_D at $\pm 45°$ and $\pm 90°$ was not significantly different suggested that the effect of the difference in frequency content is very similar at these two angles and speaker separation did not make a significant change in the gap-detection task. The very low p-value at $\pm 7.5°$ vs.

$\pm 90^{\circ}$ showed that the frequency content difference highly affected the stream segregation performance when comparing these two angles.

4.5 Analysis of the segregation task performance as a function of the modulation-rate difference

In the following, the gap-detection and segregation task performance as a function of the difference between the modulation-rate (modulating frequency) of the target and masker is analyzed. The modulation-rate of the target is represented as FMOD_T. The modulation-rate of the masker is represented as FMOD_M. The difference between the modulation-rate of the target and masker is represented as FMOD_D (as shown in equation 8). This was to evaluate how FMOD_D influenced the auditory segregation task at each angle and to compare this effect at different angles.

$$FMOD_D = FMOD_T - FMOD_M \tag{8}$$

As it was previously mentioned, the target modulating frequency range was 1.5 - 2.5 Hz. The masker modulating frequency range was 0.7 - 3.3 Hz. Hence, FMOD_D was in the range of -1.8 (3.3 - 2.5) to 1.8 (2.5 - 0.7) Hz. This range was divided into 10 bins with the bin-frequency range of 0.36 Hz and bins mid-frequency of -1.62, -1.26, -0.90, -0.54, -0.18, 0.18, 0.54, 0.90, 1.26, 1.62 Hz.

The results of the modulation-rate analysis of three subjects are shown in figure 33. The results illustrate an approximately random distribution.

The average performance of all the subjects over the 10 bins in terms of HR (top) and FAR (bottom) over the FMOD_D range is shown in figure 34. The HR distributions were approximately uniform. This illustrates that, on average, the HR was not affected by the FMOD_D at any angle.

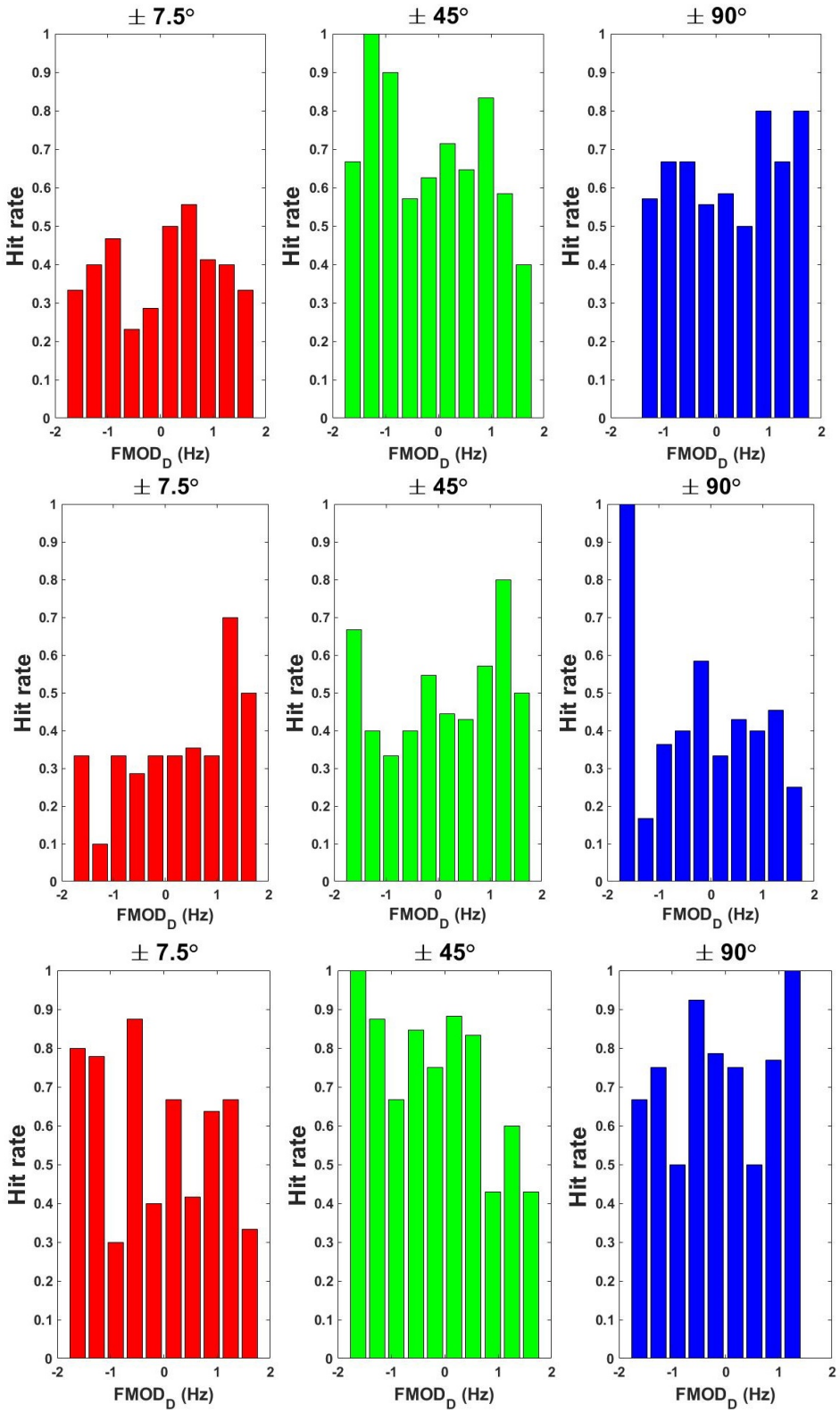

Figure 33: The distribution of HR over the FMOD$_D$ range for three subjects.

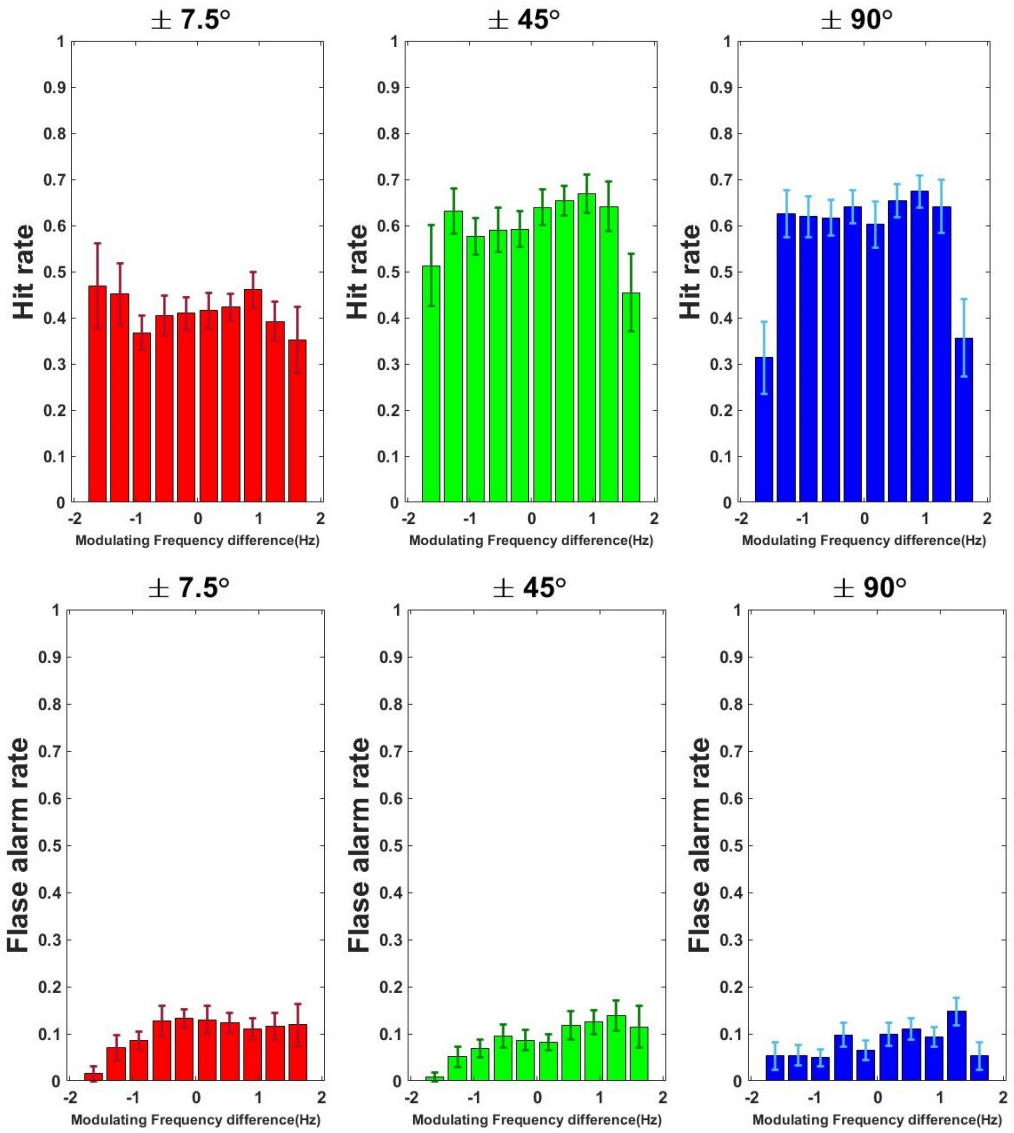

Figure 34: The average HR (top), FAR (bottom) relative to FMOD_D over all the subjects.

4.5.1 Analysis of quadratic coefficients

As it was previously mentioned, by fitting a quadratic polynomial to the HR distribution, the difference between HR over different FMOD_D from the edges to the vertex of the curve can be evaluated. If QC of a polynomial is low and close to zero, the curve shape would be instead close to a straight line, meaning the data is approximately uniformly distributed and HR at the edges and vertex are similar. In other words, FMOD_D

did not change the HR. This means that the rate difference (modulating frequency difference) between the target and masker did not affect the stream segregation performance. Figure 35 summarizes this logic flow.

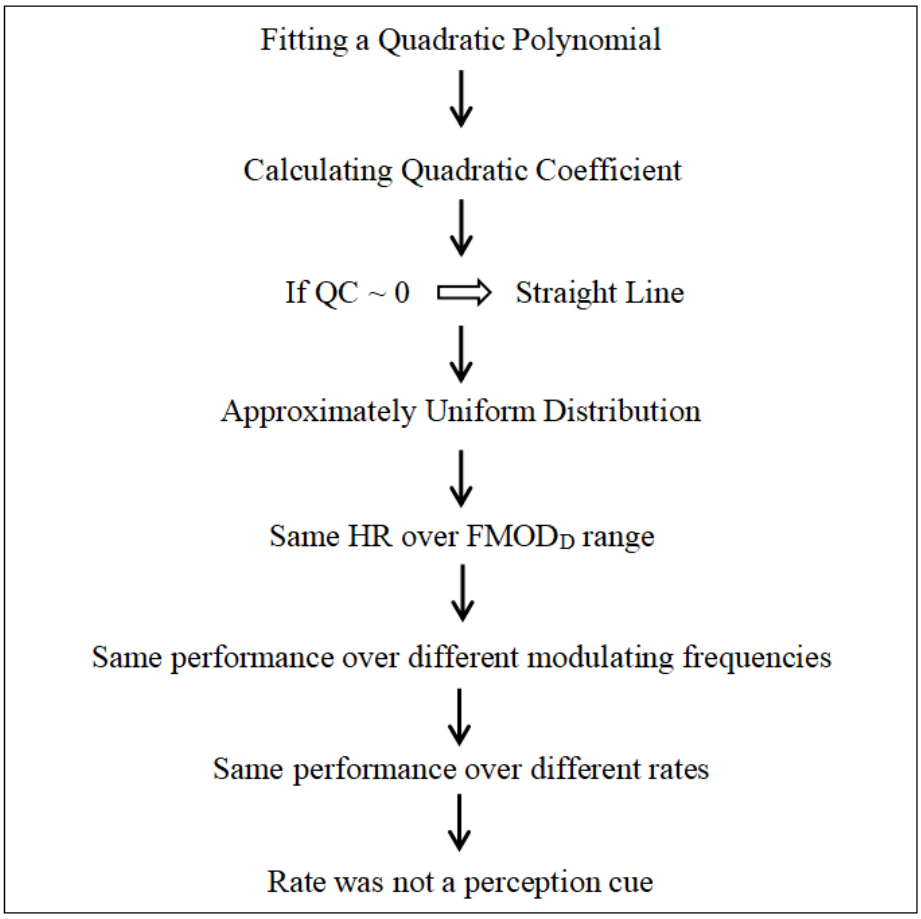

Figure 35: The logic flow of data analysis approaches applied to define the effect of rate on segregation task.

This process was done for all the 66 distributions (22 subjects and 3 angles per subject). The results are shown in figure 36. The Kolmogorov-Smirnov test was applied to these distributions and the results showed that none of them were normality distributed. Therefore, to further analyze this data, a non-parametric test was applied.

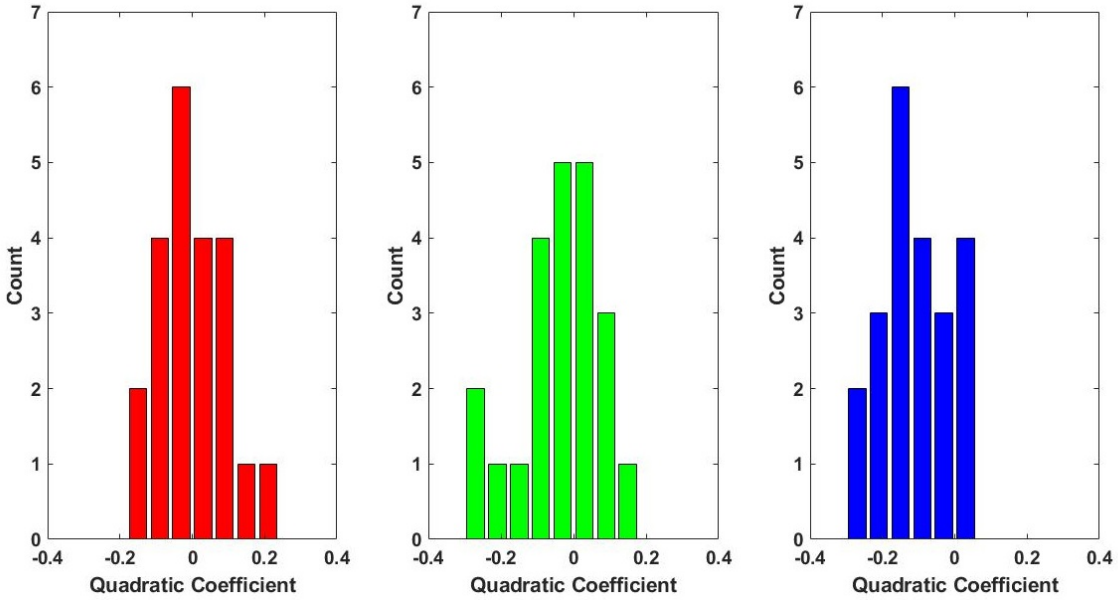

Figure 36: The distribution of quadratic coefficient over all the subjects.

To evaluate the strength of the relationship between FMOD_D and HR, the permutation test which is a non-parametric test was applied on QC of all the 66 distributions. The permutation test takes into account both the condition information and the performance information. Here, FMOD_D is the condition information and HR is the performance information. It illustrates how likely is to achieve the same performance information if the condition information changes.

To apply the test, the performance information is shuffled many times in such a way that each time, a new performance information distribution is assigned to the same condition information.

The null hypothesis of the permutation test is all of these permutations which are the results of rearranging the distribution of the collected data have the same probability to happen. In other words, if the calculated p-value of this test does not show a significant change ($p > 0.05$), it can be concluded that the test value is randomly likely to happen and the observed data does not show a significant impact of the condition information on

the performance information.

For a sample size made of n values, $n!$ permutations can be defined. For a large sample size, implementing $n!$ permutations might be computationally costly. Therefore, it is suggested that permuting the results 1000 times should be sufficient. More repetition would be needed in case the resulting p-value would be smaller than 0.001 [102]. The p-value of the test is calculated as the following

$$p = P(|T| > t_{obs}|H_0) \tag{9}$$

Where t_{obs} is the statistics of the observed data, H_0 is the null hypothesis and T is the collection of the test statistics calculated in each permutation [90, 97, 103–106]. Here, the test statistic is the quadratic coefficient at each permutation.

To apply the test, for each subject at each angle, two datasets of size 90 were defined; 1. The FMOD_D of each of 90 embedded gaps in that direction, 2. The corresponding status of the generated gaps; Hit: 1, Miss: 0 (at a speaker pair with the HR of n, 90-n gaps are not detected and missed).

Next, the second dataset was shuffled 1000 times and each time a new HR distribution over the 10 bins was achieved. This means that at each repetition, the distribution of hits and misses were changed over the same FMOD_D vector. The QC corresponding to each of the distributions was calculated, accordingly.

Using the achieved permutation distribution, the ratio of the number of the newly calculated QC that were larger than the original QC (which was calculated from the distribution of the observed data) and the total number of the calculated QC (here 1000) was calculated as a one-tail p-value of the permutation test. This process is illustrated in figure 37.

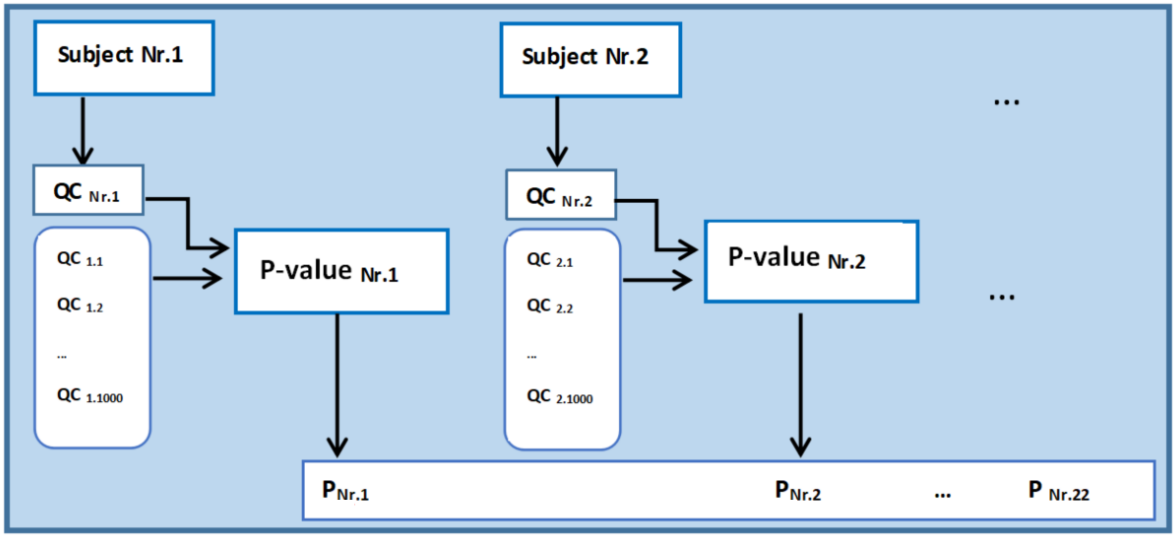

Figure 37: The summary of the procedure of applying the permutation test on QC.

Figure 38 shows the distribution of the 1000 quadratic coefficients calculated for 2 subjects. The black circle and the corresponding black line show the position of the QC of the observed data relative to the 1000 quadratic coefficients which were calculated in the test.

Figure 39 shows the wide range of p-values calculated over all the subjects at each angle. It was concluded that at all three angles, performance information (HR) did not have a strong correlation with the condition information (FMOD_D) as 64 out of 66 p-values were not significant.

In other words, the difference between the modulating frequency of the target and masker did not improve or impair the subject's segregation ability. Therefore, the rate difference between the target and masker did not play the role of a cue for auditory stream segregation task at any angle.

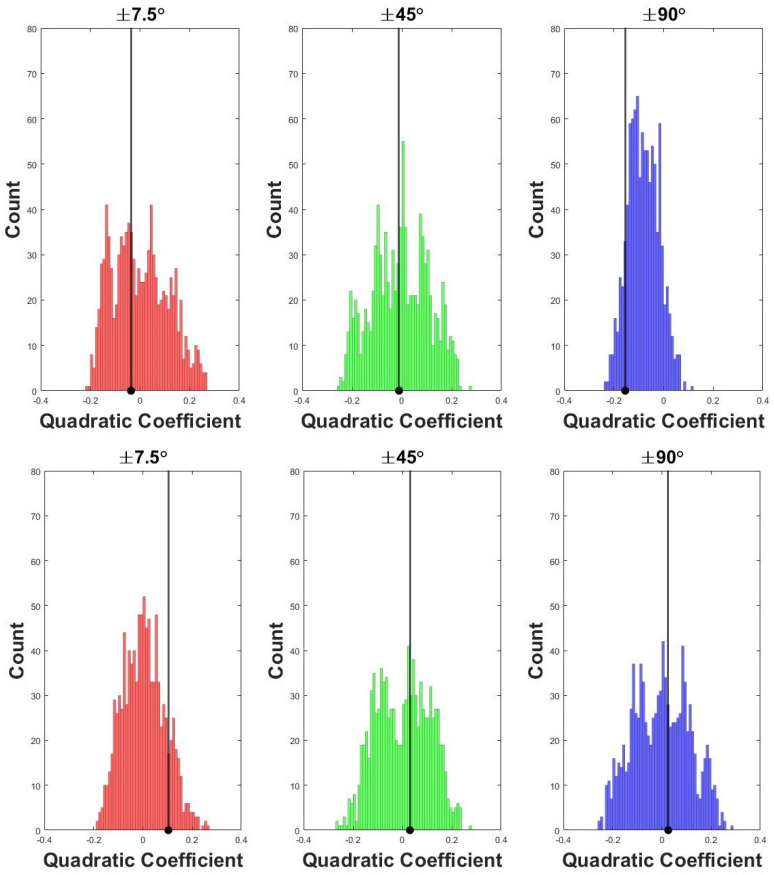

Figure 38: The results of applying the permutation test for two subjects. The distribution of the QC is shown here. The black circle and the corresponding black line show the position of the originally calculated QC of the observed data.

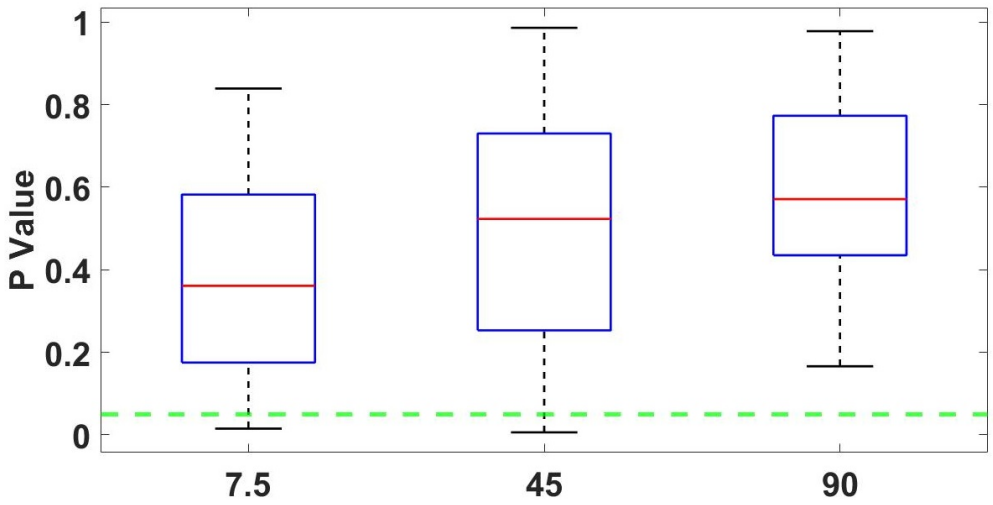

Figure 39: The distribution of p-value of the permutation test results over all the subjects. The Green horizontal line shows the p=0.05 for an easier comparison.

So far, the result of FC_D and $FMOD_D$ analysis proved that rate was not a segregation task cue, but frequency content was a strong cue for segregating the channels. Consequently, a more detailed question raised. "Would $FMOD_D$ play the role of a segregation cue in case FC_D is small?" In other words, "Do subjects rely on the change in rate to perform the task when the difference in the frequency content is not so informative and does not provide a cue?"

To answer this question, the trials of the test in which the $|FC_D|$ was a small value were needed to be analyzed. As it was mentioned, the frequency content of the target falls into critical band Nr. 9. Therefore, the trials in which their masker frequency content falls into the same critical band as target (band Nr.9) and the two neighbor critical bands (i.e. band Nr. 8 with the center frequency of 840 Hz and band Nr. 10 with the center frequency of 1175 Hz) were chosen. In these trials, $|FC_D|$ was less than 175 Hz.

For each speaker pair, these trials which all had $|FC_D| < 175$ were divided into three groups based on their $FMOD_D$ value;
Group 1: The trials with $0 \leq |FMOD_D| < 0.6$ Hz
Group 2: The trials with $0.6 \leq |FMOD_D| < 1.2$ Hz
Group 3: The trials with $1.2 \leq |FMOD_D| \leq 1.8$ Hz

Next, the distribution of HR over each group was determined (figure 40-top). The results of applying the Kolmogorov-Smirnov illustrate that none of the 9 estimated HR distributions were normal. Therefore a non-parametric test was used to analyze the data. The Kruskal Wallis test was applied to the HR of the three groups of each speaker pair.

The corresponding p-values ($p = 0.8635$ at $\pm 7.5°$, $p = 0.9027$ at $\pm 45°$, $p = 0.3006$ at $\pm 90°$) showed that there was no significant difference between the HR of different groups for any speaker pair, therefore a two by two comparison was not needed. These results

showed that in case the $|\mathrm{FC}_D|$ was a small value (here less than 175 Hz), HR was not significantly different at different groups. This means that the segregation ability was not affected by FMOD_D and the rate did not provide a segregation cue even when the difference between the frequency content of the two streams was small and did not provide a strong segregation cue. In other words, the subjects could not rely on the modulation-rate difference to perform the task regardless of the strength of the spatial and spectral cues.

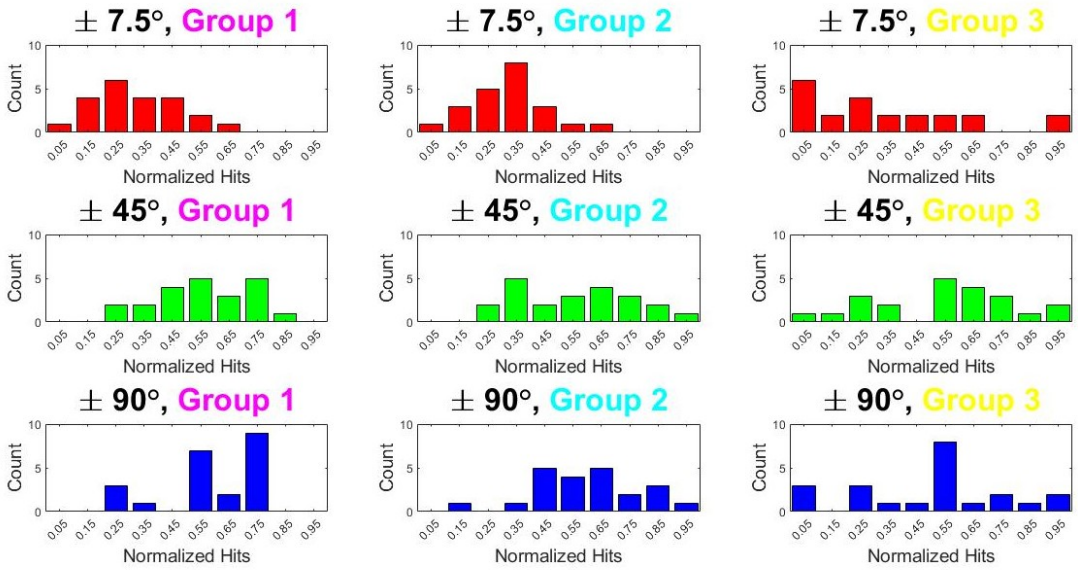

Figure 40: The distribution of HR in different groups over all the subjects. For all the groups $|\mathrm{FC}_D| \leq 175Hz$. Group 1 (magenta): $0 \leq |\mathrm{FMOD}_D| < 0.6$ Hz. Group 2 (cyan): $0.6 \leq |\mathrm{FMOD}_D| < 1.2$ Hz. Group 3 (yellow): $1.2 \leq |\mathrm{FMOD}_D| \leq 1.8$ Hz.

Table 10 illustrates the average normalized number of hits over all the subjects. The small difference between the mean HR of each group can be observed.

Table 10: The average HR over all the subjects in each group.

Speaker pair	Group 1 (μ)	Group 2 (μ)	Group 3 (μ)
±7.5°	0.3203	0.3136	0.3266
±45°	0.5566	0.5462	0.5243
±90°	0.5883	0.5887	0.4863

4.6 Analysis of reaction time

The Reaction Time (RT) is defined as the time difference between the moment that the generated gap was presented to the subject and the time that the subject pushed the silence button to report the detected gap. To evaluate the difference between the RT of the reported gaps at different angles, for each subject, this time difference was calculated for hits and false alarms. Figure 41 shows the distribution of RT of hits for three subjects.

To make an overall assessment on RT of all the subjects, the following approach was applied. A dataset was made of the median value of RT of all the detected gaps for each subject (separately for HR and FAR). Then the distribution of these median values was evaluated over all the subjects (HR: figure 42, FAR: figure 43). The Kolmogorov-Smirnov test showed none of the distributions were normally distributed.

The Kruskal-Wallis test was applied to the median RT distribution of the hits. The result was $p < 0001$ which showed a highly significant effect. The Wilcoxon signed-rank test was applied to compare the reaction times of different angles two by two. The following p-values were calculated; $p < 0001$ at $\pm 7.5^\circ$ vs. $\pm 45^\circ$, $p < 0001$ at $\pm 7.5^\circ$ vs. $\pm 90^\circ$, and $p = 0.4262$ at $\pm 45^\circ$ vs. $\pm 90^\circ$.

This analysis showed that there was no significant difference in RT at $\pm 45^\circ$ vs. $\pm 90^\circ$. While the subjects were significantly slower in responding to gaps at $\pm 7.5^\circ$ in comparison with the other two angles meaning when spatial cues were not strong (the speakers were close to each other), it took a longer time for subjects to segregate the streams and make a decision to report gaps in the target stream. The Kruskal-Wallis test p-value of FAR was $p = 0.3643$ which showed in terms of FAR, there was no significant difference between the responses at any angle.

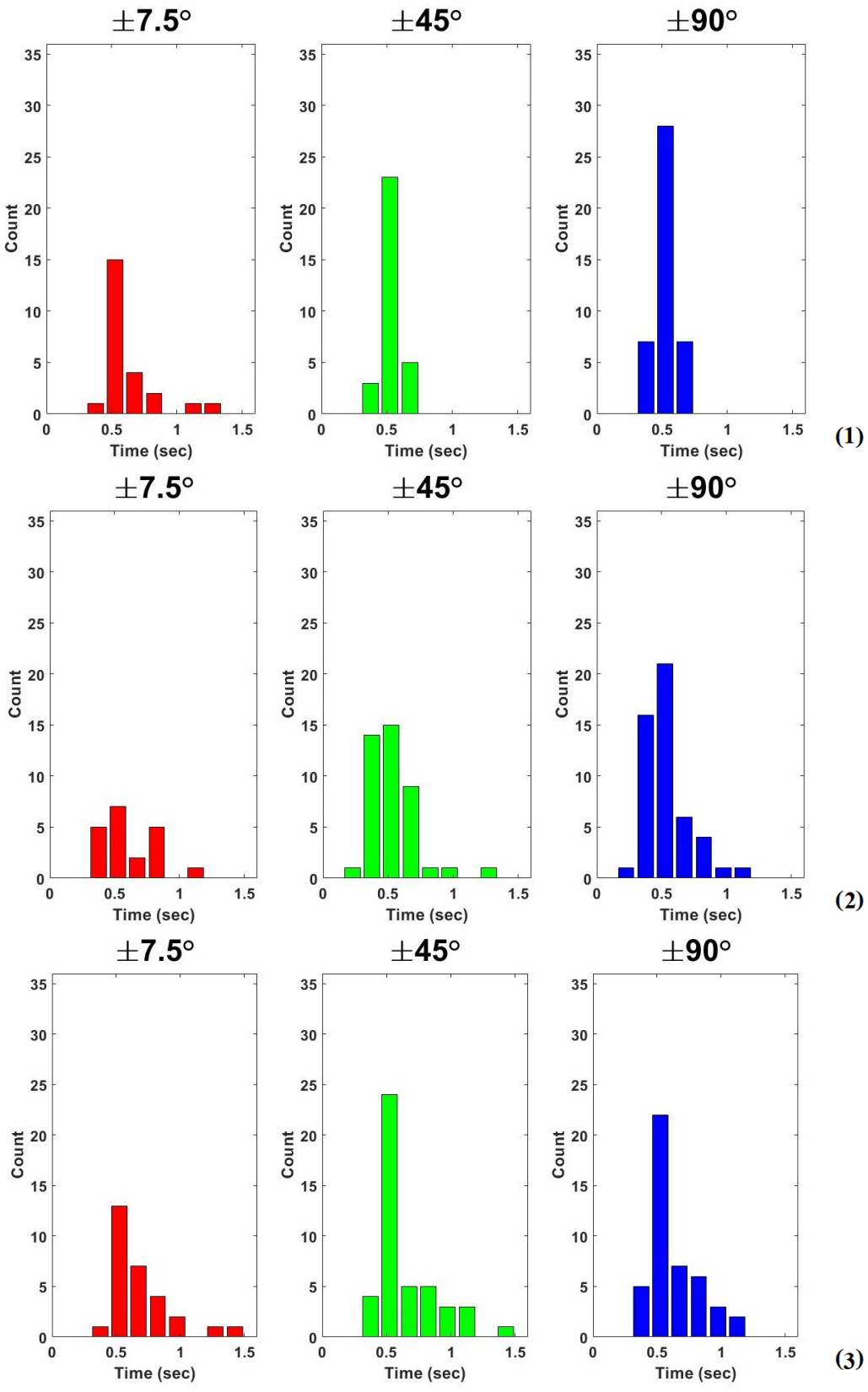

Figure 41: The distribution of RT of HR for three subjects.

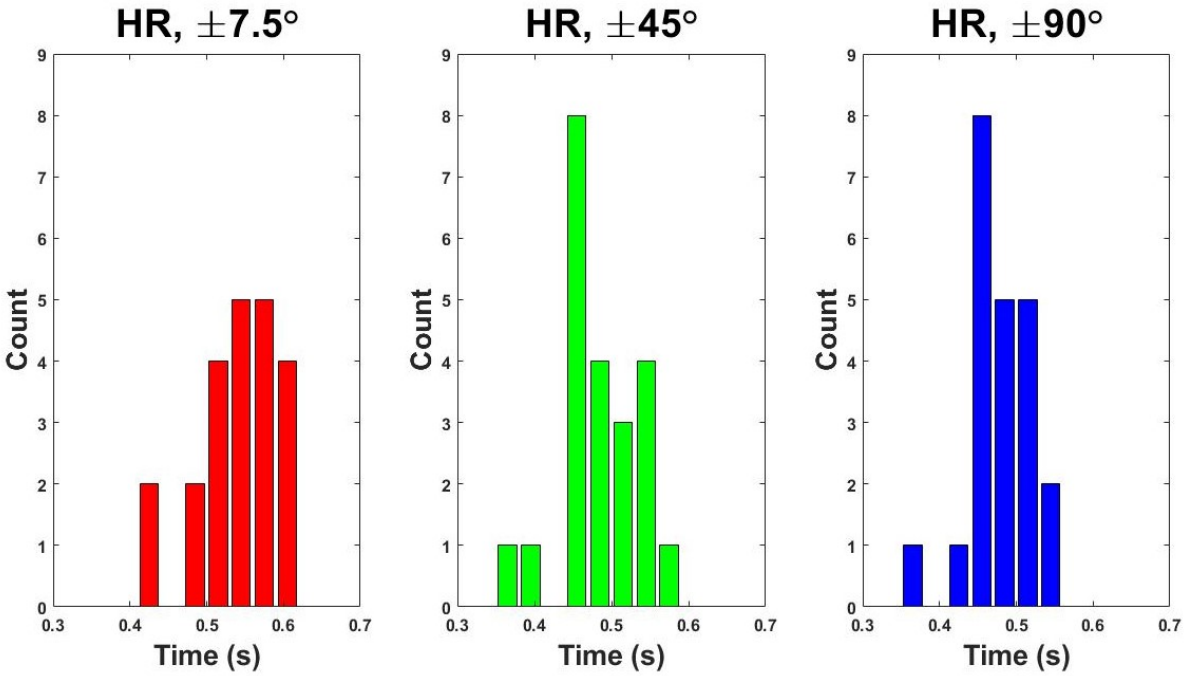

Figure 42: The distribution of median of RT of HR over all the subjects.

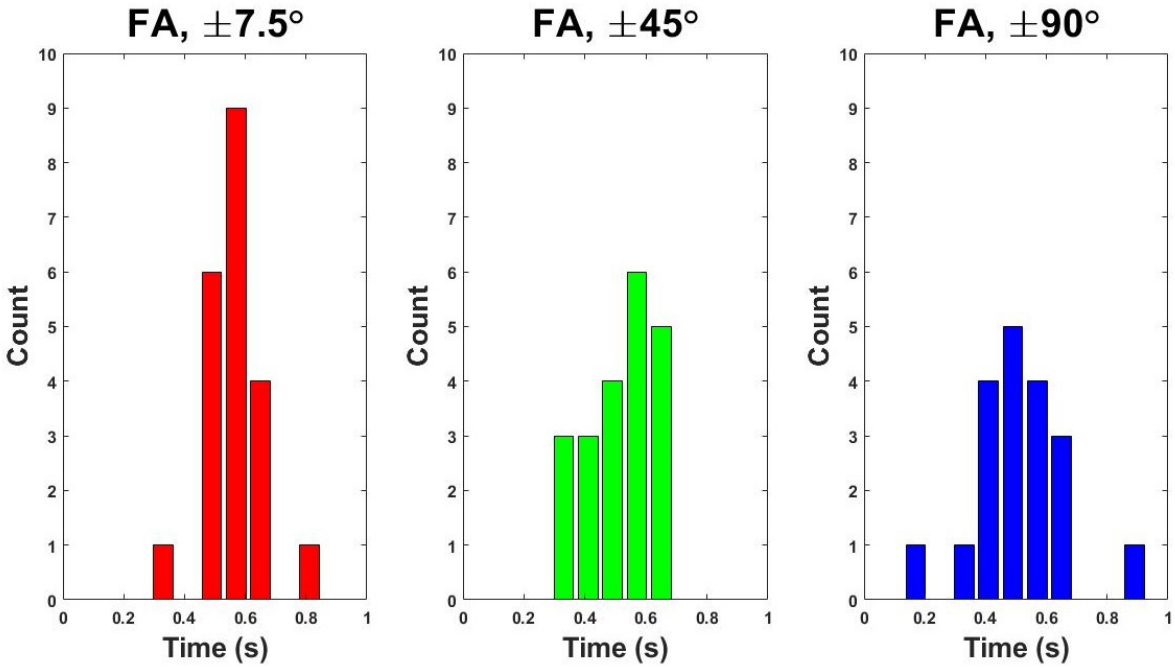

Figure 43: The distribution of median of RT of FAR over all the subjects.

5 Follow-up Experiments

The outcome of the first experiment suggested a similar pattern at ±45° and ±90° in terms of accuracy and reaction times on the stream-segregation task and the degree to which frequency acted as a cue for stream segregation. All these variables were significantly different at ±7.5° in comparison with the other two angles. To have an overall pattern of how the segregation task might be performed at other angles, a larger number of speakers were needed to be placed in the horizontal plane.

As the results of segregation task at ±45° and ±90° were similar and the results at ±45° and ±7.5° were significantly different, the new speakers had to be positioned between ±7.5° and ±45° to determine the changes in the above parameters between these two angles. To do so, a speaker setup with a greater spatial resolution (reduced angle difference between the speakers) was needed. With the new experimental setup, two follow-up experiments were designed and conducted to evaluate the changes between ±7.5° and ±45°.

5.1 Experiment Nr. 2

The first follow-up experiment was designed by adding four speakers between +7.5° and +45° and four speakers between -7.5° and -45° as shown in figure 44. The following 12 directions were considered within the range of -45° to +45° as illustrated in Figure 44: ±7.5°, ±15°, ±22.5°, ±30°, ±37.5°, and ±45°.

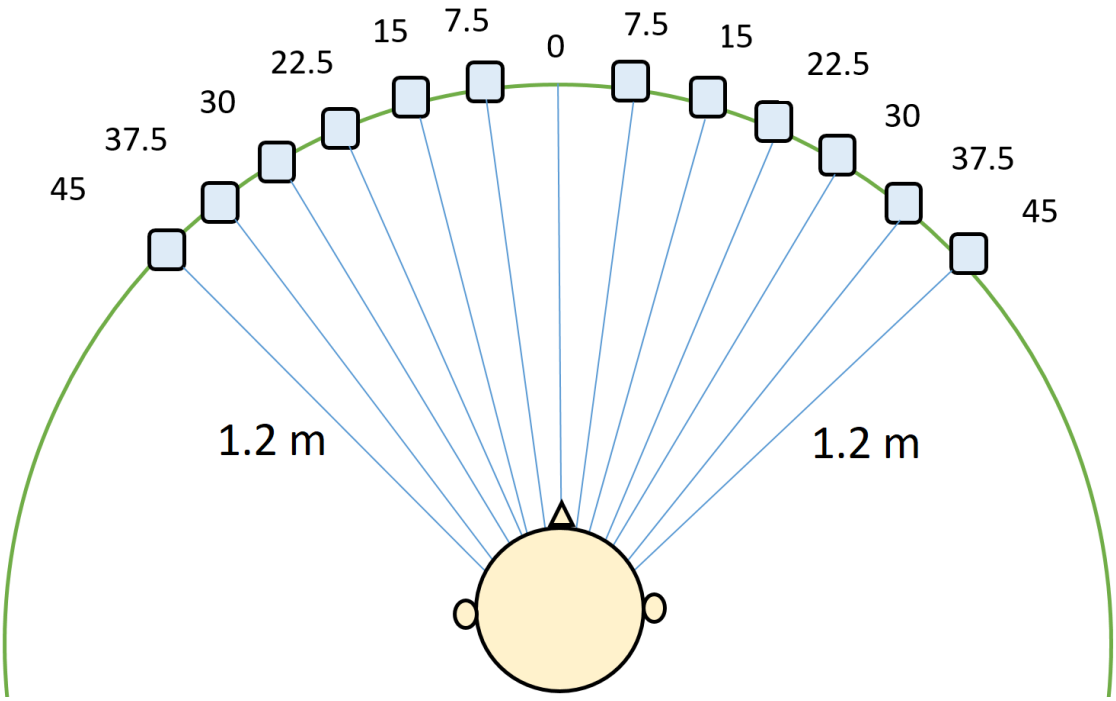

Figure 44: The speaker layout of the experiment Nr.2.

5.1.1 Experimental design

The test design was implemented in the Audio Information Processing department lab at the Technical University of Munich. The lab is equipped with a circular speaker array of 96 speakers, ±3.75° apart as shown in figure 45.

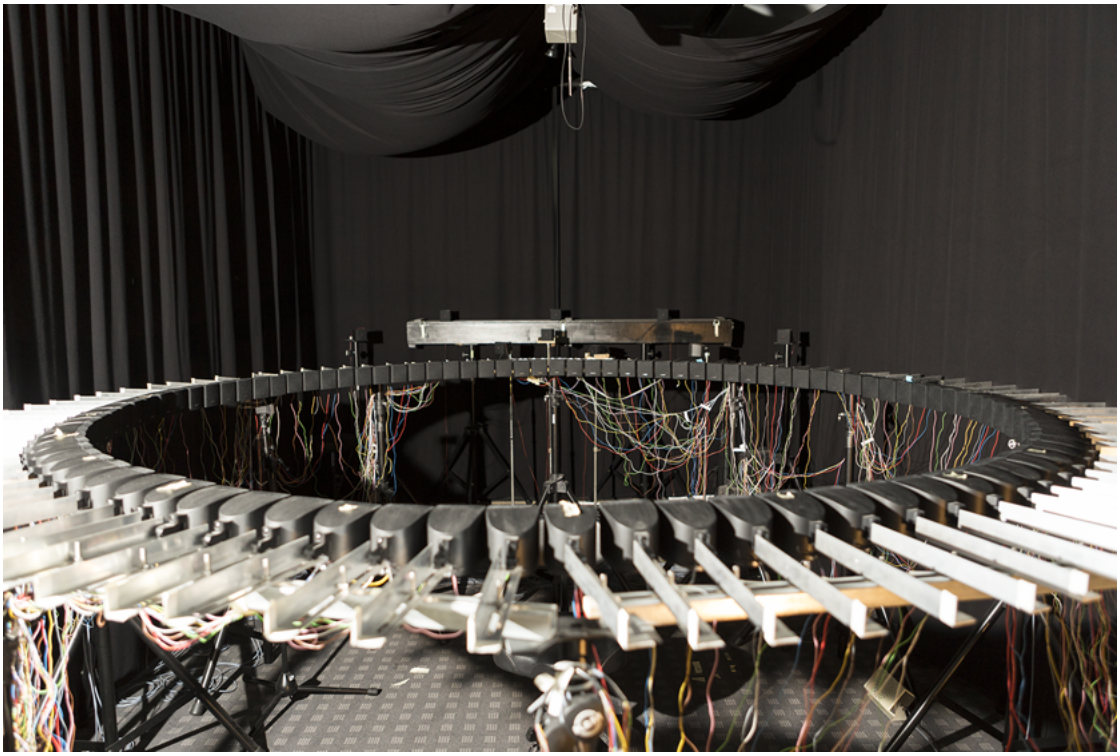

Figure 45: Audio Information Processing department lab, Technical University of Munich, adapted from [7].

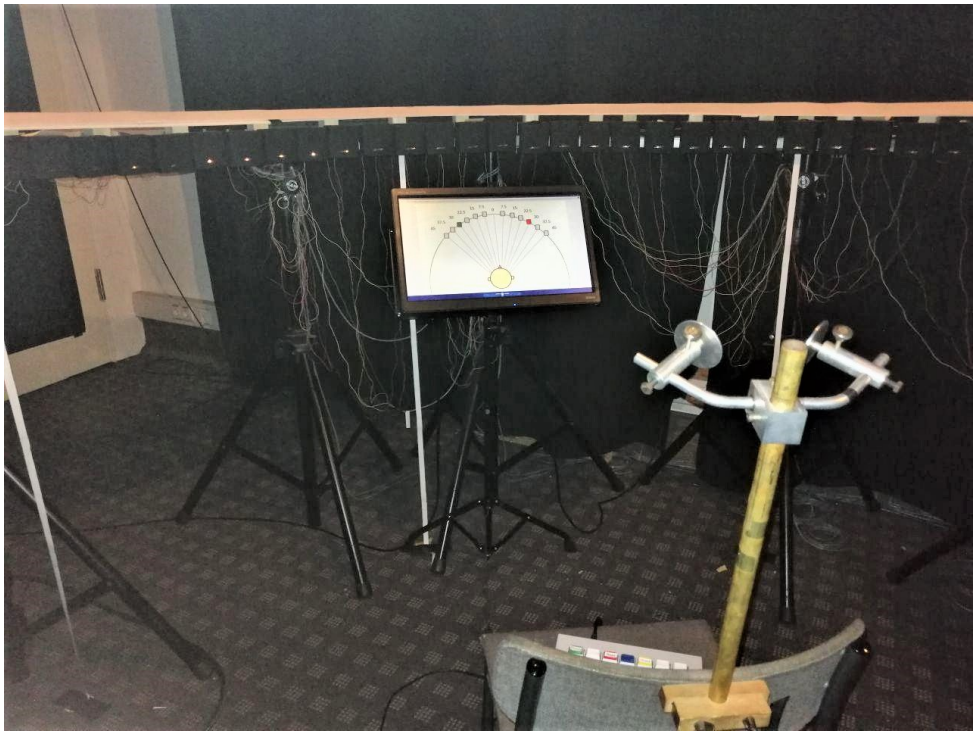

Figure 46: The relative position of the subject's location and the selected speakers.

The experimental design was similar to the first experiment except for the following parameters.

1. The subject was located in the center of the circular speaker array, 1.2 m away from the speakers.

2. A neck rest was used to keep the head of the subject fixed during the experiment.

3. Four subjects were selected for this experiment, all of them were students of the electrical and computer engineering department at the university. There were reimbursed 24 Euro for participating. The subjects were between the ages of 23 - 26 years with an average age of 24.75.

4. The gap length measured for the subjects were as the following: 45, 51, 58, 60 ms.

5. 360 trials were generated for each subject, 60 trials were played from each speaker pair.

6. In the total of 360 trials, there were 540 gaps in the target and 540 gaps in the masker. The trials of each speaker pair included a total of 90 gaps in the target and 90 gaps in the masker.

Figure 47 shows two examples of the visual cues shown on screen when the sound was played from two given speakers, simultaneously.

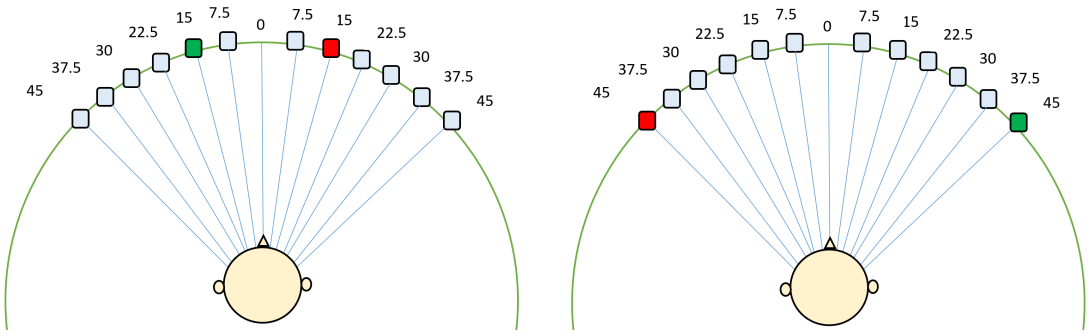

Figure 47: The visual cue of the experiment Nr. 2, right: ±45° (target:+45°), left: ±15° (target:-15°).

5.1.2 Analysis of the segregation task performance as a function of speaker separation

Figure 48 illustrates the average HR over all the subjects. It can be observed that HR increases gradually from ±7.5° to ±45°. The largest change was observed between ±7.5° and ±15°. The changes in HR were less noticeable from ±22.5° to ±45° as the subjects were able to detect 52% of the gaps at ±22.5°, 54% of the gaps at ±30°, 55% of the gaps at ±37.5°, and 59% of the gaps at ±45°. Similar to the results of the first experiment, there is a large difference between the percentage of detected gaps at ±7.5° (32%) and at ±45° (58%).

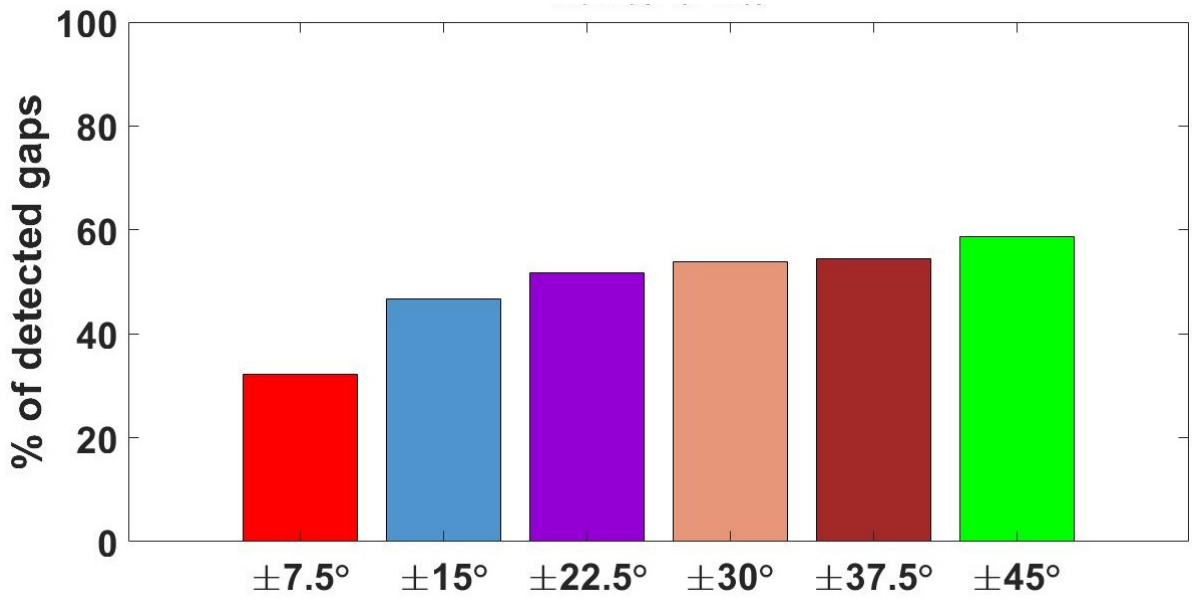

Figure 48: The average hit rate over all the subjects in experiment Nr.2.

Figure 49 shows the average sensitivity index calculated at each angle. Similar to HR, a large change in the sensitivity index was observed from ±7.5° to ±15°. The sensitivity index was 0.86 at ±7.5° and 1.43 at ±45°.

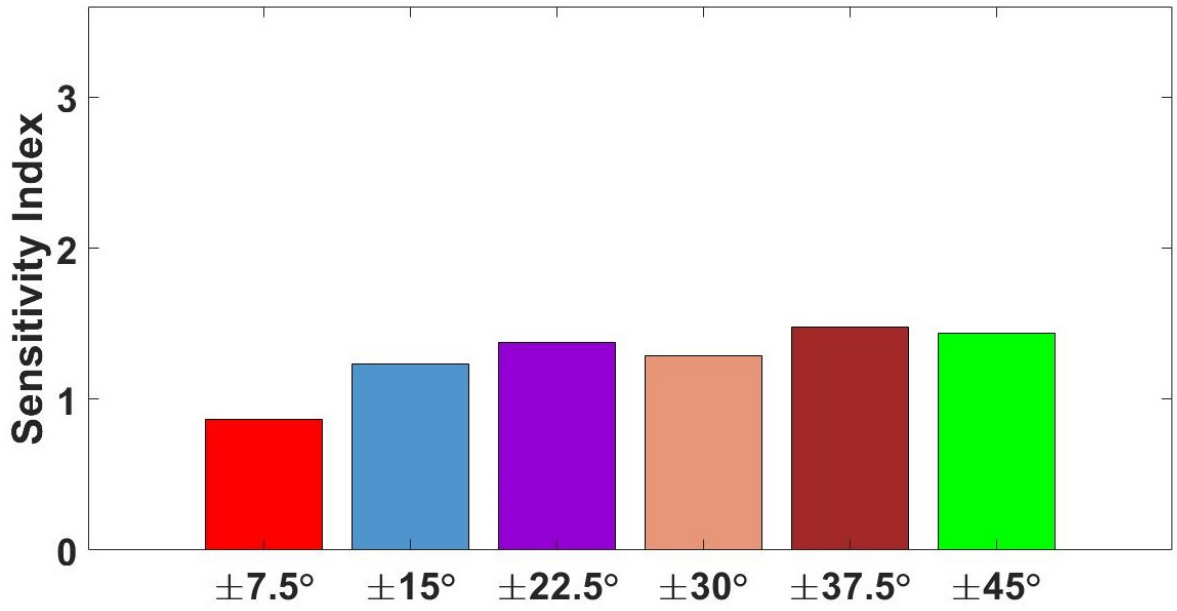

Figure 49: The average sensitivity index over all the subjects in experiment Nr.2.

5.1.3 Analysis of the segregation task performance as a function of the difference in frequency content and the modulation-rate

Figure 50 illustrates the average of the HR distributions over FC_D evaluated for all the subjects. Similar to the first experiment, a quadratic polynomial was fit to all the distributions. The overall pattern of HR distribution over FC_D is similar to the first experiment; The HR increases as FC_D becomes larger. This effect is stronger at $\pm 7.5°$ in comparison to that at other angles.

Figure 51 shows the average of the HR distributions over $FMOD_D$ over all the subjects. These results also supported the conclusion of the first experiment. At all angles, the changes in the modulation-rate of the target and masker did not improve the accuracy of the segregation task. Therefore, similar to the first experiment, the rate difference did not play the role of a segregation cue.

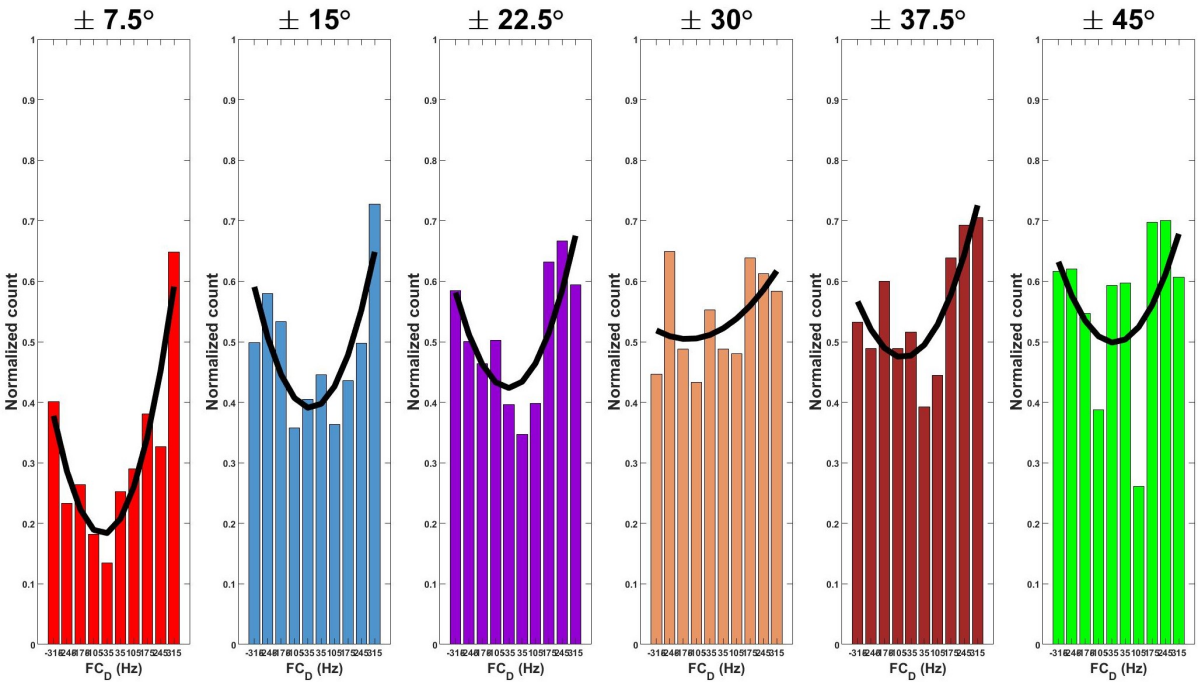

Figure 50: The distribution of HR over FC_D, experiment Nr.2.

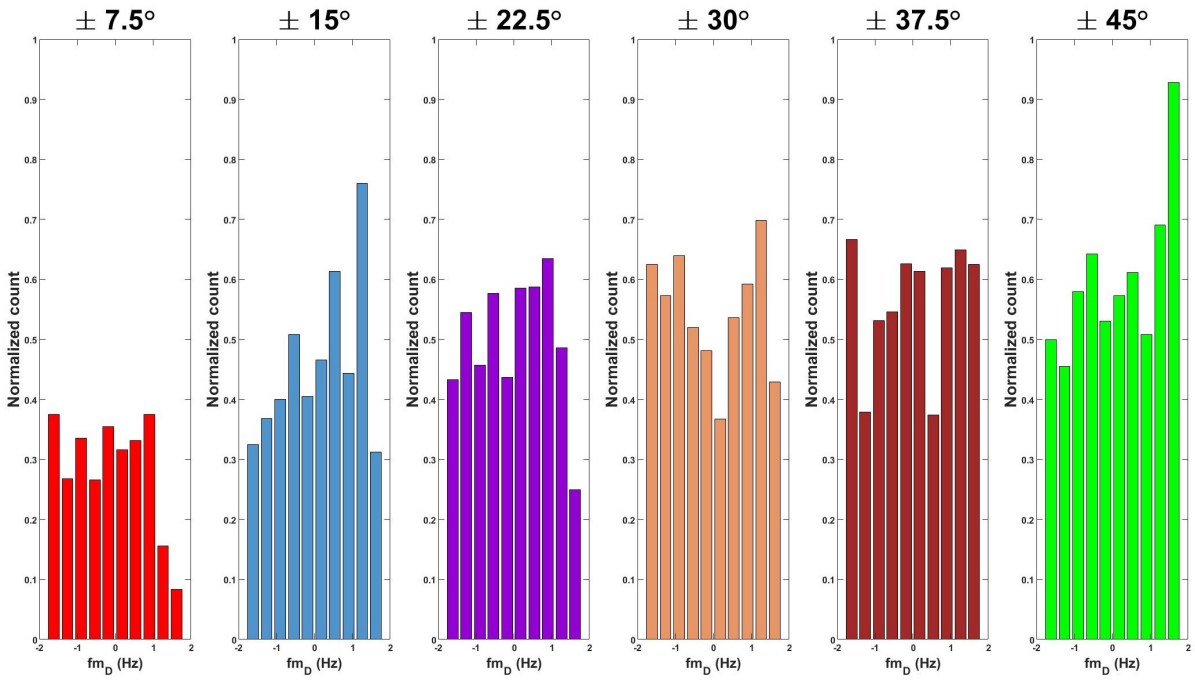

Figure 51: The distribution of HR over $FMOD_D$, experiment Nr.2.

5.2 Experiment Nr. 3

The results of the experiment Nr.2 suggested that the largest change in the HR occurred between ±7.5° and ±15°. Therefore, it was concluded that an even higher angular resolution should be considered between these two angles to be able to specifically illustrate the changes in HR below ±15°. Hence, a new experiment was designed with the highest possible resolution feasible with the setup available in the aforementioned laboratory. The new setup included the following angles: ±3.75°, ±7.5°, ±11.25°, ±15°, ±30°, and ±45° as shown in figure 52.

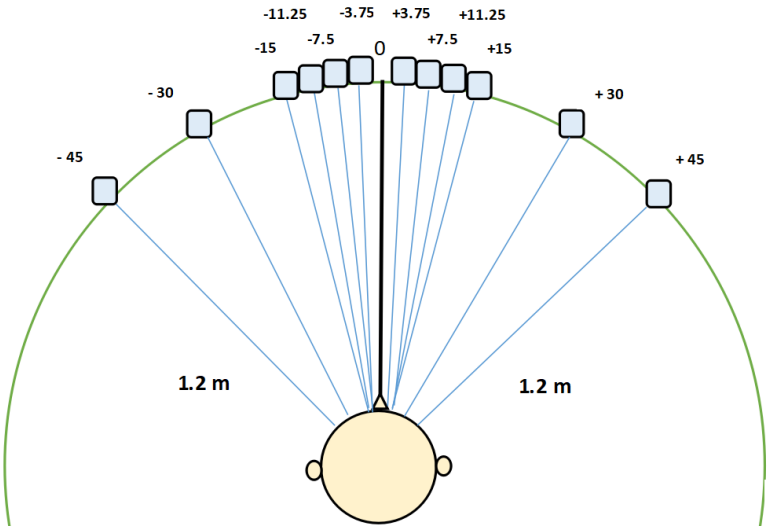

Figure 52: The physical layout of experiment Nr.3.

Figure 53 shows two examples of the visual cue shown on the screen.

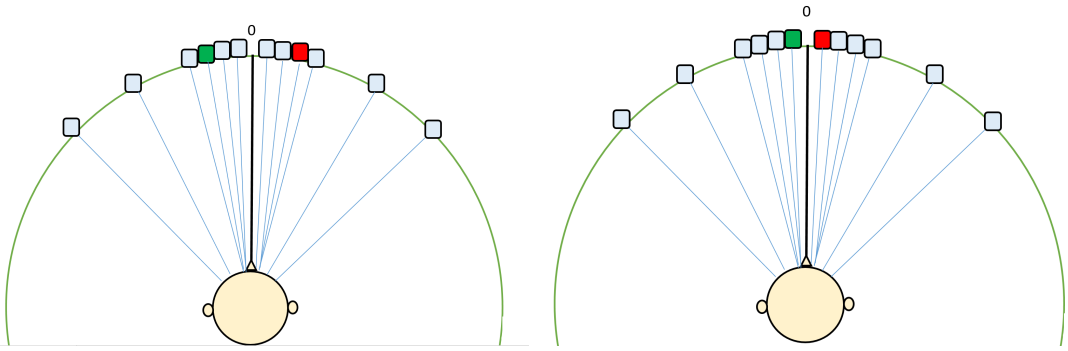

Figure 53: The visual cue of the experiment Nr. 3, Right: ±3.75° (target:-3.75°), Left: ±11.25° (target:-11.25°).

5.2.1 Experimental design

The experimental design of the third experiment, in general, was similar to the second experiment. Nevertheless, in this experiment, five new subjects were chosen to participate in the test. The subjects were in the range of 20-29 years old with an average age of 23.80. The gap length measured for the subjects were as follows: 35, 45, 53, 63, 67 ms.

5.2.2 Analysis of the segregation task performance as a function of speaker separation

Figure 54 illustrates the average HR over all the subjects. Similar to the previous experiment, HR was increased gradually. The similar HR at ±30° and ±45° was also observed in this experiment. The HR at ±11.25° and ±15° was observably very close. On average, the subjects were able to detect 31 % of the gaps at ±3.75° while they detected 60 % of the gaps at ±45°.

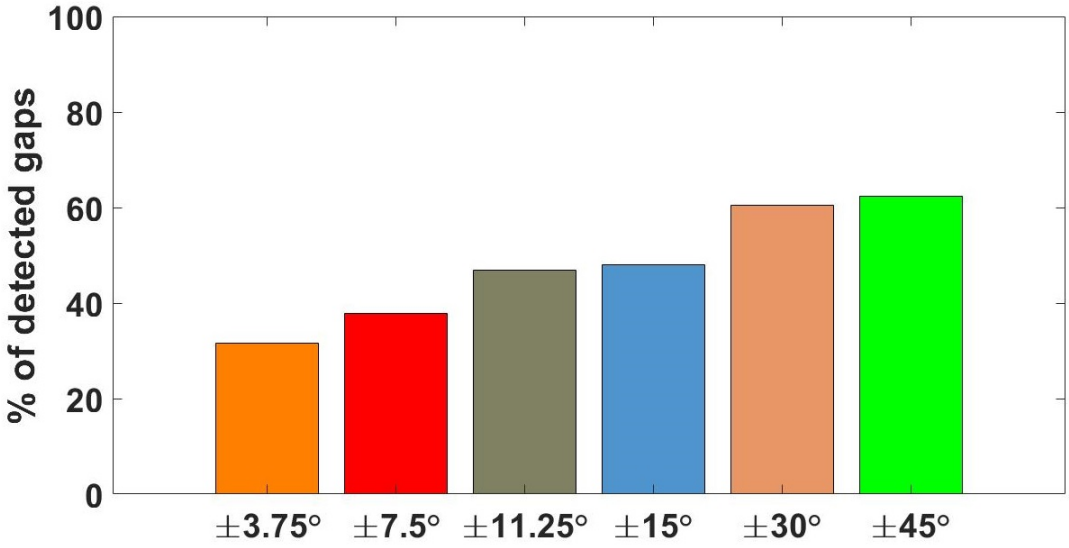

Figure 54: The average hit rate of segregation task in experiment Nr.3.

Figure 55 demonstrates the average sensitivity index over all the subjects. Similar to HR, this index was increased from ±3.75° to ±45°. The large difference in the sensitivity index can be seen between ±3.75° and ±7.5°. This is because the highest number of gaps in the masker was reported at ±3.75°. Therefore, the sensitivity index which counts for the high false alarm, decreases at this angle.

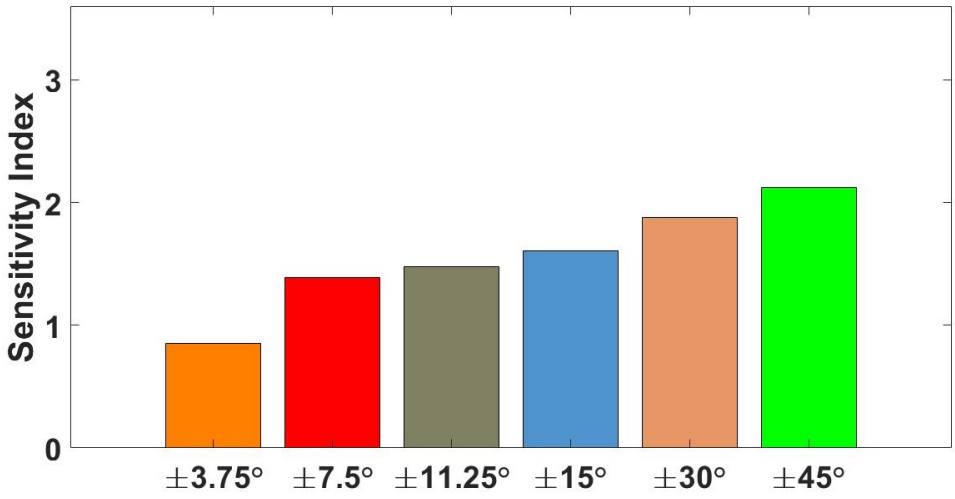

Figure 55: Sensitivity Index showing the average performance of the subjects in experiment Nr.3.

5.2.3 Analysis of the segregation task performance as a function of the difference in frequency content and the modulation-rate

Figure 56 demonstrates the average distribution of HR over FC_D over all the subjects in this experiment. Similar to the first and second experiments, a quadratic polynomial was fit to these distributions. At $\pm 3.75°$, where FC_D was the smallest (bin 5 and bin 6), on average, the subjects were able to detect 13 % of the gaps. While at $\pm 15°$, they were able to detect 27 % of the gaps and at $\pm 45°$ they detected 55 % of the gaps in bin 5 and bin 6.

Figure 57 illustrates the average distribution of HR over $FMOD_D$ in experiment Nr.3 over all the subjects. Similar to the previous experiments, the HR had a random distribution over the different modulation-rates of the target and masker. Therefore, the rate also failed to play the role of a cue in stream segregation in this experiment .

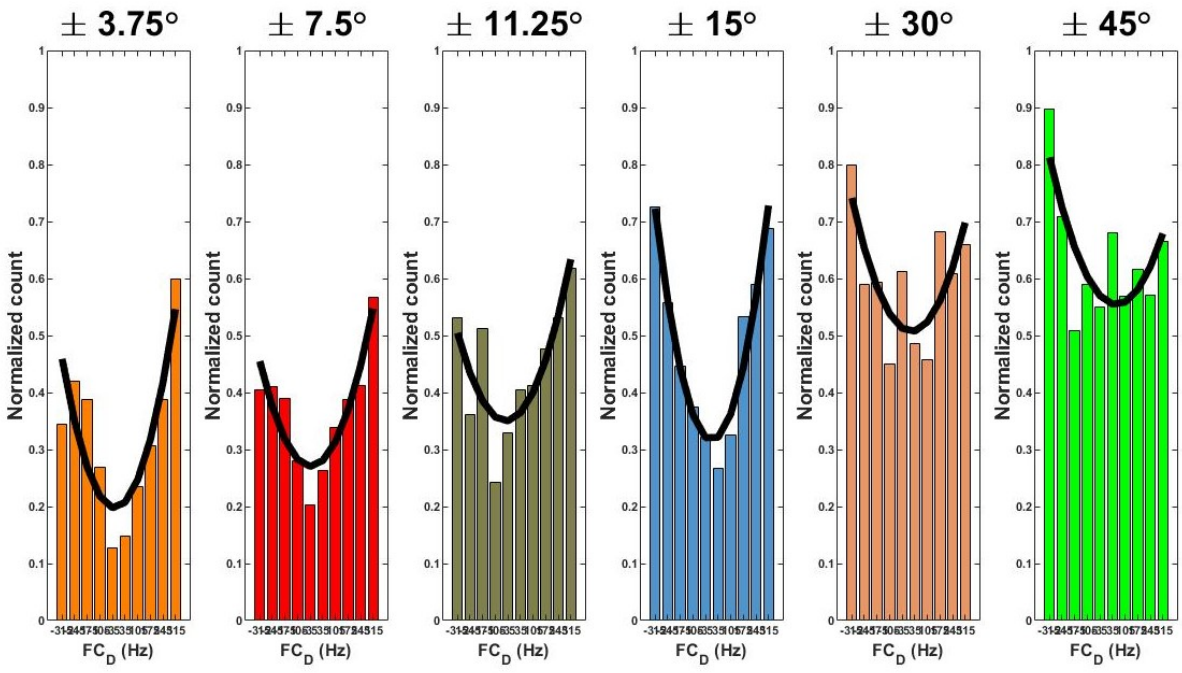

Figure 56: The distribution of HR over FC_D, experiment Nr.3.

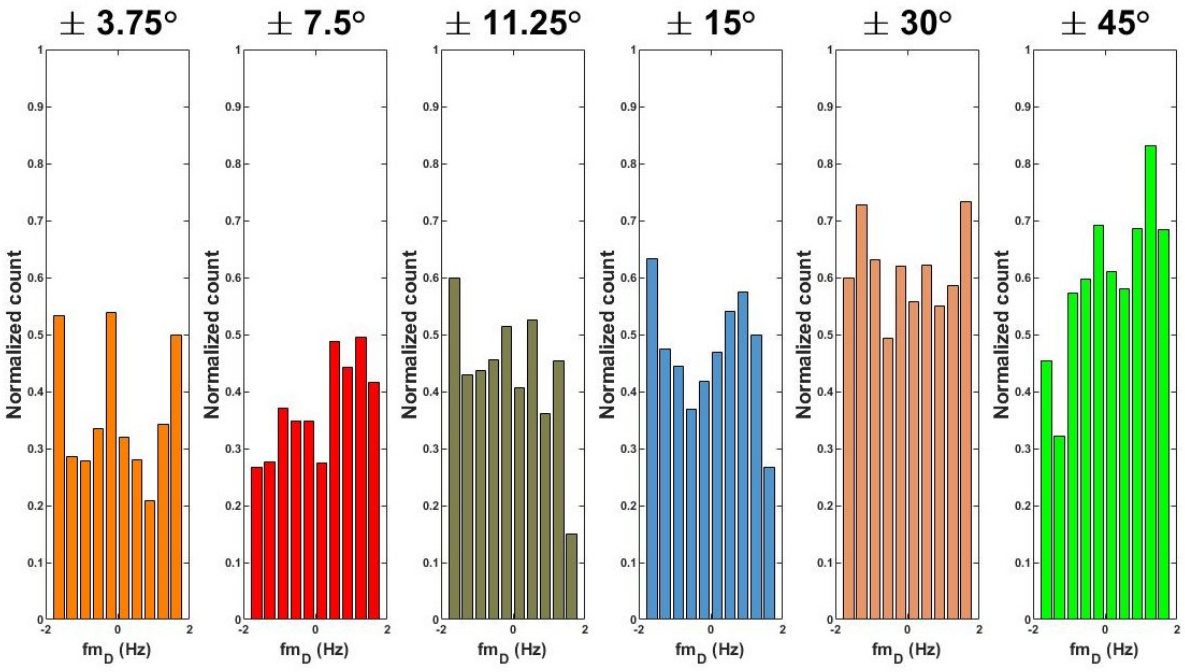

Figure 57: The distribution of HR over FMOD_D, experiment Nr.3.

5.3 Summary of three experiments

The outcome of the second and third experiments justified that which was speculated from the first experiment. As with all the experiments conducted, a positive correlation between the number of hits and angular distance between the target and masker speaker was observed. Figure 58 illustrates the average of the detected gaps in percentage over all the 9 angles that were tested in the three experiments. The sample size that the hits are averaged over is not the same as the number of subjects in different experiments was dissimilar. It can be observed that the number of hits increases from $\pm 3.75°$ to $\pm 90°$. As the spatial cues became more informative in light of the speakers being further away from each other, the stream segregation task was performed with higher accuracy.

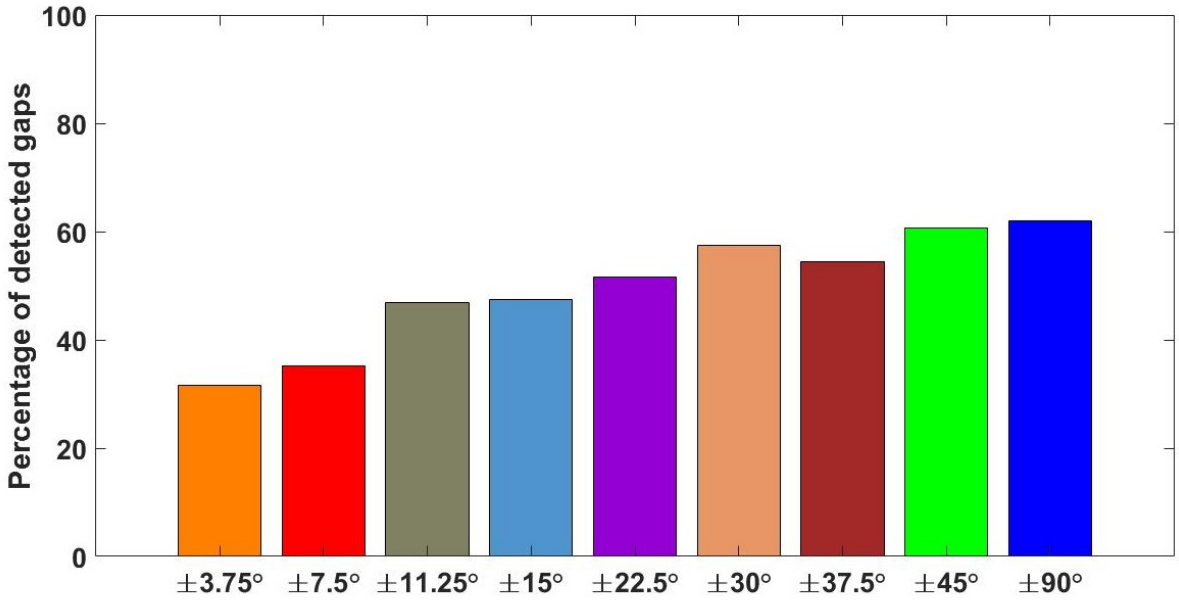

Figure 58: The average of the percentage of the detected gaps over all the 9 angles that were tested in three experiments.

The frequency difference between the target and masker at all the angles had an influence on the HR, thus played the role of a segregation cue. In general, the subjects detected more number of gaps where the difference between the frequency of the streams was larger.

Unlike the frequency content difference, there was no positive interaction between the changes in the HR and the changes in the modulation-rate difference between the two streams. This was evident at all angles. It was considered that irrespetive of the amount of angular separation of the two sound sources and the strength of spatial cues, the modulation-rate difference did not play the role of a stream segregation cue.

6 Discussion

The first research question was the following: "Does the physical location of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation ?". To answer this question, the outcome of applying the SDT model was analyzed in three steps; analysis of hits, analysis of false alarms, and analysis of the sensitivity index.

The results showed that when the speakers were close to each other and in front of the subject ($\pm 7.5^\circ$), the spatial cues were not so informative, owing to which the stream segregation task was poorly performed. However, in the cases where the speakers were further from each other ($\pm 45^\circ$ and $\pm 90^\circ$) resulting in the spatial cues being more informative, the task was performed significantly more accurately.

Therefore, the answer to the first research question is that the physical location of the desired and competing sound sources, play the role of a cue in gap-detection and auditory stream segregation. Given that the results of this study were achieved based on examining 18 directions of arrival in the horizontal plane (the results of all the three experiments combined), expanding the results to all the angles in the horizontal plane must be interpreted with care.

In the follow-up experiments, the speaker placement at the angles between $\pm 45^\circ$ and $\pm 90^\circ$ was excluded. This choice was made as the results of SDT at these angles in the first experiment were not significantly different. Therefore, it was concluded that there is no need for examining the results between $\pm 45^\circ$ and $\pm 90^\circ$. The analysis of the results of the second experiment showed a very similar number of hits and the sensitivity index at $\pm 30^\circ$, $\pm 37.5^\circ$, and $\pm 45^\circ$. These results also suggested that the changes in the SDT results are not noticeable above $\pm 30^\circ$ which supported the choice that was initially made.

However, based on the observed pattern, it is likely that no significant difference in the hits and sensitivity index between $\pm 45^\circ$ and $\pm 90^\circ$ will be apparent.

Furthermore, another reason that the outcome of the experiments fails to explain the changes in SDT results in the entire horizontal plane can be attributed to no investigation performed on the relationship between stream segregation and speaker separation, behind the head.

In all of the trials of this study, the angular distance between the target speaker and the localization reference point (0° in front of the subject) was equal to the angular distance between the masker speaker and 0°. Considering the results at $\pm 45^\circ$, the target and masker speakers were 90° apart from each other.

However, the results of this case do not necessarily reflect the general effect of 90° separation in space on the gap-detection performance. For instance, if the target sound was generated from 0° and the masker sound from -90°, although the speakers are 90° apart from each other in this case as well, the results of the segregation task performance cannot be predicted according to the results of this study at $\pm 45^\circ$.

This outcome substantiates previous findings in the literature that in general, the spatial separation of two sound sources can potentially improve the segregation task performance [48, 49, 62, 72, 77, 80, 84].

In different studies, different combinations of stimulus type, directions of arrival for the target and masker, and the assigned task were considered. These different grouping methods could potentially affect the strength of the spatial separation cue on the stream segregation. Broadbent [107] suggested that the effect of the spatial separation on the stream segregation is stronger when the subject is asked to listen to one sound and ignore the other as opposed to the case where the subject is asked to pay attention to both the streams.

The experimental design of this study allows re-evaluating this statement in the future. The current task is designed according to the selective attention logic. It means the subject must select a stream and only pay attention to that. The task in the new design would be to react to the gaps in both target and masker sounds. Two separate buttons on the response device can be considered, one to be pushed when the gap in the target stream is detected and another to report the detected gaps in the masker. The strength of the spatial separation of this new experiment can then be compared with that of the current study.

The second research question was "Does the frequency content difference of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation?" To answer the question, the hits distribution over the difference between the mean carrier frequency of the target and masker were evaluated. The results showed that the frequency difference between the streams significantly affected the distribution of hits at all the three angles.

This means that the performance of the subject was significantly affected by the changes in the frequency content difference. When the difference between the target and masker frequency was larger, subjects performed the gap-detection task with higher accuracy. When the frequency of target and masker were similar, segregating the streams was perceptually more challenging at all the angles.

Nextly, the strength of this spectral cue on stream segregation was compared at different angles. The results showed that when the speakers were closer and the spatial cue was less informative (at $\pm 7.5^\circ$) in comparison to the case where the speakers have a larger angular distance (at $\pm 45^\circ$ and $\pm 90^\circ$), the frequency difference was a significantly stronger cue.

The results of this analysis was in line with previous studies where the influence

of the frequency separation on the stream segregation ability was analysed [8, 9, 49, 58, 62, 68, 108].

Divenyi & Brandmeyer [33] evaluated the effect of prosodic fluctuation and the fundamental frequency of speech on stream segregation in a cocktail party scenario. They claimed that the fundamental frequency difference played the role of a cue, but their results showed that when the target frequency is higher than the masker frequency, the stream segregation task was performed significantly more accurately than when the masker had a higher frequency. In this study, such an effect was not observed at any angle. This is presented in figure 31-top where the average hit rate of the subjects are illustrated.

The third research question was "Does the modulation-rate of the desired and competing sound sources play the role of a cue for gap-detection and stream segregation?" The distribution of hits over the difference between the modulation-rate of the target and masker was analyzed. The single-subject based results showed a random distribution of the hit rate over the modulation-rate difference.

The average performance over all the subjects showed an almost uniform distribution of hits over the modulation-rate difference. The statistical analysis of the results showed that the interaction of the modulation-rate difference between the target and masker and the hit rate was not significant at any angle. Therefore, the answer to the third question was that the modulation-rate did not play the role of a cue for streams segregation at any angle.

In this study, the influence of the modulation-rate was also evaluated when the frequency difference of the streams was not a strong cue, meaning the target and masker had similar average carrier frequencies. Even in this case, the rate difference did not significantly affect the stream segregation task.

Devergie et.al [33] designed an auditory stream segregation experiment in which the frequency and timbral cues were not provided to the subject. The study presented in [33] was only aimed at investigating the interaction between rhythm and performing the segregation task. Here, two cases where the masker was either a familiar rhythmic or unfamiliar rhythmic stimulus were considered. Their results showed that the segregation task was performed more accurately when the masker was a familiar rhythmic stimulus. It was assumed that the rhythm was recalled by long term memory. It was concluded that when other cues are not provided, the rhythm played the role of a segregation cue when it was the rhythm of a familiar sound.

However, in this study, where the frequency difference was less than 175 Hz and therefore it was not a strong cue, it was observed that the rate of the stimulus did not play the role of a segregation cue. The different outcomes of these two studies can be explained based on the effect of familiarity and long term memory involved. Here, the stimulus was not a familiar sound to the subjects. Therefore, it was not possible for the subject to recall the rate/rhythm of the stimulus from long term memory, which in turn failed to serve as a segregation cue.

Finally, the reaction time of the subjects to make a decision and report a gap was evaluated as a function of spatial separation between the target and masker sounds. The results showed that the subjects were significantly slower in reporting a gap at $\pm 7.5^\circ$ in comparison to that observed at $\pm 45^\circ$ and $\pm 90^\circ$. There was no significant difference between the reaction time at $\pm 45^\circ$ and $\pm 90^\circ$.

This means that where the sound sources are close to each other owing to which the spatial cues are not as informative as in the case where the speakers are further away from each other, segregating the streams appears to be a more difficult task. Therefore, in this case, decision making was a relatively longer process.

As it was mentioned before, if the segregation task was a direct one, meaning the subject was only asked to report whether 1 or 2 streams are detected, the reaction time analysis would not be possible. The outcome of this study revealed a significant difference between the reaction time at $\pm 7.5^{\circ}$ in comparison to that at $\pm 45^{\circ}$ and $\pm 90^{\circ}$. This significant outcome was not possible to be observed if the task was a direct one. Therefore, this was another reason why the choice of the experimental task was appropriate.

6.1 Limitations

This study was designed based on three critical choices; The choice of stimulus, the physical layout of the test implementation, and the experimental task. This section explains possible limitations of the experimental design implemented in this thesis.

6.1.1 Stimulus

The stimulus of this study was designed in such a way that it had the frequency modulation-rate similar to that of the prosodic rate of speech. The stimulus was not directly related to a specific language. It was temporally continuous with no onset and offset in time to avoid any extra cues that these may provide. Finally, both the sound of interest and the competing sound were temporally regular. This was to simulate the speech conversation that includes two rhythmic sounds.

As it was mentioned before, for each trial, a set of carrier frequencies (30 sinusoids) were added to make a noisy stimulus. This means that in each trial, a set of neighboring frequencies stimulate a larger area of the auditory cortex as per its tonotopic organization. Therefore, this stimulus would be a suitable trigger for a follow-up Electroencephalogram study as it allows for the recording of a larger area of the auditory cortex [109–112].

In addition, embedding a temporal gap in a noise creates less auditory artifacts as compared to the case where a transient gap is embedded in a pure sine wave. Calculating the gap length threshold with a noisy sound allows for a higher measurement resolution as the gap thresholds are more variable across participants in noise.

In this study, there was only one distractor sound which was of the same type

as the sound of interest (both frequency-modulated noise with a bandwidth of 200 Hz). In a real-life scenario, multiple sound sources might exist with different frequency ranges including a broadband noise which is capable of masking a wider range of frequencies. Therefore, the outcome of this experiment potentially could be limited to the results of competing two narrow-band noise in an auditory scene.

6.1.2 Physical layout

As the goal was to recreate a real life scenario in the lab, the physical layout of the experiment had to be designed in such a way that it reflects the existence of multiple sound sources. Separate sources in space were considered to create different directions of arrival relative to the subject.

Implementing a cocktail party scenario in an acoustically well-treated lab eliminates the effect of the reverberation from the surfaces and obstacles that exist in a real-life scenario on the auditory perception. The reflections and echos can affect the perception of speech in a noisy environment [8, 113–115]. It is not feasible to recreate all the possible environments with different reverberation profiles in the lab to evaluate the effect of reverberation on speech perception in different scenarios. However, this limitation can be compensated in two ways.

If the cocktail party study is conducted with a headphone, the characteristics of the desired environments can be simulated with room simulation softwares. The calculated impulse responses of the virtually designed room, can be used to filter the target and masker(s) sounds and the result can be presented binaurally. In this case, the resulting sound carries the information regarding the room acoustical profile.

If the experiment is implemented with an array of loudspeakers in a lab, the subject can be surrounded by different acoustic panels with different absorption coefficients

[116] to be able to emulate different desired amounts of reverberation from the walls.

In this experiment, the sound source location was pre-defined for the subject with the target source was marked in Green on the visual cue shown on the screen. This means that in this design, the subject was aware of the direction of arrival and was not asked to localize the sound source.

In a real-life scenario in some cases, similar to this study, a visual cue is provided to the listener. For example, if two people are having a conversation and one is listening to another in full view of each other, a visual cue exists. Therefore, the listener can localize the speaker with the auditory localization cues along with the extra information which is provided by the visual cue. However, in some cases, no visual cue is provided. For instance, if there is an obstacle between the listener and the speaker, the listener has to localize the position of the speaker and only with the help of auditory localization cues should he or she estimate the direction of arrival. This could present itself to be a more challenging a task than that presented in this thesis.

In a real-life scenario, the sound sources and the listener might be in motion. In this study however, the loudspeakers and the subject were static throughout the experiment. Therefore, the possible effects of the acoustical cues that affect the perception of a sound coming from moving sound sources or received by a moving listener were excluded. These cues include but are not limited to motion parallax and acoustic tau [117–120].

6.1.3 Task

An engaging task was designed to indirectly evaluate their auditory stream segregation ability according to their gap-detection performance. This allowed for the investigation to employ the signal detection theory to evaluate the performance of the subjects and answer the research questions using the respective model.

In this study, the gap-detection ability was only evaluated at the presence of a distractor sound to simulate a cocktail party scenario. It could be beneficial if the gap-detection task at each angle, at first, was evaluated in a controlled manner. This means that, only the sound of interest is played and there is no simultaneous distractor in which case, the gap-detection could be evaluated as a function of source location in space. This way, an evaluation could be done on the differences between the number of detected gaps with and without a distractor. This would be beneficial as it could clarify the effect of including a masker in the scene on the gap-detection ability, at each angle.

The reaction time could be used as a measure to increase the reliability of analyzing the number of hits reported by each subject. The hits which were reported in less than 200 ms after the gap was presented could have been removed from the hits dataset. This is because, based on the calculated reaction time of all the subjects, very few gaps were reported in such a short time span.

This can potentially mean that if a subject pushed the bottom within such a short time, he/she did so before even a gap was detected. Since the acceptable time window was 1.5 seconds from the instant the gap was presented, the gaps reported very early were also considered as a hit. As it is unlikely that a gap was detected in such a short time, a more accurate time window for registering the hits could have been between 200 and 1500 ms.

7 Conclusion

The goal of this study was to evaluate the auditory stream segregation ability in a cocktail party scenario where two sounds were presented simultaneously through a gap-detection task and by applying signal detection theory. The gaps were embedded in both streams and the subjects were asked to only report the gaps in the target sound. If the subjects could detect the gaps in target and not the masker, it was considered that they were able to successfully segregate the streams. Three angles on the horizontal plane were selected to evaluate these effects; $\pm 7.5^\circ$, $\pm 45^\circ$, and $\pm 90^\circ$.

At first, the performance of the subjects in terms of the auditory stream segregation was evaluated as a function of speaker separation. The signal detection model was used to describe the performance of the subjects in terms of the hit rates, false alarms, and the sensitivity index. The results showed that the hit rates and sensitivity index at $\pm 7.5^\circ$ were significantly lower than $\pm 45^\circ$ and $\pm 90^\circ$. However, the hit rates and sensitivity index at $\pm 45^\circ$ and $\pm 90^\circ$ were not significantly different. The false alarm rate at all angles was not significantly different.

It was concluded that according to this analysis, in general, the task was significantly more difficult at $\pm 7.5^\circ$ than at the other two angles. It means that where spatial cues are less informative, segregating two auditory streams is a more difficult task and is performed with less accuracy.

Next, the results of the stream segregation performance as a function of the frequency difference between the target and masker sound were analyzed. The results showed that the influence of frequency difference on performing the task was significant at all the angles. This effect at $\pm 7.5^\circ$ was significantly stronger than at $\pm 45^\circ$. The results of comparing the effect at $\pm 7.5^\circ$ and $\pm 90^\circ$ showed that the effect was noticeably stronger

at ±7.5°. The outcome showed the usefulness of frequency content separation as a cue in stream segregation. However, the effect at ±45° and ±90° was not significantly different.

This means when the target and masker were more different from each other in frequency, participants performed the gap-detection task better, indicating that it was easier to separate the target and masker into two streams. When the target and masker were close to each other in frequency, performance was worse, suggesting that listeners had a harder time segregating the two streams.

Then, the results of the stream segregation performance as a function of the difference between the modulation-rate of the target and masker sound were analyzed. The outcome showed that at irrespective of the angle the interaction of modulation-rates difference and the stream segregation performance was not significant. Therefore, it was concluded that the subjects did not rely on the rate difference as a cue for detecting the gaps. Hence, no interaction between the rate difference of the streams and 3D space was found in this cocktail party scenario.

The results of reaction time analysis illustrated the same pattern as the hits and sensitivity index meaning the reaction time was significantly lower at ±7.5° than that observed at ±45° and ±90°. However, the reaction time at ±45° and ±90° was not significantly different. This means where stream segregation task was more difficult, the subjects needed more time to decide about reporting a gap.

Table 11 summarized the effects that were seen in the experiment at different angles.

Table 11: Data analysis summary: ✓: statistically significant, ×: not statistically significant.

Combination	Hit Rate	False Alarm	dprime	FC_D	$FMOD_D$	Reaction Time
±7.5° & ±45°	✓	×	✓	✓	×	✓
±7.5° & ±90°	✓	×	✓	✓*	×	✓
±45° & ±90°	×	×	×	×	×	×

As the changes in the above parameters were observed between the speakers located at ±7.5° and ±45°, to further investigate these parameters in the horizontal plane, two follow-up experiments were implemented to closely observe the changes between these two angles.

In the second experiment, 12 speakers were located at the following angles; ±7.5°, ±15°, ±22.5° ±30°, ±37.5°, and ±45°. The results showed that the hit rate and sensitivity index gradually increased from ±7.5° to ±45°. The largest difference was observed between ±7.5° to ±15° for all the subjects. Therefore, to investigate the performance difference below ±15°, the third experiment was designed. In this experiment, 12 speakers were located at ±3.75°, ±7.5°, ±11.25°, ±15°, ±30°, and ±45°. The largest change in the sensitivity index was observed from ±3.75° to ±7.5°.

The results of the frequency content analysis of the second and third experiments showed that as with the first experiment, there was a positive correlation between the gap-detection performance and the frequency difference of the streams. Therefore, the frequency content difference played the role of a cue for segregation.

The results of rate analysis of the second and third experiments showed that for all the angles below ±45° which were tested, the changes of hit rate over rate difference was random and the latter did not play the role of a segregation cue at any angle.

7.1 Future work

The simulation of a real cocktail party scene demands using several speakers to simulate different directions of arrival as done in this study. Ideally, a single device that is compatible with generating 3D audio and projecting a sound to different required locations in space can be used. Such a scenario is implementable by applying the enhanced pressure matching method using a compact linear speaker array [121–124]. This experiment can be followed up by using such a device to overcome the hardware limitations of the current test design to evaluate the scene at a wider range of angles.

Speech stimulus in different languages could be evaluated at the prosodic rate similar to the stimulus of this study. The effect of different prosodic rates can be compared with the effect of modulation-rate difference that was applied here. To evaluate the potential effect of familiarity and linguistic background, two stimuli with the same rate, one in the subject's native language and one language unknown to the participant be used to compare the results.

In this study, the gaps were uniformly scattered over the entire possible instantaneous phases. To compare the gap-detection performance of the subjects at different phases, a larger number of trials per phase is needed. In this case, the number of hits and false alarms can be estimated as a function of the gap's instantaneous phase to observe the phase at which the subjects are more successful in performing the segregation task.

References

[1] Hervé Abdi. Signal detection theory (sdt). *Encyclopedia of measurement and statistics*, pages 886–889, 2007.

[2] Hasomed medical chin rest. `https://www.chinrest.net/store#!/Chin-Rest/p/133920939/category=0`. Accessed: 10/07/2019.

[3] Genelec 8040a studio monitor manual. `https://www.genelec.com/sites/default/files/media/Previous%20models/2-way%20loudspeakers/8040A/om8040a_8050a.pdf`. Accessed: 10/07/2019.

[4] Genelec 8040a studio monitor. `https://www.genelec.com/support-technology/previous-models/8040a-studio-monitor`. Accessed: 10/07/2019.

[5] Nti audio, spl meter, xl2, and microphone m4261. `https://shop.nti-audio.com/pd-xl2-analyzer-with-microphone-m4260.cfm`. Accessed: 11/17/2019.

[6] Fluke 424d laser distance meter. `https://www.fluke.com/en-us/product/building-infrastructure/laser-distance-meters/fluke-424d#`. Accessed: 11/19/2019.

[7] Acoustic lab, auditory information processing department, technical university of munich. `https://www.ei.tum.de/en/aip/research/virtual-acoustics/`. Accessed: 02/01/2020.

[8] Barry Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50, 1992.

[9] Yan-min Qian, Chao Weng, Xuan-kai Chang, Shuai Wang, and Dong Yu. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, 19(1):40–63, 2018.

[10] Jens Blauert. *Communication acoustics*, volume 2. Springer, 2005.

[11] Mendel Kleiner. *Acoustics and audio technology*. J. Ross Publishing, 2011.

[12] Michael Vorländer. *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[13] William A Yost, Richard R Fay, and Arthur N Popper. *Auditory perception of sound sources*, volume 29. Springer Science & Business Media, 2007.

[14] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[15] AS Bregman. Auditory scene analysis: The perceptual organization of sound. cambridge, ma, us, 1990.

[16] Norbert M. Seel, editor. *Signal Detection Theory*, pages 3064–3064. Springer US, Boston, MA, 2012.

[17] Lewis O Harvey Jr. Detection theory: Sensory and decision processes.

[18] Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149, 1999.

[19] Maija Hausen, Ritva Torppa, Viljami R Salmela, Martti Vainio, and Teppo Särkämö. Music and speech prosody: a common rhythm. *Frontiers in psychology*, 4:566, 2013.

[20] Stuart Rosen. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1278):367–373, 1992.

[21] Sieb Nooteboom. The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673, 1997.

[22] Shinya Fujii and Catherine Y Wan. The role of rhythm in speech and language rehabilitation: the sep hypothesis. *Frontiers in human neuroscience*, 8:777, 2014.

[23] Joan H Leung, Suzanne C Purdy, Lynette J Tippett, and Sylvia HS Leao. Affective speech prosody perception and production in stroke patients with left-hemispheric damage and healthy controls. *Brain and language*, 166:19–28, 2017.

[24] Ilse Depraetere and Raphael Salkie. *Semantics and pragmatics: Drawing a line*, volume 11. Springer, 2017.

[25] Nai Ding and Jonathan Z Simon. Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, 8:311, 2014.

[26] Charles E Schroeder, Peter Lakatos, Yoshinao Kajikawa, Sarah Partan, and Aina Puce. Neuronal oscillations and visual amplification of speech. *Trends in cognitive sciences*, 12(3):106–113, 2008.

[27] Amalia Arvaniti and Tristie Ross. Rhythm classes and speech perception. *Understanding prosody: The role of context, function and communication*, 13:75, 2012.

[28] Steven Greenberg and Takayuki Arai. What are the essential cues for understanding spoken language? *IEICE TRANSACTIONS on Information and Systems*, 87(5):1059–1070, 2004.

[29] Jess R Kerlin, Antoine J Shahin, and Lee M Miller. Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience*, 30(2):620–628, 2010.

[30] Satoko Hisanaga, Kaoru Sekiyama, Tomohiko Igasaki, and Nobuki Murayama. Language/culture modulates brain and gaze processes in audiovisual speech perception. *Scientific reports*, 6:35265, 2016.

[31] Samuel Evans, Carolyn McGettigan, Zarinah K Agnew, Stuart Rosen, and Sophie K Scott. Getting the cocktail party started: masking effects in speech perception. *Journal of cognitive neuroscience*, 28(3):483–500, 2016.

[32] Ingrid S Johnsrude, Allison Mackey, Hélène Hakyemez, Elizabeth Alexander, Heather P Trang, and Robert P Carlyon. Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24(10):1995–2004, 2013.

[33] Aymeric Devergie, Nicolas Grimault, Barbara Tillmann, and Frédéric Berthommier. Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, 128(1):EL1–EL7, 2010.

[34] Philipos C Loizou, Yi Hu, Ruth Litovsky, Gongqiang Yu, Robert Peters, Jennifer Lake, and Peter Roland. Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *The Journal of the Acoustical Society of America*, 125(1):372–383, 2009.

[35] Caroline Bey and Stephen McAdams. Schema-based processing in auditory scene analysis. *Perception & psychophysics*, 64(5):844–854, 2002.

[36] Orsolya Szalárdy, Alexandra Bendixen, Tamás M Böhm, Lucy A Davies, Susan L Denham, and István Winkler. The effects of rhythm and melody on auditory stream segregation. *The Journal of the Acoustical Society of America*, 135(3):1392–1405, 2014.

[37] Vani G Rajendran, Nicol S Harper, Benjamin D Willmore, William M Hartmann, and Jan WH Schnupp. Temporal predictability as a grouping cue in the perception of auditory streams. *The Journal of the Acoustical Society of America*, 134(1):EL98–EL104, 2013.

[38] Alexandra Bendixen, Susan L Denham, Kinga Gyimesi, and István Winkler. Regular patterns stabilize auditory streams. *The Journal of the Acoustical Society of America*, 128(6):3658–3666, 2010.

[39] Lefkothea-Vasiliki Andreou, Makio Kashino, and Maria Chait. The role of temporal regularity in auditory segregation. *Hearing research*, 280(1-2):228–235, 2011.

[40] Andrew J Sach and Peter J Bailey. Some characteristics of auditory spatial attention revealed using rhythmic masking release. *Perception & psychophysics*, 66(8):1379–1387, 2004.

[41] Robert S Bolia, W Todd Nelson, and Rebecca M Morley. Asymmetric performance in the cocktail party effect: Implications for the design of spatial audio displays. *Human factors*, 43(2):208–216, 2001.

[42] Eveline Geiser, Tino Zaehle, Lutz Jancke, and Martin Meyer. The neural correlate of speech rhythm as evidenced by metrical speech processing. *Journal of Cognitive Neuroscience*, 20(3):541–552, 2008.

[43] Anastasia G Sares, Nicholas EV Foster, Kachina Allen, and Krista L Hyde. Pitch and time processing in speech and tones: the effects of musical training and attention. *Journal of Speech, Language, and Hearing Research*, 61(3):496–509, 2018.

[44] Kevin JP Woods and Josh H McDermott. Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, 115(14):E3313–E3322, 2018.

[45] Joel S Snyder, Olivia L Carter, Suh-Kyung Lee, Erin E Hannon, and Claude Alain. Effects of context on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):1007, 2008.

[46] Robert S Hong and Christopher W Turner. Sequential stream segregation using temporal periodicity cues in cochlear implant recipients. *The Journal of the Acoustical Society of America*, 126(1):291–299, 2009.

[47] Alexander J Billig, Matthew H Davis, John M Deeks, Jolijn Monstrey, and Robert P Carlyon. Lexical influences on auditory streaming. *Current Biology*, 23(16):1585–1589, 2013.

[48] Masanao Ebata. Spatial unmasking and attention related to the cocktail party problem. *Acoustical Science and Technology*, 24(5):208–219, 2003.

[49] Mark A Bee and Christophe Micheyl. The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it? *Journal of comparative psychology*, 122(3):235, 2008.

[50] Nicole Marrone, Christine R Mason, and Gerald Kidd Jr. Evaluating the benefit of hearing aids in solving the cocktail party problem. *Trends in amplification*, 12(4):300–315, 2008.

[51] Paul Avan, Fabrice Giraudet, and Béla Büki. Importance of binaural hearing. *Audiology and Neurotology*, 20(Suppl. 1):3–6, 2015.

[52] Gilles Courtois, Patrick Marmaroli, Morten Lindberg, Yves Oesch, and William Balande. Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions. In *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[53] Adam Kuklasiński, Simon Doclo, Søren Holdt Jensen, and Jesper Jensen. Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 61–65. IEEE, 2014.

[54] Xuedong Huang and Li Deng. An overview of modern speech recognition., 2010.

[55] Valeria Farinazzo Martins Salvador, Joao Soares de Oliveira Neto, and Andre Satoshi Kawamoto. Requirement engineering contributions to voice user interface. In *First International Conference on Advances in Computer-Human Interaction*, pages 309–314. IEEE, 2008.

[56] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[57] István Winkler. Predictive models in auditory stream segregation. *International Journal of Psychophysiology*, 3(77):215, 2010.

[58] Harold L Hawkins, Teresa A McMullen, and Richard R Fay. *Auditory computation*, volume 6. Springer Science & Business Media, 2012.

[59] Stephen Ed McAdams and Emmanuel Ed Bigand. Thinking in sound: The cognitive psychology of human audition. In *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society.* Clarendon Press/Oxford University Press, 1993.

[60] William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[61] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT press, 2011.

[62] John C Middlebrooks, Jonathan Z Simon, Arthur N Popper, and Richard R Fay. *The auditory system at the cocktail party*, volume 60. Springer, 2017.

[63] Timothy D Griffiths and Jason D Warren. What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887, 2004.

[64] Brian CJ Moore and Hedwig E Gockel. Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):919–931, 2012.

[65] Niels R Disbergen, Giancarlo Valente, Elia Formisano, and Robert J Zatorre. Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in neuroscience*, 12:121, 2018.

[66] Jens Blauert. *The technology of binaural listening*. Springer, 2013.

[67] Stanley Gelfand. Hearing: An introduction to psychological and physiological acoustics. 2004.

[68] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[69] William A Yost. Auditory image perception and analysis: The basis for hearing. *Hearing research*, 56(1-2):8–18, 1991.

[70] Sara Popham, Dana Boebinger, Dan PW Ellis, Hideki Kawahara, and Josh H McDermott. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature communications*, 9(1):1–13, 2018.

[71] Ze-Nian Li, Mark S Drew, and Jiangchuan Liu. *Fundamentals of multimedia*. Springer, 2004.

[72] Adelbert W Bronkhorst. The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, 77(5):1465–1487, 2015.

[73] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: facts and models*, volume 22. Springer Science & Business Media, 2006.

[74] Heather J Gilbert, Trevor M Shackleton, Katrin Krumbholz, and Alan R Palmer. The neural substrate for binaural masking level differences in the auditory cortex. *Journal of Neuroscience*, 35(1):209–220, 2015.

[75] James Pickles. *An introduction to the physiology of hearing*. Brill, 2013.

[76] Tahereh Afghah. A brief overview of 3d audio localization and lateralization cues. *University of California San Diego*, 06 2018.

[77] Daniel Pressnitzer, Alain de Cheveigne, Stephen McAdams, and Lionel Collet. *Auditory signal processing: physiology, psychoacoustics, and models*. Springer Science & Business Media, 2006.

[78] Pim Van Dijk, Deniz Başkent, Etienne Gaudrain, Emile De Kleine, Anita Wagner, and Cris Lanting. *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, volume 894. Springer, 2016.

[79] Suzuki Yoiti, Brungart Douglas, and Iida Kazuhiro. *Principles and applications of spatial hearing*. World Scientific, 2011.

[80] Barbara G Shinn-Cunningham, Antje Ihlefeld, and Eric Larson. Bottom-up and top-down influences on spatial unmasking. *Acta Acustica United with Acustica*, 91(6):967–979, 2005.

[81] Tahereh Afghah, Andrew Allen, Peter Otto, and Aravindan Joseph Benjamin. The evaluation of the effect of sound directionality in horizontal plane on the human auditory distance perception in a large reverberant room. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[82] Tahereh Afghah. Perception/evaluation of distance after a continuous change of distance. Master's thesis, Technical University of Ilmenau, 8 2015.

[83] Renita Sudirga, Margaret F Cheesman, and Ewan A Macpherson. Effects of room-acoustic exposure on localization and speech perception in cocktail-party listening situations. *The Journal of the Acoustical Society of America*, 135(4):2366–2366, 2014.

[84] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[85] Hong Zhou, Lucia Melloni, David Poeppel, and Nai Ding. Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. *Frontiers in human neuroscience*, 10:274, 2016.

[86] Gerald Langner. Neural processing and representation of periodicity pitch. *Acta Oto-Laryngologica*, 117(sup532):68–76, 1997.

[87] David H Brainard and Spatial Vision. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

[88] Carol E Osborn. *Statistical applications for health information management*. Jones & Bartlett Learning, 2006.

[89] Morten Fagerland, Stian Lydersen, and Petter Laake. *Statistical analysis of contingency tables*. Chapman and Hall CRC, 2017.

[90] James E De Muth. *Basic statistics and pharmaceutical statistical applications*. Chapman and Hall/CRC, 2014.

[91] Edwin B Wilson and Margaret M Hilferty. The distribution of chi-square. *proceedings of the National Academy of Sciences of the United States of America*, 17(12):684, 1931.

[92] Janet Peacock and Philip Peacock. *Oxford handbook of medical statistics*. Oxford University Press, 2011.

[93] HK Ramakrishna. *Medical statistics: for beginners*. Springer, 2016.

[94] Tahereh Afghah. A brief overview of the statistical data analysis methods and applying significance tests. *University of California San Diego*, 06 2018.

[95] Harvey J Motulsky and Lennart A Ransnas. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The FASEB journal*, 1(5):365–374, 1987.

[96] MartinJ Crowder. *Analysis of repeated measures*. Routledge, 2017.

[97] Wolfgang Karl Härdle, Sigbert Klinke, and Bernd Rönz. *Introduction to statistics: using interactive MM* Stat elements*. Springer, 2015.

[98] Tue Tjur. Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 63(4):366–372, 2009.

[99] JA Cornell and RD Berger. Factors that influence the value of the coefficient of determination in simple linear and nonlinear regression models. *Phytopathology*, 77(1):63–70, 1987.

[100] Olivier Renaud and Maria-Pia Victoria-Feser. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140(7):1852–1862, 2010.

[101] Dabao Zhang. A coefficient of determination for generalized linear models. *The American Statistician*, 71(4):310–316, 2017.

[102] Dario Basso, Fortunato Pesarin, Luigi Salmaso, and Aldo Solari. *Permutation tests for stochastic ordering and ANOVA: theory and applications with R*, volume 194. Springer Science & Business Media, 2009.

[103] Robert Rosenthal and Ralph L Rosnow. *Essentials of behavioral research: Methods and data analysis*, volume 2. McGraw-Hill New York, 1991.

[104] Jun Shao. Mathematical statistics. 2003.

[105] Germán Aneiros, Enea G Bongiorno, Ricardo Cao, and Philippe Vieu. *Functional statistics and related fields*. Springer, 2017.

[106] Bonnie J LaFleur and Robert A Greevy. Introduction to permutation and resampling based hypothesis tests. *Journal of Clinical Child & Adolescent Psychology*, 38(2):286–294, 2009.

[107] Donald Eric Broadbent. *Perception and communication*. Elsevier, 2013.

[108] Albert S Bregman and Steven Pinker. Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1):19, 1978.

[109] Julie G Arenberg, Shigeto Furukawa, and John C Middlebrooks. Auditory cortical images of tones and noise bands. *Journal of the Association for Research in Otolaryngology*, 1(2):183–194, 2000.

[110] David L Woods, G Christopher Stecker, Teemu Rinne, Timothy J Herron, Anthony D Cate, E William Yund, Isaac Liao, and Xiaojian Kang. Functional maps of human auditory cortex: effects of acoustic features and attention. *PloS one*, 4(4), 2009.

[111] Gian Luca Romani, Samuel J Williamson, and Lloyd Kaufman. Tonotopic organization of the human auditory cortex. *Science*, 216(4552):1339–1340, 1982.

[112] Dorea R Ruggles, Alexis N Tausend, Shihab A Shamma, and Andrew J Oxenham. Cortical markers of auditory stream segregation revealed for streaming based on tonotopy but not pitch. *The Journal of the Acoustical Society of America*, 144(4):2424–2433, 2018.

[113] Catherine L Rogers, Jennifer J Lister, Dashielle M Febo, Joan M Besing, and Harvey B Abrams. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27(3):465–485, 2006.

[114] Maria Klatte, Thomas Lachmann, and Markus Meis. Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. *Noise and Health*, 12(49):270, 2010.

[115] Anthony J Watkins. Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, 118(1):249–262, 2005.

[116] James P Cowan. *Handbook of environmental acoustics*. John Wiley & Sons, 1993.

[117] Kim F Fluitt, Timothy Mermagen, and Tomasz Letowski. Auditory perception in open field: Distance estimation. Technical report, ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING ..., 2013.

[118] Erik C Sobel. The locust's use of motion parallax to measure distance. *Journal of Comparative Physiology A*, 167(5):579–588, 1990.

[119] Jon M Speigle and Jack M Loomis. Auditory distance perception by translating observers. In *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*, pages 92–99. IEEE, 1993.

[120] Brian K Shaw, Richard S McGowan, and MT Turvey. An acoustic variable specifying time-to-contact. *Ecological Psychology*, 3(3):253–261, 1991.

[121] Tahereh Afghah, Elliot Patros, and Miller Puckette. A pseudoinverse technique for the pressure-matching beamforming method. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[122] Tahereh Afghah, Elliot Patros, and Miller Puckette. The physical evaluation of the efficiency of an enhanced pressure-matching beamforming method using eigen decomposition pseudoinverse mathematical approach. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[123] Miller Puckette, Elliot Patros, and Tahereh Afghah. Accoustic beamforming, June 2019. International Patent, International Application No.:PCT/US2018/065047, Pub. No.: WO/2019/118521.

[124] Elliot Patros, Tahereh Afghah, and Peter Otto. Reproduction and analysis of near and far stereophonic sound fields with compact uniform linear arrays. *The Journal of the Acoustical Society of America*, 146(4):3058–3058, 2019.