TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Diploma Thesis

Perceptual Modeling: Factors influencing Speech Intelligibility in a Multitalker Environment and Applications in Speech Separation

ausgeführt am
Institut für Analysis und Scientific Computing
an der Technischen Universität Wien

unter Anleitung von
a.o. Univ.-Prof. Dipl.-Ing. Dr.Dr. Frank Rattay

durch
Andrea Kainz
Herbeckstraße 47/2
1180 Wien

Wien, 12. Mai 2017 _____

Abstract

The aim of this thesis is the investigation of speech intelligibility in multitalker environments, where the challenge for the listener is to focus on one speaker in the presence of simultaneous interfering talkers or background noise in order to follow the conversation. In general, this is not a difficult task for normal hearing people, but it can be a challenge for people suffering from hearing impairment. Furthermore, it still remains a problem for machines to deal with interfering speech signals.

Within this thesis, different speech segregation algorithms and their mathematical and statistical background are presented. There are different approaches of processing interfering speech signals. Motivated by the powerful ability of the auditory system to analyze and segregate incoming sounds, Computational Auditory Scene Analysis (CASA) aims at replicating the different auditory processing stages. Another essential approach in the context of the separation of interfering speech signals which differs from CASA is Blind Source Separation (BSS) which uses results from Statistics and Information Theory to separate a signal mixture into its sources.

In the experimental part of the thesis, a speech intelligibility (SI) test was performed which was implemented in MATLAB® (R2015b). The aim was the investigation of factors affecting Speech Intelligibility where the main focus was on analyzing attributes of the masker signals and their influence on speech perception of the target signal. 12 normal hearing listeners participated in the test and the task was to determine the target signals in the presence of different masker signals. The target signals consisted of 14 nonsense-syllables (e.g. 'affa' or 'assa') from the Oldenburger Logatome Corpus (OLLO) spoken by four female persons. The masker signals included sentences from the Oldenburger Satztest (e.g. 'Britta verleiht elf alte Bilder'), the International Speech Test Signal (ISTS) and Speech Shaped Noise (SSN). The test was evaluated using a two-way repeated measures analysis of variance (ANOVA) in SPSS® Statistics (24) including the two within-subject factors "Signal-to-Noise Ratio" (SNR) and "Masker Type". The results showed a significant main effect in both factors ($p<0.001$) and in further research, ANOVA also demonstrated a significant influence of the factors "Number of Maskers" ($p<0.001$) and "Spectral Diversity of the Masker" ($p<0.001$) on speech intelligibility.

Zusammenfassung

Das Ziel dieser Arbeit ist die Untersuchung von Sprachverständlichkeit in Situationen, in denen ein Hörer vor der Aufgabe steht, ein Sprachsignal aus mehreren Sprechern und Hintergrundgeräuschen zu extrahieren, um einer Unterhaltung folgen zu können. Im Allgemeinen stellt dies kein Problem für den Menschen dar, jedoch sind solche sogenannten "Cocktail-Party Szenarien" für ältere Personen oder jene mit Schwerhörigkeit meist schwieriger zu bewältigen. Weiters ist es bis heute in der Signalverarbeitung eine Herausforderung, überlagerte Sprachsignale in ihre ursprünglichen Anteile zu zerlegen.

Im Rahmen dieser Arbeit werden verschiedene Algorithmen zur Trennung von Sprachsignalen vorgestellt und ihr mathematischer und statistischer Hintergrund wird erläutert. Basierend auf Segmentierungs- und Gruppierungsprozessen sowie Erkenntnissen aus der Gestaltspsychologie wird mittels Computational Auditory Scene Analysis (CASA) das menschliche Gehör simuliert, um eine gewünschte Quelle aus einem gemischten Signal herauszufiltern. Eine weitere Möglichkeit zur Trennung von Sprachsignalen bietet Blind Source Separation (BSS), wobei sich dieses Verfahren hinsichtlich seiner Herangehensweise grundlegend vom zuvor genannten unterscheidet. Hier besteht die Ausgangslage in mehreren Sensoren mit Aufzeichnungen eines überlagerten Sprachsignals und dieses wird mittels statistischer Verfahren oder Resultaten aus der Informationstheorie in seine einzelnen Quellen unterteilt.

Im experimentellen Teil der Arbeit wurde ein Sprachverständlichkeitstest durchgeführt mit dem Ziel, den Einfluss des Signal-Rausch-Verhältnisses (SNR) und der Eigenschaften des sogenannen Maskierers auf die Verständlichkeit des Zielsignals zu untersuchen. Der Test wurde in MATLAB® (R2015b) implementiert, 12 Personen nahmen daran teil. Die Aufgabe bestand darin, 14 Nonsens-Silben (beispielsweise ‘affa’ oder ‘assa’) aus dem Oldenburger Logatomkorpus (OLLO) in verschiedenen Hintergrundszenarien zu bestimmen, wobei insgesamt Aufnahmen von vier weiblichen Sprechern verwendet wurden. Die Maskierer beinhalteten Sätze aus dem Oldenburger Satztest (beispielsweise ‘Britta verleiht elf alte Bilder’), das Internationale Sprachtest Signal (ISTS) und Speech Shaped Noise (SSN), ein bestimmtes Rauschen mit spektralen Eigenschaften der weiblichen Stimme. Der Test wurde mit einer zweifaktoriellen Varianzanalyse mit Messwiederholung (ANOVA) in SPSS® Statistics (24) mit den Innersubjektfaktoren "SNR" und "Maskierungsart" ausgewertet. Die Resultate der Analyse zeigten signifikante Haupteffekte in beiden Faktoren ($p<0.001$) und im Rahmen weiterer Tests wurde ebenfalls ein signifikanter Einfluss der Faktoren "Anzahl der Maskierer" ($p<0.001$) und "Spektrale Vielfalt des Maskierers" ($p<0.001$) auf die Sprachverständlichkeit festgestellt.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich während des Studiums und des Verfassens dieser Diplomarbeit unterstützt haben.
In erster Linie gilt mein Dank meinem Professor ao. Univ.-Prof. Dipl.-Ing. Dr.Dr. Frank Rattay, den Betreuer und Begutachter meiner Diplomarbeit, der mir immer wieder wertvolle Tipps gab und mich bei wichtigen Fragen unterstützte.
Ein besonderer Dank gilt auch jenen Personen, die sich die Zeit nahmen, an meinem Sprachverständlichkeitstest teilzunehmen.
Weiters möchte ich mich bei meinen Freunden und Studienkollegen bedanken, die mich während des Studiums unterstützten und mir den nötigen Rückhalt gaben.
Abschließend möchte ich mich ganz herzlich bei meiner Familie bedanken, die mich immer motivierte und mir emotional zur Seite stand.

List of Abbreviations

AD Attenuated and Delayed

AMM Amplitude Modulation Masking

ANOVA Analysis of variance

ASA Auditory Scene Analysis

BSS Blind Source Separation

CASA Computational Auditory Scene Analysis

CVC Consonant-Vowel-Consonant

EM Energetic Masking

FFT Fast Fourier Transform

IBM Ideal Binary Mask

ICA Independent Component Analysis

IHC Inner Hair Cell

IM Informational Masking

HRTF Head-Related Transfer Function

LI Linear Instantaneous

LTASS Long-Term Average Speech Spectrum

MS Mean Sum of Squares

OHC Outer Hair Cell

OLLO Oldenburg Logatome Corpus

rANOVA Repeated Measures ANOVA

SAF Summary Autocorrelation Function

SI Speech Intelligibility

SNR Signal-to-Noise Ratio

SRT Speech Reception Threshold

SS Sum of Squares

TF Time-Frequency

VCV Vowel-Consonant-Vowel

Contents

1 Introduction

1.1 Motivation

During a typical conversation, there are multiple competing sounds at the same time including different speakers or noise. In general, human beings manage to focus on one sound source but these so-called "Cocktail-Party Situations" can be very difficult for people suffering from hearing impairment.

Different kinds of masking effects influence speech intelligibility (SI) in the presence of multiple talkers and factors like age and hearing impairment play an important role in masked speech perception. In order to overcome the difficulties for hearing-impaired people, the goal of hearing aid designs is the improvement of speech intelligibility. Approaches like Computational Auditory Scene Analysis and Blind Source Separation manage to improve speech intelligibility in the presence of multiple talkers significantly and as a result, the implementation of real-time digital signal processing in hearing aids has become an essential topic.

Many studies have investigated speech intelligibility in high-context speech scenarios including words or sentences. However, only few investigation has been made in the fields of intelligibility of low-context speech segments like Vowel-Consonant-Vowels (VCVs) or Consonant-Vowel-Consonants (CVCs) in maskers with varying properties. Nevertheless, the perception of high-context speech segments like words or sentences depends critically on the error rate of low-context speech segments, which underlines the necessity of studies on low-context speech segments.

1.2 Research Question

Within this thesis, different approaches for analyzing overlapped speech signals are presented and the intelligibility of low-context speech segments in the presence of different maskers is investigated. A speech intelligibility test was performed to test the following hypothesis:

- $H_{0,1}$: Signal-to-Noise Ratio (SNR) significantly influences low-context speech intelligibility
- $H_{0,2}$: Masker type significantly influences low-context speech intelligibility

1.3 Thesis Organization

In the first chapter of this thesis, the theoretical background is presented starting with the auditory processing stages and perception on a higher cognitive level like auditory scene analysis (ASA). Furthermore, speech production and perception mechanisms as well as masking effects are investigated.

The next chapter deals with the mathematical and technical background of speech separation and processing algorithms. An important result of digital signal processing, the Nyquist-Shannon Theorem, is presented and proven. Next, the focus is on speech processing methods like Computational Auditory Scene Analysis (CASA), which mimics the auditory processing stages followed by Blind Source Separation (BSS), that aims at the separation of speech signals based on statistical independence. Eventually, the statistical background of Analysis of Variance (ANOVA) is presented, which is used in the evaluation of the experimental part of the thesis.

The technical part of the thesis deals with the description of the experiment which included a speech intelligibility test in the presence of different maskers. The preparation of the material and the different target-masker configurations are described and finally, the results are presented and statistically evaluated followed by a confusion analysis of the target signals.

2 Theoretical Background

In this chapter, the theoretical background of the thesis is presented. The anatomy of the ear and the different stages of the hearing process are described, followed by a study of speech production and speech intelligibility. Finally, auditory masking effects and their contribution to speech intelligibility are investigated.

2.1 Auditory Perception

Auditory perception includes the physiological hearing process starting from the sound wave reaching the outer ear up to higher-level processing in the fields of cognitive science and psychoacoustics.

2.1.1 Anatomy of the Ear and Hearing Process

To start with, the anatomic structure of the ear is described in the following picture.

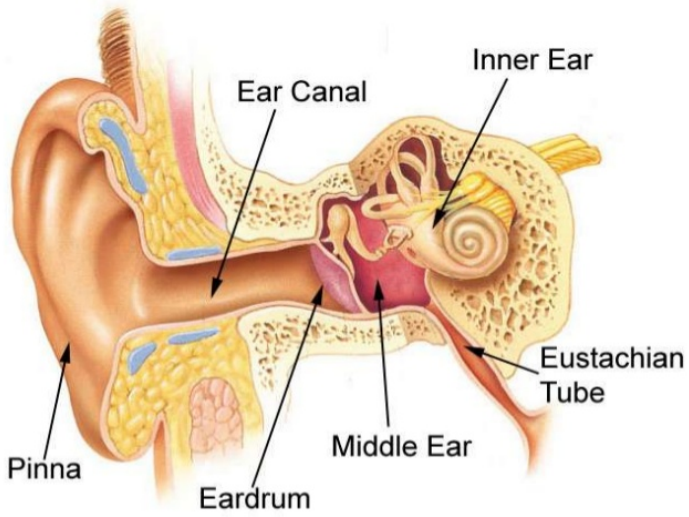


Figure 2.1: *Anatomy of the ear: The pinna and the ear canal define the main parts of the outer ear and the middle and the outer ear are separated by the tympanic membrane, which is also called eardrum. In the middle ear, the three ossicles, the hammer, the anvil and the stirrup are located and the cochlea is situated in the inner ear.*
This image is taken from https://sites.google.com/site/dranhtruong/bellevue-ruptured-eardrum-perforation

Detailed Structure and Hearing Process:

The pinna collects and channels the sound into the ear canal and the tympanic membrane vibrates in the frequency of the sound. The three ossicles, the hammer, anvil and

the stirrup conduct the sound to the inner ear. The main part of the inner ear is the cochlea which amplifies the sound and transduces the vibration into nervous impulses. Also, frequency and intensity of the sound is analyzed and informations like the sound level are transformed to the brain by the rate of nerve firing (Alberti, 2001).

Anatomy of the Inner Ear:

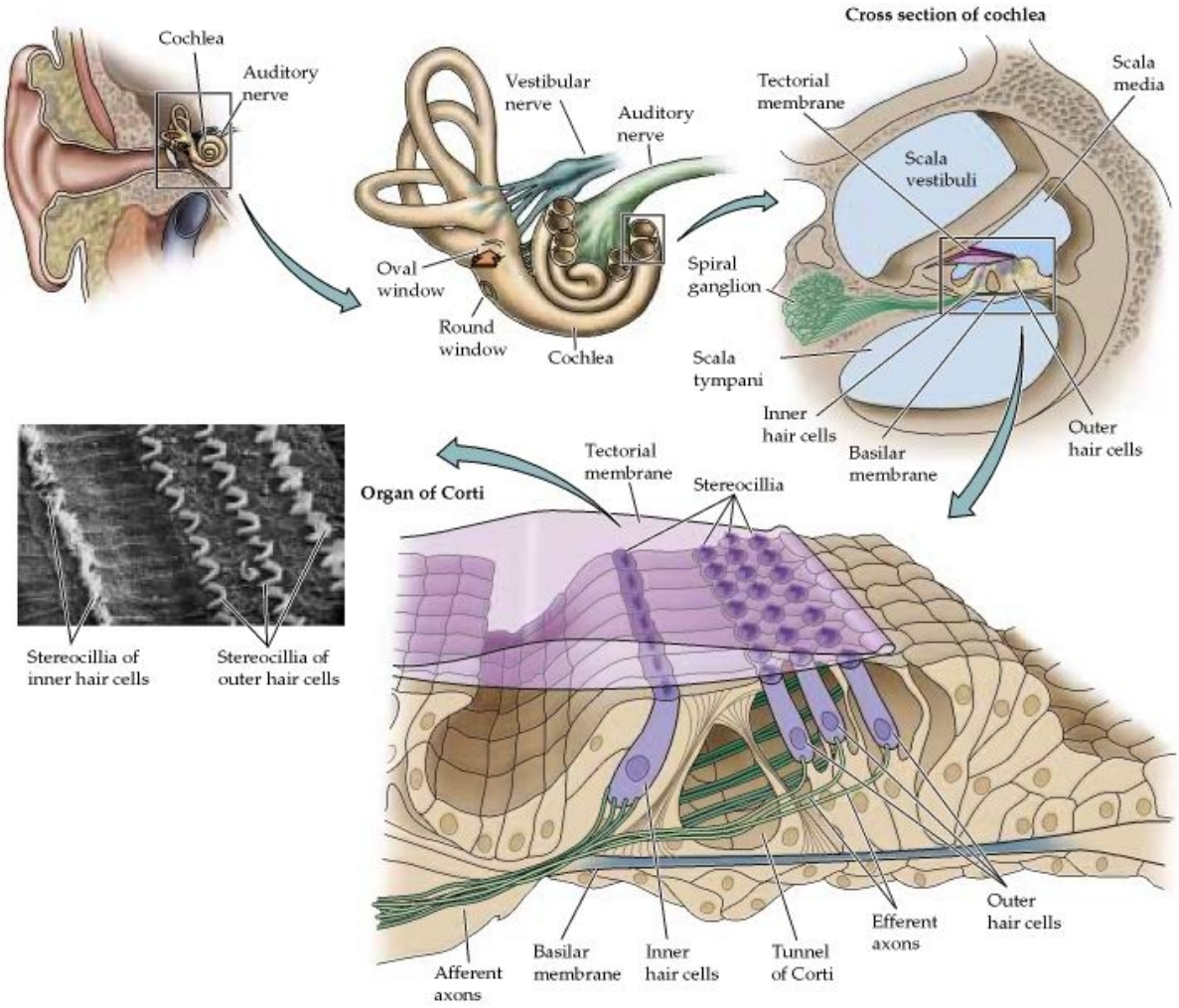


Figure 2.2: *Anatomy of the Inner Ear: The Organ of Corti is located on the basilar membrane which contains one row of inner hair cells (IHC) and three rows of outer hair cells (OHC). The stereocilia are bundles extending from the hair cells and they perform the transduction of sound into neural impulses via a shearing action (Alberti, 2001).*
The image is taken from http://flipper.diff.org/app/items/info/6238.

Place Theory of Hearing:

The place theory was first stated by Helmholtz in 1863.
According to the theory, each position on the basilar membrane is associated with a certain frequency, where the highest frequency response is at the basal end and the lowest frequency is on the apical end. This is called *tonotopic mapping of the cochlea.*

The place theory was later examined by Bekesy who observed that a traveling pressure wave reached this resonance point (Von Békésy and Wever, 1960; Rattay and Lutter, 1997).

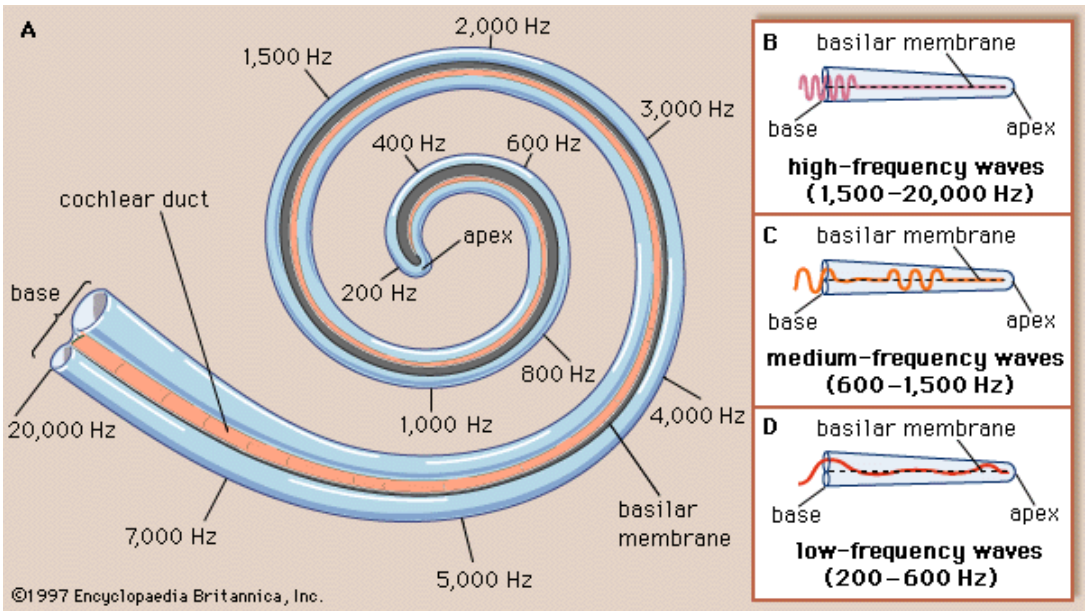


Figure 2.3: *Tonotopic Mapping of the Cochlea: the highest frequency response is situated on the basal end and the lowest on the apical end.*
This image is taken from https://www.britannica.com/science/bony-labyrinth.

Frequency Theory of Hearing:

The frequency theory of hearing was founded by Rutherford in 1886.
This theory assumes the signal is entirely described by the firing pattern of the auditory nerve fibers (Rutherford, 1886).

2.1.2 Binaural Hearing

Human beings have two ears which work more or less independently from each other. When the sound enters both ears in the hearing process, the stimulations indicate some important differences in various aspects such as time of arrival, level as well as the spectrum of the signal (Sayers and Cherry, 1957).

Binaural hearing includes various advantages in the hearing process as it significantly improves hearing quality at the sound segregation stage and the localization of sound sources.

Furthermore, it can be very difficult to understand speech in a multitalker environment, especially at a low sound intensity. Binaural hearing can be very advantageous in these scenarios and it may result in a benefit up to 7 dB in comparison with monaural hearing (Hawley et al., 2004).

Sound Source Localisation:

In general, the human being is able to localize sounds on different levels. The three most important cues for auditory localization are the following:

- **Interaural Time Difference (ITD)**: The ITD describes the difference between the times the sound reaches each ear.

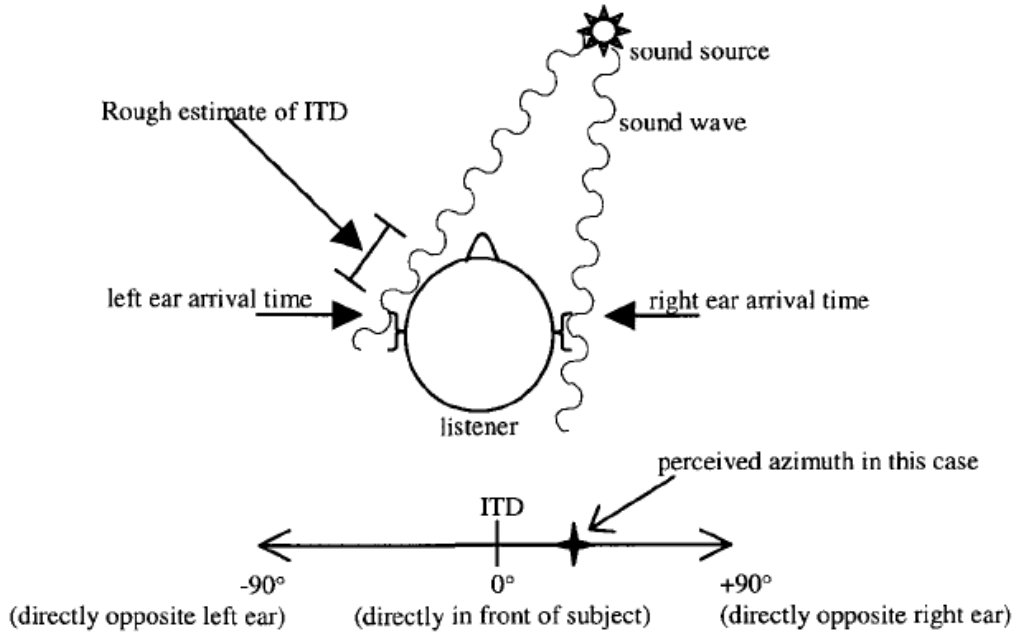


Figure 2.4: *Sound source localization: The ITD indicates the sound source location. This image is taken from http://archive.cnx.org/contents/00e2b539-92b7-4d86-9407-cca7cb190c6f@2/background.*

- **Interaural Level Difference (ILD)**: The ILD refers to the difference between the sound pressure levels that reach each ear.

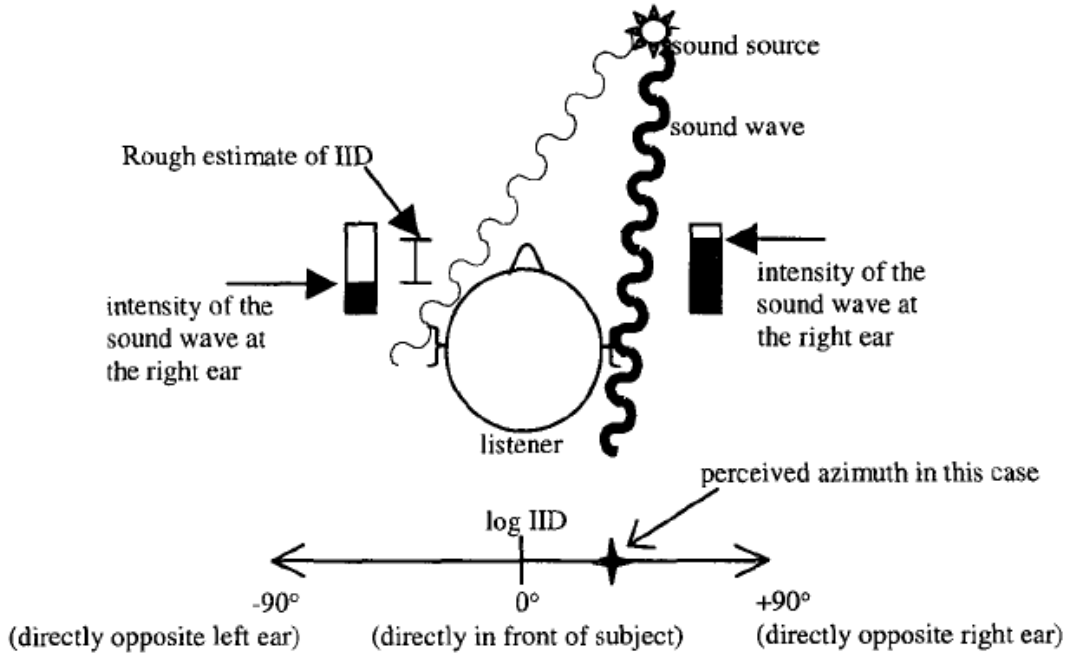


Figure 2.5: *The ILD improves sound source location as well. This image is taken from http://archive.cnx.org/resources/ 6f818c24a9a38f919bd2991831f208180b2d447e/Picture.*

- **Head-Related Transfer Functions (HRTF)**: The HRTF corresponds to diffraction and reflection effects due to the size and shape of the head, the torso and the pinna.

2.1.3 Acoustics and Psychoacoustics

The hearing process is now analyzed on a higher level. To start with, the *sound pressure level* and the *sound intensity level* are defined:

Definition 2.1 (Sound pressure level)**.**

$$L = 20\log_{10}\frac{p}{p_0}\ \text{dB} \quad \text{with } p_0 = 20\ \mu Pa \tag{2.1}$$

Definition 2.2 (Sound intensity level)**.**

$$L_I = 10\log_{10}\frac{I}{I_0}\ \text{dB} \quad \text{with } I_0 = 1\text{pW/m}^2 \tag{2.2}$$

Humans are able to hear sound waves with frequencies between about 20 Hz and 20 kHz where sound above 20 kHz is called *ultrasound* and sound below 20 Hz is called *infrasound.*

The human range of hearing is 20 Hz to 20 kHz and the threshold of audibility depends on the frequency of the sound as can be seen in Figure 2.6.

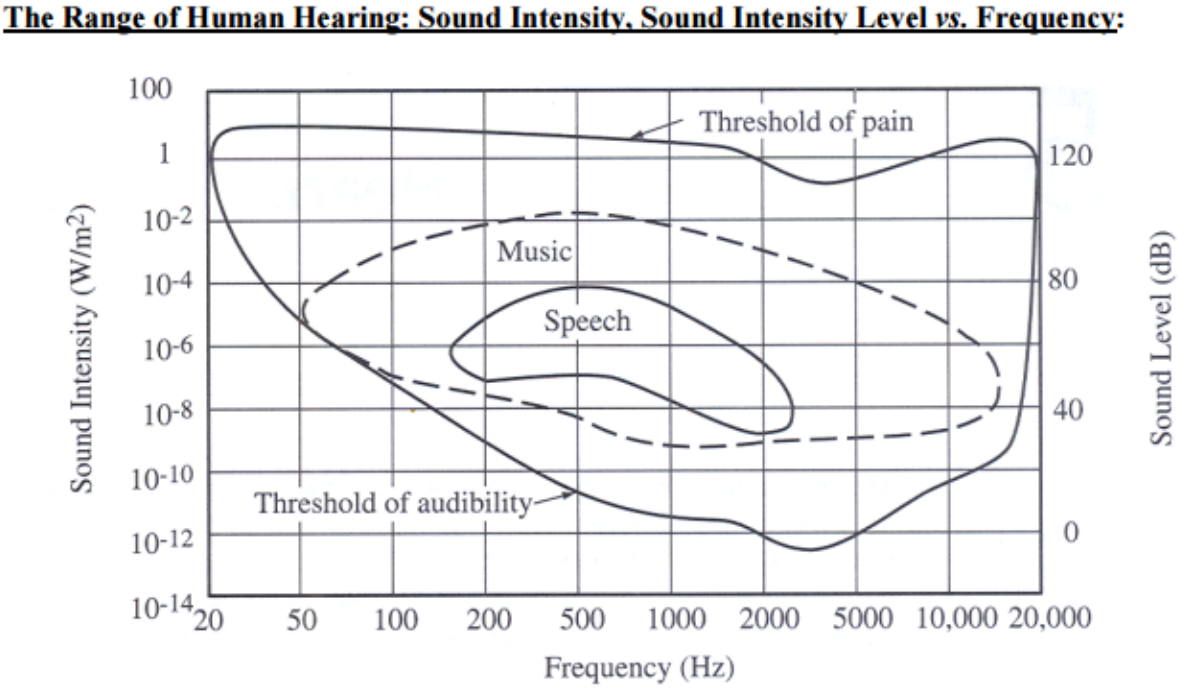


Figure 2.6: *Range of Hearing: threshold of audibility and threshold of pain dependent on the frequency.*
The image is taken from http://www.livecollectiva.com/2015/10/review-of-3m-peltor-earmuffs.html

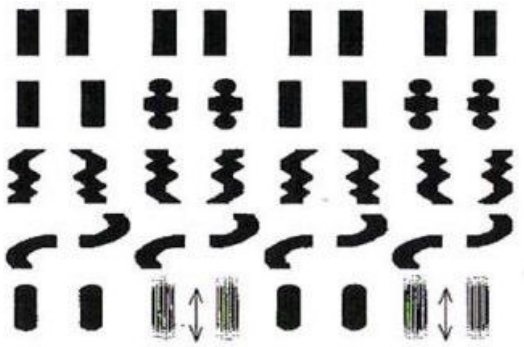
2.1.4 Auditory Scene Analysis

The results in this subsection are taken from Bregman (1994).

In a typical listening situation there are many different acoustic sound sources at the same time. The difficulty is that only a single pressure wave arrives at the ear which includes interleaved and overlapped components in time and frequency. The auditory system has to segregate and group the incoming information into separate mental descriptions. This process is called auditory scene analysis (ASA) and the separate sound patterns are called auditory streams. The term Auditory Scene Analysis was first mentioned in Bregman (1994).

In order to form the streams and to separate the incoming signal into separate segmentations, different grouping cues are used that also occur in vision.

Gestalt's Principles



- Proximity
- Similarity
- Symmetry
- Continuity
- Common fate

Figure 2.7: *Gestalt's principles. The image is taken from Shepard and Levitin (2002)*

The auditory streams are formed by two different approaches of grouping processes: *Sequential Grouping* and *Simultaneous Grouping*. Sequential Grouping refers to grouping over time and Simultaneous Grouping to grouping over frequency.

Simultaneous Grouping:

Simultaneous grouping describes the division of data arriving at the same time into different sources. There are different cues that indicate that components are coming from the same source such as synchrony of onsets and offsets, spatial location and same patterns of amplitude fluctuation.

Sequential Grouping:

Sequential grouping refers to similarities in the spectrum from one moment to the next and it connects sense data over time.

As an example, the streaming phenomenon is presented.

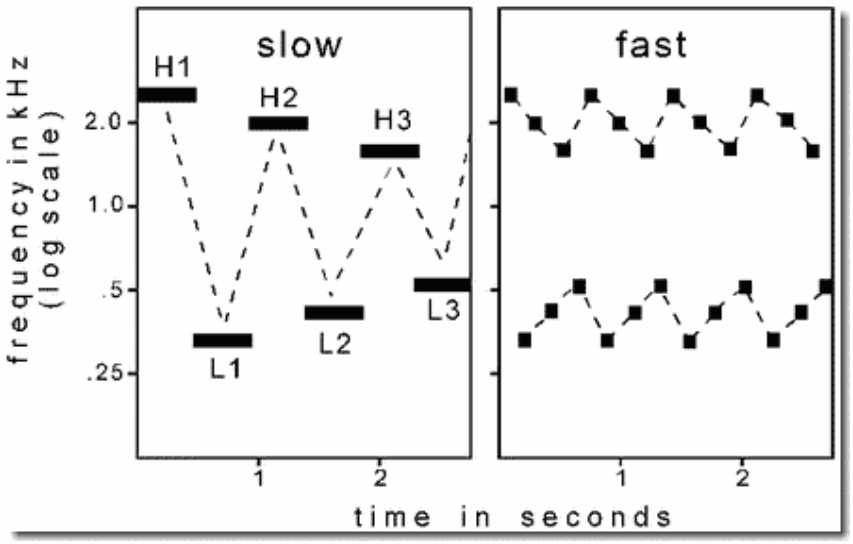


Figure 2.8: *The streaming phenomenon: There are pure tones of high frequency and of low frequency. Starting from a low speed, the tones are alternated. At a low speed, the cycles are considered as one pattern whereas at a high speed, perceptual grouping into two distinct streams is performed. The image is taken from http://webpages.mcgill.ca/staff/Group2/abregm1/web/images/Fig01.gif*

Bottom-up and top-down aspects of ASA:

In Auditory Scene Analysis, *bottom-up* (primitive) and *top-down* (knowledge-based) aspects are distinguished.

Primitive processes are subject of most of ASA research and they are defined by the acoustic properties of the input. On the other hand, knowledge-based processes include conscious attention and past experiences with different sounds.

2.2 Speech

The following chapter describes the basic mechanisms of speech production and perception and the spectral properties of speech as well as factors influencing speech intelligibility are investigated. It is based on Benesty et al. (2007).

2.2.1 Speech Production

Firstly, a brief overview of the speech production process is presented. This is of importance, because the spectral properties change in relation to different speech production mechanisms, which, in conclusion, influences speech intelligibility.

In Figure 2.9, the human speech production system is presented.

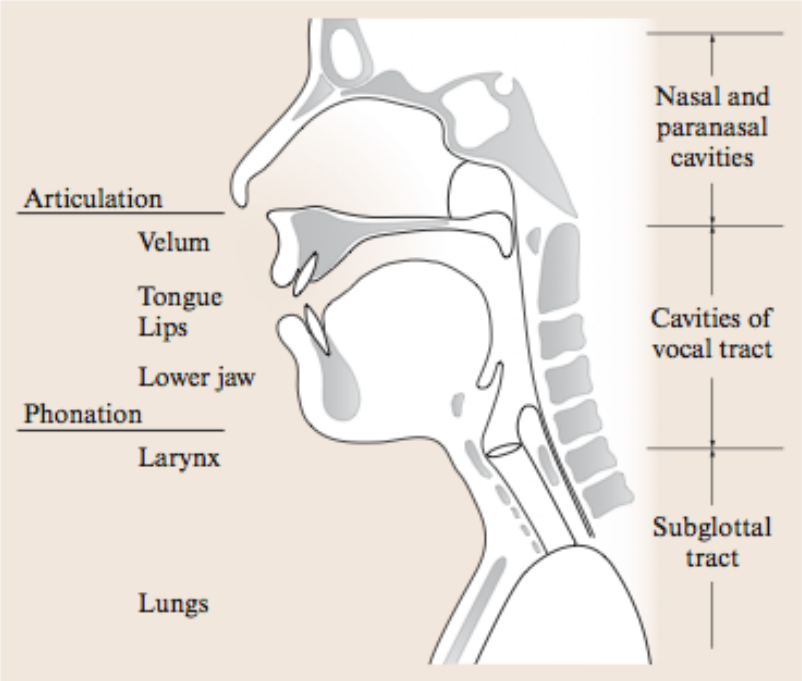


Figure 2.9: *Speech Production System: The sound propagation takes place at three levels: the subglottal tract, the vocal tract and nasal and paranasal cavities. The image is taken from Benesty et al. (2007), p.8.*

Phonation refers to the periodic vibration of vocal folds which generates voiced speech sounds whereas articulation describes the generation of voiceless sounds. All vowels are voiced whereas consonants can either be voiced or unvoiced.

2.2.2 Consonants

Consonants can be classified in the following way:

1. Voiced (vocal chords) and unvoiced
2. Place of articulation
3. Manner of articulation

The manner of articulation influences the spectrotemporal features of the signal.

Manner of Articulation:

- plosive (stops): ‘b’, ‘p’, ‘t’,‘d’,‘k’,‘g’
- fricative: ‘s’, ‘f’, ‘sh’,‘w’
- affricate: ‘s’
- nasal: ‘m’, ‘n’
- lateral: ‘l’
- glide: ‘j’

2.2.3 Vowels

In the classification of vowels, the *fundamental frequency* f_0 plays an important role which describes the lowest harmonic component of a voiced sound and corresponds to the natural frequency of vocal fold vibration.

A *formant* is the concentration of acoustic energy around a particular frequency in the speech wave. The formants f_1 and f_2 are of great importance in the fields of vowel perception. They can be seen in Figure 2.10.

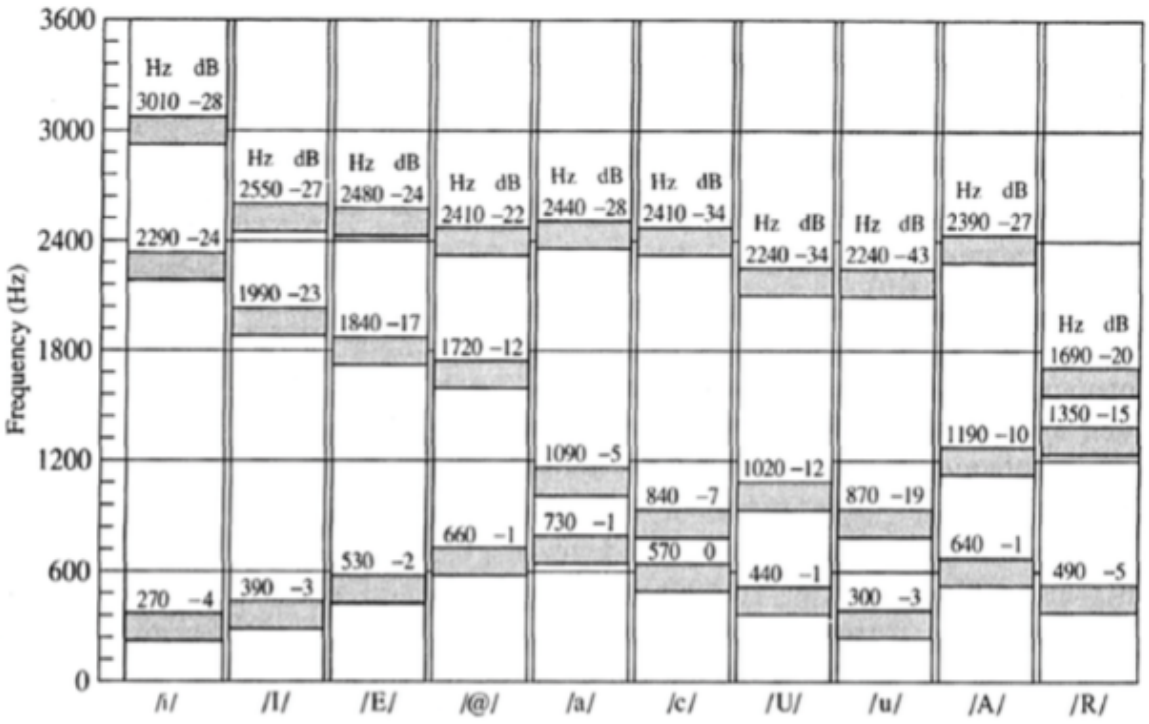


Figure 2.10: *Average formant locations for vowels in American English. The image is taken from Peterson and Barney (1952)*

Range of Fundamental Frequency in Human Speech:

- male: 80-400 Hz
- female: 120-800 Hz

The term *pitch* is a perceptual concept which is mostly used to describe the perceived fundamental frequency, an acoustical term. It should not be confused with *tone height* which coincides with pitch only in sinusoids (Benesty et al., 2007).

Different Presentations of a Speech Signal:

In the following, different presentations of the syllable 'iddi' spoken by a female speaker are presented. The syllable is taken from the Oldenburger Logatome Corpus (OLLO) which was also used in the speech intelligibility test in this thesis.

- *Time-domain Presentation* (Oscillogram):

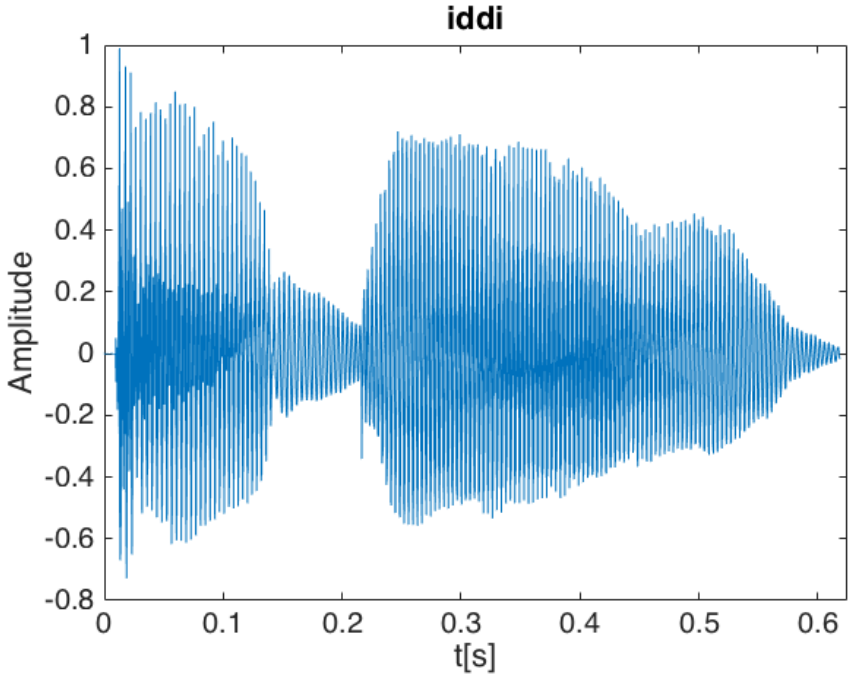


Figure 2.11: *Oscillogram of the syllable 'iddi' by a female speaker: The oscillogram shows the variations of the amplitude of the signal over time.*

- *Frequency-domain Presentation* (Power Spectral Density):

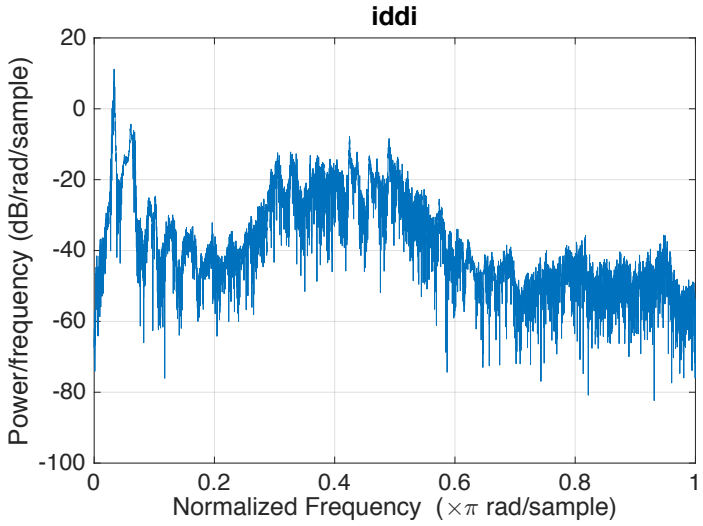


Figure 2.12: *Spectrum of the syllable 'iddi' by a female speaker: The power spectral density describes the distribution of the power over the frequency components of the signal.*

- *Time-Frequency Presentation* (Spectrogram): The spectrogram of a signal shows the spectrum of the frequencies of the signal varying in time. In general, it is a two-dimensional time-frequency presentation of a signal.

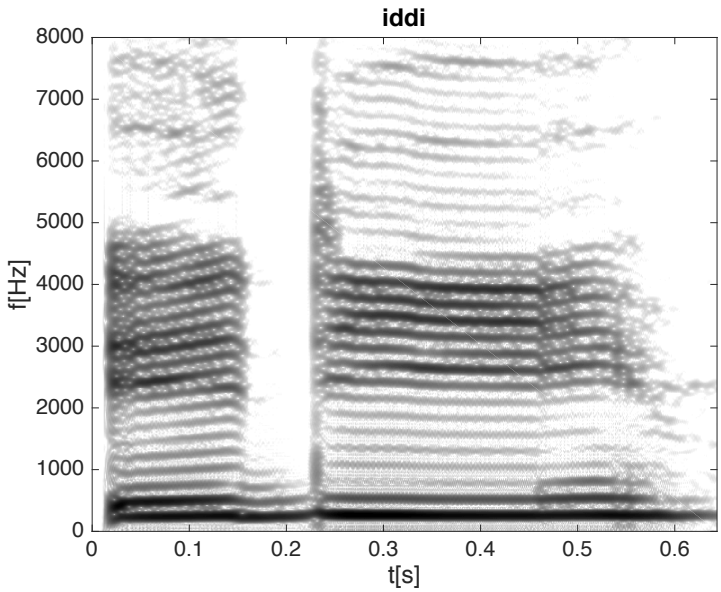


Figure 2.13: *Spectrogram of the syllable 'iddi' by a female speaker: The formants of the vowel 'i' can be seen very well.*

2.2.4 Speech Intelligibility

Speech intelligibility (SI) is an essential topic in various fields of research, for example in the design of hearing aids, because speech intelligibility still remains a problem for hearing-impaired people (Benesty et al., 2007).

In the context of speech intelligibility tests, the term "Speech Intelligibility" is defined as follows:

Definition 2.3 (Speech Intelligibility)**.** *Speech Intelligibility is the proportion of speech items (e.g. syllables, words or sentences) correctly repeated by listener for a given speech intelligibility test (Benesty et al., 2007).*

Several algorithms manage to improve speech quality. However, this does not necessarily imply the improvement of speech intelligibility, which is still an unresolved issue in research (Loizou and Kim, 2011).

Factors influencing Speech Intelligibility (Bronkhorst, 2015):

- spectral differences between target and interfering sounds
- spatial configuration of the sound sources
- reverberation
- (degree of) hearing impairment
- fluctuations in level
- masking effects

Subjective Speech Intelligibility Measurements:

Speech intelligibility tests are a common method for diagnosing hearing impairments or testing the power of hearing aids. Furthermore, they may be used in research for the investigation of speech processing and perception.

There are different forms of speech intelligibility tests including sentences, words or nonsense-syllables which are also called logatomes. Logatomes may be categorized in VCVs (Vowel-Consonant-Vowels), for example ‘adda’, or CVCs (Consonant-Vowel-Consonants), like ‘sas’ (Fellbaum, 2013).

- **Sentence Tests:** e.g. Oldenburger Satztest
- **Word Tests:** e.g. Freiburger Worttest
- **Logatome Tests:** e.g. VCV, CVC

Objective Speech Intelligibility Measurements:

There are various methods to measure speech intelligibility. Besides empirical methods using hearing tests and an afterward statistical evaluation, some objective measurements have been founded.

Two of the most common methods for the prediction of SI are the Articulation Index (AI) which was later renamed Speech Intelligibility Index (SII) and the Speech Transmission Index (STI). All of these indices make the assumption that speech is coded by several speech channels that carry independent information.

$$AI = \sum_i AI_i \tag{2.3}$$

In ANSI (1997), the SII is defined as follows:

"The SII measure is based on the idea that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands contributing equally to intelligibility and estimating the weighted average of the signal-to-noise ratios in each band".

2.3 Masking Effects

Masking effects play an important role in speech perception and speech intelligibility. In the last decades, many investigations have been made in this area.

2.3.1 Definitions and Classification of Masking Effects

In literature, different definitions and classifications of the term "masking" occur. The following basic definition is taken from Benesty et al. (2007):

Definition 2.4 (Auditory Masking Effect)**.** *Auditory masking effects describe the affection of sound perception in the presence of another sound, the masker.*

Masking effects can be classified in their temporal appearance in relation to the target.

Temporal Classification of Masking Effects:

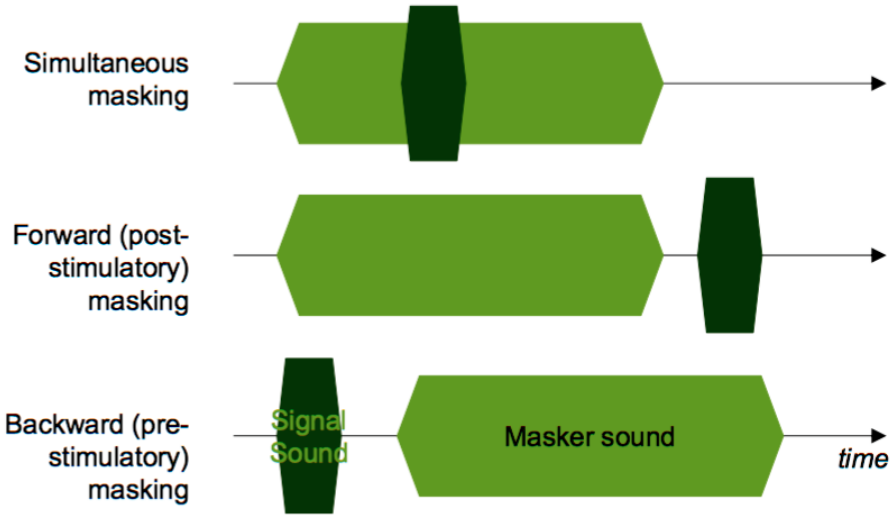


Figure 2.14: *Temporal masker and target configurations: 1) at the same time, 2) masker before target 3) masker after target signal*
The image is taken from http://neurobiologyhearing.uchc.edu

Other classifications of masking effects include neural masking and dynamic masking as well as energetic masking, informational masking and amplitude modulation effects in speech which will be defined in the next chapter (Benesty et al., 2007; Brungart, 2001).

There are different factors that influence masking effects like number of talkers or the spatial configuration. Futhermore, age and hearing impairment also influence masked speech perception (Goossens et al., 2017).

2.3.2 Release from Masking

In some scenarios, release from masking can occur which - as a result - improves audibility and speech intelligibility.

Spatial Release from Masking:

When sources are spatially separated, release from masking occurs (Kidd Jr et al., 1998; Freyman et al., 2001; Arbogast et al., 2002).

Binaural Unmasking:

Binaural hearing can lead to a considerable amount of release from masking (Levitt and Rabiner, 1967; Durlach et al., 1986).

In Figure 2.15, the binaural masking release can be observed. Hearing the pure tone and the noise in both ears leads to poor tone detection whereas inverting or removing one of the signals of one ear improves tone detection.

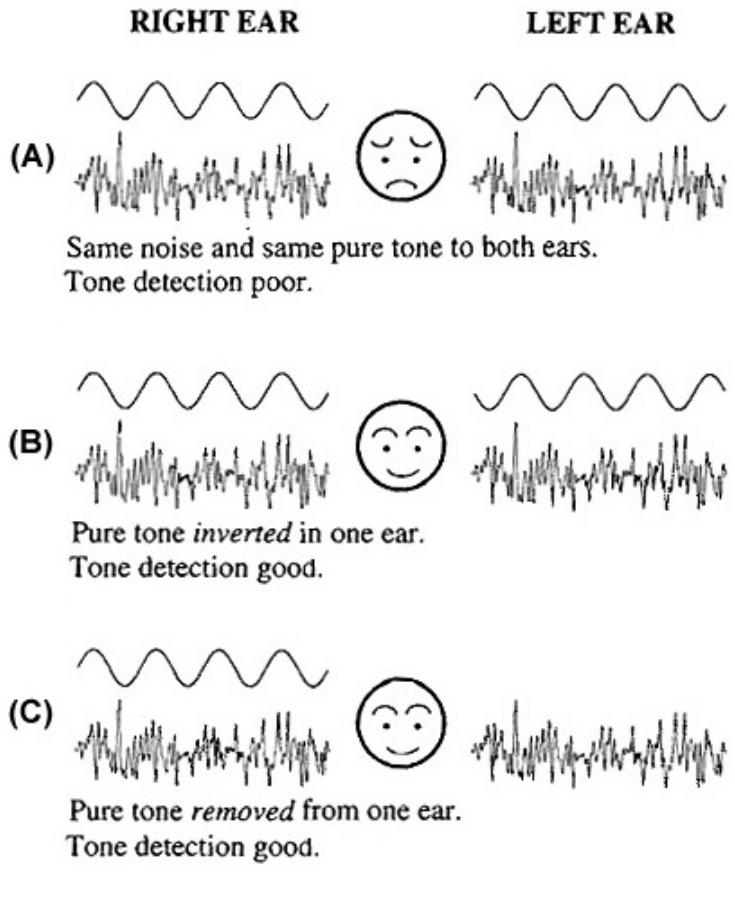


Figure 2.15: *Binaural masking release: Improvement in tone detection. The image is taken from http://acousticslab.org/psychoacoustics/PMFiles/Module07a.html*

2.3.3 Masking Effects in Speech

In the following, masking effects which especially occur in interfering speech signals are presented and analyzed.

In comparison with pure tone masking, higher-level masking effects like informational masking and modulation masking effects play an important role in speech processing. The following definitions are taken from Brungart (2001) and Brungart et al. (2001).

Energetic Masking:

In 1940, H. Fletcher introduced the concept of the critical band. He stated, that the auditory system contains a bank of overlapping bandpass filters, where a stronger signal masks a weaker signal within a critical band (Fletcher, 1940).

Energetic masking (EM) during the perception of one or multiple competing talkers occurs when the utterances of target and masker contain energy in the same critical bands at the same time or, in other words, the concurrent sounds overlap in time and frequency, and, as a result, portions of the target speech signals become inaudible (Greenwood, 1961).

Informational Masking:

Informational masking (IM) occurs on a higher cognitive level. Although target and masker signals are audible, the listener is unable to distinguish elements of the target from the similar-sounding masker signal.

The term *informational* underlines the interference of the informational component in comparison with the energetic masking effect (Evans et al., 2016; Brungart et al., 2001; Srinivasan and Wang, 2008; Goossens et al., 2017).

There are different scenarios leading to IM (Ihlefeld and Shinn-Cunningham, 2008; Durlach et al., 2003; Cooke et al., 2008):

- similarity between target and masker regarding perceptual or linguistic attributes
- uncertainty about either target or masker
- failures in segregation or attention

Amplitude Modulation Masking:

Amplitude modulation masking (AMM) occurs when there is an interaction between the temporal modulations in the target signal and the masker (Schubotz et al., 2016; Dubbelboer and Houtgast, 2008).

2.3.4 Masking Effects in the Presence of Multiple Talkers

The number of competing talkers plays an important role in the analysis of masking effects. It is not fully understood, how the overall amount of masking relates to the single masking effects.

In some multitalker experiments, the performance significantly decreases as the number of masker decreases.

Investigating additivity of masking effects:

The following results are taken from Durlach (2006). Let T be the target signal, M_1 and M_2 be the maskers and m the amount of masking.

In Durlach (2006), it is stated that there is still uncertainty about the additivity of masking effects. It is still unclear, how $m(M_1 + M_2, T)$ is related to $m(M_1, T)$ and $m(M_2, T)$.

Although, in some cases, additivity is satisfied,

$$m(M_1 + M_2, T) = m(M_1, T) + m(M_2, T) \tag{2.4}$$

so-called *excess masking* can occur:

$$m(M_1 + M_2, T) >> m(M_1, T) + m(M_2, T). \tag{2.5}$$

If the maskers M_1 and M_2 also mask each other in addition to the target T, the following relation is possible:

$$m(M_1 + M_2, T) << m(M_1, T) + m(M_2, T). \tag{2.6}$$

In general, there is not even evidence, that the masking effects satisfy a combination law in the following form including any function F:

$$m(M_1 + M_2, T) = F(m(M_1, T) + m(M_2, T)). \tag{2.7}$$

Multimasker Penalty:

The term *Multimasker Penalty* refers to the phenomenon that the extraction of the target signal becomes significantly more difficult if more than one masking signals are present. It may be explained by limited attentional resources of the listener (Iyer et al., 2010).

3 Mathematical and Technical Background

In this chapter, the mathematical and technical background of different approaches in speech processing are presented.

As already mentioned, it can be a difficult task for hearing-impaired people to understand a target voice in the background of competing talkers. So there is need of developping algorithms that improve speech intelligibility and which can be implemented in digital hearing aids. In general, there are two different approaches in speech processing:

Algorithms for SI Improvement in Multitalker Environments

1. Imitation of the auditory processing stages (e.g. Computational Auditory Scene Analysis)
2. Separation of target signal and interfering sounds using source separation algorithms (e.g. Blind Source Separation)

3.1 Basic Definitions of Signal Processing

In signal processing, *analog signals* refer to time-continuous and *digital signals* to time-discrete representations. Analog signals can be converted to digital signals via Analog-to-Digital Converters (ADC).

The process of sampling is defined as follows:

Definition 3.1 (Sampling:)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous signal and let $(t_j)_{j\in\mathbb{Z}}$ be a sequence in $\mathbb{R}$. Then, the procedure*

$$f_d : \mathbb{Z} \to \mathbb{R} \tag{3.1}$$

$$j \mapsto f(t_j)$$

is called Sampling of the continuous signal f into the discrete signal f_d.

In the following, the major results of signal processing are presented including the Nyquist-Shannon Sampling Theorem, which states, that continuous signals that contain frequencies below a certain value f_{max} can be exactly reconstructed by a series of equidistant samples. To start with, some basic mathematical definitions are mentioned.

Definition 3.2 (Exponential Function)**.** *For $z \in \mathbb{C}$, the exponential function* $\exp(z)$ *is defined as follows:*

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{3.2}$$

Based on the exponential function, sine and cosine are defined.

Definition 3.3 (Sine and Cosine Function)**.** *For $z \in \mathbb{C}$, the trigonometric functions* sin(z) *and* cos(z) *are defined as follows:*

$$\sin(z) = \frac{\exp(iz) - \exp(-iz)}{2i} \tag{3.3}$$

$$\cos(z) = \frac{\exp(iz) + \exp(-iz)}{2} \tag{3.4}$$

For $1 \leq p < \infty$ and $\Omega \subseteq \mathbb{R}$, the space $L^p(\Omega)$ can be defined as follows:

$$L^p(\Omega) = \Big\{ f : \Omega \to \mathbb{C} : \|f\|_p := \Big(\int_\Omega |f(x)|^p dx \Big)^{1/p} < \infty \Big\} \tag{3.5}$$

A 2π-periodic function in $L^2([-\pi, \pi])$ can be expanded into its Fourier series in the following way:

Definition 3.4 (Fourier Series)**.** *Let f be a function in $L^2([-\pi, \pi])$ which is periodic with period 2π. Then, the Fourier series of f is defined as*

$$f(x) = \sum_{n \in \mathbb{Z}} c_n \exp(inx) = a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + i b_n \sin(nx) \tag{3.6}$$

with the Fourier coefficients

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \exp(-int) dt \tag{3.7}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(nt) dt, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(nt) dt, n \in \mathbb{Z}. \tag{3.8}$$

For a periodic function with period T and $f \in L^2([-T, T])$, the Fourier series can be calculated via a transformation:

$$f(x) = \sum_{n \in \mathbb{Z}} c_n \exp\Big(in \frac{\pi x}{T}\Big) = a_0 + \sum_{n=1}^{\infty} a_n \cos\Big(n \frac{\pi x}{T}\Big) + b_n \sin\Big(n \frac{\pi x}{T}\Big) \tag{3.9}$$

with the coefficients

$$c_n = \frac{1}{2T} \int_{-T}^{T} f(t) \exp\Big(\frac{-in\pi t}{T}\Big) dt, \tag{3.10}$$

$$a_n = \frac{1}{T} \int_{-T}^{T} f(t) \cos\Big(n \frac{\pi t}{T}\Big) dt, \quad b_n = \frac{1}{T} \int_{-T}^{T} f(t) \sin\Big(n \frac{\pi t}{T}\Big) dt. \tag{3.11}$$

Now, the *Fourier transform* is presented which is the foundation of various applications in Time-Frequency Analysis, Digital Signal Processing or Audio Processing and enables the decomposition of a function of time into frequency components (Sager, 2012).

Definition 3.5 (Fourier transform)**.** *Let $f \in L^1(\mathbb{R})$. Then, the Fourier transform $\hat{f}$: $\mathbb{R} \to \mathbb{C}$ is defined as follows:*

$$\hat{f}(x) = \int_{\mathbb{R}} f(t)\exp(-2\pi i x t)dt \tag{3.12}$$

The following theorem states that under certain conditions it is possible to recover a function from its Fourier transform.

Theorem 3.1 (Fourier Inversion Theorem)**.** *Let f and $\hat{f} \in L^1(\mathbb{R})$ and let $\hat{f}$ be the Fourier transform of f. Then there holds the inversion formula*

$$f(t) = \int_{\mathbb{R}} \hat{f}(x)\exp(2\pi i t x)dx. \tag{3.13}$$

For a proof, see Mattila (2015).

3.2 The Nyquist-Shannon Sampling Theorem

In this subsection, one of the most essential results of signal processing and communication and information theory, the *Nyquist-Shannon Sampling Theorem*, is presented.

There are different proofs of the theorem and one of them is presented and adapted from "The Shannon Sampling Theorem and Its Implications" (Lerman). Prior to that, some additional terms are presented.

The *support* of a function f in $\mathbb{R}$ or $\mathbb{C}$ is defined as follows:

$$\operatorname{supp}(f) := \overline{\{x : f(x) \neq 0\}} \tag{3.14}$$

Let $f \in L^1(\mathbb{R})$ and $\hat{f} \in L^1(\mathbb{R})$ be the fourier transform of f. f is called *bandlimited* if

$$\exists B \in \mathbb{R} : \operatorname{supp}(\hat{f}) \subseteq [-B, B]. \tag{3.15}$$

In the following, the Nyquist-Shannon Sampling Theorem is presented. It states, that a function f with bandlimit B can be exactly reconstructed by sampling at the rate of $1/(2B)$ without loss of information.

Theorem 3.2 (Nyquist-Shannon Sampling Theorem)**.** *Let $f \in L_1(\mathbb{R})$ and $\hat{f}$ be the Fourier transform of f with* $\operatorname{supp}(\hat{f}) \subseteq [-B, B]$*. Then,*

$$f(x) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big)\operatorname{sinc}\Big(2B\Big(x - \frac{n}{2B}\Big)\Big). \tag{3.16}$$

in the L_2 sense meaning that the series converges to f in $L_2(\mathbb{R})$.

For $x \neq 0$, the cardinal sine function is defined as

$$\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{3.17}$$

Proof. $\text{supp}(\hat{f}) \subseteq [-B, B]$ and $\hat{f} \in L_2([-B, B])$. As a result, $\hat{f}$ can be expanded into its Fourier series:

$$\hat{f}(\xi) = \sum_{n\in\mathbb{Z}} c_n e^{\frac{\pi i n \xi}{B}} \tag{3.18}$$

where the coefficients c_n can be rewritten as follows according to the bandlimitation of f and the inversion formula of the Fourier transform in Theorem 3.1:

$$c_n = \frac{1}{2B}\int_{-B}^{B} \hat{f}(x) e^{\frac{-\pi i n \xi}{B}} dx = \frac{1}{2B}\int_{-\infty}^{\infty} \hat{f}(x) e^{\frac{-\pi i n \xi}{B}} dx = \frac{1}{2B} f\Big(\frac{-n}{2B}\Big). \tag{3.19}$$

In the next step, the reformulation of the coefficients c_n is inserted in the Fourier series in 3.18:

$$\hat{f}(\xi) = \sum_{n\in\mathbb{Z}} \underbrace{\frac{1}{2B} f\Big(\frac{-n}{2B}\Big)}_{c_n} e^{\frac{\pi i n \xi}{B}} = \sum_{n\in\mathbb{Z}} \frac{n}{2B} f\Big(\frac{n}{2B}\Big) e^{\frac{-\pi i n \xi}{B}}. \tag{3.20}$$

Now, $\hat{f}$ is inverted using Theorem 3.1:

$$f(x) = \int_{-B}^{B} \hat{f}(\xi) e^{2\pi i x \xi} d\xi = \int_{-B}^{B} \underbrace{\sum_{n\in\mathbb{Z}} \frac{n}{2B} f\Big(\frac{n}{2B}\Big) e^{\frac{-\pi i n \xi}{B}}}_{\hat{f}(\xi)} e^{2\pi i x \xi} d\xi$$

Simplification and integration leads to the following result:

$$\begin{aligned} f(x) &= \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big) \frac{1}{2B} \int_{-B}^{B} e^{2\pi i \xi (x - \frac{n}{2B})} d\xi = \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big) \frac{1}{2B} \frac{e^{2\pi i \xi (x-\frac{n}{2B})}}{2\pi i (x - \frac{n}{2B})}\Bigg|_{\xi=-B}^{B} \\ &= \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big) \frac{1}{2\pi B (x - \frac{n}{2B})} \frac{e^{2\pi i B (x-\frac{n}{2B})} - e^{-2\pi i B (x-\frac{n}{2B})}}{2i} \end{aligned}$$

Finally, the definitions of sine (3.3) and arcsine are used:

$$f(x) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big) \frac{sin(2\pi B (x - \frac{n}{2B}))}{2\pi B (x - \frac{n}{2B})} = \sum_{n\in\mathbb{Z}} f\Big(\frac{n}{2B}\Big) sinc\Big(2B\Big(x - \frac{n}{2B}\Big)\Big).$$

□

3.3 Computational Auditory Scene Analysis

The human auditory system has the powerful ability of analyzing and segregating incoming sounds. Therefore, the goal of different computational approaches in speech processing is to mimic the different auditory processing stages.

This procedure is called *Computational Auditory Scene Analysis* (CASA) and it is described as the "study and use of ASA in computers using different algorithms" (Brown and Wang, 2005).

The implementations are based on the auditory processing stages and principles of perception and organisation of sound by humans which have already been described in the first chapter (Brown and Cooke, 1994).

3.3.1 Stages of CASA

In this section, the typical stages of CASA implementations are described. They are based on Wang and Brown (2006a). In Figure 3.1, an overview of CASA algorithms is presented.

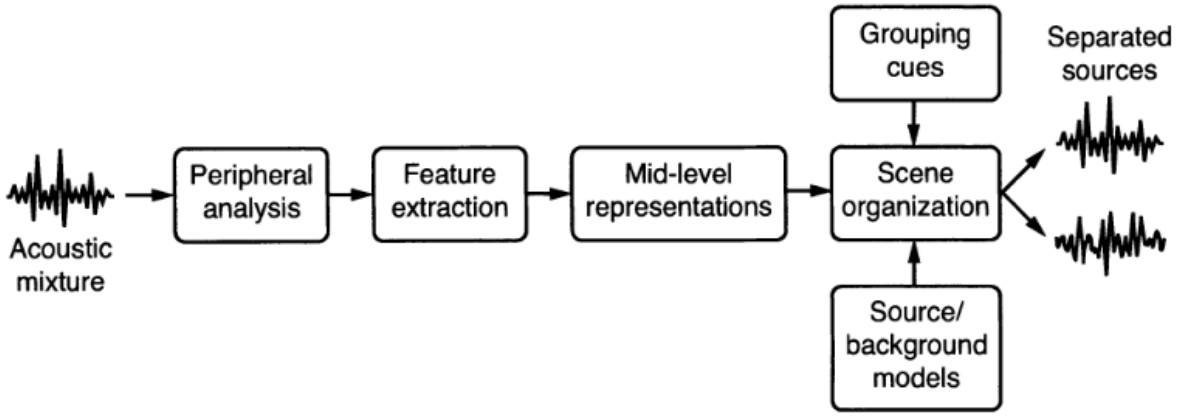


Figure 3.1: *Different stages of CASA:*
Peripheral analysis: Time-Frequency representation of the signal
Feature extraction: e.g. on- and offsets, periodicity
Mid-level representations: segments are formed using the features
Scene organization: primitive grouping cues produce separate streams
The image is taken from Wang and Brown (2006b).

1) Cochleagram

The first part of CASA includes models of the outer and middle ear as well as the frequency selectivity of the cochlea and the transduction by the inner hair cells which can either be described by a transfer function or by a linear filter (Brown and Cooke, 1994).

A bank of auditory bandpass filters simulates the frequency response at a certain basilar membrane position. The *gammatone filter* is a widely-used approach in auditory filter models and describes an analytical approximation of physiologically-recorded impulse responses of auditory nerve fibres by De Boer and Kuyper (1968).

Definition 3.6 (Gammatone Filter)**.** *The impulse response of the gammatone filter is*

$$g_i(t) = t^{n-1}\exp(-2\pi b_i t)\cos(2\pi f_i t + \phi_i), \quad 1 \leq i \leq N \tag{3.21}$$

where n is the filter order, ϕ is the phase, b_i is the bandwidth and f_i the center frequency.

2) Correlogram

The next stage of CASA includes autocorrelation of the simulated auditory fibres for the different frequency channels and it is computed over a window shaped by a window function w.

Definition 3.7 (Correlogram)**.** *The Correlogram is computed by autocorrelation of the simulated auditory nerve firing activity of each cochlear filter channel*

$$A(t, f, \tau) = \sum_{n=0}^{N-1} h(t-n, f)h(t-n-\tau, f)w(n) \tag{3.22}$$

where $h(t, f)$ is the cochlear filter response for frequency f at time t, τ is the autocorrelation delay and w is a window function.

The correlogram is a useful approach for detecting periodicities and therefore estimating the fundamental frequency f_0, which is one of the most important cues in monaural sound segregation (Brown and Wang, 2005).

Definition 3.8 (Summary Autocorrelation Function)**.**

$$S(t, \tau) = \sum_{f=1}^{M} A(t, f, \tau) \tag{3.23}$$

The summary autocorrelation function (SAF) sums up the channels of the correlogram over frequency. It has a peak at the period of each f_0 period and is used in multipitch analysis.

3) Cross-Correlogram

The Cross-Correlogram is based on Jeffress (1948) and uses ITD between the two ears. The interaural cross-correlation can be modeled as follows:

Definition 3.9 (Cross-Correlogram)**.** *The cross-correlogram is computed by cross-correlating the delays of left and the right ear.*

$$C(t, f, \tau) = \sum_{n=0}^{N-1} a_l(t-n, f) a_r(t-n-\tau, f) w(n) \tag{3.24}$$

where $a_l(t, f)$ and $a_r(t, f)$ correspond to the cochlear filter response for frequency f at time t for the left and the right ear, τ is the cross-correlation delay and w represents a window function.

4) Feature Extraction

In the next stage, the signal components are split into groups based on different features. Sequential grouping describes grouping across time whereas simultaneous grouping means grouping across frequency which have already been described in the second chapter.

The most common features are

- onset and offset synchrony
- fundamental frequency (f0)
- harmonicity
- amplitude and frequency modulation

5) Time-Frequency Masking

The last stage of a CASA system includes the computation of a time-frequency (T-F) mask which weights the T-F representation obtained by a cochleagram, as an example. The main approach of the T-F mask is to to emphasize target-dominated regions and to suppress regions dominated by other sources.

In general, the masks can be real-valued in which they correspond to a probabilty of target-dominated sources.

Binary masks aim at keeping T-F regions of the target that are stronger than the interferer and delete those which are weaker. This concept is motivated by the auditory masking effect in which one sound is inaudible in the presence of another.

In Wang (2005), an *ideal binary mask* (IBM) has been proposed as the goal of CASA:

Definition 3.10 (Ideal Binary Mask)**.** *The ideal binary mask (IBM) is defined as follows:*

$$i(t,f) = \begin{cases} 1 & \text{if } s(t,f) > n(t,f) \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

where $s(t,f)$ is the target energy in a T-F unit of the speech signal and $n(t,f)$ is the energy of the noise signal.

3.4 Blind Source Separation (BSS)

In the last subsection, the stages of CASA algorithms were described. Now, a second computational approach of dealing with interfering speech signals is presented which is not based on the human auditory processing stages but, as an alternative, analyzes the speech mixture signals in regards of statistical measures.

In this context, *Blind Source Separation* (BSS) is a powerful method which is defined as follows:

Definition 3.11 (Blind Source Separation)**.** *Blind source separation (BSS) is the recovering of a set of source signals from a set of mixed signals without prior information about the signals or the mixing process (Naik et al., 2014).*

A classical example of a BSS problem is a cocktail party scenario where different people are speaking at the same time. A set of microphones is used to record these speech signals each of which including a mixture of the signals.

Now, the goal is to separate the mixture without any prior information.

Further applications of BSS problems are teleconferencing scenarios with different overlapping speech signals but also medical applications like the electroencephalogram (EEG) where brain waves are recorded by multiple microphones or sensors.

In general, there are different methods for dealing with BSS problems and the most common is Independent Component Analysis (ICA) which will be explained in this section. Prior to that, the BSS problem including the mixing as well as the demixing process will be defined in a mathematical way.

3.4.1 Formulation of the Problem

In a BSS problem, there are N unknown source signals S and M observed signals X that are a mixture of the sources with the global relation $X = A \cdot S$. The goal is to estimate the source signals.

On the whole, one distinguishes the following cases:

- Determined: $N = M$: equal number of sensors and signals
- Overdetermined: $N < M$: more sensors than sources
- Underdetermined: $N > M$: more sources than sensors

In many real-world scenarios, the BSS problem tends to be underdetermined.

Example of a BSS problem:

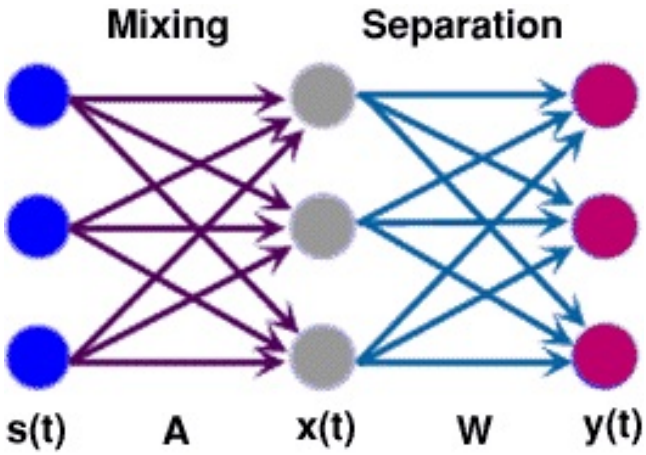


Figure 3.2: *Determined BSS problem* ($N = M = 3$)
The image is taken from http://www.huginn.com/knuth/bse.html.

3.4.2 Modeling the Mixing Process

For $1 \leq i \leq N$ and $1 \leq j \leq M$ let $x_i(n)$ be the measured sensor or microphone signals, $s_j(n)$ be the unknown source signals and a_{ij} be the coefficients of the mixing matrix. In the following, different mixture models are described.

Linear Instantaneous (LI) Mixture:

The first proposed model describes the simplest form of a linear mixing process.

$$x_i(n) = \sum_{j=1}^{M} a_{ij} s_j(n) \tag{3.26}$$

where a_{ij} are the coefficients of the mixing matrix A (Makino et al., 2007).

Attenuated and delayed (AD) mixtures:

The following extension of the mixture process takes into account attenuation and delay of the sound propagation (Puigt and Deville, 2005):

$$x_i(n) = \sum_{j=1}^{N} a_{ij} s_j(n - n_{ij}) \tag{3.27}$$

where n_{ij} correspond to the time shifts.

Convolutive mixtures:

Convolutive mixing takes into account time delays and multipath propagation due to reverbation effects:

$$x_i(n) = \sum_{j=1}^{N} a_{ijk} s_j(n - k) \tag{3.28}$$

where a_{ijk} correspond to the coefficients of the linear time-invariant mixing system $\{\mathcal{A}_k\}_{k=-\infty}^{\infty}$.

In the following, only the linear instantaneous mixture process (3.26) is considered. Its demixing process is modeled as follows:

Demixing process:

The subsequent procedure is the determination of the coefficients w_{ji} of a *demixing* matrix W for $1 \leq i \leq N$ and $1 \leq j \leq M$ to recover an estimation of the source signals $y_i(n)$. The demixing process is modelled as follows:

$$y_j(n) = \sum_{i=1}^{N} w_{ji} x_i(n) \tag{3.29}$$

3.4.3 Independent Component Analysis

Independent Component Analysis (ICA) is a widely used approach for solving BSS problems. It describes a statistical method that separates a multidimensional random vector into independent components, respectively, into maximally independent sources.

The results in this subsection are taken from Hyvärinen et al. (2004).

In the following, the central limit theorem will be described, which states that a series of identically distributed random variables converges in distribution to the standard normal distribution. This is of great importance as it will lead to the result that independence can be set equally to non-gaussianity in the fields of ICA as will be described in this chapter.

In order to formulate the central limit theorem, some further definitions are necessary. To start with, independence of random variables, which is the major concept in ICA, is described.

Definition 3.12 (Independence)**.** *Let $X_1, \ldots, X_n$ be random variables. They are called independent if the joint distribution function equals the product of the marginal distribution functions:*

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \cdots \cdot F_{X_n}(x_n) \tag{3.30}$$

Now, let X be a random variable with the distribution function $f(x)$. The *expected value* and the *variance* of X are defined as follows:

Definition 3.13 (Expected value)**.** *The expected value of a random variable X is defined as follows*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx. \tag{3.31}$$

Definition 3.14 (Variance)**.** *The variance of X is defined as the expected value of the squared deviation of μ:*

$$\mathbb{V}[X] = \mathbb{E}[(X - \mu)] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx. \tag{3.32}$$

The *characteristic function* will be essential in the proof of the central limit theorem.

Definition 3.15 (Characteristic Function)**.** *Let X be a random variable. Then, the characteristic function is defined as follows:*

$$\phi_X(t) = \mathbb{E}(e^{iXt}) \tag{3.33}$$

If $X_1, \ldots, X_n$ are independent random variables, the characteristic function of $\sum_{i=1}^{n} X_i$ can be calculated as:

$$\phi_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} \phi_{X_i}. \tag{3.34}$$

In the following, the density function of the normal distribution and the special case of the standard normal distribution are described.

The parameters of the normal distribution are $\mu \in \mathbb{R}$ and $\sigma > 0$ and its density function is defined as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big). \tag{3.35}$$

The standard normal distribution defines the special case of $\mu = 0$ and $\sigma = 1$ with the density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big) \tag{3.36}$$

and the distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\Big(-\frac{1}{2}t^2\Big).dt \tag{3.37}$$

The central limit theorem states that a series of identically distributed independent random variables converges *in distribution* to the standard normal distribution which is defined as follows:

Definition 3.16 (Convergence in Distribution)**.** *Let F_n be a sequence of distribution functions. F_n converges in distribution to F, $F_n \to F$, if for every bounded continuous function f:*

$$\lim_{n\to\infty} \int f dF_n = \int f dF. \tag{3.38}$$

A sequence of random variables X_n converges in distribution to a random variable X, if the sequence of distribution functions converges to the distribution function of X.

The following result relates convergence in distribution to convergence of the characteristic functions:

Theorem 3.3 (Continuity Theorem)**.** *Let F_n be a sequence of distribution functions and ϕ_n be the corresponding characteristic functions.*
If ϕ_n converges to a function ϕ which is continuous at 0, then ϕ is the characteristic function of a distribution function F and $F_n \to F$.

A proof of the Continuity Theorem can be seen in Grill (2017).

Now, one of the most important results in statistics is presented. The proof is taken and adapted from Grill (2017).

Theorem 3.4 (Central Limit Theorem)**.** *Let $X_1, \dots X_n$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2. Then,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sqrt{n}\sigma} \leq x\Big) = \Phi(x) \tag{3.39}$$

Proof. It is sufficient to prove the result for $\mu = 0$ and $\sigma^2 = 1$, because otherwise, the transformation

$$Y = \frac{X - \mu}{\sigma} \tag{3.40}$$

can be used. Then, $\mathbb{E}(Y) = 0$ and $\mathbb{V}(Y) = 1$ and

$$\frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}. \tag{3.41}$$

Now, the taylor expansion at zero is used:

$$\phi_X(t) = 1 + it\mathbb{E}(X) - \frac{t^2}{2}\mathbb{E}(X)^2 + o(t^2) \text{ for } t \to 0 \tag{3.42}$$

which results for the characteristic function ϕ_n of $\sum_{i=1}^{n} X_i/\sqrt{n}$ in

$$\phi_n(t) = \Big(\phi_X\Big(\frac{t}{\sqrt{n}}\Big)\Big)^n = \Big(1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big)\Big)^n \tag{3.43}$$

using $\mathbb{E}(X) = 0$. For fixed t, this implicates

$$\Big(1 - \frac{t^2}{2n} + o\Big(\frac{1}{n}\Big)\Big)^n \to e^{-t^2/2} \tag{3.44}$$

which is the characteristic function of the standard normal distribution. The Continuity Theorem 3.3 implies, that the distribution function of $\sum_{i=1}^{n} X_i/\sqrt{n}$ converges to Φ which proves the theorem. □

Now, the central limit theorem can be used in ICA to relate independence to gaussianity. In the following, let $\mathbf{x}$ be a random vector with elements $x_1, \ldots, x_n$ which define the mixtures of the signals and $\mathbf{s}$ be a random vector with elements $s_1, \ldots, s_n$ corresponding to the sources. $\mathbf{A}$ denotes to the mixing matrix.

The mixing model is now written as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3.45}$$

Assumptions:

1. Linear mixture of source signals $\mathbf{s}$
2. Non-gaussianity of $\mathbf{s}$
3. All source signals $\mathbf{s}$ are statistically independent
4. $M = N$, the number of observed signals must be equal to the number of sources

Maximization of Non-Gaussianity leads to Independent Components:

Maximization of non-gaussianity can be used to obtain independence of the source signals which will be derived in the following.

Let $\mathbf{x} = \mathbf{As}$. As an assumption, the mixing process can be inverted:

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} \tag{3.46}$$

The approach of estimating the independent components is to find linear combinations of the mixture variables $\mathbf{x}$, respectively, a linear combination of the independent components $\mathbf{s}$:

$$\mathbf{y} = \sum_i b_i x_i = \sum_i q_i s_i \tag{3.47}$$

where $\mathbf{q} = \mathbf{b}^T\mathbf{A}$.

Now, the coefficients of $\mathbf{q}$ are varied. The central limit theorem states that the sum of two independent random variables is usually more gaussian than the original variables. As a matter of fact, it becomes least gaussian, when it equals one of them. Since the values of q are unknown, the non-gaussianity of $\mathbf{b}^T\mathbf{x}$ is maximized resulting in one of the independent components.

Measures of Non-gaussianity:

In order to obtain the independent components, different measures of non-gaussianity are used which lead to equivalent results.

1) Higher-order statistics:

The first approach is the use of higher-order statistics like the kurtosis.

Definition 3.17 (Kurtosis)**.** *The kurtosis of a random variable X is defined as*

$$k = \frac{E(X - E(X)^4)}{(E(X)^2)^2} - 3 \tag{3.48}$$

Since the normal distribution has kurtosis 0 by this definition, maximization of non-gaussianity is measured by the absolute value of the kurtosis.

2) Information-theoretic Approach

Another possibility of determining independent components via non-gaussianity originates from information theory, another research area.

In information theory, the term *entropy* is defined as follows:

Definition 3.18 (Entropy)**.** *Let X be a random variable with density function f_X. Then, the entropy of X is defined as*

$$H(X) = -\int_{\mathbb{R}} f_X(y)\log(f_X(y))dy \tag{3.49}$$

Satz 3.19. *The entropy of the gaussian distribution is*

$$H(X_{gauss}) = \frac{1}{2}(1 + \log(2\pi\sigma^2)) \tag{3.50}$$

Proof. The probability density function of the normal distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} \tag{3.51}$$

Inserting in definition (3.18), one gets

$$H(X_{gauss}) = -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}}\Big) dx$$

Now, the logarithm is separated in two parts:

$$= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} \Big[\underbrace{\log(\frac{1}{\sqrt{2\pi\sigma^2}})}_{-\frac{1}{2}\log(2\pi\sigma^2)} + \underbrace{\log\Big(e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\Big)}_{\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\Big] dx$$

Finally, the relation $\int_{\mathbb{R}} \phi(x)dx = 1$ and the definition of σ^2 are used:

$$= \frac{1}{2}\log(2\pi\sigma^2) \underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} dx}_{1} + \frac{1}{2\sigma^2} \underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} (x-\mu)^2 dx}_{\sigma^2}$$

$$= \frac{1}{2}(1 + \log(2\pi\sigma^2)).$$

□

Now, it can be proven that the gaussian distribution maximizes the entropy (Conrad, 2013; Shannon, 2001).

Satz 3.20. *Let $f(x)$ be a continuous probability distribution on $\mathbb{R}$ with mean μ and variance σ^2. Then,*

$$H(f) \leq \frac{1}{2}(1 + \log(2\pi\sigma^2). \tag{3.52}$$

That means, among random variables with fixed variance, the gaussian has the largest entropy.

Proof. The goal is to maximize $-\int_{\mathbb{R}} f(x)\log f(x)dx$ under the constraints

1. $\int_{\mathbb{R}} f(x)dx = 1$
2. $\int_{\mathbb{R}} xf(x)dx = \mu$
3. $\int_{\mathbb{R}}(x-\mu)^2 f(x)dx = \sigma^2$.

The problem is maximized using Lagrange multipliers.

$$H(x,\lambda) = f(x) + \sum_{j=1}^{3} \lambda_j F_j(x)$$

Inserting into the function, one gets

$$\begin{aligned} H(x,\lambda_1,\lambda_2,\lambda_3) = &-\int_{\mathbb{R}} f(x)\log f(x)dx + \lambda_1\Big(\int_{\mathbb{R}} f(x)dx - 1\Big) + \lambda_2\Big(\int_{\mathbb{R}} xf(x)dx - \mu\Big) \\ &+ \lambda_3\Big(\int_{\mathbb{R}}(x-\mu)^2 f(x)dx - \sigma^2\Big) \\ = &\int_{\mathbb{R}}(-f(x)\log f(x) + \lambda_1 f(x) + \lambda_2 x f(x) + \lambda_3(x-\mu)^2 f(x))dx \\ &- \lambda_1 - \mu\lambda_2 - \sigma^2\lambda_3 \end{aligned}$$

This problem may be solved using variation of calculus and requires

$$\frac{\partial H}{\partial f} = -1 - \log f(x) + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 \stackrel{!}{=} 0$$

for the maximum entropy which results in

$$f(x) = e^{\lambda_1 - 1 + \lambda_2 x + \lambda_3(x-\mu)^2}$$

Now, the constraints are inserted into the formula. At first, to ensure the existence of the integral, the following is required:

$$\int_{\mathbb{R}} f(x)dx < \infty : \lambda_2 = 0, \lambda_3 < 0.$$

which leads to

$$f(x) = e^{\lambda_1 - 1} e^{-\lambda_3(x-\mu)^2}.$$

Using the constraints, the results are

$$\lambda_3 = -\frac{1}{2\sigma^2} \quad \text{and} \quad e^{\lambda_1 - 1} = \frac{1}{\sqrt{2\pi\sigma^2}}. \tag{3.53}$$

which finally results in the probability density function of the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}}$$

□

As a result, a measure of non-gaussianity is the so-called *negentropy*.

Definition 3.21 (Negentropy)**.** *Let Y be a random variable with the same variance as X. Then, the negentropy of X is defined as*

$$J(X) = H(X_{gauss}) - H(X) \tag{3.54}$$

Finally, gradient methods and fast fixed-point algorithms are used in ICA to derive the solutions.

3.5 Statistical Background of Analysis of Variance (ANOVA)

In this chapter, the statistical background of Analysis of variance (ANOVA) is presented which is a widely spread method of evaluating speech intelligibility and masking experiments (Brungart et al., 2001; Schubotz et al., 2016).

To start with, some important statistical definitions are explained followed by the presentation of the basic one-way ANOVA, the two-way ANOVA and the repeated measures ANOVA which will be used in the next chapter.

This section is based on Fahrmeir et al. (2016); Backhaus et al. (2015); Rasch et al. (2010).

3.5.1 Notation and Definitions

ANOVA is a statistical method to analyze differences between group means and it describes an extension of the t-test. In general, the following hypothesis is tested:

Null Hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \tag{3.55}$$

Alternative Hypothesis:

$$H_1 : \exists i, j : \ \mu_i \neq \mu_j \tag{3.56}$$

In the procedure, the variation within and between the groups are calculated.

Now, let $x_1, \ldots, x_n$ be a sample data of size n. Then, the *sample mean*, the *sample variance* and the *sum of squares* (SS) are defined as follows:

Definition 3.22 (Sample Mean)**.** *The sample mean of a sample data $x_1 \ldots x_n$ is defined as*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.57}$$

Definition 3.23 (Sample Variance)**.** *The sample variance of a sample data $x_1 \dots x_n$ is defined as*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3.58}$$

Definition 3.24 (Sum of Squares)**.** *The sum of squares (SS) of a sample data $x_1 \dots x_n$ is defined as*

$$SS = \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3.59}$$

In general, the following assumptions are made in ANOVA:

Assumptions:

- Normal distribution of the dependent variable
- Homogeneity of variance
- Independence of errors

3.5.2 One-way ANOVA

The one-way ANOVA investigates the influence of one variable or factor A, the *independent variable*, on another variable B, the *dependent variable*. The different levels of A are also called factor levels.

The one-way ANOVA is based on the following a linear model.

Definition 3.25 (Linear Model of the One-Way ANOVA)**.** *The linear model for a one-way ANOVA is*

$$x_{ij} = \overline{x} + \alpha_i + \epsilon_{ij} \tag{3.60}$$

$$\text{with } \epsilon \sim \mathcal{N}(0, \sigma^2) \text{ and } \sum_{i=1}^{k} \alpha_i = 0 \text{ for } i = 1 \dots k, j = 1 \dots n_i$$

where $\overline{x}_i = \overline{x} + \alpha_i$.
x_{ij} represents the dependent variable, $\overline{x}$ the mean, $\overline{x_i}$ the mean in condition i, α_i the effect of factor i which describes the systematical variance and ϵ_{ij} denotes to the residual error, also called error variance.

In this section, the following notation is used.

Notation:

- $K \ldots$ number of the conditions (factors),
- $N_k \ldots$ number of the subjects taking part in condition k
- $N \ldots$ overall number of measure points
- $\overline{x}_k \ldots$ mean in condition k
- $\overline{x} \ldots$ overall mean.

In the following, the different SS which are used in ANOVA are presented.

SS_{total} describes the overall SS:

$$SS_{total} = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (x_{kn} - \overline{x}) \tag{3.61}$$

$SS_{between}$ denotes to the SS caused by the groups/between the groups:

$$SS_{between} = \sum_{k}^{K} N_k (\overline{x}_k - \overline{x})^2 \tag{3.62}$$

SS_{within} refers to the SS caused by random effects/within the groups:

$$SS_{within} = \sum_{k=1}^{K} \sum_{n=1}^{N_k} k(x_{kn} - \overline{x}_n) \tag{3.63}$$

Partitioning of Variance:

The partitioning of variance describes a major step in ANOVA and is also called the *Fundamental Theorem of ANOVA.*
The total sum of squares SS_{total} can be split up as follows:

$$SS_{total} = SS_{within} + SS_{between} \tag{3.64}$$

The next step in ANOVA is the calculation of the mean sum of squares (MS) of SS_{within} and $SS_{between}$.

$$MS_{between} = \frac{SS_{between}}{K - 1} \tag{3.65}$$

$$MS_{within} = \frac{SS_{within}}{N - K} \tag{3.66}$$

Finally, the test statistic F is calculated as follows:

$$F = \frac{MS_{between}}{MS_{within}} \tag{3.67}$$

3.5.3 Two-way ANOVA

The two-way ANOVA describes an extension to the one-way ANOVA and investigates the influence of *two* independent variables A and B on one dependent variable.

In this approach, the *main effects* of A and B as well as the *interaction effects* $A \times B$ are analyzed. In general, the procedure is very similar to the 1-way ANOVA.

The two-way ANOVA is based on the following linear model:

Definition 3.26 (Linear Model of the two-way ANOVA)**.** *The linear model of the two-way ANOVA is*

$$x_{ijk} = \overline{x} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{3.68}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

$$i = 1 \dots k, j = 1 \dots n_i$$

where $\overline{x}_i = \overline{x} + \alpha_i$ *and* $\overline{x}_j = \overline{x} + \beta_j$.
The new term $(\alpha\beta)_{ij}$ *describes the interaction effect between factor A and factor B and* ϵ_{ijk} *denotes to the residual error.*

As before, the partitioning of the variance is used, where the term $SS_{between}$ splits up in three parts:

Partitioning of Variance:

The total sum of squares (SS_{total}) can be split up in the following way:

$$SS_{total} = SS_{within} + \underbrace{SS_A + SS_B + SS_{A\times B}}_{SS_{between}} \tag{3.69}$$

Finally, the F-statistics for both factors A and B as well as for the interaction $A \times B$ are calculated.

$$F_A = \frac{MS_{betweenA}}{MS_{withinA}} \tag{3.70}$$

$$F_B = \frac{MS_{betweenB}}{MS_{withinB}}$$

$$F_{A\times B} = \frac{MS_{betweenA\times B}}{MS_{withinA\times B}}$$

3.5.4 Repeated Measures ANOVA (rANOVA)

The next extension of the ANOVA describes the *Repeated Measures ANOVA* (rANOVA). In this approach, the same testing persons take part in the whole procedure including all scenarios or conditions which leads to *dependent measurements* or *repeated measurements*.

In this context, the error variance can be reduced because the subject-specific variance can be eliminated, which underlines the advantages of rANOVA.

Again, the corresponding linear model is presented:

Definition 3.27 (Linear Model of the Two-Way Repeated Measures ANOVA)**.** *The linear model of the two-way rANOVA is*

$$x_{ijk} = \overline{x} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{3.71}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

$$i = 1 \ldots k, j = 1 \ldots n_i$$

where $\overline{x}_i = \overline{x} + \alpha_i$, $\overline{x}_j = \overline{x} + \beta_j$ *and* $\overline{x}_k = \overline{x} + \gamma_k$.
The error effects of the model are described by $\epsilon_{ijk} = (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$ *which include the new term of the interaction effects corresponding to the subjects factor.*

In the rANOVA, a new term, the subject-specific SS, occurs:

$$SS_{subjects} = \sum_{i=1}^{k} k \cdot (\overline{x}_i - \overline{x})^2 \tag{3.72}$$

Again, the partitioning of the variance is used:

Partitioning of Variance:

The total sum of squares SS_{total} may be split up in the following way:

$$SS_{total} = SS_{within} + SS_{between} \tag{3.73}$$

where

$$SS_{within} = SS_{subjects} + SS_{error} \tag{3.74}$$

In this case, the term SS_{error} can be reduced in rANOVA.

As before, the mean sum of squares $MS_{betweenA}$, $MS_{betweenB}$ and $MS_{betweenA\times B}$ as well as the F-Statistics F_A, F_B and $F_{A\times B}$ are calculated.

$$\begin{aligned} F_A &= \frac{MS_{betweenA}}{MS_{withinA}} \\ F_B &= \frac{MS_{betweenB}}{MS_{withinB}} \\ F_{A\times B} &= \frac{MS_{betweenA\times B}}{MS_{withinA\times B}} \end{aligned} \tag{3.75}$$

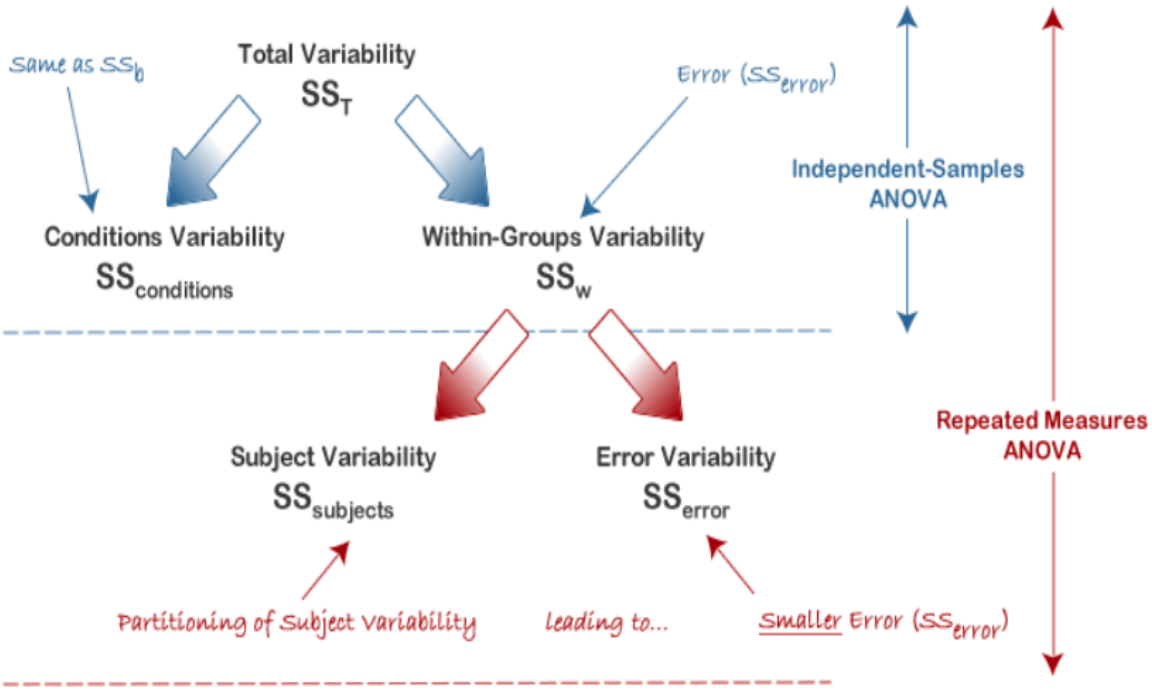


Figure 3.3: *Partitioning of Variance: In this figure, one can see that in the rANOVA, the error term SS_{error} can be reduced. The image is taken from https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide-2.php*

Mauchly's Test on Sphericity

In rANOVA, the sphericity assumption states that the variance of the differences of all combinations of groups have to be equal. Compound symmetry implies sphericity and states that all response variables have the same variance and each pair of values share a common correlation. This results in the following covariance matrix Σ:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \tag{3.76}$$

The most important test for sphericity is the *Mauchly's test on sphericity* (Mauchly, 1940). If the symmetry assumption is not fulfilled, correction factors ϵ can be applied to adjust the degrees of freedom in order to obtain an F-statistic which is approximately F-distributed.

In the following, let k be the number of repeated measures and n be the number of subjects. The three most important correction factors which occur in rANOVA are the following:

- **Greenhouse-Geisser Correction:**

$$\epsilon_{GG} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{(k-1)\sum_{i=1}^{k} \lambda_i^2} \tag{3.77}$$

  where $\lambda_i, i = 1 \dots k$ describe the eigenvalues of Σ.

- **Huynh and Feldt Correction:**

$$\epsilon_{HF} = \frac{n(k-1)\epsilon_{GG} - 2}{(k-1)(n-1) - (k-1)\epsilon_{GG}} \tag{3.78}$$

- **Lower bound:**

$$\epsilon_{LB} = \frac{1}{k-1}. \tag{3.79}$$

Measures of Effect Size:

In ANOVA, the effect size of a variable can be calculated using η^2 or $\eta^2_{partial}$ which are defined as follows:

$$\eta^2_{partial} = \frac{SS_{conditions}}{(SS_{conditions} + SS_{error})} \quad \text{and} \quad \eta^2 = \frac{SS_{conditions}}{SS_{error}}. \tag{3.80}$$

4 Technical Part: Speech Intelligibility Test

In the technical part of the diploma thesis, a study on speech intelligibility in the background of different maskers was performed.

A speech intelligibility test was implemented in Matlab and intelligibility of VCVs (Vowel-Consonant-Vowel) was investigated. On the whole, 14 syllables spoken by four female speakers were used as target signals. They were taken from the *Oldenburger Logatome Corpus* (OLLO).

The masker conditions included sentences from the *Oldenburger Satztest* and varied in number of speakers, gender, intelligibility and spectral properties. Also, female speech-shaped noise (SSN) was used, which has a long-term average spectrum similar to that of female speech.

4.1 Description of the Experiment

Many studies have investigated speech intelligibility in the presence of multiple simultaneous talkers (Brungart et al., 2001; Iyer et al., 2010; Srinivasan and Wang, 2008; Schubotz et al., 2016).

However, only few investigation has been made in the fields of intelligibility of low-context speech segments like VCVs or CVCs in the presence of multiple simultaneous talkers. In Boothroyd and Nittrouer (1988), it is stated, that the perception of high-context speech segments like words depends critically on the error rate of low-context speech segments, which indicates the necessity of studies on low-context speech segments.

In the study, different masking effects are expected to influence speech intelligibility.

Expected Masking Effects:

- Energetic Masking (EM)
- Amplitude Modulation Masking (AMM)
- Informational Masking (IM)

4.2 Material

4.2.1 Target Signals

For the target signals, parts the *Oldenburg Logatome Corpus* (OLLO) were used which consists of female and male recordings of different logatomes (VCVs and CVCs) in various emotional states.

In this investigation, only the female recordings of VCVs including the vowel /a/ in a normal state were used in order to focus on intelligibility depending on the variability of the consonants. The syllables were recorded by four different female speakers which led to 56 different target signals.

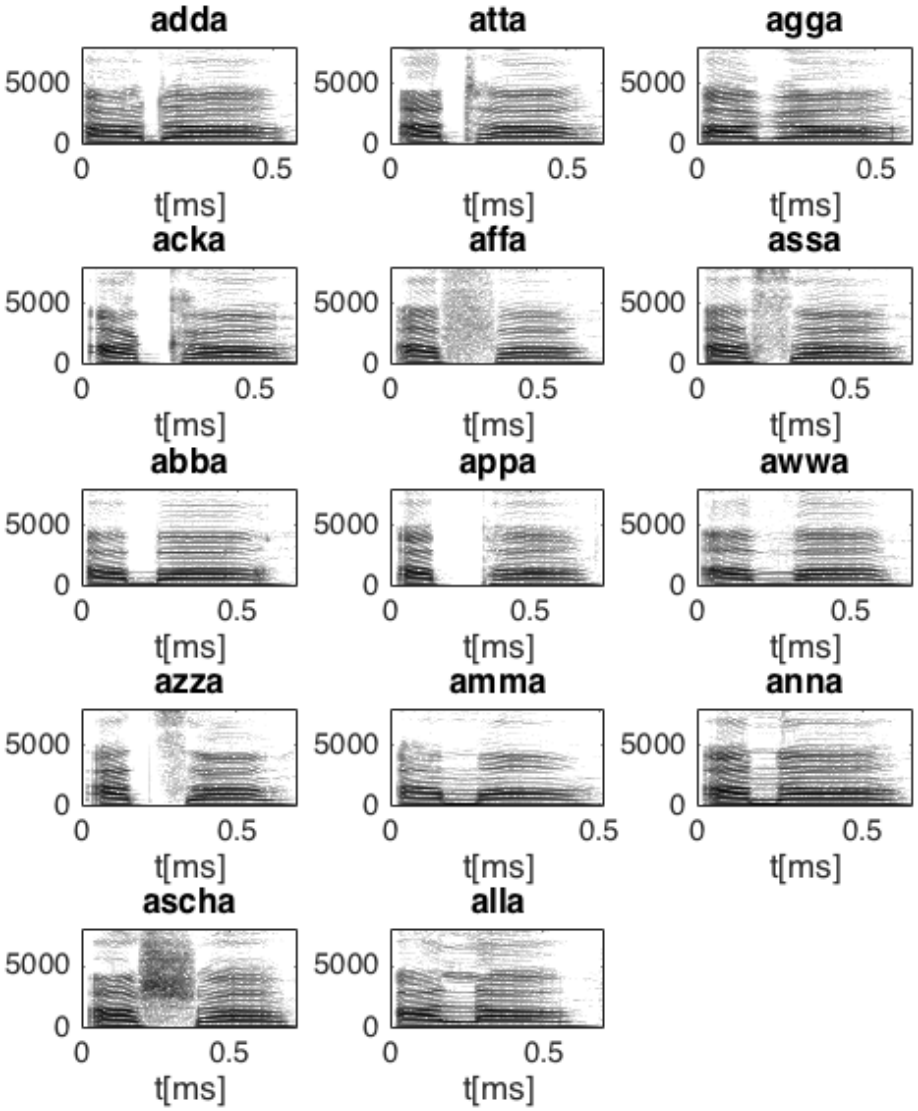


Figure 4.1: *Spectograms of the 14 target syllables 'adda', 'atta', 'agga', 'acka', 'affa', 'assa', 'abba', 'appa', 'awwa', 'azza', 'amma', 'anna', 'ascha', 'alla' spoken by one of the four female speakers*

4.2.2 Masker Signals

The masker signals were for the most part taken from recordings of the *Oldenburger Satztest*.

The following three sentences from the Oldenburger Satztest were used in the experiment:

1. 'Britta verleiht elf alte Bilder.'
2. 'Ulrich hat fünf kleine Dosen.'
3. 'Tanja kauft acht nasse Messer.'

The sentences were spoken by one female and one male speaker. In some scenarios, the recordings were edited using the program Audacity® in order to manipulate the pitch.

Also, a part of the International Speech Test Signal (ISTS) was used, which is an unintelligible mixture of six different languages including Arabic, Chinese, French, American English, German and Spanish. It is an internationally used signal in the evaluation of hearing aids and corresponds to the long term average speech spectrum standards (LTASS).

Eventually, one scenario included Speech Shaped Noise which is also a widely-used masker in speech intelligibility tests. The signal is based on a Fast Fourier Transform (FFT) of the ISTS, a randomization of the coefficients and a concluding Inverse Fourier Transform (IFT) which results in an equal long-term spectrum of both signals. It is taken from Schubotz et al. (2016).

4. International Speech Test Signal (ISTS)
5. Speech Shaped Noise (SSN)

The target signals were mixed with the masker signals at the same point of time in each scenario in order to obtain reliable results since the length of the target signals were only about 0.5 s.

4.2.3 Scenarios

The target signals were mixed with eight different masker types at four Signal-to-Noise ratios (SNRs).

Masker types:

1. 1 female: 'Britta verleiht elf alte Bilder'
2. 2 female: 'Britta verleiht elf alte Bilder'
   'Ulrich hat fünf kleine Dosen'.

3. 3 female: ‘Britta verleiht elf alte Bilder’
   ‘Ulrich hat fünf kleine Dosen’
   ‘Tanja kauft acht nasse Messer’.
4. 3 male: ‘Britta verleiht elf alte Bilder’
   ‘Ulrich hat fünf kleine Dosen’
   ‘Tanja kauft acht nasse Messer’.
5. 3 female (manipulated pitch):
   ‘Britta verleiht elf alte Bilder’
   ‘Ulrich hat fünf kleine Dosen’
   ‘Tanja kauft acht nasse Messer’.
6. 2 female (manipulated pitch) and 1 male:
   ‘Britta verleiht elf alte Bilder’
   ‘Ulrich hat fünf kleine Dosen’
   ‘Tanja kauft acht nasse Messer’.
7. 1 female ISTS: unintelligibile speech
8. Speech Shaped Noise (SSN)

In masker type 5, two of the three recordings were pitch-manipulated in Audacity®. The program uses the open-source SoundTouch™ Audio Processing Library for pitch modification without changing the duration of the signal. The pitch control combines time-stretching and sample rate transposing algorithms where the former describes a change in signal duration without pitch change and the latter refers to a linear interpolation of the signal changing pitch as well as duration.

In order to diversify the spectrum of the female masker signals, the pitch of the speech signal ‘Ulrich hat fünf kleine Dosen’ was increased by three half steps and the pitch of the signal ‘Tanja kauft acht nasse Messer‘ was decreased by three half steps.

Signal-to-noise ratios:

The *signal-to-noise ratio* (SNR) of the target and the masker signal is also called *target-to-masker ratio* (TMR) in the context of masking experiments in speech.

Definition 4.1 (Signal-to-noise ratio)**.** *Let x be a signal (the target signal) and y be the noise or masker signal. Then, the signal-to-noise ratio (SNR) of x and y is defined as follows:*

$$SNR(x,y)[dB] = 20 \cdot \log_{10}\left(\frac{RMS(x)}{RMS(y)}\right) \tag{4.1}$$

where

$$RMS(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i|^2} \tag{4.2}$$

for $x = (x_1, \ldots, x_n)$.

The following four SNRs were used in the experiment: -3,-6,-9,-12 [dB]

In the following, the spectrograms of the different masker scenarios are presented. The RMS of each masker signal was set to a fixed value before the mixing process.

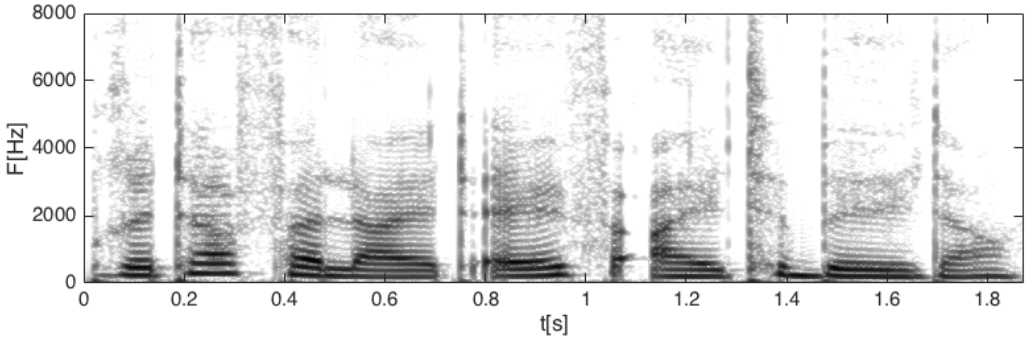


Figure 4.2: *Masker scenario 1: 1 female speaker*

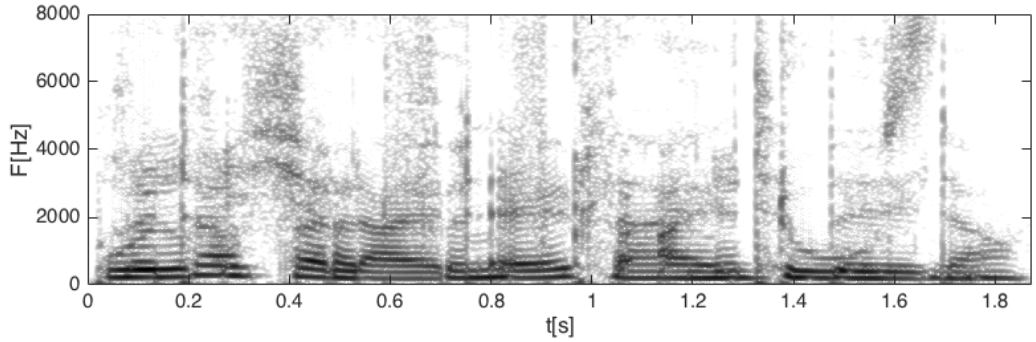


Figure 4.3: *Masker scenario 2: 2 female speakers*

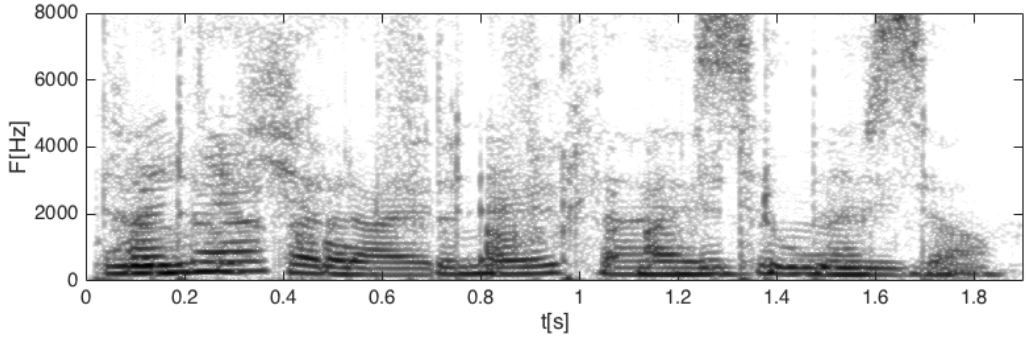


Figure 4.4: *Masker scenario 3: 3 female speakers*

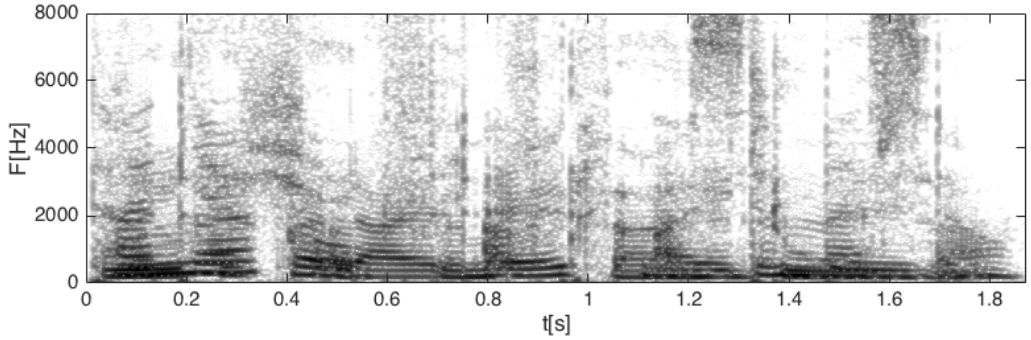


Figure 4.5: *Masker scenario 4: 3 male speakers*

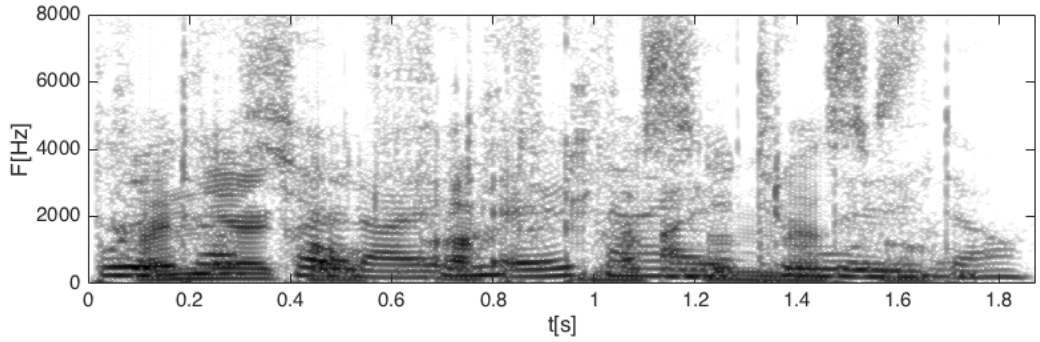


Figure 4.6: *Masker scenario 5: 3 female speakers (pm)*

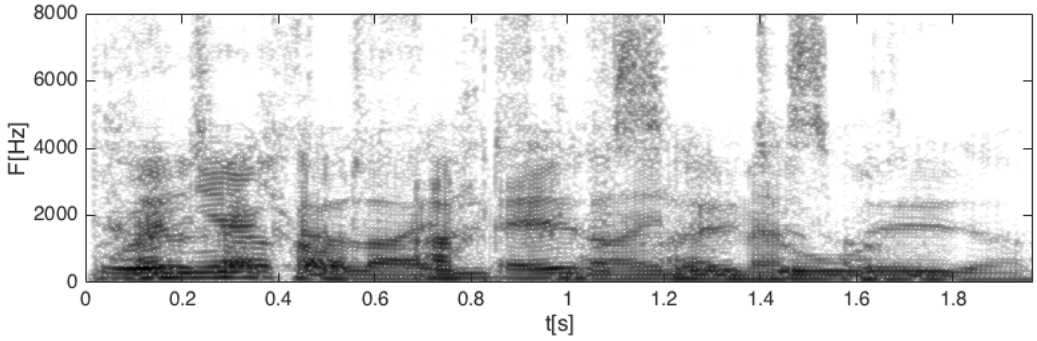


Figure 4.7: *Masker scenario 6: 1 male and 2 female (pm) speakers*

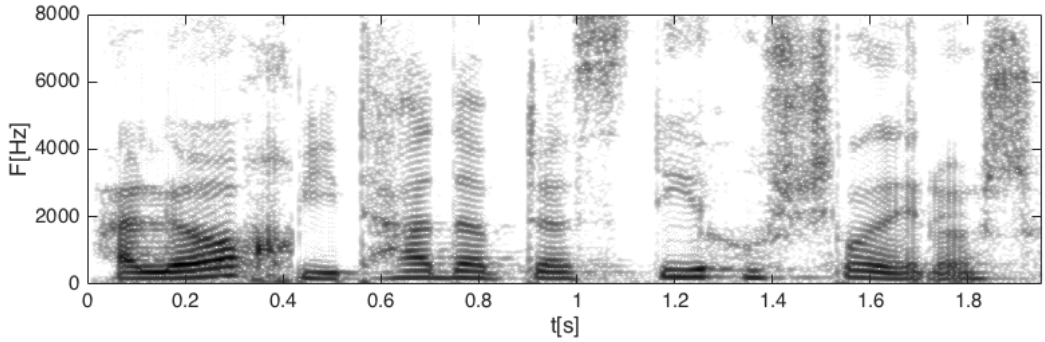


Figure 4.8: *Masker scenario 7: ISTS*

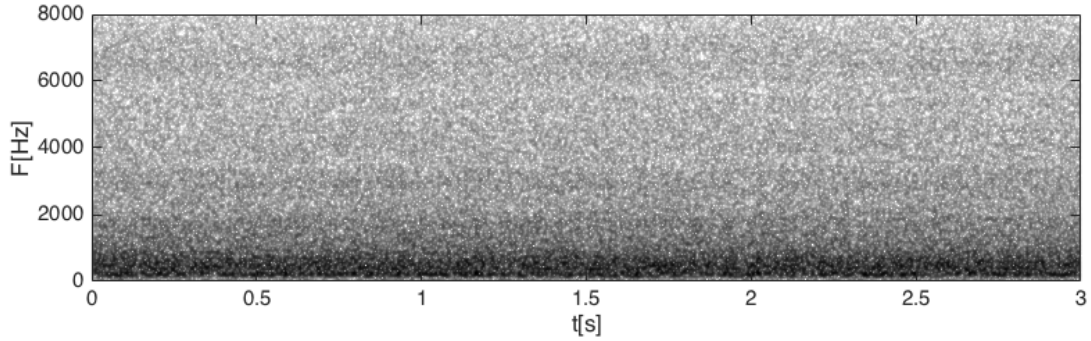


Figure 4.9: *Masker scenario 8: SSN*

The main part of the research question was to test whether masker type and SNR significantly influence speech intelligibility. Furthermore, it was investigated whether the following factors affect SI:

- gender
- number of maskers
- mixed genders
- differences in pitch
- intelligibility

4.2.4 Preparation of Material

Before the test procedure, the signals were prepared using Audacity® and MATLAB®.

The Signal-to-Noise Ratios were calculated using the root-mean-square (RMS) of the signal. The masker signal y was predefined and set at a certain RMS and the RMS

of the target x was adjusted in order to obtain the mixture of target and masker at a defined SNR. The corresponding factor was calculated via reformulation of the formula:

$$k = \frac{RMS(y) * 10^{\frac{SNR}{20}}}{RMS(x)} \tag{4.3}$$

Finally, the signals x and y were linearly mixed:

$$z = y + k \cdot x \tag{4.4}$$

4.2.5 Test Procedure

12 subjects took part in the experiment. During the test procedure, the target signals were randomly selected and mixed with the the masker signals at four different SNRs.

The subject was asked to type the heard syllable in a command-line interface in MATLAB®. Each Masker-SNR Scenario was tested 20 times and subsequently, the percentage of correct answers was calculated.

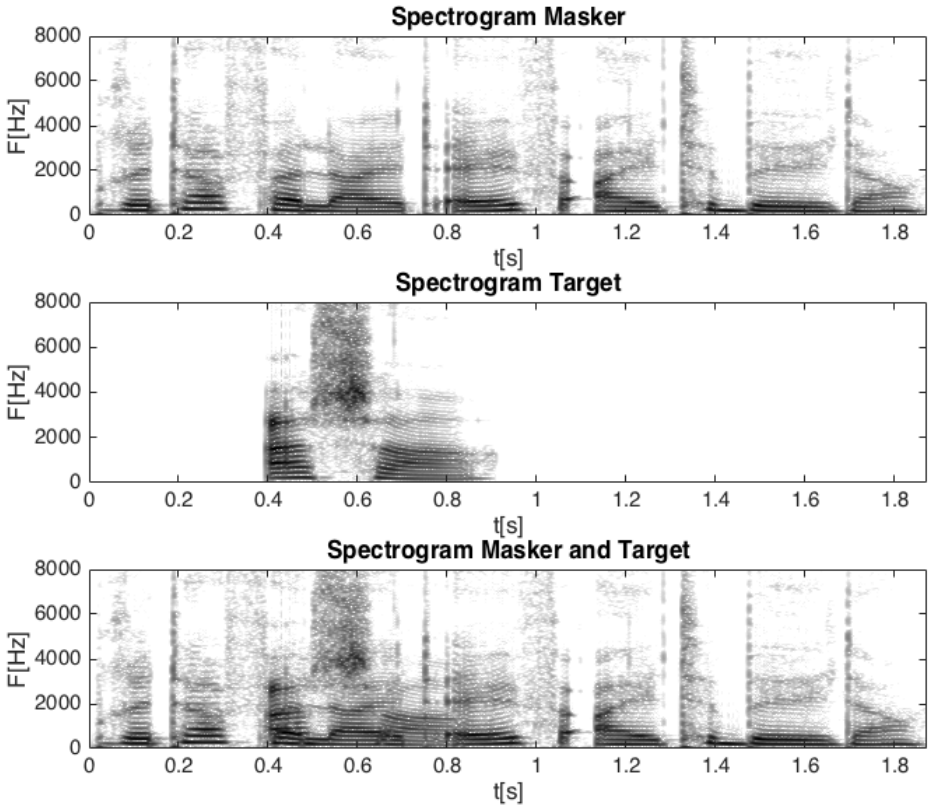


Figure 4.10: *Example of a test signal including the target signal 'ascha' during the masker scenario '1 female'.*

4.3 Results

In this section, the results of the study are presented. On the whole, there were 32 measurements for each subject corresponding to eight different maskers and four different SNR conditions. The proportion of right answers was used in statistical evaluation.

4.3.1 Graphical Representation of the Results

To start with, the overall results are presented in boxplots followed by the results in the different masker scenarios and SNR values.

Overall Results:

Figure 4.11 presents the boxplots of all 32 scenarios. One can observe that the fourth scenario has the best performance. On the other hand, Scenario 25, which corresponds to the mixed gender and pitch-manipulated maskers, has the worst result.

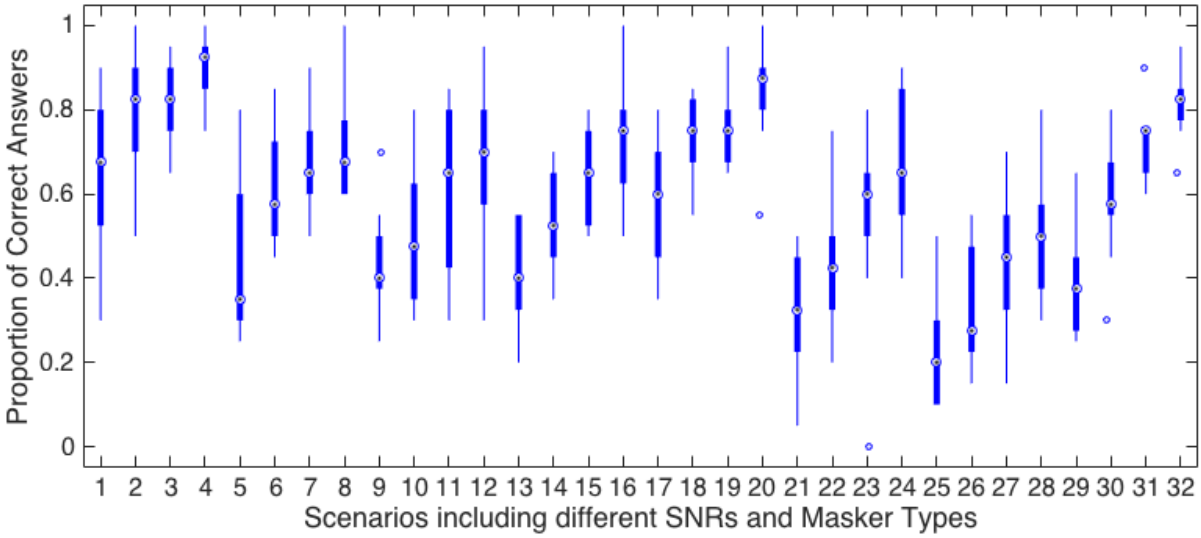


Figure 4.11: *Overall Results:*
Each masker type was tested at four SNRs -12,-9,-6 and -3 dB.
1-4: 1 female
5-8: 2 female
9-12: 3 female
13-16: 3 male
17-20: 1 ISTS female
21-24: 3 female pitch-manipulated
25-28: 2 female and 1 male
29-32: Speech Shaped Noise

Results of different masker types:

In figure 4.12, the box plots of the different masker types are presented. Masker type 1 shows the best result followed by masker type 5 which corresponds to the unintelligible ISTS. The values in masker type 3 show the largest variance among all masker types.

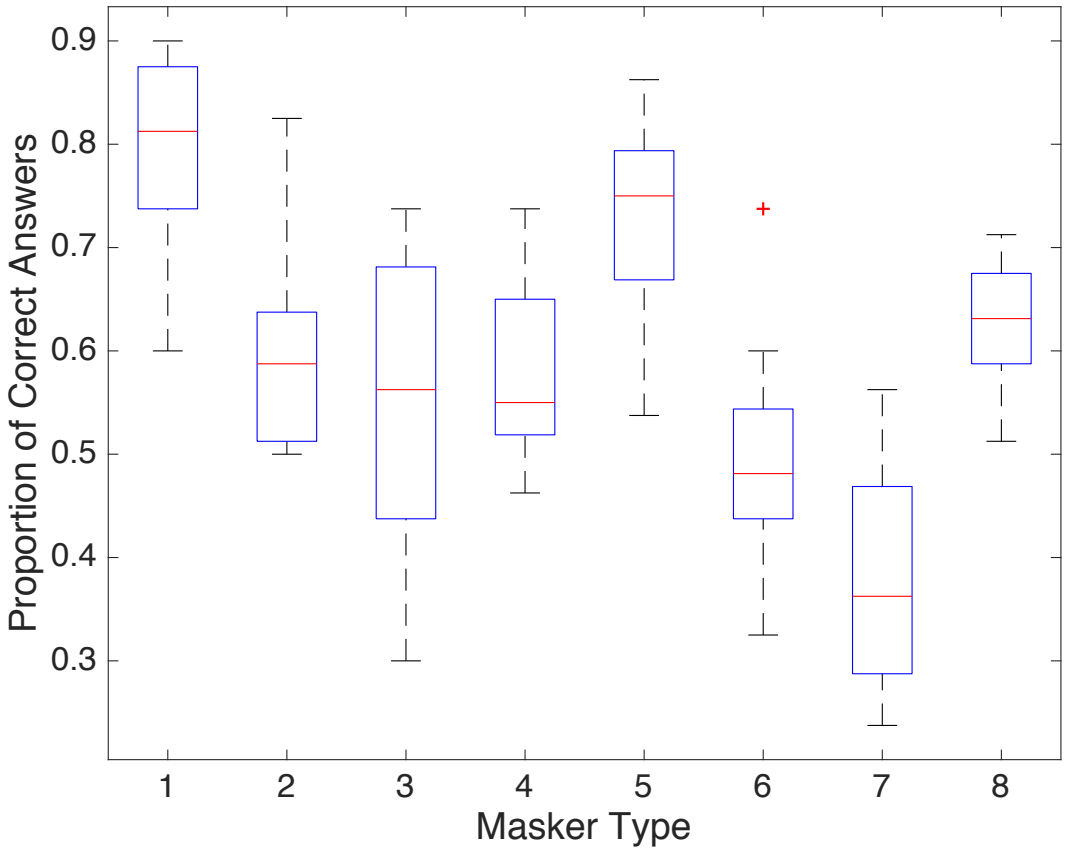


Figure 4.12: *Box plots of different masker types:*
1: 1 female
2: 2 female
3: 3 female
4: 3 male
5: 1 ISTS female
6: 3 female pitch-manipulated
7: 3 different gender
8: Speech Shaped Noise

Results at different SNRs:

In Figure 4.13, the box plots of the different SNRs are shown. As SNR increases, the proportions of right answers is increasing as well.

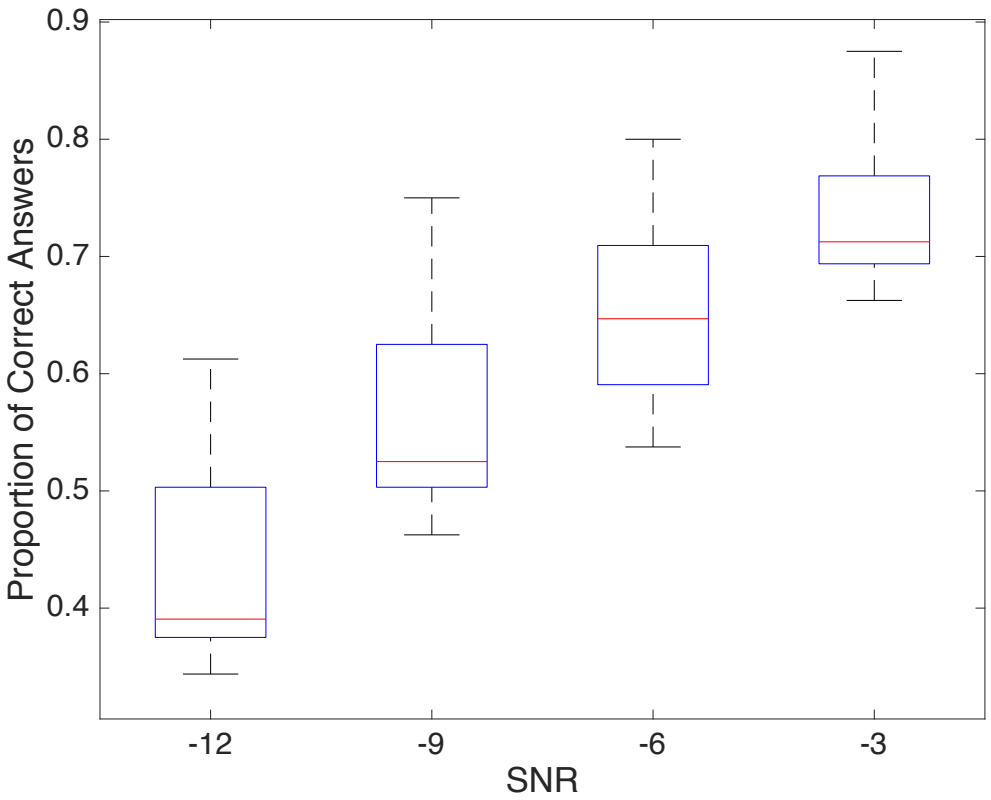


Figure 4.13: *Box plots of different SNRs*

4.3.2 Statistical Evaluation

A two-way rANOVA was performed in SPSS® Statistics (24) using the within-subject variables "SNR" and "Masker Type" in order to test their influence on Speech Intelligibility.

Independent Variable: Speech intelligibility (Proportion of correct answers)
Dependent Variables: Type of masker and SNR

In the following, the SPSS® outputs of the ANOVA are presented. They include the p-values of the different factors as well as the effect size $\eta^2_{partial}$. The results in the case of assumed sphericity as well as results including the correction factors are shown.

Test Scenario 1: Influence of masker type and SNR on speech intelligibility

Source		Sig.	Partial Eta Squared
Masker	Sphericity Assumed	,000	,761
	Greenhouse-Geisser	,000	,761
	Huynh-Feldt	,000	,761
	Lower-bound	,000	,761
Error(Masker)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
SNR	Sphericity Assumed	,000	,931
	Greenhouse-Geisser	,000	,931
	Huynh-Feldt	,000	,931
	Lower-bound	,000	,931
Error(SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
Masker * SNR	Sphericity Assumed	,161	,107
	Greenhouse-Geisser	,259	,107
	Huynh-Feldt	,198	,107
	Lower-bound	,274	,107
Error(Masker*SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		

Figure 4.14: *Results: Repeated measures ANOVA in SPSS®, in the Column Sig. one can see that the value is* < 0.05 *which states that masker type and SNR have a significant influence on speech intelligibility. There are no significant interaction effects.*

Test Scenario 2: Influence of Number of Maskers and SNR on SI

Source		Sig.	Partial Eta Squared
Number	Sphericity Assumed	,000	,658
	Greenhouse-Geisser	,000	,658
	Huynh-Feldt	,000	,658
	Lower-bound	,001	,658
Error(Number)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
SNR	Sphericity Assumed	,000	,776
	Greenhouse-Geisser	,000	,776
	Huynh-Feldt	,000	,776
	Lower-bound	,000	,776
Error(SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
Number * SNR	Sphericity Assumed	,397	,088
	Greenhouse-Geisser	,391	,088
	Huynh-Feldt	,397	,088
	Lower-bound	,326	,088
Error(Number*SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		

Figure 4.15: *Results: In the column Sig. the value is* < 0.05 *and so the number of maskers and SNR have a significant influence on speech intelligibility. Again, there are no significant interaction effects.*

Test Scenario 3: Influence of Gender and SNR on SI

Source		Sig.	Partial Eta Squared
Gender	Sphericity Assumed	,568	,030
	Greenhouse-Geisser	,568	,030
	Huynh-Feldt	,568	,030
	Lower-bound	,568	,030
Error(Gender)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
SNR	Sphericity Assumed	,000	,726
	Greenhouse-Geisser	,000	,726
	Huynh-Feldt	,000	,726
	Lower-bound	,000	,726
Error(SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
Gender * SNR	Sphericity Assumed	,504	,068
	Greenhouse-Geisser	,488	,068
	Huynh-Feldt	,504	,068
	Lower-bound	,391	,068
Error(Gender*SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		

Figure 4.16: *Results: Gender of speaker has no significant influence on speech intelligibility.*

Test Scenario 4: Influence of Mixed Gender and SNR on SI

Source		Sig.	Partial Eta Squared
Mixed	Sphericity Assumed	,000	,790
	Greenhouse-Geisser	,000	,790
	Huynh-Feldt	,000	,790
	Lower-bound	,000	,790
Error(Mixed)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
SNR	Sphericity Assumed	,000	,647
	Greenhouse-Geisser	,000	,647
	Huynh-Feldt	,000	,647
	Lower-bound	,001	,647
Error(SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
Mixed * SNR	Sphericity Assumed	,688	,043
	Greenhouse-Geisser	,599	,043
	Huynh-Feldt	,629	,043
	Lower-bound	,496	,043
Error(Mixed*SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		

Figure 4.17: *Results: The Factor "Mixed Gender" or "Spectral Diversity" of the speakers has a significant influence on speech intelligibility. There are no significant interaction effects with SNR.*

Test Scenario 5: Influence of Intelligibility and SNR on SI

Source		F	Sig.	Partial Eta Squared
Intelligibility	Sphericity Assumed	3,687	,081	,251
	Greenhouse-Geisser	3,687	,081	,251
	Huynh-Feldt	3,687	,081	,251
	Lower-bound	3,687	,081	,251
Error(Intelligibility)	Sphericity Assumed			
	Greenhouse-Geisser			
	Huynh-Feldt			
	Lower-bound			
SNR	Sphericity Assumed	27,814	,000	,717
	Greenhouse-Geisser	27,814	,000	,717
	Huynh-Feldt	27,814	,000	,717
	Lower-bound	27,814	,000	,717
Error(SNR)	Sphericity Assumed			
	Greenhouse-Geisser			
	Huynh-Feldt			
	Lower-bound			
Intelligibility * SNR	Sphericity Assumed	,069	,976	,006
	Greenhouse-Geisser	,069	,937	,006
	Huynh-Feldt	,069	,962	,006
	Lower-bound	,069	,797	,006
Error(Intelligibility*SNR)	Sphericity Assumed			
	Greenhouse-Geisser			
	Huynh-Feldt			
	Lower-bound			

Figure 4.18: *Results: Intelligibility of the masker has no significant influence on speech intelligibility.*

Test Scenario 6: Influence of Pitch and SNR on SI

Source		Sig.	Partial Eta Squared
Pitch	Sphericity Assumed	,087	,243
	Greenhouse-Geisser	,087	,243
	Huynh-Feldt	,087	,243
	Lower-bound	,087	,243
Error(Pitch)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
SNR	Sphericity Assumed	,000	,704
	Greenhouse-Geisser	,000	,704
	Huynh-Feldt	,000	,704
	Lower-bound	,000	,704
Error(SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		
Pitch * SNR	Sphericity Assumed	,348	,094
	Greenhouse-Geisser	,343	,094
	Huynh-Feldt	,348	,094
	Lower-bound	,309	,094
Error(Pitch*SNR)	Sphericity Assumed		
	Greenhouse-Geisser		
	Huynh-Feldt		
	Lower-bound		

Figure 4.19: *Results: Pitch has no influence on speech intelligibility.*

Summary of the Statistical Evaluation:

Summary

	Independent Variables		Results		
Test Scenario 1	Masker type	SNR [dB]	Masker	SNR	Masker*SNR
	1 female 2 female 3 female 3 male 1 female ISTS 3 female pitch manipulated 3 mixed gender Speech Shaped Noise	-3 -6 -9 -12	p<0.05	p<0.05	p>0.05
Test Scenario 2	Number of Maskers	SNR [dB]	Number	SNR	Number*SNR
	1 female 2 female 3 female	-3 -6 -9 -12	p<0.05	p<0.05	p>0.05
Test Scenario 3	Gender	SNR [dB]	Gender	SNR	Gender*SNR
	3 female 3 male	-3 -6 -9 -12	p>0.05	p<0.05	p>0.05
Test Scenario 4	Mixed or Unmixed Gender	SNR [dB]	Mixed	SNR	Mixed*SNR
	3 female 3 mixed gender	-3 -6 -9 -12	p<0.05	p<0.05	p>0.05
Test Scenario 5	Intelligibility	SNR [dB]	Intelligibility	SNR	Intelligibility*SNR
	1 female 1 female ISTS	-3 -6 -9 -12	p>0.05	p<0.05	p>0.05
Test Scenario 6	Pitch	SNR [dB]	Pitch	SNR	Pitch*SNR
	3 female 3 female pitch-manipulated	-3 -6 -9 -12	p>0.05	p<0.05	p>0.05

Figure 4.20: *Summary of Results: the red-coloured p-values highlight the scenarios in which the null hypothesis was rejected.*

4.4 Interpretation

On the whole, the results were in accordance with other publications on the topic. However, some interesting findings were achieved in the fields of the influence of spectrotemporal diversity of the masker on SI.

Surprisingly, the results were significantly worse when the maskers had different spectrotemporal features like in the scenarios with masker type 7 that included two female speakers, one of them pitch-manipulated, and one male speaker. One possible explanation for the decrease in performance could be the argument that if the maskers are spectrotemporally similar, they might mask each other and so the overall masking effect on the target is smaller which was stated in section 2.3.4. Therefore, the amount of masking within the masker signal is low in the case of spectrally diverse speech signals.

Furthermore, speech intelligibility decreased with increasing number of maskers and the multimasker penalty can be observed in these scenarios.
Intelligibility of Masker did not influence speech intelligibility in a significant way which may be due to the low context of the target signals.

The pitch-manipulated maskers did not significantly change the speech intelligibility in the case of only female speakers which may be explained by the fact that the spectral components of the maskers still resembled and as a consequence, they masked each other to a certain extent.

Finally, there were no main effects of gender which can also be explained by the masking effects which occur within the masker signals if the maskers themselves are spectrotemporally similar.

4.5 Ranking of Syllables and Confusion Analysis

In the following, the ranking and the confusion of the target signals are investigated.

In speech perception which underlies masking effects, sounds may be confused with other related sounds. *Confusion analysis* investigates clusters of these confused sounds and analyzes underlying perceptual features (Phatak and Allen, 2007).

A common used tool for analyzing confusion in a closed-set experimental recognition task is the *Confusion Matrix* (CM), where each entry corresponds to a certain probability $P_{ab}(SNR)$ of a spoken sound a which was reported as sound b during the task.

In a consonant recognition task, clusters of sounds which are likely to be confused may include different manners of articulation like plosives, fricatives, nasals, affricates and laterals, that are grouped in different ways. In varying SNR, these clusters may change.

A *Confusion Pattern* (CP) is a common used method in confusion analysis to graphically present a row of the CM as a function of SNR.

4.5.1 Ranking of the Syllables

To start with, the overall results are presented including a ranking of the 14 syllables.

Ranking and Percentage rate of correctly identified syllables:

In Figure 4.21, it can be seen that the syllable 'ascha' had the best recognition rate followed by 'anna'. On the other hand, 'appa' had the highest error and confusion rate.

Overall Results

VCV	Proportion of right answers
'ascha'	0.8765
'anna'	0.7588
'agga'	0.7413
'atta'	0.7049
'acka'	0.6654
'assa'	0.6458
'alla'	0.6092
'awwa'	0.6004
'azza'	0.5760
'affa'	0.5178
'amma'	0.4712
'adda'	0.4311
'abba'	0.4216
'appa'	0.3739

Figure 4.21: *Results of VCVs*

Proportion of correct answers depending on SNR:

On the whole, the proportion of correct answers increases with increasing SNR except for a slight decrease in performance of 'awwa' and 'azza' at the SNR -6 dB to the SNR -3 dB as one can see in Figure 4.22.

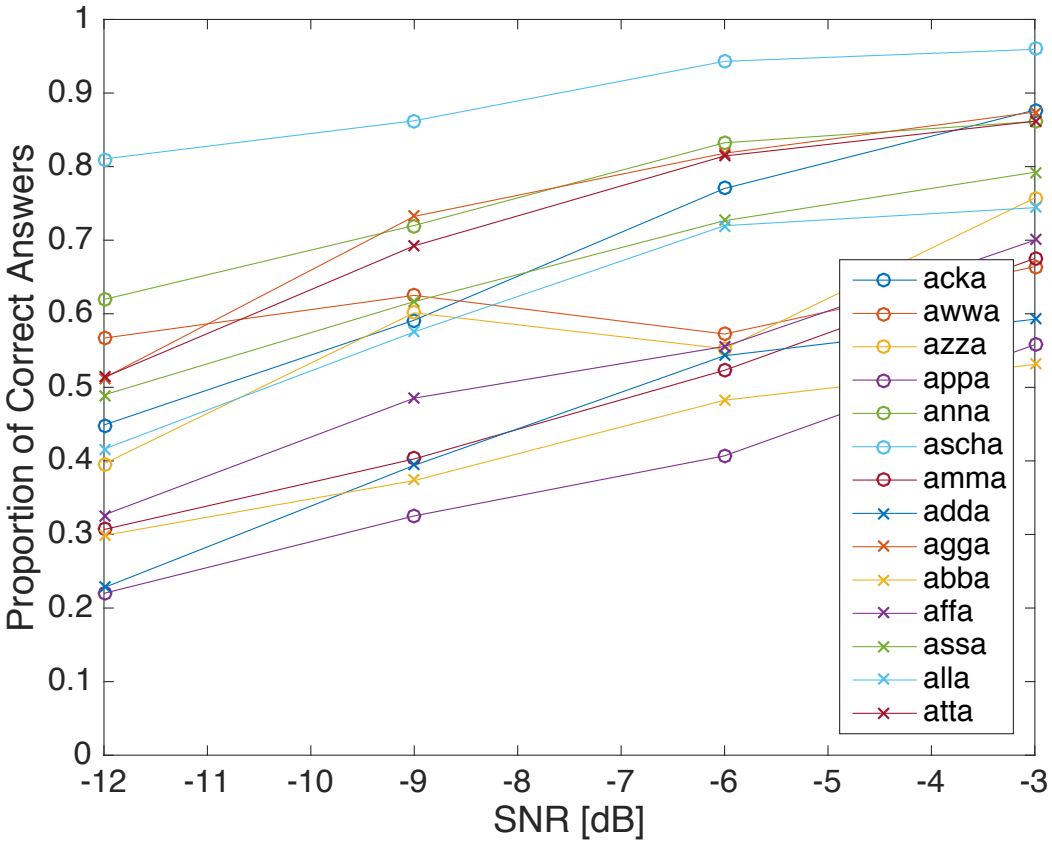


Figure 4.22: *Proportion of correct answers depending on different SNRs*

Proportion of correct answers depending on Maskers:

Figure 4.23 demonstrates that the proportion of correct answers of the syllable 'ascha' tends to be constantly at a high level with only small changes except for the seventh scenario in which the performance strongly decreases.

The progress of the proportion of correct answers of the syllable 'awwa' seems to be more random in comparison with that of the other syllables. Furthermore, the variation of proportion is smaller in the scenarios with only one masker voice (scenario 1 and scenario 5).

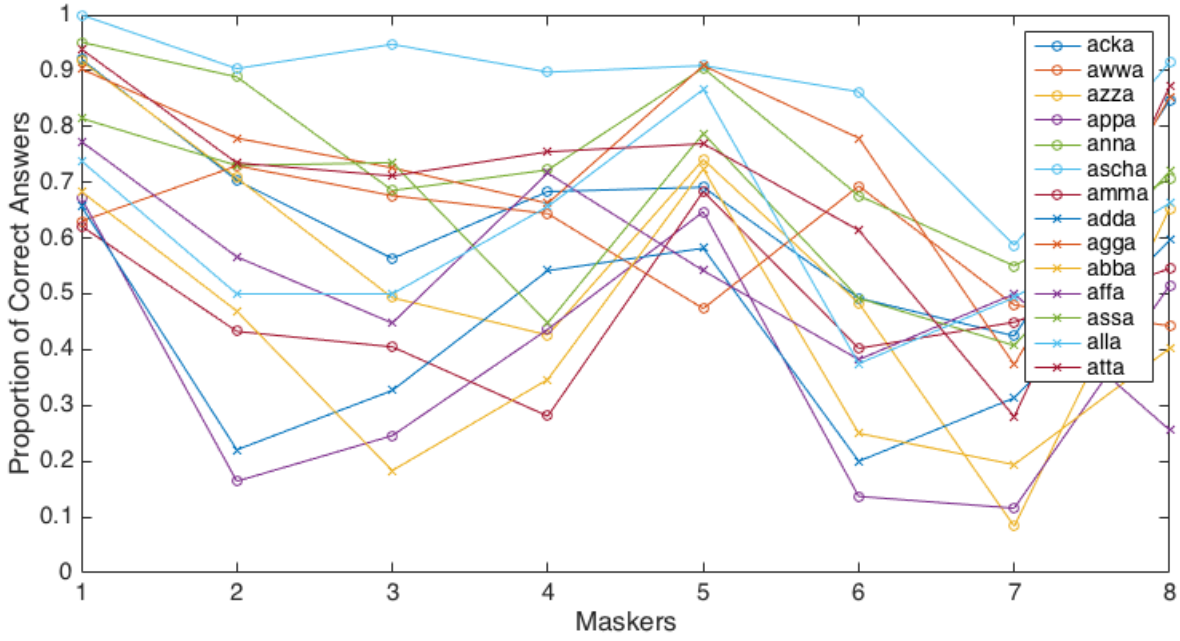


Figure 4.23: *Proportion of correct answers depending on Masker type:*
1: 1 female
2: 2 female
3: 3 female
4: 3 male
5: 1 ISTS female
6: 3 female pitch-manipulated
7: 3 different gender
8: Speech Shaped Noise

4.5.2 Confusion Analysis

Confusion Analysis is a widely used task in closed speech intelligibility tests. The Confusion Matrix (CM) and the Confusion Pattern (CP) are possibilities to demonstrate differences as well as patterns in the confusion rates.

Confusion Matrix:

The Confusion Matrix (CM) is the most common method in confusion analysis. The rows of the CM correspond to the actual syllable and the columns to the assumed syllable. In the diagonal entries of the matrix, one can observe the amount of correct answers.

In Figure 4.24, the overall CM is shown.

Confusion Matrix

	acka	awwa	azza	appa	anna	ascha	amma	adda	agga	assa	affa	atta	abba	alla
acka	360	9	13	7	7	3	2	10	60	14	22	11	8	2
awwa	1	341	2	7	22	0	50	16	6	3	25	2	77	12
azza	15	7	360	15	5	3	4	14	11	74	55	49	2	5
appa	34	32	26	209	10	3	9	15	14	15	135	20	33	3
anna	2	32	2	1	453	0	30	17	10	0	2	1	2	42
ascha	3	4	3	2	0	447	0	7	6	1	4	10	10	3
amma	1	115	2	6	129	0	295	15	7	4	9	4	14	19
adda	5	19	2	1	7	2	4	197	157	12	7	9	6	21
agga	24	21	3	2	8	2	2	30	407	8	8	7	14	11
assa	5	14	19	7	10	2	6	18	8	341	75	10	8	2
affa	11	58	23	34	12	2	18	15	18	14	306	8	64	6
atta	19	10	20	7	2	5	3	8	6	9	15	289	6	7
abba	1	236	1	5	8	0	8	21	20	5	19	1	242	5
alla	0	33	0	1	73	4	18	37	25	6	3	2	10	332

Figure 4.24: *Overall Confusion Matrix: $CM(i, j)$ denotes to the true syllable i and the heard syllable j*

In the following, differences in Confusion Matrices of different SNRs are shown and analyzed.

The CM in Figure 4.25 corresponds to the results of the scenarios including the SNR -3 dB which had the best results. As one can see, the values apart from the diagonal are very low and there are many zero entries which underline the low confusion rates.

Confusion Matrix at SNR -3

	acka	awwa	azza	appa	anna	ascha	amma	adda	agga	abba	affa	assa	alla	atta
acka	121	0	0	0	0	0	1	1	12	0	1	1	1	0
awwa	0	83	0	1	5	0	13	0	1	16	3	0	3	0
azza	3	0	128	6	1	0	1	0	0	0	8	13	1	8
appa	8	5	2	77	2	0	2	3	2	6	24	1	1	5
anna	0	2	0	1	118	0	7	0	1	2	1	0	4	1
ascha	0	0	0	0	0	118	0	0	1	0	2	0	0	2
amma	0	18	0	0	22	0	106	2	0	2	0	1	6	0
adda	1	0	0	0	1	1	1	64	35	0	1	2	1	1
agga	9	1	0	0	0	0	1	1	125	2	1	1	1	1
abba	0	52	0	1	2	0	0	1	3	77	6	0	1	0
affa	1	10	0	7	3	1	1	4	2	17	112	0	1	1
assa	0	3	2	0	0	0	1	0	0	1	18	103	1	1
alla	0	7	0	0	12	0	2	3	3	2	1	1	96	2
atta	1	2	2	1	0	1	0	1	0	1	2	1	1	81

Figure 4.25: *Confusion Matrix at SNR -3 dB*

The next CM in Figure 4.26 describes the confusion rates of the scenarios including the SNR -12 dB, which was the lowest in this experiment.

One can clearly recognize that the values apart from the diagonal are larger in comparison with the previous CM of the SNR -3 dB which demonstrates the high confusion rates.

Confusion Matrix at SNR -12

	adda	atta	anna	awwa	acka	azza	amma	assa	abba	ascha	appa	alla	affa	agga
adda	23	3	3	8	3	2	2	8	3	1	0	6	3	36
atta	7	57	2	3	8	10	1	4	3	2	2	4	4	4
anna	11	0	83	11	1	0	8	0	0	0	0	13	1	6
awwa	8	1	8	89	1	0	9	1	18	0	4	4	13	1
acka	4	4	6	5	57	5	1	6	5	3	4	0	8	19
azza	5	15	2	4	6	61	3	21	2	3	5	3	15	9
amma	7	3	41	39	0	0	50	2	5	0	3	8	1	4
assa	13	5	4	5	5	9	1	71	4	1	2	1	18	6
abba	10	0	5	55	1	0	5	4	43	0	2	3	9	7
ascha	4	6	0	2	1	2	0	0	3	94	0	1	0	3
appa	8	5	3	10	17	6	2	6	10	3	33	2	39	8
alla	19	0	31	9	0	0	6	4	2	1	0	57	0	8
affa	7	5	5	14	5	14	5	7	15	1	11	3	48	7
agga	14	5	6	9	7	2	1	3	2	0	1	5	3	61

Figure 4.26: *Confusion Matrix at SNR -12 dB*

Confusion Pattern:

The order of a CM is of big importance in analyzing the behaviour of consonant confusion and formation of certain clusters.

The concept of Confusion Patterns (CP) overcomes the difficulty of recognizing perceptual clusters among consonants (Allen, 2005). A Confusion Pattern describes the graphical representation of a CM row as a function of SNR (Phatak and Allen, 2007).

As an example, the CP of the syllable 'affa' is presented. In Figure 4.27, it can be seen that the proportion of confusion of the syllables is more likely to decrease with increasing SNR because the proportion of right answers increases in this case.

It is noteworthy that the proportion of confusion with the most similar sounding syllable 'awwa' increases from the SNR -12 dB to -6 dB and then strongly decreases at the SNR -3 dB.

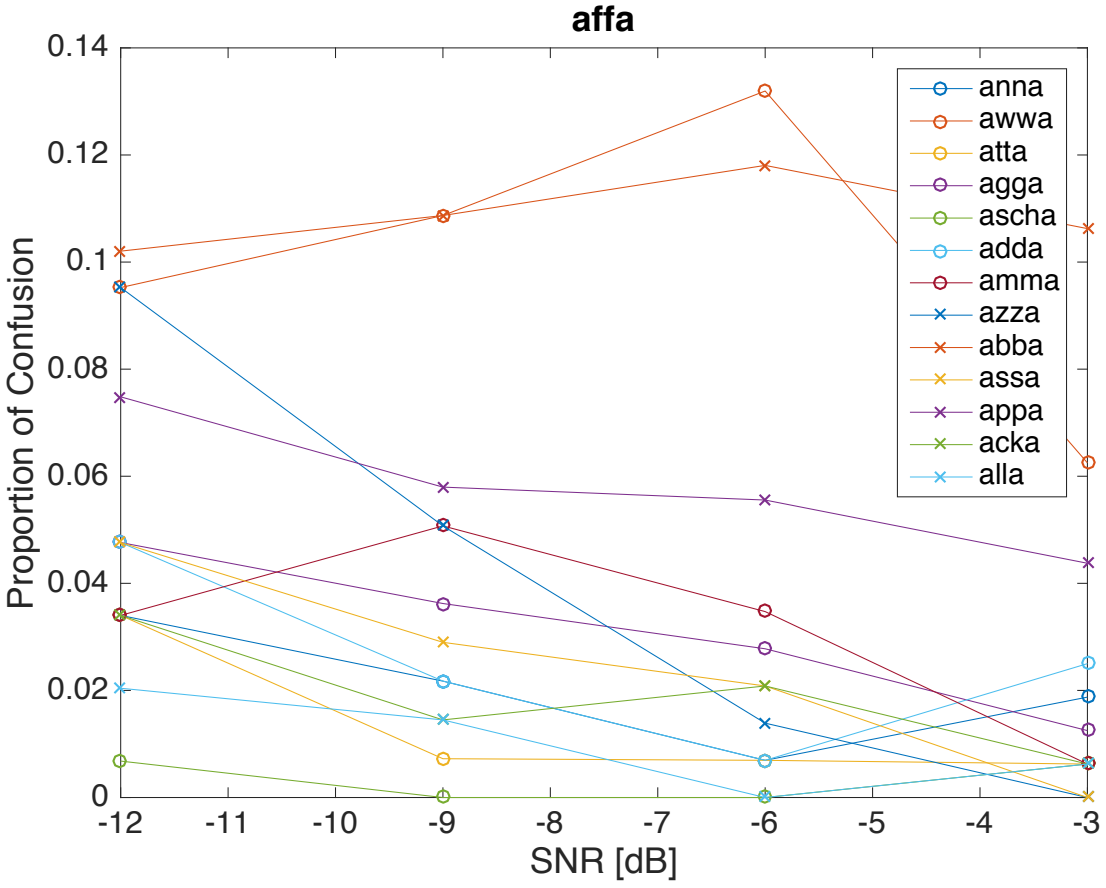


Figure 4.27: *Confusion Pattern (CP) of the syllable 'affa'*

Another option in Confusion Analysis is the investigation of the CP by clustering the consonants. As mentioned in Section 2.2.2, there are different articulation manners of the consonants which lead to similar sounds that can be clustered in groups which include consonants that are most likely to be confused with each other.

Again, the CP of the syllable 'affa' is shown. In order to obtain homogeneous groups, nasal, affricate and lateral consonants are clustered in this case. In Figure 4.28, the syllable 'affa' is included in the group of fricatives. The proportion of confusion within this group is increasing as the SNR increases. In general, 'affa' tends to be confused with plosives rather than with the third group including nasals, affricates and laterals.

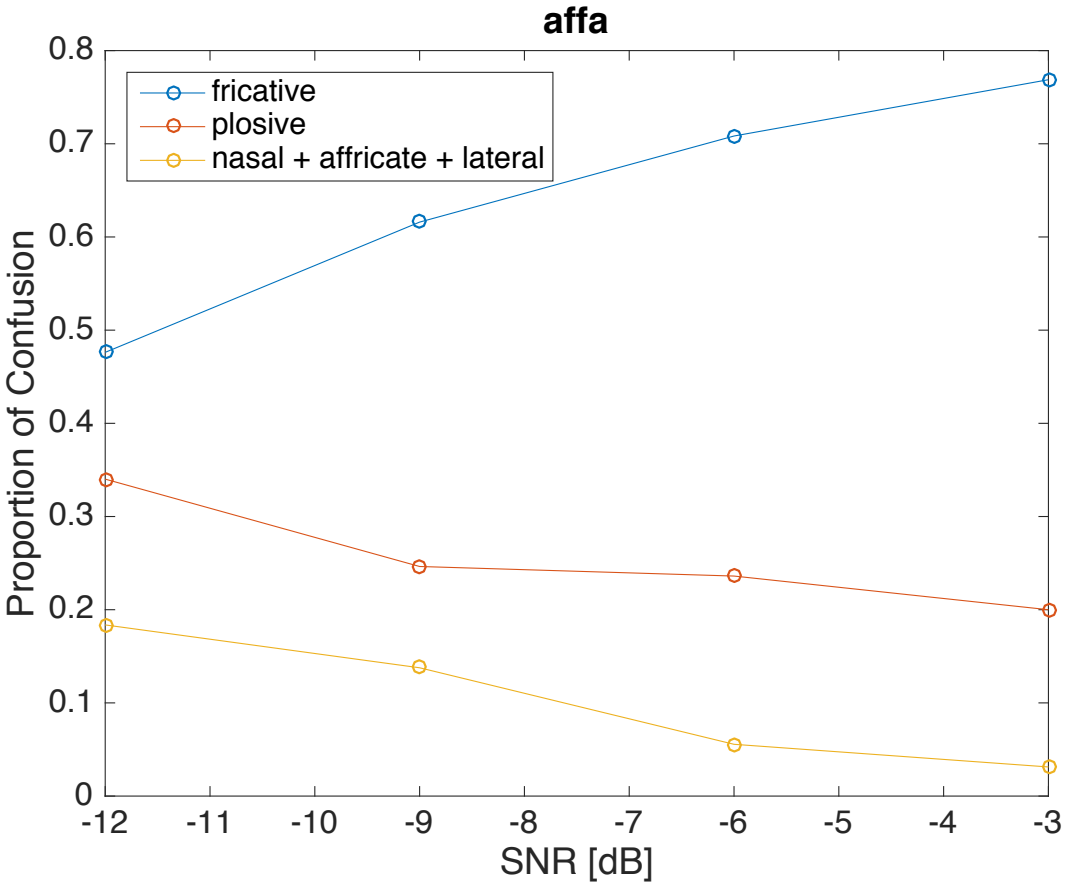


Figure 4.28: *CP of the syllable 'affa' including clustering of the consonants:*
fricative: 'assa', 'affa', 'ascha', 'awwa'
plosive: 'appa', 'abba', 'atta', 'adda', 'acka', 'agga'
nasal: 'amma', 'anna'
affricate: 'azza'
lateral: 'alla'

5 Conclusion

In this thesis, different approaches of analyzing speech signals in multitalker environments were presented and a speech intelligibility test was performed in order to analyze masking effects in speech perception.

Firstly, Computational Auditory Scene Analysis aims at imitating the different auditory processing stages to separate a mixture of speech signals and CASA algorithms and ideal binary mask (IBM) approaches manage to improve Speech Intelligibility significantly (Kim et al., 2009; Brown and Wang, 2005).

Secondly, Blind Source Separation (BSS) is a powerful technique to separate signal mixtures by statistical means without any information about the mixing process. The common-used approach Independent Component Analysis (ICA) manages to improve speech intelligibility in multitalker environments as well (Brown and Wang, 2005).

The experimental part of the thesis demonstrated a significant main effect of "SNR" ($p<0.001$) and "Masker Type" ($p<0.001$) on low-context speech perception and furthermore, a significant main effect of the factors "Number of Maskers" ($p<0.001$) and "Spectral Diversity of the Masker" ($p<0.001$).

The key finding of the experiment was the strong impact of spectral diversity of the masker signals on speech intelligibility performance which can be partly explained by a lack of masking effects within the masker signal due to a low amount of overlapping T-F regions.

The increasing number of maskers significantly affected SI which is in accordance with former investigation and indicates the existence of a multimasker penalty in multitalker environments including low-context target signals.
Confusion analysis demonstrated the importance of spectrotemporal attributes of the target signal in addition to that of the masker signal.

In further studies, the Speech Reception Threshold (SRT) may be used in statistical evaluation instead of the proportion factor because the latter may lead to problems in ANOVA. The SRT is the sound intensity in dB at which 50% of the syllables are correctly defined by the subject. Another possibility to overcome the problem of proportions in ANOVA is to perform data transformations like the arcsine transformation which is a common used method for dealing with proportional or percentage data in advance of ANOVA.

Also, the target signals can be extended by increasing the number of VCVs because in this thesis, only the VCVs including the vowel /a/ have been used. Also, recordings of male speakers of the target signals should be included and the Confusion Patterns (CPs) can be investigated in more detail.

Furthermore, some masker types may be added which may include a larger number of speakers or mixtures of unintelligible speech signals. Moreover, scenarios varying in the constellation of genders can be added and the number of speakers in the recordings of the masker signals should be increased for further analysis.

Finally, the impact of spectral diversity of the masker can be investigated in more detail by using different half tone steps at the pitch-manipulation stage and the test may also be extended in regards of the investigation of speech intelligibility of high-context speech like words or sentences.

References

Peter W Alberti. The anatomy and physiology of the ear and hearing. *Occupational exposure to noise: Evaluation, prevention, and control*, pages 53–62, 2001.

Jont B Allen. Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America*, 117(4):2212–2223, 2005.

ANSI. American national standard methods for calculation of the speech intelligibility index. *ANSI S3.5-1997*, 1997.

Tanya L Arbogast, Christine R Mason, and Gerald Kidd Jr. The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, 112(5):2086–2098, 2002.

Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weiber. *Multivariate analysemethoden: eine anwendungsorientierte einführung*. Springer-Verlag, 2015.

Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

Arthur Boothroyd and Susan Nittrouer. Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, 84 (1):101–114, 1988.

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

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

Guy Brown and DeLiang Wang. Separation of speech by computational auditory scene analysis. *Speech enhancement*, pages 371–402, 2005.

Guy J Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

Douglas S Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3): 1101–1109, 2001.

Douglas S Brungart, Brian D Simpson, Mark A Ericson, and Kimberly R Scott. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, 110(5):2527–2538, 2001.

Keith Conrad. Probability distributions and maximum entropy. *retrieved November*, 14: 2013, 2013.

Martin Cooke, ML Garcia Lecumberri, and Jon Barker. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, 123(1):414–427, 2008.

Egbert De Boer and Paul Kuyper. Triggered correlation. *IEEE Transactions on Biomedical Engineering*, (3):169–179, 1968.

Finn Dubbelboer and Tammo Houtgast. The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *The Journal of the Acoustical Society of America*, 124(6):3937–3946, 2008.

Nat Durlach. Auditory masking: Need for improved conceptual structure a. *The Journal of the Acoustical Society of America*, 120(4):1787–1790, 2006.

Nathaniel I Durlach, Kaigham J Gabriel, H Steven Colburn, and Constantine Trahiotis. Interaural correlation discrimination: Ii. relation to binaural unmasking. *The Journal of the Acoustical Society of America*, 79(5):1548–1557, 1986.

Nathaniel I Durlach, Christine R Mason, Gerald Kidd Jr, Tanya L Arbogast, H Steven Colburn, and Barbara G Shinn-Cunningham. Note on informational masking (l). *The Journal of the Acoustical Society of America*, 113(6):2984–2987, 2003.

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*, 2016.

Ludwig Fahrmeir, Christian Heumann, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag, 2016.

Klaus Fellbaum. *Sprachverarbeitung und Sprachübertragung*. Springer-verlag, 2013.

Harvey Fletcher. Auditory patterns. *Reviews of modern physics*, 12(1):47, 1940.

Richard L Freyman, Uma Balakrishnan, and Karen S Helfer. Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 109(5):2112–2122, 2001.

Tine Goossens, Charlotte Vercammen, Jan Wouters, and Astrid van Wieringen. Masked speech perception across the adult lifespan: Impact of age and hearing impairment. *Hearing research*, 344:109–124, 2017.

Donald D Greenwood. Auditory masking and the critical band. *The journal of the acoustical society of America*, 33(4):484–502, 1961.

K. Grill. *Mass- und Wahrscheinlichkeitstheorie*. Karl Grill, 2017.

Monica L Hawley, Ruth Y Litovsky, and John F Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843, 2004.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Antje Ihlefeld and Barbara Shinn-Cunningham. Spatial release from energetic and informational masking in a selective speech identification task a. *The Journal of the Acoustical Society of America*, 123(6):4369–4379, 2008.

Nandini Iyer, Douglas S Brungart, and Brian D Simpson. Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *The Journal of the Acoustical Society of America*, 128(5):2998–3010, 2010.

Lloyd A Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

Gerald Kidd Jr, Christine R Mason, Tanya L Rohtla, and Phalguni S Deliwala. Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, 104(1):422–431, 1998.

Gibak Kim, Yang Lu, Yi Hu, and Philipos C Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–1494, 2009.

Gilad Lerman. The shannon sampling theorem and its implications. *Lecture Notes in Mathematics*, 467.

H Levitt and LR Rabiner. Binaural release from masking for speech and gain in intelligibility. *The Journal of the Acoustical Society of America*, 42(3):601–608, 1967.

Philipos C Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE transactions on audio, speech, and language processing*, 19(1):47–56, 2011.

Shoji Makino, Te-Won Lee, and Hiroshi Sawada. *Blind speech separation*, volume 615. Springer, 2007.

Pertti Mattila. *Fourier analysis and Hausdorff dimension*, volume 150. Cambridge University Press, 2015.

John W Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

Ganesh R Naik, Wenwu Wang, et al. *Blind source separation*. Springer, 2014.

Gordon E Peterson and Harold L Barney. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184, 1952.

Sandeep A Phatak and Jont B Allen. Consonant and vowel confusions in speech-weighted noise a. *The Journal of the Acoustical Society of America*, 121(4):2312–2326, 2007.

Matthieu Puigt and Yannick Deville. Time–frequency ratio-based blind separation methods for attenuated and time-delayed sources. *Mechanical Systems and Signal Processing*, 19(6):1348–1379, 2005.

Björn Rasch, Malte Friese, Wilhelm Hofmann, and Ewald Naumann. Quantitative methoden 2. einführung in die statistik für psychologen und sozialwissenschaftler. 3., erweiterte auflage, 2010.

Frank Rattay and Petra Lutter. Speech sound representation in the auditory nerve: computer simulation studies on inner ear mechanisms. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 77(12):935–943, 1997.

W Rutherford. A new theory of hearing. *Journal of anatomy and physiology*, 21(Pt 1): 166, 1886.

Herbert Sager. *Fourier-Transformation: Beispiele, Aufgaben, Anwendungen.* vdf Hochschulverlag AG, 2012.

Bruce McA Sayers and E Colin Cherry. Mechanism of binaural fusion in the hearing of speech. *The Journal of the Acoustical Society of America*, 29(9):973–987, 1957.

Wiebke Schubotz, Thomas Brand, Birger Kollmeier, and Stephan D Ewert. Monaural speech intelligibility and detection in maskers with varying amounts of spectrotemporal speech features. *The Journal of the Acoustical Society of America*, 140(1): 524–540, 2016.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Roger N Shepard and Daniel J Levitin. *Cognitive psychology and music.* MIT Press, 2002.

Soundararajan Srinivasan and DeLiang Wang. A model for multitalker speech perception. *The Journal of the Acoustical Society of America*, 124(5):3213–3224, 2008.

Georg Von Békésy and Ernest Glen Wever. *Experiments in hearing*, volume 8. McGraw-Hill New York, 1960.

Deilang Wang and Guy J Brown. Fundamentals of computational auditory scene analysis. *Computational auditory scene analysis: Principles, Algorithms, and Applications*, pages 1–44, 2006a.

DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pages 181–197. Springer, 2005.

DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE Press, 2006b.