Technische Universität Berlin

Masterarbeit

Speech Intelligibility and Reverberation at Low Frequencies

Vorgelegt von: Amirmohammad Ala
[redacted]

Gutachter: Prof. Dr. Stefan Weinzierl
Betreuer: Dr. Fabian Brinkmann, Dr. Paul Luizard

Technische Universität Berlin, Fakultät für Geistes- und Bildungswissenschaften
Institut für Sprache und Kommunikation,
Fachgebiet Audiokommunikation und -technologie,
Wintersemester 2020/21

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Berlin, den 28.10.2020
Amirmohammad Ala

Acknowledgement

Foremost, I would like to express my sincere gratitude to my supervisor, Prof. Stefan Weinzierl for suggesting the topic, at the very beginning, and his valuable and consistent guidance during the rest of the project.
I would also like to thank my advisors, Dr. Fabian Brinkmann and Dr. Paul Luizard for their useful instructions and giving me always a helping hand, whenever I ran into a problem.
Last but not least, I would like to thank my mother, father and brother, for their support and continuous encouragement through the process of writing this thesis.

Abstract

Reverberation time (RT) plays an essential role in speech intelligibility in rooms. In few studies, the effect of low-frequency RT on speech intelligibility was investigated, in most of which the results of room acoustical simulations were used for auralization and performing listening tests to assess the speech intelligibility. The results show a deteriorating effect of increasing the low-frequency RT on the speech intelligibility where either Bass Ratio or merely RT at 250-Hz octave band was considered as the independent variable. However, a systematic study investigating the independent alteration of RT at 125-Hz and 250-Hz octave bands, together with a wide range of the other essential parameters, including broadband RT, room volume and room shape, was missing. Moreover, no study investigated the effect of these parameters on single-channel measures (such as STI), binaural measures (such as Binaural Benefit) and their combination, simultaneously.

In this contribution, input variables, including four broadband RT's (0.5, 1, 2 and 4 *sec*), four typical room shapes (Fan shaped, Horseshoe, Shoebox and Vineyard) scaled in four volumes (250, 1000, 4000 and 16000 m^3), together with RT alterations in seven steps at 125-Hz and 250-Hz octave bands, were investigated. Results of room acoustical simulations based on a combination of ray tracing and image sound source methods are fed into speech intelligibility algorithms to calculate the output variables STI, Binaural Benefit and Binaural STI (according to the method introduced in this study).

Finally, the interaction of the input and output variables is statistically analysed using a Generalized Linear Model (GLM).

The results show that the alteration of RT at 125-Hz or 250-Hz octave band has a statistically significant effect on Binaural STI. Specifically, an increase of RT at 125-Hz or 250-Hz octave band (by a scaling factor of 0.7 to 1.3) results in an average drop of Binaural STI by 0.01 or 0.02, respectively. These findings suggest taking measures to reduce the RT at low frequencies, besides doing so at the well-known range of 500 Hz - 4 kHz, to improve the speech intelligibility.

Zusammenfassung

Die Nachhallzeit (RT) spielt eine wesentliche Rolle für die Sprachverständlichkeit in Räumen. Der Effekt der tieffrequenten Nachhallzeit auf die Sprachverständlichkeit wurde erst in wenigen Studien untersucht. In den meisten dieser Studien wurden die Ergebnisse raumakustischer Simulationen für die Auralization und die Durchführung von Hörversuchen benutzt. Die Ergebnisse zeigen eine negative Wirkung der Erhöhung der tieffrequenten Nachhallzeit auf die Sprachverständlichkeit, wobei entweder das Bassverhältnis oder nur die Nachhallzeit im 250-Hz-Oktavband als unabhängige Variable betrachtet wurde. Es fehlte jedoch eine systematische Studie, die die unabhängige Veränderung der Nachhallzeit im 125-Hz- und 250-Hz-Oktavband zusammen mit einer breiten Palette der anderen entscheidenden Parameter, einschließlich breitbandiger Nachhallzeit, Raumvolumen und Raumform, untersuchte. Außerdem betrachtete keine Studie gleichzeitig die Wirkung dieser Parameter auf einkanalige Maße (wie STI), binaurale Maße (wie 'Binaural Benefit') und deren Kombination.
In dieser Masterarbeit wurden Eingangsvariablen, u. a. vier breitbandige Nachhallzeiten (0,5, 1, 2 und 4 *sec*), vier typische Raumformen (Fächer, Hufeisen, Schuhbox und Weinberg) in vier Volumina (250, 1000, 4000 und 16000 m^3) skaliert und zusammen mit RT-Änderungen in sieben Schritten bei 125-Hz- und 250-Hz-Oktavbändern untersucht. Ergebnisse von raumakustischen Simulationen, die auf einer Kombination von Strahlverfolgung (Englisch: Ray Tracing) und Spiegelschallquellen basieren, wurden in Sprachverständlichkeitsalgorithmen eingespeist, um die Ausgangsvariablen 'STI', 'Binaural Benefit' und 'Binaural STI' (nach der in dieser Studie vorgestellten Methode) zu berechnen.
Schließlich wurde die Interaktion der Eingang- und Ausgangsvariablen mit Hilfe eines allgemeinen linearen Modells (Englisch: Generalized Linear Model (GLM)) statistisch analysiert.
Die Ergebnisse zeigen, dass die Änderung der Nachhallzeit im 125-Hz- oder 250-Hz-Oktavband eine statistisch signifikante Auswirkung auf die 'Binaural STI' hat. Genauer erklärt, führt eine Erhöhung der Nachhallzeit im 125-Hz- oder 250-Hz-Oktavband (mit einem Skalierungsfaktor von 0,7 bis 1,3) zu einer durchschnittlichen Abnahme der 'Binaural STI' von 0,01 bzw. 0,02. Diese Ergebnisse legen nahe, Maßnahmen zu ergreifen, um die Nachhallzeit bei tiefen Frequenzen zu reduzieren, neben der RT-Reduzierung in dem schon bekannten Frequenzbereich von 500 Hz - 4 kHz, um die Sprachverständlichkeit zu verbessern.

Contents

List of Figures

1 Introduction

There are many rooms designed for the purpose of music performance or speech communication or in some cases for both of them. These rooms are designed in a wide range of sizes which can vary from a small classroom or meeting room to large concert halls, auditoria or theatres. In this thesis, the influence of low-frequency reverberation time on the speech intelligibility is investigated.
Good speech intelligibility plays an important role in different room acoustical applications. For instance, some studies have shown that the learning process is slower in classroom environments that have higher reverberation times and as a consequence of that lower speech intelligibility (Lubman and Sutherland, 2007, [1]).
Generally, there is a tendency to neglect the effect of room acoustical parameters (specially for speech) including reverberation time at low frequencies. Some typical arguments for this are the low sensitivity of human's auditory system and low excitation of the room by speech in low frequencies. There are, however, a number of studies which stress the significant effect of low frequency reverberation time on speech intelligibility (See chapter 2), most of which emphasize on the fact that high frequency details of a signal can be masked by the low frequency part (Fuchs, 2017, [2]).
The investigations of this contribution are done based on performing room acoustical simulations followed by assessment of speech intelligibility using speech intelligibility prediction algorithms in different scenarios. The modeling and simulation tools and the algorithms are introduced in chapter 3. In this chapter, a number of fundamental technical terms are introduced which are necessary for a comprehensive understanding of the concepts presented in the next chapters.

1.1 Room Acoustical Parameters

1.1.1 Reverberation Time T

The reverberation time T is the traditional objective measure of *reverberance* in a room invented by *W.C. Sabine*. It is defined as the time it takes for the sound level in a room to decrease by 60 dB after a continuous sound source has been shut off (Rossing, 2007, [3]). This sound pressure level drop corresponds approximately to the dynamic range of a large orchestra (Sabine, 1923, [4]). Since this dynamic range is hardly reachable, the reverberation time is in practice determined by measuring a range of $-5\ dB$ up to $-35\ dB$. Then the doubled value is referred to as $T30$ (Ahnert and Tennhardt, 2008, [5]).
Typical values of reverberation time vary from 0.3 *sec* (living rooms) up to 10 *sec* (large churches or reverberation chambers). Most large halls have a reverberation time of 0.7 *sec* up to 2 *sec* (Kuttruff, 2001, [6]).
From the subjective point of view, the listener can follow the decay process only until the noise level in the room is reached. Therefore the subjectively perceived reverberation depends, in addition to the reverberation time, on the excitation and the noise level. This subjective perception is specially at low sound levels and during running programs rather consistent with the so-called Early Decay Time than with the reverberation time (Ahnert and Tennhardt, 2008, [5]).

Statistical Calculation of the Reverberation Time

Sabine Method

The Sabine formula in 22° C is shown in the equations 1.1 and 1.2.

$$T60 = 0.161\frac{V}{A}\ \ [sec]. \tag{1.1}$$

$$A = 4mV + \alpha.S\ \ [m^2]. \tag{1.2}$$

The variables of the equations above are listed in the following:

V: Room Volume [m^3], S: Room Surface Area [m^2]
α: The Average Diffuse-Field Absorption Coefficient of the Room Surfaces
m: Energy Air Absorption Constant [$\frac{1}{m}$]

The Sabine equation assumes a diffuse sound field in the room where a sound wave encounters surfaces one after another, as it travels around the room. Furthermore, it assumes that the absorbing power in the room is almost uniformly distributed so that the resulting reverberation time of the formula is nearly independent of the room shape (Hodgson, 1993, [7] and Beranek, 2006, [8]).

Eyring Method

The statistical formula for calculation of the reverberation time developed by Eyring assumes the absorbing power in the room to be nearly uniformly distributed over all the surfaces. The other assumption is that the sound field is almost diffuse so that the values of the reverberation time are

almost independent of a room's shape (Beranek, 2006, [8]). The Eyring formula is shown in the equation 1.3 (Ahnert and Tennhardt, 2018, [5]).

$$T60 = 0.163 \frac{V}{-Ln(1-\alpha)S_{total} + 4mV} \quad [sec]. \tag{1.3}$$

The variables of the equations above are listed in the following:
$\alpha = \frac{A_{total}}{S_{total}}$
A_{total}: Total Absorption Surface
S_{total}: Total Room Surface Area
m: Energy Air Absorption Constant $[\frac{1}{m}]$

$$A_{total} = \sum_n \alpha_n . S_n + \sum_k A_k \quad [m^2]. \tag{1.4}$$

Where α_n denotes the absorption coefficients of the surfaces S_n and A_k shows the absorption of parts which do not make a surface, such as audience or items of furniture.
For small absorption values (α<0,25), the logarithm in the equation 1.3 can be approximated by a line and this equation returns to the Sabine formula (equations 1.1 and 1.2). The relation between the reverberation time $T60$, room volume V and the equivalent sound absorbing surface A_{total} is shown in the figure 1.1 (Ahnert and Tennhardt, 2018, [5]).

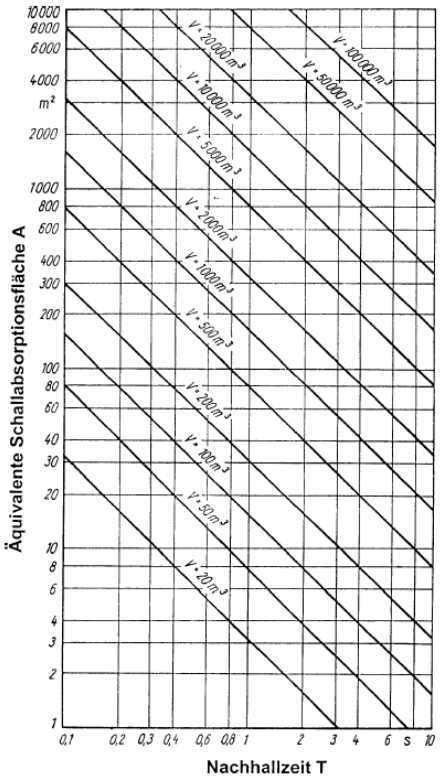


Figure 1.1: The relation between reverberation time T (hor. axis), equivalent sound absorbing surface A and room volume V, (after Ahnert and Tennhardt, 2018)

1.1.2 Early Decay Time (EDT)

The decay rate of the sound can also be defined by the parameter Early Decay Time as proposed by Jordan (1970, [9]). This is the time in which the first 10 dB drop of the decay process occurs, multiplied by a factor 6. This could be larger or shorter than the reverberation time defined by Sabine. Moreover, recent studies have shown that the subjectively perceived reverberance in a room is closely correlated with EDT (Kuttruff , 2001, [10]).

1.1.3 Sound Strength

The parameter sound strength G shows the influence of the room on the perceived loudness. One method to measure this parameter is measuring the difference in dB of a continuous, calibrated sound source in the room and the level of the same source in 10 m distance measured in an anechoic room. Another method is measuring the total energy of the impulse response and that of the direct sound in 10 m distance from the source (Rossing, 2007, [3]). Equation 1.5 shows the formula for the calculation of G where $g(t)$, $g_{10m}(t)$ and t_{dir} are impulse response, impulse response in 10 m distance and the duration of direct sound, respectively.

$$G = 10 \log \frac{\int_0^\infty g(t)^2 dt}{\int_0^{t_{dir}} g_{10m}(t)^2 dt} \; dB. \tag{1.5}$$

1.1.4 Bass Ratio (BR)

The bass ratio (BR) is the proportion of the reverberation time in the frequency range of 125Hz and 250 Hz to that of 500 Hz and 1000 Hz, the formula of which is shown in equation 1.6.

$$BR = \frac{T_{125} + T_{250}}{T_{500} + T_{1000}} \tag{1.6}$$

1.1.5 Treble Ratio (TR)

Similar to the bass ratio, treble ratio (TR) can be formed as shown in the equation 1.7. BR and TR can also be calculated based on G (see 1.5) rather than T values. (Rossing, 2007, [3]).

$$TR = \frac{T_{2000} + T_{4000}}{T_{500} + T_{1000}} \tag{1.7}$$

1.1.6 Speech Intelligibility

Before explaining the concept of speech intelligibility, it is instructive to explain the fundamental characteristics of speech as a sound source which is done in the following section.

Speech as a Sound Source

When the vocal cords vibrate, they contact each other and produce pulses of air pressure in a repeating rhythm with a frequency of 70-250 Hz (in some literature 50-350 Hz, Kuttruff, 2001, [11]). This rapid rhythm yields a sound in the air having the vibration rate as its fundamental frequency, which is heard as a pitched tone. Furthermore, the mechanical nature of the vocal cords produces a series of harmonic frequencies which are integer multiple of the lowest frequency (Fulop, 2011, [12]).
A speech sound can generally be described as created from a sound sound source the output of which is modified by the vocal tract which can be seen as a resonating chamber or resonant filter. Vowels and many other voiced sounds are mainly produced by vocal cords as a source, and the phonation output is then filtered through the prominent vocal tract resonances called formants (Fulop, 2011, [13]).
The total frequency range of conversational speech is shown in figure 1.2 (Kuttruff, 2001, [11]).

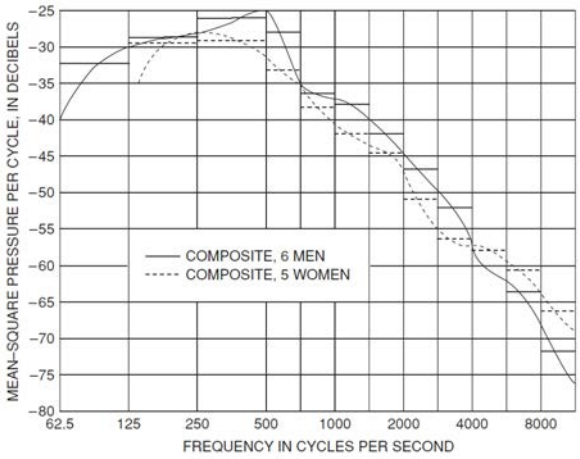


Figure 1.2: Long-time power density spectrum for continuous speech 30 cm from the mouth

The high-frequency content is mostly due to consonants including fricatives like «s» and «f» or pulsives like «p» and «t». Consonants are very important for the speech intelligibility, therefore, any system designed for speech transmission (rooms and halls can also be viewed as such systems) should transmit the high frequencies with good fidelity. The low frequency content is less important since the auditory system is able to reconstruct it if the periodic signal is rich in higher harmonics (Kuttruff, 2001, [11]).

Definition of Speech Intelligibility

Speech recognition is the end product (output) of a complex communication channel whose input is the message conceived by the talker (Pavlovic, 1987, [14]). Speech intelligibility is the probability of correct recognition of words (sounds having meaning), (Allen, 1994, [15]) at the receiver's (listener's position) in a transmission line (a room or an enclosure can be also considered as a transmission line between a talker and a listener).
Speech intelligibility depends on the relative strength of early and late reflections. Studies have shown that reflection arriving within 40 ms after direct sound (early reflections) are not perceived as echo but play a role in supporting the direct sound and improving the intelligibility. These findings led to definition of the parameter *Deutlichkeit* $D50$ (See the section 1.1.8) and in parallel *Clarity* $C50$ (See the section 1.1.8) (Nijs and Rychtáriková, 2011, [16]).
Generally, there are two approaches for assessing the speech intelligibility. The first one is based on using listening tests and the second one is based on signal measurements or acoustical simulations, both of which will be introduced in the following.

1.1.7 Speech Intelligibility Measures Based on Listening Tests

Speech Reception Threshold

The Speech Reception Threshold (SRT) is usually defined as the level at which the psychometric function intersects the 50% intelligibility line (Psychometric function gives the probability of getting a correct answer to a word as a function of the sound level), Hagerman (1979, [17]). In other words, it is the signal-to-noise ratio needed to achieve 50% sentence intelligibility (Rherbergen and Versfeld, 2005, [18]).

Subjective Syllable intelligibility V for Speech

This procedure for measuring the speech intelligibility measures the detection rate of clearly-spoken logatomes based on a frequency dictionary and a speech-related distribution of phonemes. Logatomes are monosyllabic constant-vowel-constant groups which make no recognisable or inferable sense, so that no logical estimation of not correctly understood logatomes during the test is possible (Ahnert and Tennhardt, 2008, [5]).
For each test, 200-1000 logatomes should be used. The amount of correctly understood letter sequences results in syllable intelligibility V in percent. Values over 70% imply on a good speech intelligibility, whereas values under 35% indicate a poor speech intelligibility (Ahnert and Tennhardt 2008, [5]).

1.1.8 Speech Intelligibility Measures Based on Measurements or Simulations

Speech Clarity C_{50}

C_{50} (Deutlichkeitsmaß) is a measure of intelligibility of speech or vocals[1]. It is based on the assumption that the sound energy portion within 50 ms after the arrival of the direct sound supports the clarity of the speech, while the later portion deteriorates it. Generally, C_{50} is calculated (See the equation 1.9) in four octave bands from 500 to 4000 Hz (Ahnert and Tennhardt, 2008, [5]).

$$C_{50} = 10 \log \frac{E_{50}}{E_{\infty} - E_{50}} \; dB. \tag{1.9}$$

A sound transmitter with a human speech directivity is used for the measurement of C_{50}. A value of $C_{50} = -2 \; dB$ is considered as the lower limit of a good speech or text intelligibility at which syllable intelligibility and text intelligibility are not under 80% and 95%, respectively (As a result of the context, text intelligibility is higher than syllable intelligibility). The limit of perception of the difference of this parameter is $\Delta C_{50} \approx \pm 2.5 \; dB$ (Ahnert and Tennhardt, 2008, [5]). It is also noteworthy, that the energy calculation in the equation 1.9 is done based impulse response measurement. A more detailed

[1] There is also a similar measure called Clarity Index C_{80} (originally, Klarheitsmaß) introduced by Reichardt et al. (1975, [19]) which is used to determine the transparency of music in concert halls. It is calculated by the equation 1.8, where $g(t)$ denotes the impulse response (Kuttruff , 2001, [10]).

$$C_{80} = 10 \log \frac{\int_0^{80ms} g(t)^2 dt}{\int_{80ms}^{\infty} g(t)^2 dt} \; dB. \tag{1.8}$$

The higher limit of time (80 ms) in comparison to C_{50} implies the fact that the reflections are less detectable in music compared to the speech. The value $C_{80} = 0 \; dB$ has been proven to be appropriate even for fast music passages. According to the investigations on concert halls in Europe and the USA, its typical range is between $-5 \; dB$ and $+3 \; dB$ (Kuttruff , 2001, [10]). The just noticeable difference of this measure is about $\Delta C_{80} = \pm 3 \; dB$ (Höhne and Schroth, 1995, [20]).

representation of equation 1.9 is shown in the equation 1.10 (Kuttruff , 2001, [10]) where $g(t)$ denotes the impulse response.

$$C_{50} = 10 \log \frac{\int_0^{50ms} g(t)^2 dt}{\int_{50ms}^{\infty} g(t)^2 dt} \; dB. \tag{1.10}$$

One limitation of C_{50} is the assumption of the absence of background noise. Bradley, 1986, [21] introduced another parameter U_{50} where not only the effect of the reverberation time but also that of the background noise is taken into account (Nijs and Rychtáriková, 2011, [16]).

Definition D_{50}

Another parameter as a measure of clarity of sound is called *definition* (originally, Deutlichkeit) developed by Thiele, 1953, [22] which is shown by D_{50}. It considers the proportion of the energy within the first 50 ms and to the whole energy (equation 1.11, where $g(t)$ denotes the impulse response) which is given in percent (Kuttruff , 2001, [10]).

$$D_{50} = \frac{\int_0^{50ms} g(t)^2 dt}{\int_{0ms}^{\infty} g(t)^2 dt} \; 100\%. \tag{1.11}$$

In order to give a better understanding of the effect of D_{50} on the speech intelligibility, the relation between syllable intelligibility and 'Definition' is shown in the figure 1.3 which shows a good correlation between these measures, (Boré, 1956, [23] and Kuttruff, 2001, [10]).

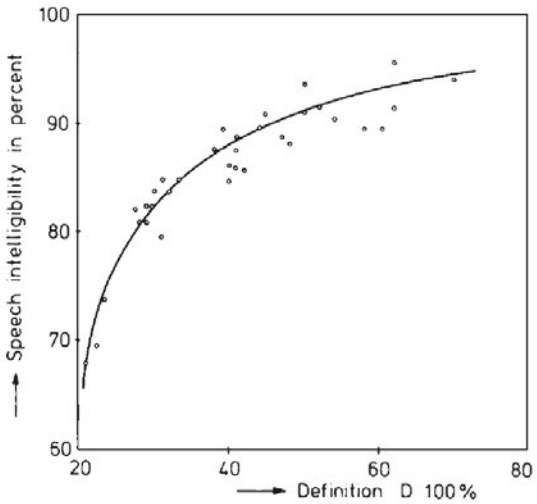


Figure 1.3: Relation between syllable intelligibility and 'Definition' after Boré, 1956

Mathematically, the relation between D_{50} and C_{50} can be shown by the equation 1.12.

$$C_{50} = 10 \log \frac{D_{50}}{1 - D_{50}} \; dB. \tag{1.12}$$

Al_{cons}	**Subjective evaluation**
$\leq 3\%$	ideal intelligibility
$3-8\%$	very good intelligibility
$8-11\%$	good intelligibility
$>11\%$	satisfactory intelligibility
$>20\%$	useless intelligibility

Table 1.1: Subjective evaluation of Al_{cons} values after Ahnert and Tennhardt, 2008

Articulation Loss

The articulation loss is another measure of evaluating the speech intelligibility in a room which is based on the investigations of Peutz, 1971, [24]. Based on an empirical formula shown in 1.13, the articulation loss of the pronounced consonants (Al_{cons}) can be calculated as a function of the following parameters (Ahnert and Tennhardt, 2008, [5].

T: Reverberation time
r_{QH}: Distance from the sound source
r_H: Critical distance

$$Al_{cons} \approx 0.625 \, \frac{r_{QH}}{r_H} \, T\%. \tag{1.13}$$

Al_{cons} can also be calculated based on the room impulse response. In order to do that, the amount of the energy up to 35 ms and the rest of the energy should be calculated and then based on the formula 1.14, Al_{cons} can be computed (Ahnert and Tennhardt, 2008, [5]).

$$Al_{cons} \approx 0.625 \, \frac{E_\infty - E_{35}}{E_{35}} \, T\%. \tag{1.14}$$

Al_{cons} values are usually reported in 1000 or 2000 Hz octave band. Table 1.1 gives a good overall understanding of the values of Al_{cons} (after Ahnert and Tennhardt, 2008, [5]).

Speech Intelligibility Index

Speech intelligibility index (SII) is one of the other room acoustical parameters used to asses the speech intelligibility. In the SII model, the average amount of speech information which is available to the listener is calculated. To do this calculation, the long-term averaged speech and noise spectrum are used in the model as input. The spectrum levels are defined as levels at the eardrum of the listener (in dB/Hz). Then the spectrum levels are calculated separately for speech and noise, computed in individual frequency bands (These can be octave, third-octave or critical bands).
In the next step, correction factors are considered to incorporate the phenomena such as upward spread of masking, inaudibility due to the auditory threshold for pure tones and distortion caused by very high speech or noise levels. Consequently, the signal-to-noise ratio is calculated for speech and noise in each frequency band and is multiplied with the band-importance function (As an example, figure 1.4 shows

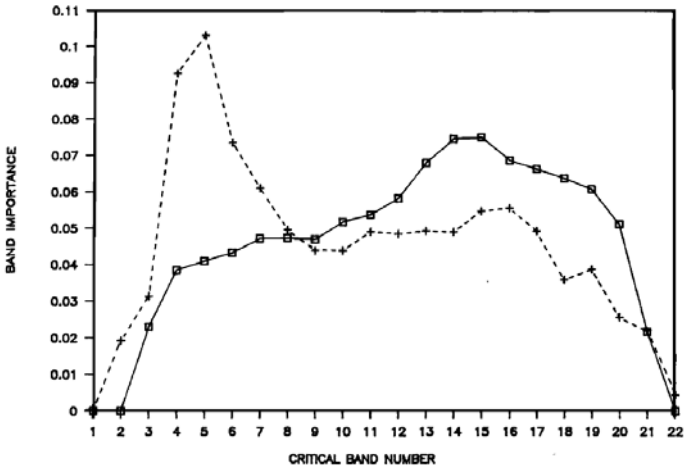


Figure 1.4: Importance functions for nonsense syllables (solid line) and for easy-running speech (dashed line), after Pavlovic, 1987

importance functions for nonsense syllables after Pavlovic, 1987, [14] which can be dependent on the type of the speech material. At the last stage, these values are added resulting in the SII value which shows the amount of speech information available to the listener (Rherbergen and Versfeld, 2005, [18]). For the case of a stationary noise masker and normal-hearing listeners, the SII is closely related to the average intelligibility (Pavlovic, 1987, [14]).

Speech Transmission Index

One of the well-established measures of predicting the speech intelligibility at the receiver's (listener's) position is the Speech Transmission Index (STI) introduced by Houtgast and Steeneken, 1971, [25]. Their fundamental idea was based on the fact that the effect of a transmission channel (e.g. a room or an enclosure) on intelligibility is strongly related to the degree to which the spectral differences on the transmitter's (talker's) side are preserved in the receiver's side. On the basis of this idea, they suggested an artificial test signal which includes a standard spectral difference (introduced at the talker's side) and also an analysis procedure to be applied on the signal of the listener's side to evaluate the degree of preservation of the spectral difference and consequently calculating the STI. Later, they introduced the concept of Modulation Transfer Function which was used by them as new basis of STI calculation.
In order to calculate the STI, the decrease in the signal modulation between the sound source's and listener's position is measured in octave-band frequencies from 125 Hz to 8000 Hz (Ahnert and Tennhardt, 2018, [5]).
The main idea of the process of calculating the SIT is that not only reverberation and noise but also any signal changes in the path between the source and receiver reduce the speech intelligibility. In order to determine this effect, modulation transfer function or MTF (See section 1.1.8) is applied. Based on the formula 1.17, 14 modulation frequencies from 0,63 Hz to 12,4 Hz[2] are used to calculate the MTF's. Consequently, a weighted modulation transfer function (WMTF) is applied to the calculated MTF's to achieve a higher correlation with the speech intelligibility. Then the modulation transfer functions are divided into seven frequency bands, each of which is applied to the modulation frequency resulting in 14x7 = 98 modulation reduction factors (m_i). The effective signal-to-noise ratios X_i's can be then calculated based on modulation reduction factors, as shown in the formula 1.15 after Ahnert and Tennhardt, 2018, [5].

[2] 0.63, 0.8, 1, 1.25, 1.6, 2, 2.5, 3.15, 4, 5, 6.3, 8, 10, 12.4 Hz

Syllable intelligibility	STI value
poor	0-0.3
weak	0.3-0.45
appropriate	0.45-0.6
good	0.6-0.75
excellent	0.75-1

Table 1.2: STI values evaluation, after Ahnert and Tennhardt, 2018

$$X_i = 10 \log \frac{m_i}{1 - m_i} \; dB. \tag{1.15}$$

Consequently, these values are averaged and Modulation Transfer Indices (MTI, see the formula 1.16 after Ahnert and Tennhardt, 2018, [5]) are calculated in octave bands.

$$MTI = \frac{X_{average} + 15}{30} \; dB. \tag{1.16}$$

Finally, the STI values are calculated after a frequency weighting in seven octave bands which is done in some cases for male and female speakers, separately. The table 1.2 shows some STI values with their respective evaluation, after Ahnert and Tennhardt, 2018, [5].

Modulation Transfer Function

The concept of Modulation Transfer Function (MTF) was introduced as a measure in room acoustics for assessing the effect of an enclosure on speech intelligibility (Houtgast and Steeneken, 1973, [26]).
Based on the contribution of Houtgast and Steeneken, 1973, [26], When a signal is transmitted through an enclosure, the envelope of the received signal is a smoothed version of the transmitted one. As if the enclosure would function as a low-pass filter. This filter characteristic (Modulation Transfer Function or MTF) is a feature of the enclosure. From MTF, a single value can be calculated which highly correlates with the speech intelligibility. Therefore, MTF can be used as a convenient predictor of Speech Intelligibility.

MTF can be mathematically described, based on the definition of T. Houtgast et al., 1980, [27], as the Fourier Transform of the squared Impulse Response of the room. Ahnert and Tennhardt, 2018, [5] have introduced the formula 1.17 for the MTF.

$$m(F) = \frac{1}{\sqrt{1 + (2\pi F \frac{T}{13.8})^2}} \cdot \frac{1}{1 + 10^{-\frac{S/N}{10dB}}}. \tag{1.17}$$

The variables in the formula 1.17 are defined as follows:
F: Modulation frequency in Hz
T: Reverberation time in sec
S/N: Signal-to-noise ratio in dB

T. Houtgast et al., 1980, [27] presented a model of calculating Speech Transmission Index (STI) based on MTF using assumptions of statistical acoustics. As shown in the figure 1.5, they took as input the volume

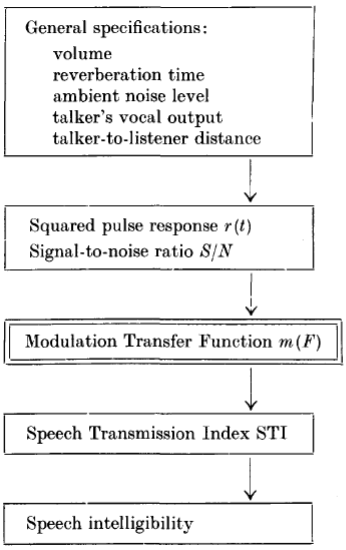


Figure 1.5: General scheme of calculating the STI based on MTF, after T. Houtgast et al., 1980

and the reverberation time of the room, as well as the ambient noise level, talker's vocal output and talker-to-listener distance. Then, based on an assumption of a purely exponential reverberation process, they calculated the pulse response (r(t)) and consequently the MTF. They concluded, depending on the nature of interfering noise, a reverberation time of less than 0.8 s to be an optimum value regarding the speech intelligibility.
In another approach to calculate STI from MTF, Plomp et al., 1980, [28] performed some calculations based on geometrical acoustics. They made some simplifying assumptions, such as perfectly reflecting planes and non-directivity of the sound source to calculate STI. Their results seemed to be more accurate in terms of predicting STI as a function of distance compared to the statistical-acoustics approach of T. Houtgast et al., 1980, [27].

Speech Interference Level

Another measure to evaluate the quality of speech intelligibility used in the international community is based on speech interference level (SIL). First, SIL is defined as the arithmetic average of sound pressure levels in the ambient environment at the 500, 1000, 2000, and 4000 Hz octave-band frequencies. It can be also calculated by subtracting 8 dB from the overall dBA ambient level. Then, the difference between the SIL and the A-weighted sound pressure level at the listener's location is used as a measure to rate the speech intelligibility (Rossing et. al, 2007, [29]). Table 1.3 shows general characteristics of evaluating speech intelligibility against this signal-to-noise ratio, after Rossing et. al, 2007, [29].
These calculations are based on an assumption of steady-state noise. However, attention must be paid to the fact that this effect is not equal for steady-state and fluctuating noise. More precisely, in almost all cases when normal-hearing listeners are concerned, listeners perform better in conditions with fluctuating noise compared to those with stationary noise of the same RMS level (Rhebergen and Versfeld, 2005, [18]). This finding has been phenomenologically explained by referring to the fact that listeners are able to catch "glimpses of speech during the noise silence periods". Some other studies reveal that non-linear property of the basilar membrane enables the listener to benefit from an increased gain during the silence period (Rhebergen and Versfeld, 2005, [18]).

Signal-to-noise ratio at listener's position (dBA-SIL)	Speech intelligibility rating
<-6	Insufficient
-6 to -3	Unsatisfactory
-3 to 0	Sufficient
0 to 6	Satisfactory
6 to 12	Good
12 to 18	Very good
>18	Excellent

Table 1.3: speech intelligibility ratings based on SIL, after Rossing et. al, 2007

1.2 Room Acoustical Simulation Methods

1.2.1 Geometrical acoustics

In geometrical acoustics, the concept of a wave is replaced by the concept of a sound ray, which is an idealisation. Similar to the geometrical optics, a sound ray is meant to be a small portion of a spherical wave with vanishing aperture which originates from a certain point (Kuttruff, 2001, [30]).
This approach requires that the wavelength of sound be relatively small in comparison to the geometrical dimensions of the room and large relative to the roughness and curvature of the room's walls (Schröder and Lentz, 2006, [31]).

Image Source Model

One of the models used in the field of geometrical acoustics is the so-called image source model the fundamentals of which are described in the following.

Fundamentals of the model

When a sound ray hits a wall, it gets reflected based on the well-known rule in optics. It gets reflected so that the reflected ray is in the plane including the incident ray and the normal to the surface where the angle of reflection is equal to the angle of reflection (Kuttruff, 2001, [30]). This is shown in the figure 1.6 where the vector are defined as follows:
u: incident ray
u': reflected ray
n: normal vector to the wall

If we assume boundaries made of plane and uniform walls then we can use the notion of image sources. As shown in the figure 1.7 (after Kuttruff, 2001, [30]), the exact incident ray direction originating from point A which leads to a reflection direction passing through the point B can be found using image source technique. The virtual source A' is located behind the wall, on the line perpendicular to the wall and placed at the same distance from it as the original source A. This shows the fundamental idea of the image source method. Based on this idea, one can assume each ray reflected from the one as originating from the point A'. By doing this, the wall can be neglected as its effect can now be considered by the image source A'.

One of the fundamental studies of this method was done by Allen and Berkly, 1978, [32] where they calculated the exact impulse response of a small rectangular room using the image source method. They

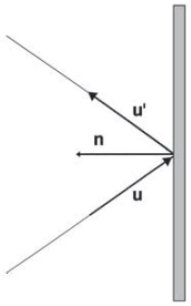


Figure 1.6: Specular reflection after Kuttruff, 2001

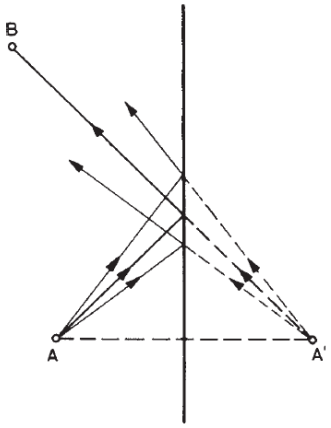


Figure 1.7: Construction of an image source after Kuttruff, 2001

began their calculation with the pressure field of a single-frequency point source in free field the formula of which is shown in the equation 1.18.

$$P(\omega,X,X') = \frac{\exp(i\omega(\frac{R}{c} - t))}{4\pi R} \tag{1.18}$$

The variables in 1.18 are defined as follows:

P = pressure,
$\omega = 2\pi f$
f = frequency,
t = time,
$R = |X - X'|$
X = vector of talker's location (x,y,z)
X' = vector of microphone's location (x',y',z')
$i = \sqrt{-1}$
c= speed of sound

Consequently, they incorporated the effect of the boundary condition (rigid wall) by placing an image symmetrical source on the other side of the wall which yields the equation 1.19.

$$P(\omega,X,X') = [\frac{\exp(i(\frac{\omega}{c})R_+)}{4\pi R_+} + \frac{\exp(i(\frac{\omega}{c})R_-)}{4\pi R_-}] \exp(-i\omega t). \tag{1.19}$$

R_+ and R_- in the equation 1.19 are defined as follows (The wall is placed at $x = 0$.):

$R_-^2 = (x - x')^2 + (y - y')^2 + (z - z')^2$
$R_+^2 = (x + x')^2 + (y - y')^2 + (z - z')^2$

Finally, considering all six rigid walls, calculating the frequency response and taking the Fourier transformation, they calculated the impulse response, shown in equation 1.20.

$$P(t,X,X') = \sum_{p=1}^{8} \sum_{r=-\infty}^{\infty} \frac{\delta[t - (|R_p + R_r|/c)]}{4\pi|R_p + R_r|} \tag{1.20}$$

The variables in equation 1.20 are defined as follows:

$R_p = (x \pm x', y \pm y', z \pm z')$,
$R_r = 2(nL_x, lL_y, mL_z)$

Equation 1.20 shows the room impulse response function (time domain Green's function) assuming rigid walls for a point source at $X = (x,y,z)$ and receiver at $X' = (x',y',z')$. The two-dimensional visualization of the equation 1.20 is shown in figure 1.8 where the source and its images are depicted (after Allen and Berkly, 1978, [32]).
It can bee seen that the image source method enables us to represent a boundary-value problem in terms of an equivalent problem containing multiple sources without any boundaries (Borish, 1984, [33]).

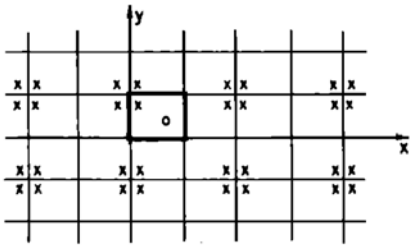


Figure 1.8: A 2D-representation of a room with a source and its image sources (The solid box represents the main room), after Allen and Berkly, 1978

Moreover, it should be taken into consideration that the number of image sources grows rapidly with the order of the reflections. If we assume an enclosure with N-plane walls, there will be $N(N-1)$ second-order image sources (Kuttruff, 2001, [30]). Consequently, the total number of the image sources up to a reflection order of i_0 will be as shown in the equation 1.21 (after Kuttruff, 2001, [30]).

$$N(i_0) = N\frac{(N-1)^{i_0} - 1}{N - 2} \tag{1.21}$$

While the contribution of Allen and Berkly, 1978, [32] was constrained to a rectangular room, one extension of the image source model to arbitrary polyhedra is provided by Borish, 1984, [33].

Ray Tracing

Ray tracing is a method of geometrical representation of the spherical sound wave propagation in a closed space (Kulowski, 1985, [34]). In this technique, the energy of a single spherical wave is divided into elements which are assumed to be discrete objects where each ray is traced until it reaches a certain negligible value (Kulowski, 1985, [34]).
This approach is simply shown in the figure 1.9 (after Kuttruff, 2001, [35]). In this method, a big number of particles are sent out of a sound source to different directions at a given time point.

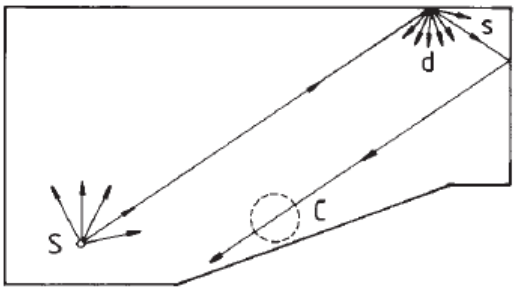


Figure 1.9: Principle of digital ray tracing. S = sound source, C = counting sphere, s = specular reflection and d = diffuse reflection, (After Kuttruff, 2001)

Then straight travelling path of each ray is pursued until it hits a wall. Consequently, it will be either specularly or diffusely reflected. In the case of specular reflection, the reflection angle is obtained based on the law of geometrical reflection and in the case of diffuse reflection, two random numbers are generated to determine the azimuth and angle of the diffuse reflection. On the other hand, the effect of the acoustical absorption of the surface materials are taken into account so that either the energy of the ray is reduced by the factor of $1 - \alpha$, or α is used as a probability factor to decide whether the particles goes on or should be absorbed (α is the absorption coefficient of the surface.).
The results of this process are gathered by counting surfaces or volumes. It means, each time a ray hits one of these counter, its energy and arrival time are stored. Then the energy of the rays (or particles) arriving within a defined time interval (bin) is added. The resulting diagram is called a histogram (See figure 1.10). Histogram can be used as a short-time averaged impulse response. These are also used in this thesis to extract the values of the reverberation time based on the simulation results. It is important to consider the effect of the length of the time interval. If it is to long, then the resulting histogram will be a rough approximation of the real-life impulse response. On the contrary, if it is too short, it will contain a lot of random time alternations. As a reasonable value, a time interval of 5-10 ms is recommended, as it also corresponds approximately the resolution of human's auditory system (Kuttruff, 2001, Room Acoustics, [35]). On the other hand, Allen and Berkly, 1979, [32] propose that the choice of the time interval be governed by the application. It means, if the speech is going to be studied in small rooms, a time interval of 0.1 *ms* should be used, while if reverber-

ation times of large enclosures are being studied higher time intervals (lower sampling rates) can be useful.

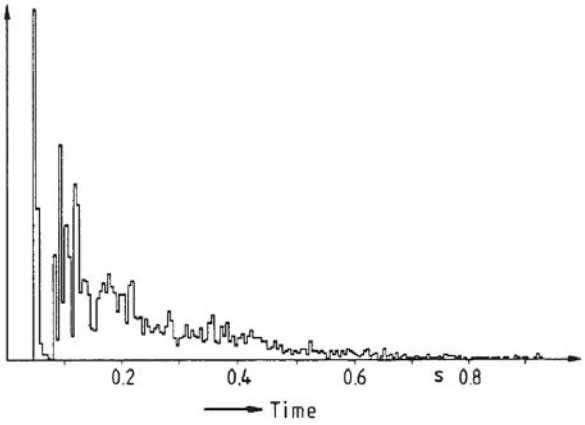


Figure 1.10: Time histogram of received particle energy (the interval width is 5 ms), (Kuttruff, 2001)

1.2.2 Wave theory

In this section, a very short introduction to wave theory is given (It is kept short, as this thesis does not use the methods based on wave theory, but those based on geometrical acoustics).
The Wave-based method, as the name suggests, are based on solving the wave equation 1.22 (Kuttruff, 2001, [11]).

$$c^2 \Delta p = \frac{\partial^2 p}{\partial t^2} \tag{1.22}$$

The variables in the equation 1.22 are defined as follows:
p: Sound pressure
t: Time
c: Sound velocity

The wave equation governs the propagation of sound waves in any lossless fluid and holds for sound pressure, density and temperature variations (Kuttruff, 2001, Room Acoustics, [11]).
The equation 1.22 can be written in a time-independent form (assuming a harmonic time law for pressure, particle velocity etc.) which yields the equation 1.23 known as the Helmholtz equation.

$$\Delta p + k^2 p = 0. \quad k = \frac{\omega}{c} \tag{1.23}$$

The new variables in the equation 1.23 are as follows:
k: Wave number
ω: Angular frequency
The wave equation results in non-zero solutions which satisfy the boundary conditions only for certain discrete values of k called *eigenvalues* . Each of these eigenvalues are related to a solution of the wave

equation representing a standing wave or a so-called «normal mode» of the room.
The rooms and concert halls which we deal with in every-day life have significant irregularities in shape resulting in complexity of formation of the boundary conditions. Moreover, the solution of the wave equation in such cases require intensive numerical calculations which limits the application of wave theory in typical rooms. However, this theory is the most reliable theory from a physical perspective and essential for more than superficial understanding of sound propagation in enclosures (Kuttruff, 2001, [6]).

1.3 Human's Auditory Perception

1.3.1 A-weighting

A-weighting is a common method of weighting the spectrum of sound, as introduced in American tentative standards for sound level meters (Z24.3-1936, 1936, [36]) for measurement of noise and other sounds, to account for the difference of Human's auditory system sensitivity at different frequency ranges.

1.3.2 Equal-Loudness Contours

In many cases, an assessment of the subjective perception of the loudness of a signal is desired by engineers and acousticians. One of the quite well-accepted methods of this assessment is based on using the equal-loudness contours to obtain the loudness level.
In order to obtain the loudness level of a given sound, a subject is asked to adjust the level of a 1000-Hz pure tone until it has the same loudness as the test sound. The level of the 1000-Hz tone that gives equal loudness is the loudness level or the test sound, measured in 'phons'. By definition, the loudness level of a 1000-Hz tone is equal to its sound pressure level in dB SPL (Moore, 2013, [37]).
Figure 1.12 shows the equal-loudness contours for binaural listening for loudness levels from 10 to 100 phones, according to Moore, 2013, [37].

1.3.3 Characteristic Frequency

Sounds of different frequencies produce maximum displacement at different places along the basilar membrane (Basilar membrane is a part of human's peripheral auditory system, which functions as a Fourier analyzer and decomposes the sounds into their component frequencies). The frequency which gives maximum response at a particular point on the basilar membrane is known as the characteristic frequency (CF) for that point. In other words, each point on the basilar membrane can be seen as a bandpass filter with a center frequency which corresponds to the CF. For instance, figure **??** shows the response of a point on the basilar membrane with $CF = 10\ kHz$ to short impulses at different sound pressure levels (Moore, 2013, [38]).

Excitation Patterns

Studies have shown that a single low-level excitation signal causes a response with a high spike rate in neurons which have critical frequencies close to the frequency of the excitation signal. These spike rates drop off in critical frequencies on both sides of the tone frequency. It is noteworthy, that this behaviour of the auditory system is more complicated in higher levels as a result of neural saturation. The distribution of these spikes as a function of critical frequency is sometimes called the *excitation pattern*. The excitation pattern is plotted as effective level in dB and as a function of critical frequency. Figure 1.13 shows the excitation pattern for a 1-kHz sinusoid from the responses of the auditory system of a cat (after Moore, 2013, [38]).

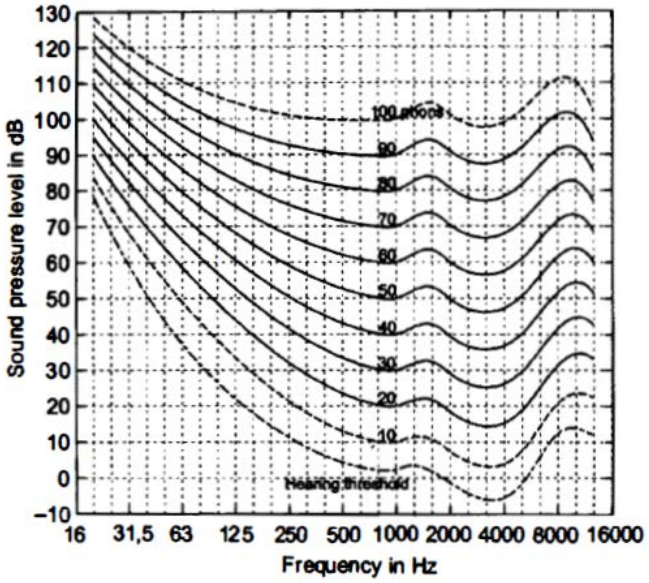


Figure 1.11: Equal-loudness contours for loudness levels from 10 to 100 phons for sounds presented binaurally from the frontal direction, after Moore, 2013 (The lowermost curve depicts the absolute hearing threshold). The curves for loudness levels of 10 and 100 phons are dashed, as they are based on interpolation and extrapolation, respectively.

1.3.4 Masking Patterns

Wegel and Lane, 1924, [39] published the first systematic paper regarding the phenomenon of masking of a pure tone as a result of the presence of another one (Moore, 2013, [38]). A graph that depicts this phenomenon, i.e. it shows the masked threshold as a function of the signal frequency is called *masking pattern*. A plot of masking patterns for a narrow-band 410 Hz noise is shown in the figure 1.14 after Moore, 2013, [38].

Masking Generally, masking occurs when the reception of a number of acoustic stimuli (targets) is degraded by other stimuli (maskers) (Durlach, 2006, [40]).

1.3.5 Upward Spread of Masking

A deep inspection of the figure 1.14 shows that the slope of the curves on both sides of the center frequency is not the same. It is shallower on the high frequency side and this becomes more noticeable at higher levels. More precisely, the amount of masking increases nonlinearly on the high-frequency side. This has been called the *upward spread of masking* (Moore, 2013, [38]).
J. M. Pickett, 1959, [41] has investigated the effect of upward spread of masking under low-frequency noise conditions and has shown the deteriorating effect of that on speech intelligibility specially in high noise levels.

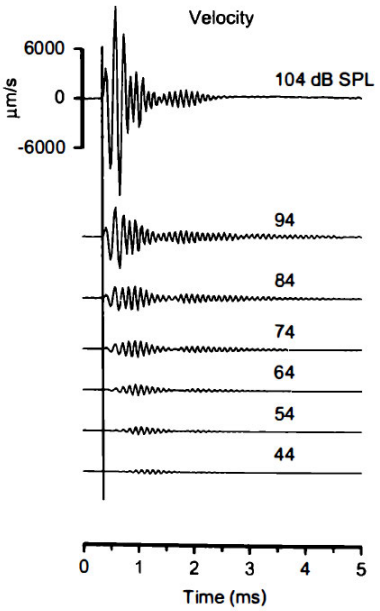


Figure 1.12: The response of a point on the basilar membrane with CF=10 kHz to short impulses (clicks) at various levels (after Moore, 2013)

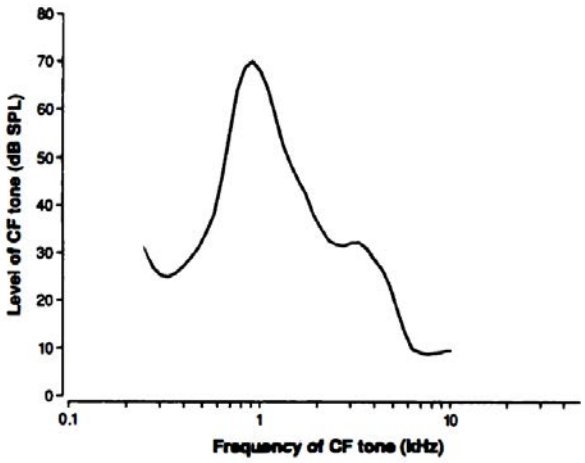


Figure 1.13: The excitation pattern for a 1-kHz sinusoid with a level of 70 dB SPL drawn based on the responses of single neurons in the auditory nerve of the cat (after Moore, 2013)

1.3.6 Better-Ear Listening

The Better-Ear rule suggests that the listeners are able to attend to the ear providing the best overall target-to-interferer ratio (Edmonds and Culling, 2006, [42]). This phenomenon is called better-ear listening.

Spatial Unmasking

Listeners often have the impression of being able to pick up any of a range of voices around them by focusing their attention on the appropriate direction (Cherry, 1953, [43]). Furthermore, Edmonds and Culling, 2006, [42] have shown that the listeners are able to obtain a better recognition of concurrent

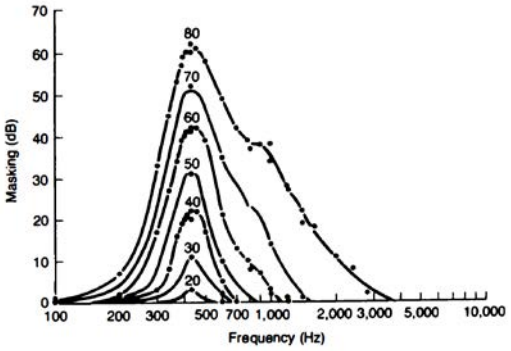


Figure 1.14: Masking patterns for a narrow-band noise with a center frequency of 410 Hz (after Moore, 2013)

voices when these voices come from different directions . This phenomenon is known as *spatial unmasking*. Spatial unmasking is in some literature also called Spatial Release form Masking (SRM). SRM takes place as a consequence of two phenomena, namely, better-ear listening and binaural unmasking. As these components are usually presented in dB, SRM is also presented in dB.

As the phenomena mentioned so far affect our perception of sound in the reality and they happen as a consequence of binaural listening, the author of this contribution is motivated to use algorithms of calculating the speech intelligibility which take into account the binaural effects. These algorithms will be discussed in chapter 3.

1.3.7 Binaural Unmasking

Hirsh, 1948, [44] discovered a phenomenon called binaural unmasking. In binaural unmasking, the auditory system combines the two signals of ears to improve the detection of the signal in the presence of noise. He found out when the interaural phase angles between the signals of the two ears are opposite for the tone (pure tone) and the noise, the binaural threshold is lower. It means, a lower sound pressure level is needed to detect the signal. Moreover, if these interaural phase angles are the same for the tone and the noise, the binaural threshold is higher than the monaural one. Hirsh also found out that this change of the masking threshold in the two cases is most marked at low frequencies and as the intensity of the masking noise is increased. Moreover, he concluded that the difference between binaural and monaural threshold is not frequency dependent when pure-tone thresholds are measured in quiet. However, if these thresholds are measured when a pure tone is presented against background noise, the binaural-monaural difference is frequency dependent.

1.3.8 Binaural Masking-Level Differences

The improvement of the masking threshold in case of binaural listening is known as Binaural Masking-Level Difference (BMLD). Some studies have used BMLD's for calculating speech intelligibility which will be discussed in chapter 3 and will be used as a basis of calculating the speech intelligibility in the master thesis.
In accordance with the results of Hirsh, 1948, [44], Licklider, 1948, [45] studied the influence of the

interaural phase differences of the signal and noise on speech intelligibility. He concluded that the speech intelligibility is higher when the speech signal is in phase in the two ears and the noise is out of phase or when the speech is out of phase and the noise is in phase (antiphasic condition) than when both the speech and the noise are either in or out of phase (homophasic condition). The intelligibility is in case of two independent noise sources (heterophasic condition) between that of the homophasic and heterophasic condition.
He also showed that the intelligibility in binaural homophasic condition is lower than in monaural listening and that the intelligibility in binaural antiphasic condition is better than that of monaural listening.

Model of Lavandier and Culling (2010)

Lavandier and Culling, 2010, [46] introduced a method of predicting the Speech Reception Threshold (SRT) based on a theory of binaural unmasking together with a model of better-ear-listening. Jelfs et. al, 2011, [47] introduced the *Cardiff binaural intelligibility model* (based on that of Lavandier and Culling) which is computationally more efficient and also shows that the model of Lavandier and Culling accurately predicts the effect of headshadow. Both of those models use a method which is shown in figure 1.15 (Figure from Jelfs et. al, 2011, [47]).

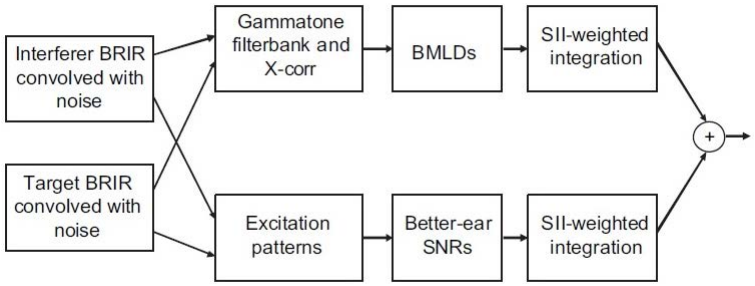


Figure 1.15: Schematic image of the method used in Lavandier and Culling, 2010 and Jelfs et. al, 2011

In the first path shown in the figure 1.15, the binaural advantage which is caused by binaural unmasking (by prediction of BMLD) (Licklider, 1948 [45] and Hirsh, 1948 [44]), is calculated. The sequence in the second path is used to calculate the benefit of better-ear listening.
In order to calculate the binaural unmasking, Lavandier and Culling, 2010, [46]) took, as the input of the model, the speech-shaped noise convolved by Binaural Room Impulse Response (BRIR) of the target and interferer to create reverberant target and interferer. In th first path of the figure 1.15, the reverberant target and interferer signals are filtered through a gammatone filterbank (Patterson et. al, 1987, [48]). Then, these filtered signals are divided into 320-ms sections for each left and right channel and then cross-correlated in order to calculate the 1. interaural coherence of the interferer, 2. interaural phase of the target and 3. interaural phase of the interferer (The cross-correlation was done by *Wave software suit* (Culling, 1996, [49])). Subsequently, based on the method and the formula given in (Culling et. al 2004, 2005 [50], [51]), and the three mentioned values, the BMLD was calculated. Finally, the broadband binaural advantage for speech is calculated based on an integration of the channels across the frequency using the Speech Intelligibility Index (SII) weighting method (ANSI, 1997, [52]).
The second path in the figure 1.15 shows the calculation process of the better-ear listening effect which is done by calculating the better-ear target-to-interferer ratio (TIR) at each frequency band. In order to calculate this, the cochlear excitation patterns (Moore and Glasberg, 1983, [53]) are computed for each

section of the target and interferer signal between 0 and 33.25 ERBs (0-10 kHz). Then the difference between the target and interferer excitation pattern (in dB) of each ear is assumed as the TIR. Then the better TIR between the left and right ear in each frequency is integrated over frequency using the SII weighting method. Finally, the values of binaural advantage and better-ear TIR are combined to obtain a single effective speech-to-noise ratio.
When the Speech Transmission Index (STI) is used to predict the speech intelligibility in binaural listening conditions, the intelligibility is underestimated (Van Wijngaarden and Drullman, 2008, [54]). For this reason, the effect of binaural listening is also taken into account in the calculations of this thesis, along with the STI method.

1.3.9 Equalization Cancellation Theory of Binaural Masking-Level Differences

According to the Equalization Cancellation (EC model) presented by Durlach, 1963, [55], when a binaural-masking stimulus is presented to a subject, subject's auditory system tries to eliminate the masker (masking component) by transforming the total signal in one ear relative to the signal in the other ear until the masking components in both ears are the same (the equalization process), and subsequently, subtracting the total signal in one ear from the total signal in the other ear.
Depending on the interaural relations of the masking signal compared to those of the target signal, if the EC process is performed with complete precision, the masking signal will be completely eliminated (Durlach, 1963, [55]).

2 State of the Art

2.1 Influence of low-frequency reverberation time on speech intelligibility

Although, there are relatively many studies investigating the effect of mid and high frequency reverberation time on the speech intelligibility, there are few studies which have dealt with the effect of low frequency reverberation time. Nevertheless, some of these studies are discussed in this section.
It is well established that reverberation time and background noise are two dominating factors[1] in speech intelligibility. Generally speaking, in order to improve the intelligibility, the reverberation time and background noise should be reduced.
One of the comprehensive studies regarding the predictors of the speech intelligibility comes back to the year 1986. Bradley, 1986, [21] performed acoustical measurements and speech intelligibility tests in five rooms with volumes from 362-20000 m^3 and 1-kHz reverberation time values from 0.8-3.8 *sec*. He obtained a wide range of acoustical measures from pulse recordings (using pistol shots) at 40 source-receiver combinations in the rooms, with 6-14 source-receiver positions in each room. At each point, he calculated certain predictors of speech intelligibility, including, steady-state signal-to-noise measures, measures derived from early/late-arriving sound ratios, speech and background noise levels and STI values derived from modulation transfer functions. Subsequently, he computed regression coefficients of different pairwise combinations of speech intelligibility and reverberation measures.

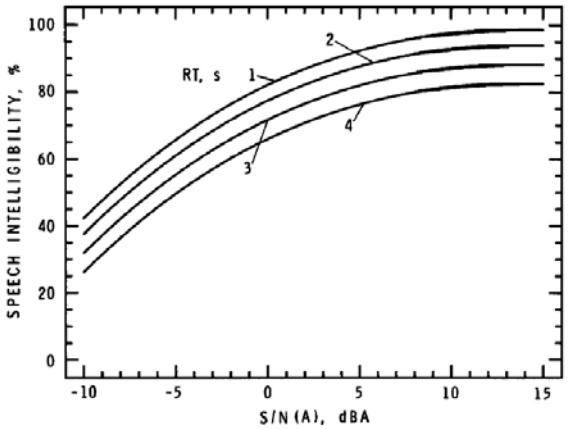


Figure 2.1: Best-fit curves of speech intelligibility versus overall signal-to-noise ratio for four RT values of 1,2,3 and 4 seconds, after Bradley, 1986.

Figure 2.1 plots curves corresponding to the regression coefficients for a case, where a combination of the A-weighted signal-to-noise ratio with the RT is considered (RT values of 1, 2, 3, and 4 s). He took

[1] It is also noteworthy to say that the effect of the early reflections is also significant on speech intelligibility. For instance, Bradley et. al (2003) [56] have studied the influence of early reflections on the speech intelligibility. They have shown that early reflections can increase the speech intelligibility significantly, especially when the talker's head is turned away from the listener or the listener is located far from the talker which is mostly the case at the positions in the rear of the rooms.

the point where the best-fit curves are closest to 100% speech intelligibility as an ideal design goal. According to the figure 2.1, this points appears to be close to $RT = 1\ sec$ and signal-to-noise ratio equal to 15 dB, which, as he indicates, is in agreement with the conventional acoustical wisdom.

On closer inspection, it should be studied how reverberation time (and also background noise which is not the area of this contribution) affect the intelligibility in different frequency ranges.
As the consonants playing the dominant role in the speech intelligibility are within the mid-high frequency, the most of recent studies have concentrated on mid-high reverberation time (Wu, et al., 2014, [57]). Nevertheless, there has been a number of studies regarding the influence of reverberation time in low frequencies on the speech intelligibility which will be the topic of this master thesis.

Kuttruff, 2001, [10] describes low frequencies as contributing little to speech intelligibility and suggests, therefore, to apply suitably designed sound absorbers to the walls of the room to reduce the reverberation time and hence the stationary sound level at low frequencies. However, he does not describe to what extent low-frequency reverberation time could affect the speech intelligibility. On the other hand, reverberation times of some renowned opera theatres presented by Kuttruff, 2001, [10], shown in table 2.1, reveal that the reverberation time in all of these rooms increases towards the low frequencies (Specifically, here, towards 125 Hz) which indicates that the possible deteriorating effect of low frequency reverberation time on speech intelligibility appears to be neglected. However, concerning the subjective sensation in case of music (and not speech), it is often believed that increasing the reverberation time towards low frequencies is responsible for what is called the 'warmth' of musical sounds (Kuttruff, 2001, [10]) which is not only not deteriorating, but also desired.

Opera house	Volume (m^3)	Number of seats/standings	Year of completion/ reconstruction	T_{125}	$T_{500-1000}$
La Scala, Milano	10,000	2290/400	1778 (1946)	1.2	0.95
Covent Garden, London	10,100	2180/60	1858	1.2	1.1
Festspielhaus, Bayreuth	11,000	1800	1876	1.7	1.5
National Theatre, Taipei	11,200	1522	1987	1.6	1.4
Staatsoper, Wien	11,600	1658/560	1869 (1955)	1.5	1.3
Staatsoper, Dresden	12,500	1290	1878 (1985)	2.3	1.7
Neues Festspielhaus, Salzburg	14,000	2158	1960	1.7	1.5
Metropolitan Opera House, New York	30,500	3800	1966	2.25[a]	1.8[a]

Table 2.1: Reverberation time of occupied opera theatres, after Kuttruff, 2001. T_{125} and $T_{500-1000}$ indicate the reverberation time at 125 Hz and 500-1000 Hz frequency ranges, respectively. a: with 80% occupancy only.

Furthermore, Polich and Segovia, 1999, [58] have proposed a more holistic consideration of reverberation time over the entire frequency range. They indicated that for a good intelligibility the reverberation time should be consistent across the frequency band and not differ more than 0.1 sec between the adjacent frequencies.
In another study, Wu et. al, 2014, [57] investigated the influence of noise and reverberation time (with different frequency characteristics) on Mandarin Chinese speech intelligibility (MCSI). For this purpose, they simulated room impulse responses at the listener's position of a room with the dimensions $8.4 \times 7.7 \times 3.9\ m^3$. The simulations were done in different reverberation time characteristics (and also different noise characteristics). Three different low frequency reverberation scenarios ($Bass\ Ratio <$

1, *Bass Ratio* ≈ 1 *and Bass Ratio* > 1) were created by adjusting the absorption and scattering coefficients of the surface materials of the model. Subsequently, they convolved the calculated impulse responses of each scenario with signals of test word lists specified by GB 4959-85, 1985, [59], in order to evaluate the speech intelligibility by means of listening tests. Ten word lists containing 25 five-word rows of similar-sounding Chinese monosyllabic words, similar to modified Rhyme Test of English in ANSI S3.2,1989 (R1999), [60], were used in the test. The MCSI scores under different Bass Ratio and noise conditions are shown in the figure 2.2.

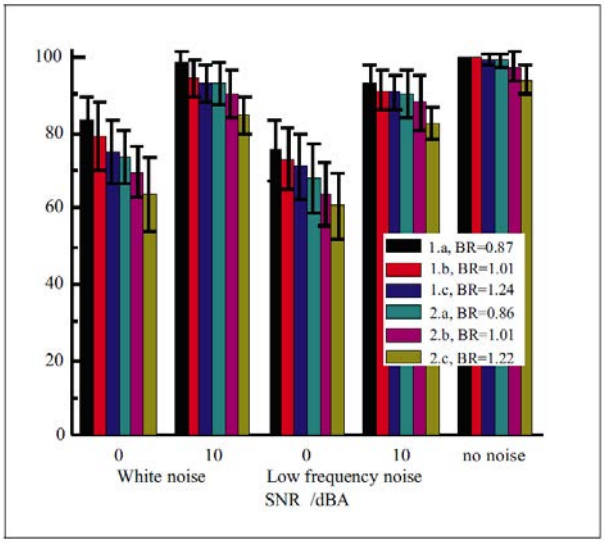


Figure 2.2: MCSI scores in different low frequency noise, RT's and bass ratios, after Wu et. al, 2014

Based on the results of this study, it turned out that low frequency reverberation time has a significant effect on Mandarin Chinese speech intelligibility. More precisely, they observed that speech intelligibility is better when low frequency reverberation time is lower than that of mid-frequency as compared to the cases where low frequency reverberation time was higher than the mid-frequency one. In other words, they showed that lower bass ratio indicates a better score in Mandarin Chinese speech intelligibility (See figure 2.2).
In another study, Mommertz et. al, 2006, [61] investigated the effect of low frequency reverberation time on the speech intelligibility in a classrooms with the dimensions 7 x 10 x 3.3 m^3. They simulated and also measured binaural room impulse responses in different listener positions under different reverberation time conditions. Then, they convolved the calculated impulse responses with an anechoic speech signal of a male speaker. Subsequently, they evaluated the speech intelligibility in different conditions based on listening tests using Bradely-Terry-Luce model (BTL model). As a result of their study, it came out that a low amount of reverberation time in 250 Hz octave band is necessary to achieve an acceptable speech intelligibility.
Moreover, a research by Cha and Fuchs, 2008, [62] has introduced a rule of thumb to reach an appropriate speech intelligibility. It is indicated that the reverberation time in the range of 125 up to 4000 Hz should not rise, in order to reduce the masking effect and improve the speech intelligibility.

Furthermore, Fuchs, 2019, [63] has referred to the different contributions of frequencies to the speech intelligibility. He points out that the frequency range of 500 - 4000 Hz contributes about 90% to speech intelligibility while the effect of low frequencies only amounts to about 10% (See figure 2.3). But these low frequencies of speech have much more energy than the middle and high frequencies. Based on the fact of masking of middle and high frequencies by the low frequencies in human's auditory system (Upward

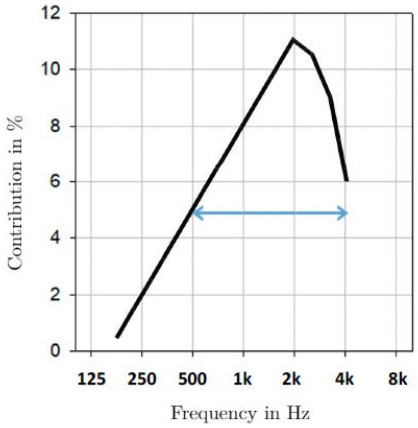


Figure 2.3: Relative contribution of frequencies to speech intelligibility (After Fuchs, 2019)

spread of masking), he recommends strongly to damp the low frequencies. Therefore, he suggests that one does everything regarding the acoustical design of the rooms to reduce the bass ratio in order to get a better speech intelligibility[2].

2.2 The Effect of Early Reflections on Speech Intelligibility

In order to investigate another room acoustical parameter which plays a crucial role in speech intelligibility, Bradley et. al, 2003, [56] studied the influence of early reflections on the speech intelligibility. They showed that early reflections can increase the speech intelligibility significantly, especially when the talker's head is turned away from the listener or the listener is located far from the talker which is mostly the case at the positions in the rear of the rooms. They investigated S/N(A) values (A-weighted signal-to-noise ratio, see section 1.3.1) for different listener's head orientation towards the talker with and without early reflections. The results, in case of no early reflections, showed a decrease in S/N(A) from -2.2 dB to -17.5 dB for the case when listener directly looks at the talker and when he/she has turned the head 180° away from the talker, respectively. While this decrease was in the range of -0.6 dB to -5.4 dB in case of presence of early reflections which supports the idea that early reflections can be beneficial for speech intelligibility.
Furthermore, They calculated the values of ERB (Early Reflection Benefit) [3] as a function of distance from the source and concluded that ERB can reach up to 9 dB for relative big distance of the listener from the source which again implies on the constructive effect of early reflections on the speech intelligibility. Regarding the importance of early reflections for the speech intelligibility, Fuchs, 2019, [63] also suggests not to mount mid- and high-frequency-range absorption material on the whole ceiling of a room, as the ceiling reflections play a vital role in supporting the direct sound and consequently the speech intelligibility.

[2]It is not also desirable to reduce the low frequency reverberation time to decrease the upward spread of the masking effect by low frequencies but it is also desired to reduce the low frequency noise (specially below 300 Hz), as it degrades the intelligibility significantly, specially in high sound pressure levels (Pickett, 1951 [64]).

[3]The energy ratio of early reflections to the direct sound. In [56] , it is calculated by $10\log_{10}\frac{E_{50}}{E_{10}}$ dB. Where E_{50} and E_{10} indicate the energy arriving within the first 50 ms and 10 ms, respectively.

2.3 The Effect of Binaural Listening on the Intelligibility

Listeners often have the impression of being able to pick up any of a range of voices around them by focusing their attention on the appropriate direction. The benefit of listening to speech with two ears instead of one in conditions including background noise is known as cocktail party effect (Cherry, 1953, [43]). The cocktail party problem is determined by several factors, including the location of speech and interferer sources, room acoustics, the type of interferer, and a potential hearing impairment of the listener (Beutelmann et al., 2010, [65]).
Moreover, Bronkhorst, 2000, [66] has shown that the binaural hearing decreases the signal-to-noise ratio (SNR) which is necessary to achieve 50% speech intelligibility in noise by up to 12 dB.
This benefit encourages one to take into account the effect of binaural listening when investigating the speech intelligibility.

Binaural Speech Intelligibility

To incorporate the effect of binaural hearing, different models and approaches has been introduced. One of the models has been proposed by Beutelmann et al., 2010, [65]. Figure 2.4 shows a schematic diagram of their model. Speech and noise signals of the left and ear are fed into the model as input. Additionally, independent masking noises simulating the internal noise in the auditory system are added to the left and right noise signals. Then, the input signals are filtered into 30 frequency bands using a gammatone filter bank (Hohmann, 2002, [67]). Subsequently, the filtered signals are processed using an EC stage (Equalization Cancellation model, according to Durlach, 1963, [55], see section 1.3.9), which models the binaural release from masking and searches for the maximal possible SNR with the given interaural differences of speech and noise in each frequency band. After that, applying some other calculation processes, the frequency-dependent SNR's are fed into single-channel SII calculation stage and finally, the broadband SNR of the input signals that result in a speech intelligibility of 50% is calculated as SRT (A detailed explanation of the process is provided by Beutelmann et al. 2010, [65]).
In this thesis, however, a different approach is used to combine the monaural and binaural effect of intelligibility together which is explained in section 3.3.2.

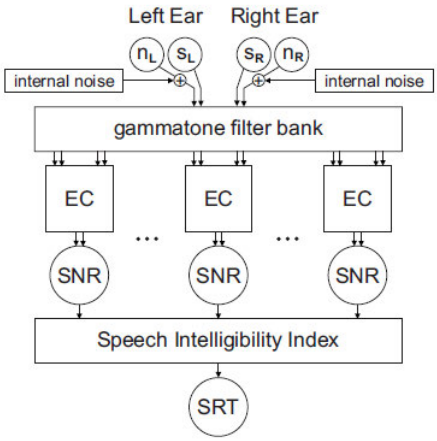


Figure 2.4: Schematic diagram of the binaural speech intelligibility model, after Beutelmann et al., 2010. s_L and n_L: Speech and noise signal at the left ear, s_L and n_L denote the same for the right ear. EC: Equalization cancellation process

2.4 Room Acoustical Simulation

The room acoustical simulations are done using a room acoustical simulation framework developed by Institute of Technical Acoustics, RWTH Aachen University, called RAVEN (Schröder and Vorländer, 2011, [68]).
In computational room acoustics, the ray tracing method has been used since computers became available where the first comfortable computer programs were introduced in 1980's in Europe (Stephenson, 2010, [69]).
The hybrid room acoustics simulation RAVEN (Room Acoustics for Virtual ENvironments) combines a deterministic image source method and a stochastic ray tracing algorithm and is capable of simulating the wave phenomena such as diffraction at low frequencies, scattering at high frequencies and specular reflections (Schröder and Vorländer, 2011, [68]).

Moreover, the stochastic ray tracing method is also extended in the RAVEN framework where a new type of particle detector called deflection cylinder is used. In addition to that, the uncertainty-based method of Stephenson, 2010, [69] is used to deflect the particles around an edge (identical with the deflection cylinder's axis) as a function of the shortest fly-by distance. An in-depth explanation of the methods and algorithms used in RAVEN framework can be found in the contribution of Schröder, 2011, [70].

2.5 Research Questions of the Thesis

Few studies which have focused on the effect of low frequency (Specifically, the octave bands 125 Hz and 250 Hz are investigated in this thesis) characteristics of the reverberation time on speech intelligibility have been pointed out in section 2.1. On closer inspection, it is noticeable that they have not covered a wide range of variety of parameters, specially in terms of the overall reverberation time (RT) and also size and form of the room. Overall RT is defined in this contribution as the constant value of the RT in all frequency bands except for the band at which RT is altered.
All of the mentioned studies in the previous section have investigated the room size of roughly 200-300 m^3 with an average reverberation time of 0.5 - 1.4 seconds. Therefore, the question is still to be answered, if and how the degrading effect of low frequency reverberation time on speech intelligibility comes into play for greater room sizes and reverberation times and also for different room shapes. In this regard, different combinations of room size and reverberation time are possible which will be discussed in section 3.1.1.
Due to the large number of simulation cases in this contribution, using speech intelligibility assessment algorithms is preferred performing listening tests.
The other aspect which is investigated in this thesis is taking into account the effect of binaural listening in the process of speech intelligibility assessment.
The research question of this thesis can be formulated, as follows:

To what extent does the low frequency reverberation time (125 Hz and 250 Hz octave bands) affect the single-channel speech intelligibility (specifically STI) considering the parameters, *average reverberation time* , *room volume* and *room shape* and to what extent does this affect the speech intelligibility taking into account the *binaural listening*?

3 Methods

The methods which will be applied in the investigations of the master thesis can be divided into two main parts. First, the 3D-modeling software and the room acoustical simulation tool and second, the algorithm used for the prediction of the speech intelligibility. These topics will be covered in this chapter.

3.1 Room Acoustic Simulations

3.1.1 Modelling Parameters

As already mentioned in section 2.5, the parameters investigated in the master thesis are overall RT, room size (room volume) and room shape. Regarding this, four overall RT's (0.5, 1, 2 and 4 sec), four room volumes (250, 1000, 4000 and 16000 m^3) and also four room shapes (Fan shaped, Horseshoe, Shoebox and Vineyard, see figure 3.2) will be taken into account in calculations. In case of the volumes 250 m^3 and 1000 m^3, only the room shapes *Fan shaped* and *Shoebox* are considered, since the other two room shapes are unrealistic in such low volumes. The amount of the change of the reverberation time in each frequency band (125 and 250 Hz) is taken into account as the variable of this study. In each case, the value of RT in the octave frequency bands 125 Hz and/or 250 Hz is multiplied by a factor denoted by ${RT^*}_{125}$ and ${RT^*}_{250}$, respectively (in some cases, RT in one of these bands is changed while the other one is kept constant and in other cases both are changed simultaneously, so that all of the possible changes are taken into account).
Figure 3.1 shows a sample target RT curve where *Overall RT* equals to 1 *sec* which is multiplied by factors ${RT^*}_{125} = 1$ and ${RT^*}_{250} = 1.2$ in 125 Hz and 250 Hz octave bands, respectively.

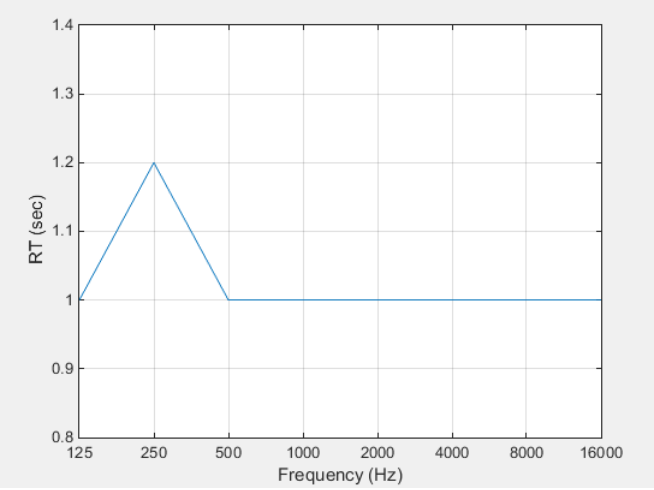


Figure 3.1: A sample target RT curve with *Overall RT* = 1 *sec*, ${RT^*}_{125} = 1$ and ${RT^*}_{250} = 1.2$

The mentioned parameters and variables including their range of change are shown in the table 3.1. Table 3.2 shows the number of states based on the room shape and volume which leads to a total number of 2352 simulations.

Parameters/Variables	Values/Cases	Number of values/cases
Overall RT	0.5, 1, 2, 4 (sec)	4
Volume	250, 1000, 4000, 16000 (m^3)	4
Shape	Fan shaped, Horseshoe, Shoebox, Vineyard	4
RT^*_{125}	0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3	7
RT^*_{250}	0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3	7

Table 3.1: The parameters and variables investigated in this study

Shape	Volume (m^3)	Total number of the states
Fan shaped	250,1000,4000,16000	784
Horseshoe	4000,16000	392
Shoebox	250,1000,4000,16000	784
Vineyard	4000,16000	392
Total		2352

Table 3.2: The number of states based on room shape and volume

3.1.2 Number of Measurement points

The norm ISO 3382-1, 2009, [71] defines the minimum number of microphone positions to be used in the measurement of room acoustical parameters as a function of the number of the seats which is shown in the table 3.3.

Number of seats	Minimum number of microphone positions
500	6
1000	8
2000	10

Table 3.3: Minimum number of receiver positions as a function of room capacity (According to ISO 3382-1, 2009), left column: number of seats, right column: number of microphones

As our investigations are based on the room volume (and not the number of seats), we need a rule of thumb to relate the room capacity and volume. According to (Holden, 2016, [72]), a value of 11 m^3 (400 ft^3) per person is reasonable for a concert hall. Based on this, the resulting typical room capacities corresponding to our defined room volumes will be as shown in table 3.4. The minimum number is kept to 6 microphone positions, based on the table 3.3
Based on the norm ISO 3382-1, 2009, [71], the minimum number of sound sources should be equal to two. Therefore, this has been kept as a minimum and increased with the increasing of the room volume, as shown in the table 3.4.

Number of seats	Room Volume (m^3)	Number of Microphones	Number of Sound Sources
23	250	6	2
91	1000	6	3
364	4000	6	4
1455	16000	10	5

Table 3.4: Number of sound sources and receivers used in the simulations based on room volume

Table 3.5 shows the number of calculated single-channel and binaural impulse responses per volume and

also the total number of them where the values listed in the table are as follows:
N_S: Number of sources
N_R: Number of receivers
N_C: Number of cases
N_{IR}: Number of single-channel impulse responses
N_{BRIR}: Number of binaural impulse responses

Volume (m^3)	N_S	N_R	N_C	N_{IR}	N_{BRIR}	Total number of IR's and BRIR's
250	2	6	392	4,704	4,704	9,408
1000	3	6	392	7,056	7,056	14,112
4000	4	6	784	18,816	18,816	37,632
16000	5	10	784	39,200	39,200	78,400
Total			2,352	69,776	69,776	139,552

Table 3.5: Number of the calculated single-channel and binaural impulse responses

3.1.3 ISO Conformity

The norm ISO 3382-1, 2009, [71] recommends the values shown in the table 3.6 as allowed deviations of the levels of the source compared to an omnidirectional one. In the simulations of this thesis, all of these deviations are theoretically equal to zero, as an ideal omnidirectional sound source is used.
Furthermore, The minimum distance between sound sources and receivers should be 1,5 meters where the acoustic center of the sound source should be 1,5 m over the floor. Moreover, the microphone positions should be a half of the wave length far apart from each other (As recommended in the norm, it should be about 2m in the usual frequency range). On the other hand, the distance of each microphone to the nearest reflecting surface should be at least a quarter of wavelength (1m) (Source and receiver positions are shown in appendix 5.2).
Additionally, in practical measurements, as required by the norm ISO 3382-1, 2009, [71], the sound source should produce a sound pressure level at least 45 dB over the noise floor of the room in each frequency band. This is not of course a problem in the case of the simulations of this thesis as there is no noise signal assumed in the simulations.
All the requirements mentioned above are satisfied in the course of preparing the models for acoustical simulation.

Frequency, Hertz	125	250	500	1000	2000	4000
Max. Deviation, dB	±1	±1	±1	±3	±5	±6

Table 3.6: Allowed directional deviations of SPL of a sound source

3.2 Room Acoustical Simulation

3.2.1 3D Models

Greif et al., 2020, [73] investigated to what extent the room acoustic signature of different room shapes of concert halls can be identified audibly, and to what extent this recognition is dependent on the room volume, reverberation time and the scattering of the walls. Their investigation was done based on

listening tests in audio-visual virtual environments of four classical room shapes of concert halls. The SketchUp[1] models used in the contribution of Greif et al., 2020, [73], as shown in the figure 3.2, are used in this study as the input 3D models of the room acoustical simulations.

3.2.2 Simulation Software

The room acoustical simulations software RAVEN (Room Acoustics for Virtual ENvironments), developed by Institute of Technical Acoustics, RWTH Aachen University (Schröder and Vorländer, 2011, [68]) (See section 2.4) is used to perform the simulations. The methods ray tracing (See section 1.2.1) and image source (See section 1.2.1) are combined in the simulation process.

3.2.3 Model Calibration

RAVEN Model Calibration Algorithm

In order to calibrate the models to achieve the desired frequency-dependent reverberation time, a built-in function in RAVEN is used. Firstly, this function calculates the reverberation time at each receiver position and the averages over them. This average value is shown as RT_{old} in the equation 3.1. Then, it extracts the equivalent absorption area of the model (A) and the total room surface (S) and based on them calculates an overall value of the absorption coefficients ($alpha_{old}$ in the equation 3.2) and then using the equation (3.1) and the defined target reverberation time (RT target) calculates the new absorption coefficients of the materials of the room surfaces. Then, the new absorption coefficients are applied to the materials of the model and the simulation is run again to calculate the resulting reverberation time. This process is repeated iteratively to come closer to the target curve in each iteration.

$$alpha_{new} = 1 - (1 - alpha_{old})^{\frac{RT_{old}}{RT_{target}}}. \tag{3.1}$$

$$alpha_{old} = A/S. \tag{3.2}$$

In order to control the loop of these iteration, a Matlab script is developed. This script compares the reverberation time values of each frequency band (octave bands) with the target. If the calculated value differs less than 5% from the target, the value is verified unless it is rejected (Based on the norm ISO 3382-1, 2009, [71], 5% of difference is the just noticeable difference of the perceived reverberation). This loop continues as long as all of the bands (except for 8kHz and 16kHz) are in the acceptable range of deviation. The reason why the 8kHz and 16 kHz bands are excluded is that the simulations are done with taking into account the air absorption. This absorption increases in a non-linear way in high frequencies (especially from 8 kHz upwards.). This causes a significant drop of the reverberation time which makes it almost impossible to achieve high reverberation times in high frequencies even when the respective absorption coefficient is almost zero. This is most considerable in the rooms with a large volume. For this reason, these two bands are excluded from the verification process of the reverberation time (The *alpha* values of these two bands which are used in the simulations are those which are calculated at the last run of the iteration after which the *alpha*s of the other bands are verified and the iteration is terminated).

[1] SketchUp is 3D-modelling software by *Trimble*. `https://www.sketchup.com/de`

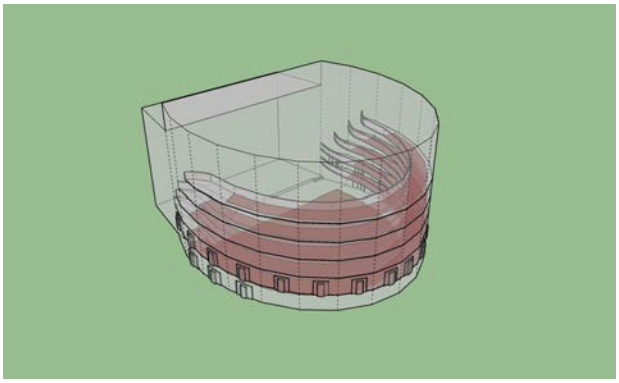
3.2.4 Impulse Response Calculation Flowchart

Figure 3.3 shows a flowchart of the process of impulse response calculation. The stages of this calculations are performed as follows:
1. Exporting the .ac file from SketchUp. This file is used by RAVEN as an input containing the geometrical information of the model and also the boundary conditions (Absorption and scattering coefficients of the room materials).
2. Exporting the .rpf file form SketchUp. This file is used by RAVEN as an input containing the source and receiver positions and the simulation settings.
3. Setting the simulation parameters. The atmospheric acoustical parameters used in the simulations are as follows:

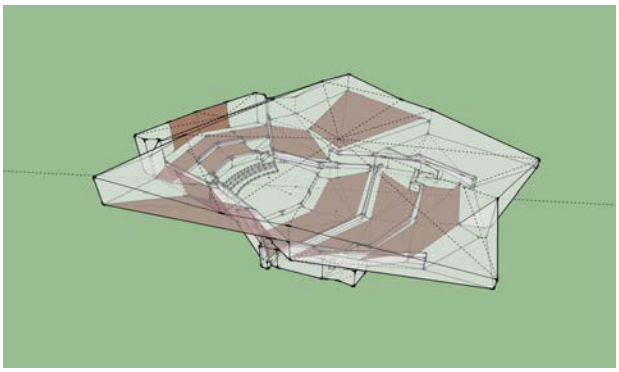
* Room temperature: 20° C
* Room humidity: 50%
* Room air pressure: 101325 Pa

Moreover, the air absorption has been enabled during all of the simulation runs. Additionally, an omnidirectional pattern for source and receiver (in case of single-channel impulse response) was used. In case of binaural impulse response, head-related transfer functions (HRTF's) of *«The Fabian head-related transfer function database»*[2] (Brinkmann et al., 2017, [74]) were used. This database contains head-related impulse responses (HRIR's) of 11 head-above-torso orientations (HATO) (The database also contains the frequency domain equivalent of the HRIR's, namely, HRTF's).
In this study, the HRTF's of the HATO equal to zero (The head is not moved to the left or right, relative to the torso) are used. The dummy heads in the 3D models are oriented so that they look at the direction of the middle of the stage.
4. Running the simulation and calculating the mean RT out of the impulse responses.
5. Comparing the calculated RT with the target RT. The calculated RT values are compared with the target curve in octave bands 31 Hz up to 4 kHz (As all of the target curves maintain a constant RT value towards the high frequencies, the octave bands 8 kHz and 16 kHz are excluded from this process, since they reveal significant drops due to the air absorption and therefore, making them constant becomes unrealistic, especially in higher RT values).
6. In this stage, the room materials are adapted if the calculated curve doesn't match target curve (admitting a threshold of 5% difference (the just noticeable difference of perceived reverberation) in each frequency band). Otherwise, the IR's and BRIR's are extracted in each measurement position and stored in the database for the further calculations.

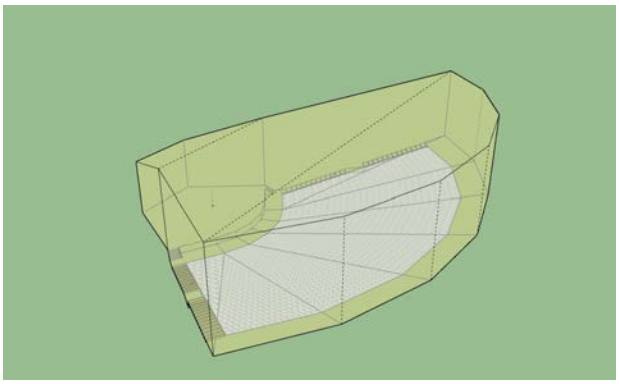
[2] `http://dx.doi.org/10.14279/depositonce-5718`



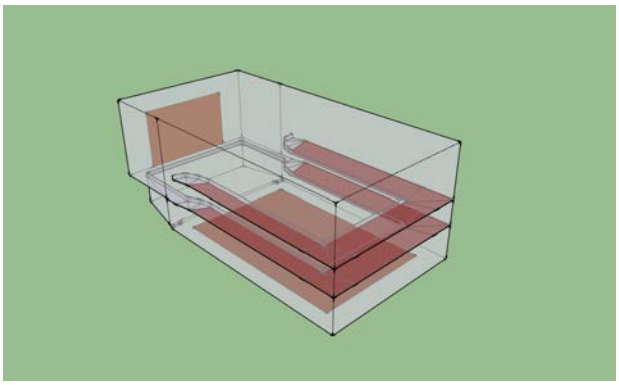
(a) Horseshoe



(b) Vineyard



(c) Fan shaped



(d) Shoebox

Figure 3.2: SketchUp models used in the simulations after Greif et al., 2020

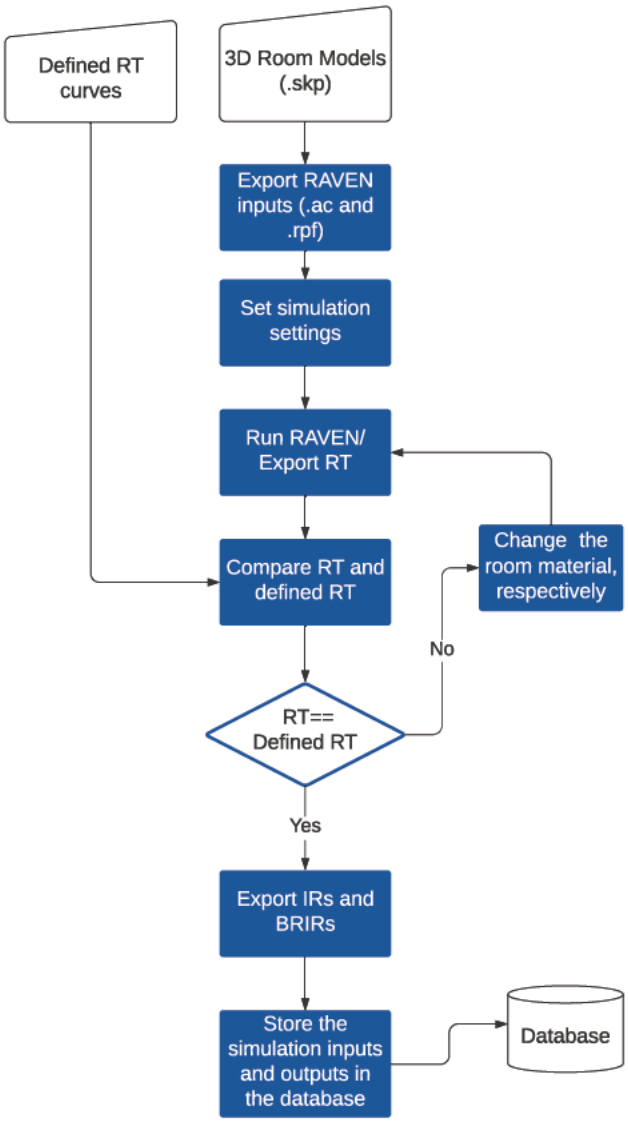


Figure 3.3: Flowchart showing the process of calculating the impulse responses out of the models

3.3 Calculation of the Speech Intelligibility

3.3.1 Speech Transmission Index (STI)

The STI value is calculated for each single-channel impulse response using a Matlab script[3]. The script calculates, first, the modulation transfer function (MTF) for octave-band-filtered impulse responses from 125 Hz up to 8 kHz. Subsequently, the magnitude of each MTF is calculated at each of the 14 modulation frequencies (See section 1.1.8) based on which an effective SNR value is computed.

3.3.2 Calculating the Effect of Binaural Listening

Binaural Benefit

The binaural benefit is calculated according to the model of Jelfs et. al, 2011, [47]. The calculation is based on the Matlab implementation of this model presented in the *Auditory Modeling Toolbox*[4], (Søndergaard and Majdak, 2013, [75]). The Matlab script calculates the increase of the speech intelligibility as a consequence of spatially separation of the target and the interferer which are described by means of their impulse responses in the script. The script takes as input the binaural impulse response of the target and also that of the interferer. Then, it gives the spatial release from masking (SRM) in dB as output. It can also calculate the component of SRM due to better-ear listening and the one due to binaural unmasking in dB, separately. Apparently, this model needs an binaural impulse response of an interferer which is not available in the investigations of this master thesis. In this contribution, no direct sound source exists in the room which interferes with the target. The only factor deteriorating the speech intelligibility are the reflections. In order for these reflection to be taken into account in the model, a model extension is needed which segments the impulse response into two parts, namely useful and detrimental parts. This task is done using the model developed by Kokabi et. al, 2018, [76].
Kokabi et. al tested the prediction accuracy with a binaural intelligibility model with an extension to model the influence of reverberation for four different virtual rooms. They have executed SRT measurements in quiet, in the absence of masking sound sources where the stimuli of the listening tests where provided by simulating binaural room impulse responses using RAVEN (Schröder and Vorländer, 2011, [68]).

Then they calculated the SNR by applying the BRIR's to the Cardiff binaural Model (Jelfs et al., 2011 , [47]) and extended the model using the UD classification (useful to detrimental) suggested by Rennies et al., 2011, [77]. In order to do the impulse response segmentation, they used two methods, namely, fixed and fitted UD limits. Two fixed limits showing the least mean absolute error (MAE), namely 50 and 100 ms and three fitted limits based on room acoustical parameters clarity (C80), direct-to-reverberant energy (D/R) and interaural cross correlation (IACC) were used for the SRT predictions. Figure 3.4 shows the measured SRT's versus the predicted ones where RS11 indicates another dataset compared to that of the «current study (The contribution of Kokabi et al., 2018, [76])» and condition S0 and S90 indicate «source in front of the listener» and «source to the right of the listener», respectively. Their work reveals that the fitted U/D-limits significantly outperforms the fixes limits, keeping in mind that IACC shows slightly better results than the other two room acoustical parameters. These leads to an improvement of 1 dB in the accuracy of the SRT prediction which can result in 17% difference in the absolute intelligibility. A sample output of this segmentation is shown in the figure 3.5 which depicts the segmentation of a BRIR into an early (useful) and late (detrimental) window. The Matlab

[3] https://www.soundzones.com/2015/10/16/the-speech-transmission-index-sti-for-matlab/
[4] http://amtoolbox.sourceforge.net/

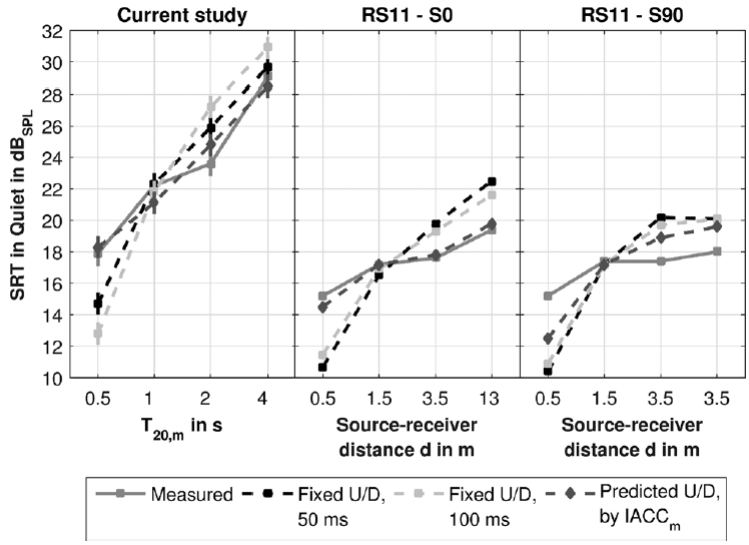


Figure 3.4: Measured and predicted SRT's with fixed and predicted U/D-limits averaged across participants after Kokabi et al., 2018

implementation[5] of this contribution is used to segment the impulse responses of this thesis in order to feed them in Jelfs model, 2011, [47].

Combining the STI with binaural benefit

In order to investigate the effect of binaural benefit which can be compared directly with STI values, a combined value is needed which also lies between zero and unity, as it is the case for STI.
Since the binaural benefit is a dB value, the relation of STI and a dB value like SNR can be used to establish a relation between STI and binaural benefit. Rhebergen and Verfsfeld, 2005, [18] have investigated the relation of Speech Intelligibility Index (SII) and Speech Reception Threshold (SRT). Figure 3.6 shows a part of this contribution where SII is depicted as a function of SNR. In this figure, the filled symbols indicate the condition of a stationary noise masker with the long-term spectrum of a female speaker. Open symbols denote calculations with fluctuating noise masker with the long-term spectrum of a female speaker and a speech-like modulation spectrum where the noise level is equal to 60 dBA.
An average curve of these two conditions is used in our calculation to relate the SNR and STI. However, the figure 3.6 shows the SII values and not the STI ones. These SII values are used as a representative for STI based on the fact that, these are almost the same when they are calculated based on MTF method (Larm and Hongisto, 2006, [78]).

[5]Link to the institutional repository for research data and publications of TU-Berlin: `http://dx.doi.org/10.14279/depositonce-6725.5`

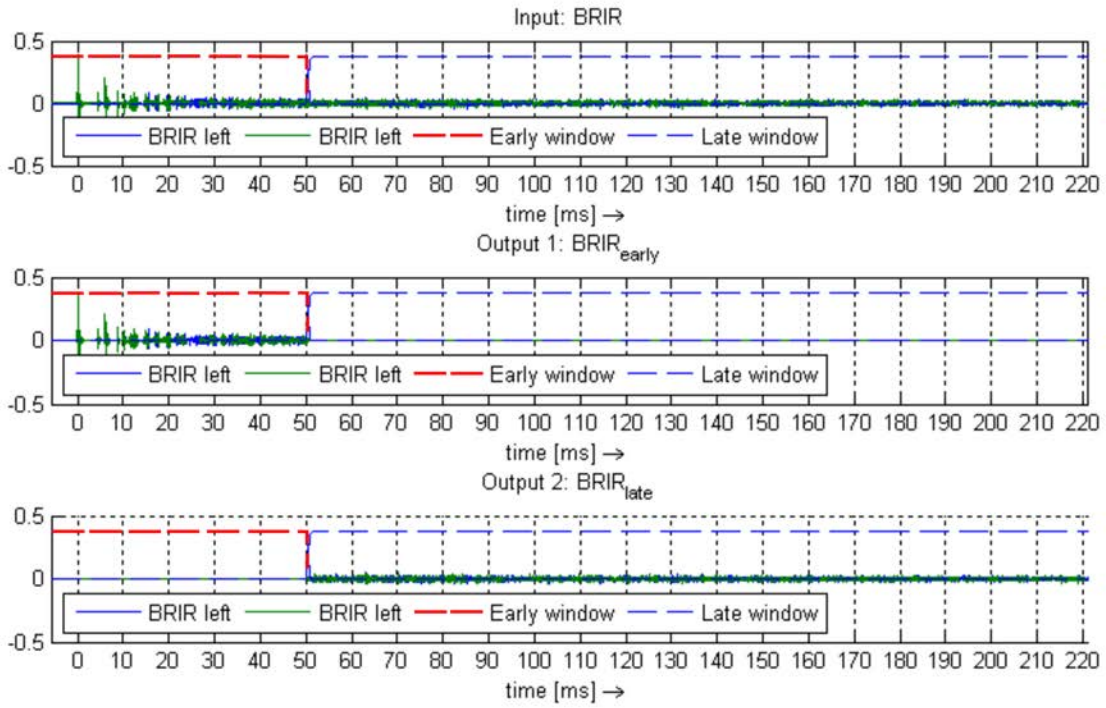


Figure 3.5: A sample of useful/detrimental-segmentation of a BRIR after Kokabi et al., 2018 (Top: BRIR, middle: early (useful) part, bottom: late (detrimental) part)

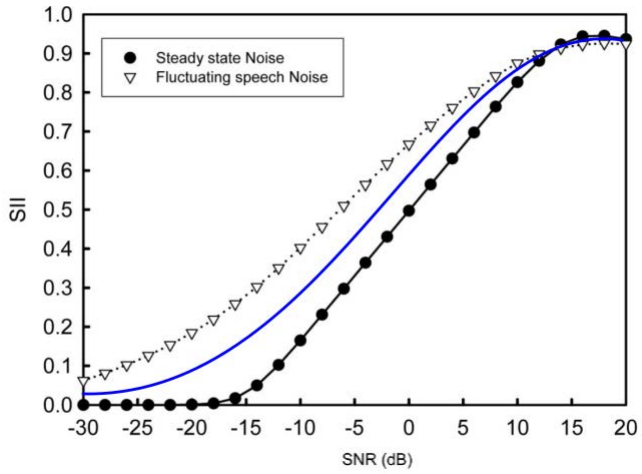


Figure 3.6: SII as a function of SNR, after Rhebergen and Verfsfeld, 2005 (Blue curve: A visual representation of the average values used in this thesis to relate SNR and STI)

The figure 3.7 shows the difference of STI and SII in different locations in an auditorium, after Larm and Hongisto, 2006, [78]. It can be seen that this difference is mostly between 1-2% and at the highest point reaches 4%. Therefore, using the curve in figure 3.6, a Matlab script is developed which takes the STI and binaural benefit values as input and gives a combined value (using *polyfit* function) between zero and unity as output denoted by Binaural STI.

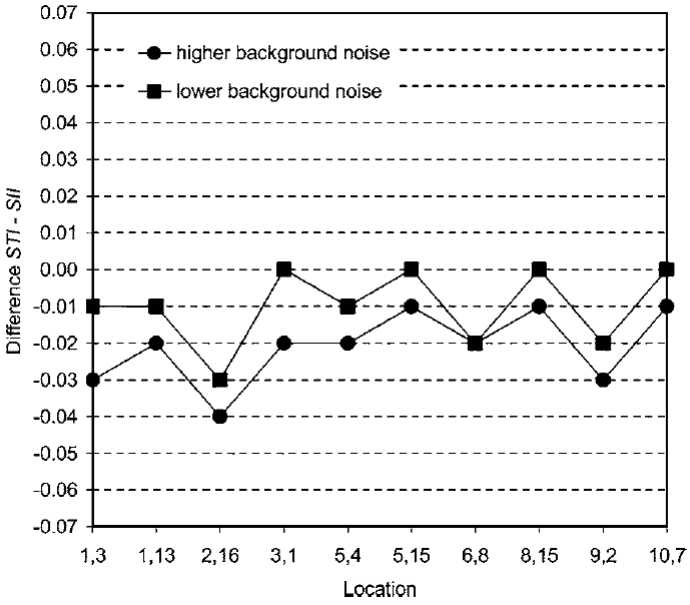


Figure 3.7: The difference between SII and STI in an auditorium, after Larm and Hongisto, 2006

A sample calculation of SNR-STI combination

Based on the previous section (See section 3.3.2), the SII values can be considered almost the same as STI. Therefore, the vertical axis in figure 3.6 can be well considered as STI, as shown in figure 3.8. According to the figure 3.8, if an example STI value of 0.60 (P1) is considered, the algorithm would first find the respective SNR value (P3) using the intersection point (P2) with the average curve (blue curve). Subsequently, the amount of change of SNR (As a representative of binaural benefit in dB) should be given as input in the algorithm. Assuming an SNR equal to +5 dB, the algorithm would move +5 dB from P3 and find the point P4. Finally, the intersection point with the average curve (P5) and the respective new STI value (P6) will be calculated which would result in a value of 0.74.
It is noteworthy that the change of STI as a function of SNR in the areas of STI almost higher than 0.8 would be lower, as a kind of saturation can be see in the figure 3.8. For instance, the same input SNR value of +5 dB would result in an increase of STI from 0.80 to 0.89 while the same SNR value would raise an STI of 0.60 to 0.74, as explained in the previous paragraph. As a result of this saturation, the maximum output of the algorithm would be equal to 0.94 which could be well used for the dataset of this study, as the maximum calculated STI over all the cases is equal to 0.91 (See table 4.1).

3.3.3 Speech Intelligibility Calculation Flowchart

The figure 3.9 shows the flowchart of calculation process of the speech intelligibility values out of the impulse responses.
The process can be explained as follows:

1. The STI values are calculated out of each single-channel impulse response.
2. The calculated STI values are averaged and the average value is stored in the database.
3. The binaural benefit value using the room-dependent segmentation is calculated out of each binaural impulse response. Then the averaged value is stored in the database. (The algorithm for calculating

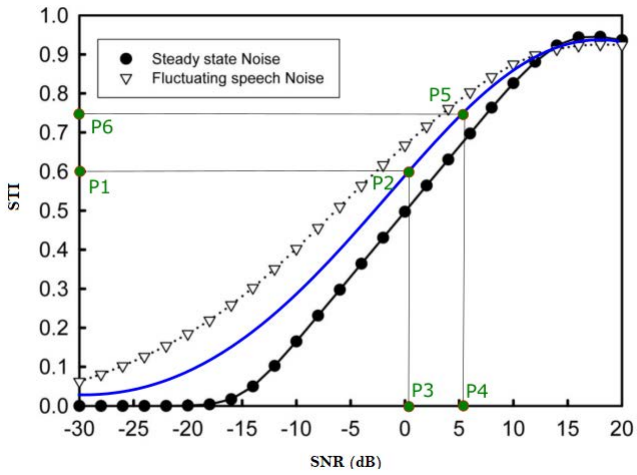


Figure 3.8: Schematic of a sample calculation of SNR-STI combination, using the SNR-SII curve (here: SNR-STI) after Rhebergen and Verfsfeld, 2005 (Blue curve: A visual representation of the average values used in this thesis to relate SNR and STI)

the binaural benefit seemed to deliver unrealistic values for the extreme case of RT=0.5 *sec* and volume=16000 m^3), therefore the results in this case are not considered in the further analyses and are considered as missing values in the statistical analysis.
4. The average STI and binaural benefit values of each scenario are combined (denoted by Binaural STI) and the values is stored in the database.

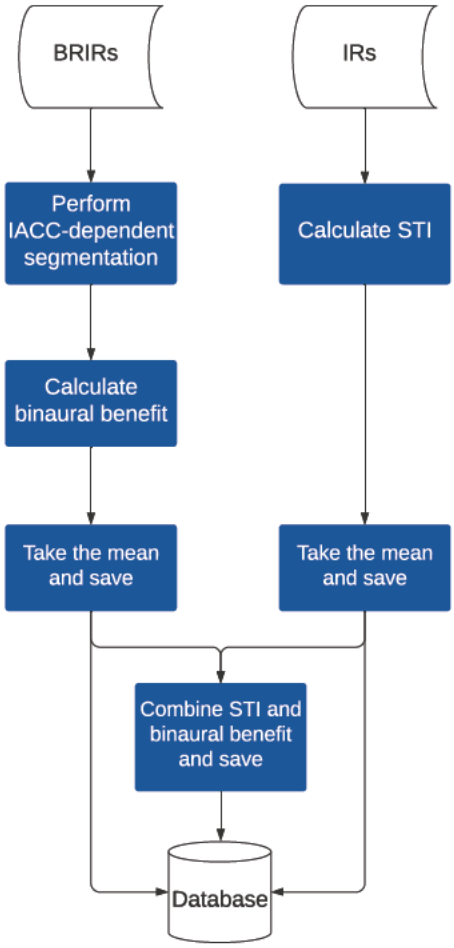


Figure 3.9: Flowchart showing the process of calculating the speech intelligibility values

4 Results

In the first section of this chapter (Section 4.1), descriptive statistics of the results are shown, the statistical analysis of which, according to a univariate Generalized Linear Model (GLM), is presented in section 4.2.

The statistical analyses of the sections 4.1 and 4.2 are done using the software IBM SPSS Statistics[1].

4.1 Descriptive Statistics

Table 4.1 provides a holistic overview of the dependent variables of the study. A total amount of 2,352 STI values are calculated where this equals to 2,156 data points in case of Binaural Benefit and Binaural STI. The 196 data points missing in case of Binaural Benefit and Binaural STI belong to the cases of Overall RT = 0.5 sec, Volume = 16000 m^3.
The mean STI is increased from 0.55 to 0.61 considering the binaural benefit where the average binaural benefit itself amounts to 3.90 dB.

	STI	Binaural Benefit	STI_Benefit
N_{Valid}	2352	2156	2156
$N_{Missing}$	0	196	196
Mean	0.55	3.90	0.61
Std. Deviation	0.17	4.30	0.22
Minimum	0.32	-6.73	0.27
Maximum	0.91	14.10	0.94

Table 4.1: Descriptive statistics of the dependent variables (N_{Valid} and $N_{Missing}$ denote the number of valid and missing data points, respectively)

Table 4.2 shows the type (independent or dependent), type of scaling and range of the variables used in the statistical analysis. As it can be seen, except for the parameter Shape which is scaled categorically, the other variables/parameters are all scaled as metric ones which are taken into the general linear model as covariates. The parameter Shape is considered as a fixed factor in the model.

Table 4.3 and 4.4 show the average and standard deviation of STI and Binaural STI, respectively, separated by shape. These variables are depicted in figure 4.1 to make a one-to-one comparison easier. It can be seen that the change of intelligibility values as a result of change of the room shape is more significant when the binaural effect is taken into account (Binaural STI).

Table 4.5 shows the average and standard deviation of Binaural Benefit. It should be considered that these values vary from -6.73 to 14.10 dB making a range of 20.83 dB (See table 4.1). The overall average and standard deviation of this variable equal to 3.90 and 4.30 dB, respectively, where the highest mean value (4.87 dB) belongs to the shape Shoebox.

[1] `https://www.ibm.com/products/spss-statistics`

Variable	Unit	Type	Scaling	Range
STI	-	Dependent	Metric	0.32 - 0.91
Binaural Benefit	dB	Dependent	Metric	-6.73 - 14.10
Binaural STI	-	Dependent	Metric	0.27 - 0.94
Overall RT	sec	Independent	Metric	0.5,1,2,4
RT^*_{125}	-	Independent	Metric	0.7,0.8,0.9,1,1.1,1.2,1.3
RT^*_{250}	-	Independent	Metric	0.7,0.8,0.9,1,1.1,1.2,1.3
Volume	m^3	Independent	Metric	250,1000,4000,16000
Shape	-	Independent	Categorical	«Four shapes»

Table 4.2: Type, scaling and the range of the variables. RT^*_{125} and RT^*_{250} indicate the scaling factor of RT in 125 Hz and 250 Hz octave bands, respectively.

Shape	STI (Mean)	STI (Std.)	Number of Cases
Fan shaped	0.54	0.16	784
Horseshoe	0.56	0.16	392
Shoebox	0.55	0.17	784
Vineyard	0.56	0.17	392
Total	0.55	0.17	2352

Table 4.3: Mean and standard deviation of STI

Figure 4.2 shows the mean Binaural Benefit values separated by room shape where Shoebox and Fan shaped seem to be the best and worst cases, respectively.

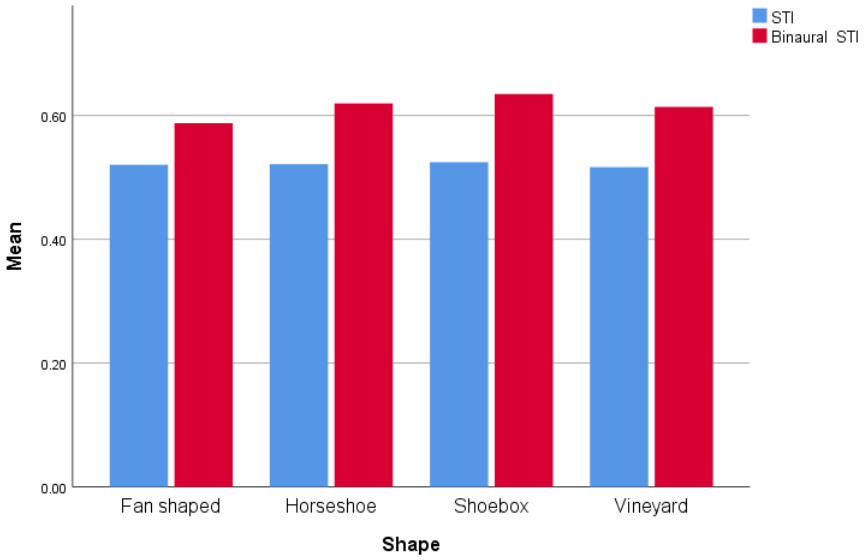


Figure 4.1: Mean STI and Binaural STI values separated by shape

Shape	Binaural STI (Mean)	Binaural STI (Std.)	Number of Cases
Fan shaped	0.59	0.22	735
Horseshoe	0.62	0.21	343
Shoebox	0.63	0.23	735
Vineyard	0.61	0.22	343
Total	0.61	0.22	2156

Table 4.4: Mean and standard deviation of Binaural STI

Shape	Binaural Benefit (Mean)	Binaural Benefit (Std.)	Number of Cases
Fan shaped	2.86	4.20	735
Horseshoe	3.94	3.65	343
Shoebox	4.87	4.63	735
Vineyard	4.00	3.84	343
Total	3.90	4.30	2156

Table 4.5: Mean and standard deviation of Binaural Benefit

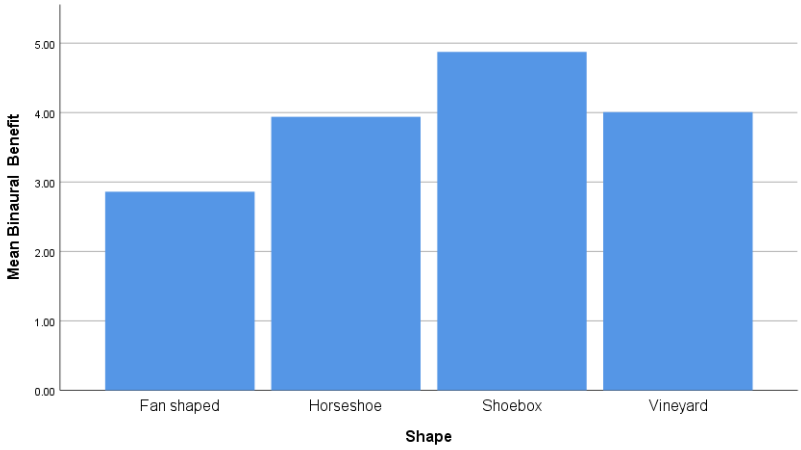


Figure 4.2: Mean Binaural Benefit values separated by shape

4.2 Univariate Generalized Linear Model (GLM) Analysis

In order to investigate the interaction of the variables, a univariate (i.e. including one dependent variable) generalized linear model (GLM) analysis (For an introduction to GLM, see Myers and Montgomery, 1997, [79]) is performed.
In all of the GLM analyses, the dependent variables are chosen as shown in table 4.2 where the independent metric variables and the one independent categorical variable shown in the table are considered as covariates and a fixed factor, respectively.
The bold Numbers in tables of this section indicate statistical significance, assuming a threshold of $p = 0.05$. The presented parameters shown in the tables of this section which show the test of the between-subject effects, are as follows:

df: Degree of freedom
Mean Square: Sum of Squares divided by their respective df.
F: F-value: Mean square regression divided by the mean Square residual.
Sig.: Statistical significance (For an introduction to GLM, see Myers and Montgomery, 1997, [79])

	df	Mean Square	F	Sig.
RT^*_{125}	1	0.015	3.852	0.050
RT^*_{250}	1	0.048	12.063	**0.001**
Overall RT	1	53.390	13370.319	**0.000**
Volume	1	1.985	497.095	**0.000**
Shape	3	0.013	3.199	**0.023**

Table 4.6: Test of the between-subject effects for the dependent variable STI (Adjusted R squared = 0.855). RT^*_{125} and RT^*_{250} indicate the scaling factor of RT in 125 Hz and 250 Hz octave bands, respectively. [Statistical significance is indicated by bold numbers.]

The results of GLM for the dependent variable STI are shown in table 4.6 where it can be seen that STI is significantly affected by RT^*_{250} while this effect is marginally significant in case of RT^*_{125}. Moreover, the parameters Overall RT, Volume and Shape show a significant effect on STI. In this model, about 85% of the variance in STI can be predicted by the independent variables (See the adjusted R squared.) Figure 4.3 depicts the decrease of average STI by increasing the RT^*_{250}, where the effect is also divided by the room shape.

	df	Mean Square	F	Sig.
RT^*_{125}	1	1.317	0.359	0.549
RT^*_{250}	1	42.393	11.556	**0.001**
Overall RT	1	28449.565	7755.126	**0.000**
Volume	1	373.335	101.768	**0.000**
Shape	3	568.973	155.098	**0.000**

Table 4.7: Test of the between-subject effects for the dependent variable Binaural Benefit (Adjusted R squared = 0.802). RT^*_{125} and RT^*_{250} indicate the scaling factor of RT in 125 Hz and 250 Hz octave bands, respectively. [Statistical significance is indicated by bold numbers.]

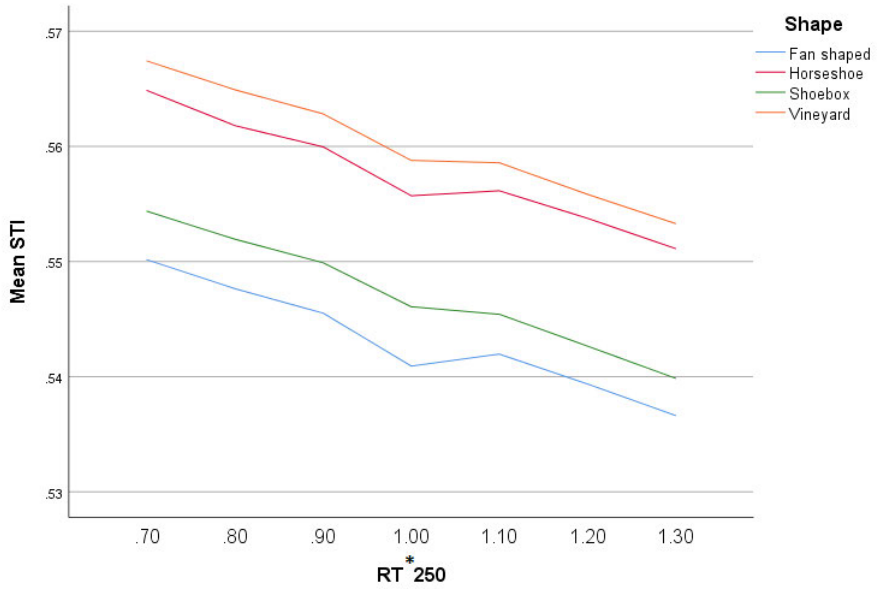


Figure 4.3: Mean STI values as a function of RT^*_{250} separated by shapes.

Table 4.7 shows the results of GLM analysis for the dependent variable Binaural Benefit where about 80% of the variance in this variable can be predicted by the independent variables in the model. Similar to the case of STI, the effect of RT^*_{250} is significant where RT^*_{125} is still statistically not significant. Moreover, Overall RT, Volume and Shape are still significant.

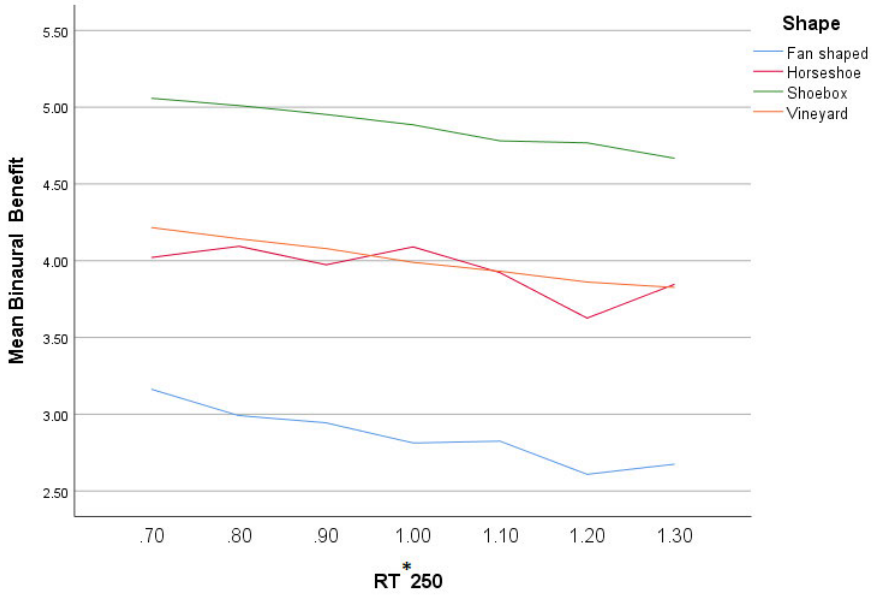


Figure 4.4: Mean Binaural Benefit values as a function of RT^*_{250} separated by shapes (The vertical axis is scaled in dB).

Table 4.8 shows the same analysis for the variable Binaural STI. In this case both RT^*_{125} and RT^*_{250} are statistically significant. Therefore, it can be concluded that the frequency band 125 Hz (besides the

250 Hz band) has a significant effect on STI when the effect of binaural hearing is incorporated in the calculation of STI. It is also noteworthy, that in this case the highest value of R-squared is reached, where about 94% of the variance of the dependent variable can be predicted by the independent variables.

	df	Mean Square	F	Sig.
RT^*_{125}	1	0.20	6.409	**0.011**
RT^*_{250}	1	0.094	30.264	**0.000**
Overall RT	1	98.814	31667.751	**0.000**
Volume	1	0.010	3.134	0.077
Shape	3	0.333	106.840	**0.000**

Table 4.8: Test of the between-subject effects for the dependent variable Binaural STI (Adjusted R squared = 0.938). RT^*_{125} and RT^*_{250} indicate the scaling factor of RT in 125 Hz and 250 Hz octave bands, respectively. [Statistical significance is indicated by bold numbers.]

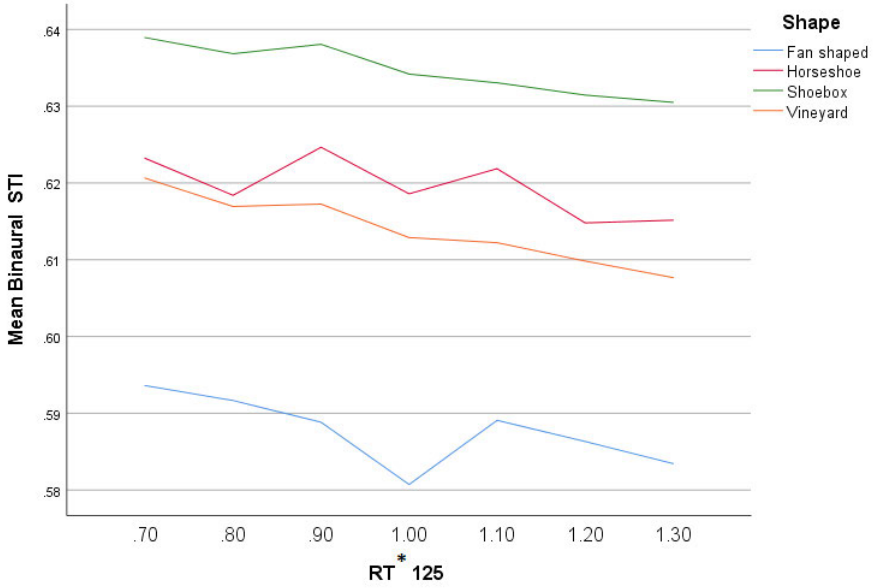


Figure 4.5: Mean Binaural STI values as a function of RT^*_{125} separated by shapes.

The average Binaural Benefit values as a function of RT^*_{250} are shown in the figure 4.4 showing a drop of about 0.5 dB with the increase of RT^*_{250} from 0.7 to 1.3. It can be seen that Shoebox and Fan shaped show the best and worst cases, respectively.

Figures 4.5 and 4.6 show the average values of Binaural STI as a function of RT^*_{125} and RT^*_{250}, respectively. The figures show that Shoebox and Fan shaped are in both cases at the first and last place of Binaural STI, where Horseshoe and Vineyard take the places in between.
It can be seen that the behaviour of the average Binaural STI is more uniform as a function of RT^*_{250} in comparison with RT^*_{125} which suggests that RT^*_{250} has a more significant affect on Binaural STI, as it is also noticeable in Sig. values in table 4.8.

Figure 4.7 shows the effect of RT^*_{125} on Binaural STI separated by parameters Overall RT and Volume. It is noteworthy that there is no graph for the case Overall RT = 0.5 sec and Volume = 16000 m^3, as there is no Binaural Benefit calculated for this case and therefore no Binaural STI could be calculated. The effect of RT^*_{250} on STI, Binaural Benefit and Binaural STI, separated by Overall RT and Volume, is shown in the figures 4.8, 4.9, and 4.10, respectively.

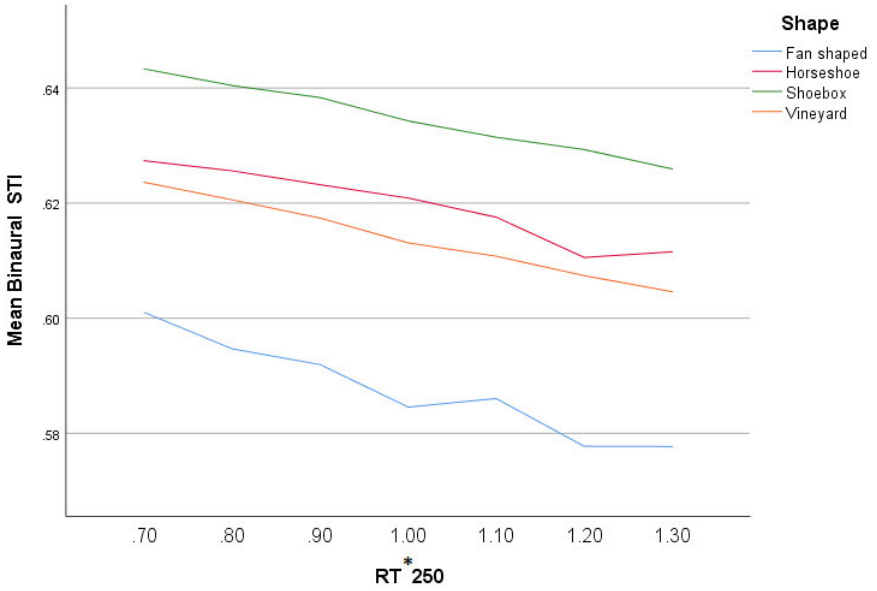


Figure 4.6: Mean Binaural STI values as a function of RT^*_{250} separated by shapes.

The behaviour of the curves presented in this section are investigated in the following chapter (See chapter Discussion and Conclusion 5).

The presented results of this section are discussed in the following chapter Discussion and Conclusion (See chapter 5).

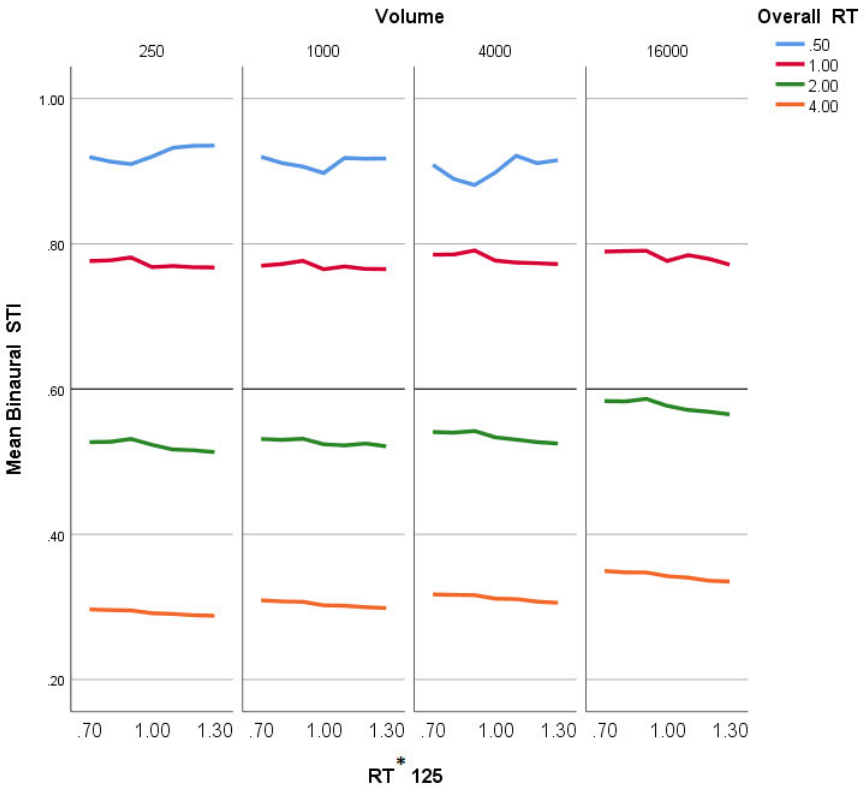


Figure 4.7: Effect of RT^*_{125} on Binaural STI separated by Overall RT and Volume

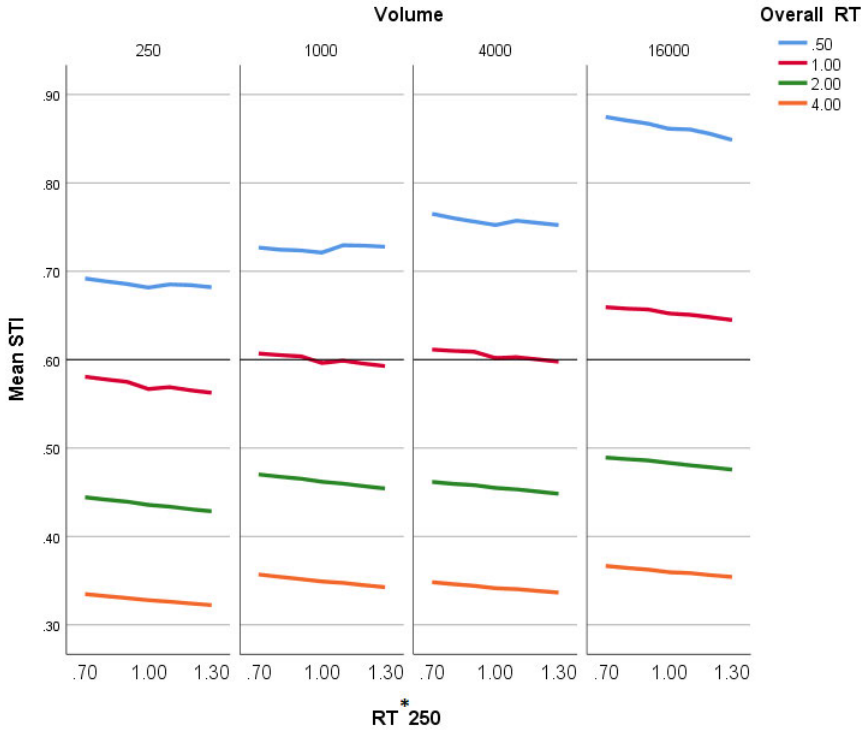


Figure 4.8: Effect of RT^*_{250} on STI separated by Overall RT and Volume

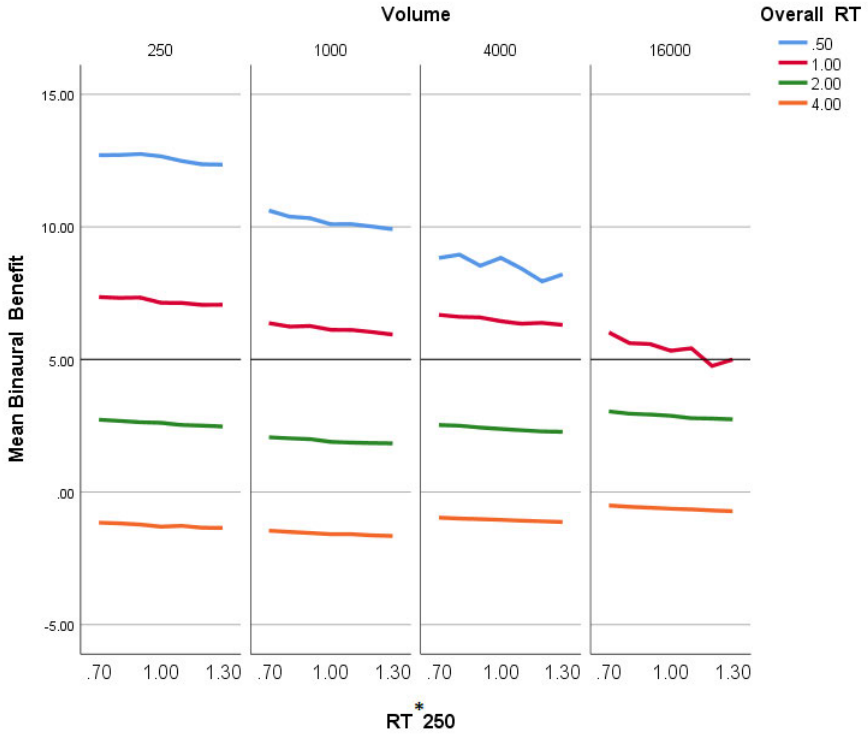


Figure 4.9: Effect of RT^*_{250} on Binaural Benefit separated by Overall RT and Volume (The vertical axis is scaled in dB).

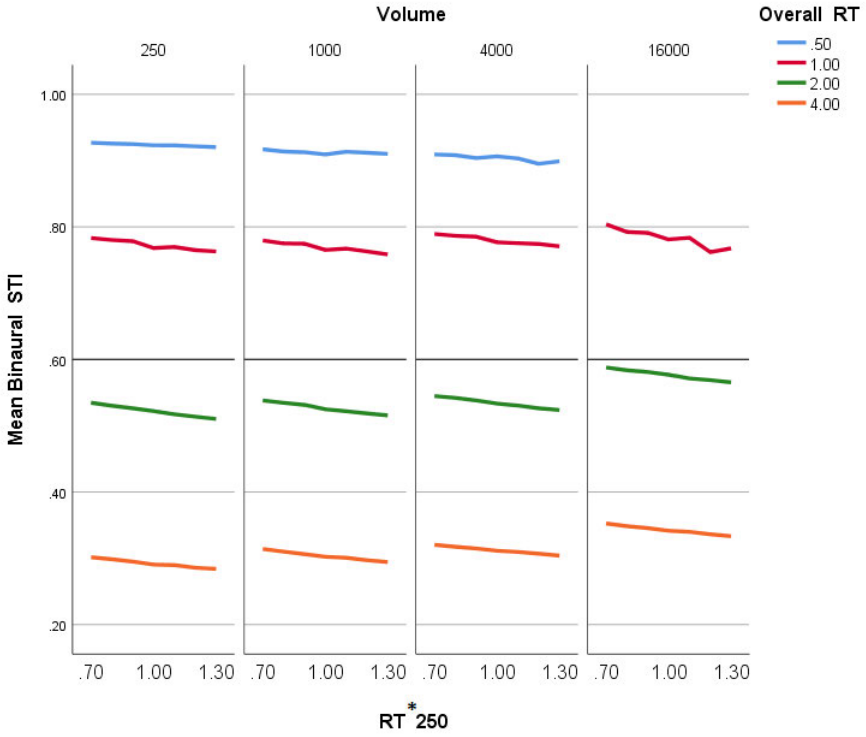


Figure 4.10: Effect of RT^*_{250} on Binaural STI separated by Overall RT and Volume

5 Discussion and Conclusion

5.1 Discussion

5.1.1 Effect of RT^*_{125} and RT^*_{250}

In section 4.2, it is shown that RT^*_{125} has no significant effect on STI or Binaural Benefit while a significant effect is seen when Binaural STI is investigated. It may be due to the fact that, when STI and Binaural Benefit are combined, the resulting measure consists of more detailed information about the level of intelligibility in humans auditory system taking into account two different approaches simultaneously (MTF in STI (See section 1.1.8) and BMLD and Binaural hearing (See sections 1.3.8 and 2.3) in Binaural Benefit). This leads to both a more realistic assessment of intelligibility and a wider range of change in resulting data and also a bigger variance (See table 4.1), making it possible to reveal a statistical significance.
RT^*_{250} shows in any case significant effect on speech intelligibility measures. One of the reasons of this influence compared to the less effectiveness of RT^*_{125} could be that the speech signal has more energy in 250 Hz octave band compared to the 125 Hz one (See figure 1.2). It could be also explained by the higher sensitivity of humans auditory system in 250 Hz octave band compared to the 125 Hz one (See section 1.3.2).

5.1.2 Effect of Volume and Overall RT

Figure 4.7 shows how Binaural STI is affected by RT^*_{125}. Considering the behaviour of the curves and their overall slope, there is no significant moderating effect by Volume, as all of the curve have an almost similar slope when Overall RT is constant and Volume changes.
Considering the Overall RT, there is also no moderating effect based on the overall slopes. However, for the case of Overall RT = 0.5 *sec*, there is a sudden decrease in Binaural STI an then an increase, as RT^*_{125} increases. This is theoretically not expected and could be a result of an artefact in STI algorithm at very low reverberation times. On the other hand, there is an overall increase in Binaural STI values as the Volume increases. Keeping the Overall RT constant and increasing the Volume, the deteriorating late reflections have to travel a longer way and therefore arrive later to the listener after the arrival of the direct sound. Hence they are more easily distinguishable and consequently less deteriorating. This could be possible explanation of this behaviour. This idea can be supported considering the figure 4.8 where this effect is the most significant in case of Overall RT = 0.5 sec and Volume = 16000 m^3.
Furthermore, figures 4.8, 4.9 and 4.10 show no significant change in their slope over Volume or Overall RT. Therefore, it can be concluded that Volume and Overall RT have no moderating effect on the change of STI, Binaural Benefit or Binaural STI as a result of change of RT^*_{250}.

5.1.3 Effect of Shape

According to the figure 4.4 the overall slope of the STI curves are very similar to each other suggesting that the effect of RT^*_{250} on STI is not dependent on the room shape, as the amount of the alteration of

STI is almost the same in all of the four cases and therefore, no moderating effect can be concluded. However, the overall value of STI is clearly different. Vineyard and Fan shaped show the best and worst performances, respectively. In case of fan shaped, the low STI could be due to the open-angled side walls (See figure 3.2) towards the back of the room which could cause a high amount of reflected energy coming from the stage to the rear areas of the room and consequently a destructive superposition with the direct sound coming from the stage. This could potentially deteriorate the intelligibility and consequently the STI in rear areas and therefore decrease the average STI value of this room shape compared to the others. On the contrary, Vineyard's special geometric design could probably provide most of the audience with useful early reflections due to numerous local side walls on different tiers which leads to the best overall STI value.
Figure 4.4 shows the mean Binaural Benefit curves as a function of ${RT^*}_{250}$. In this case, Fan shaped shows still the worst performance while Shoebox comes to the first place, compared to the case of STI. It can be expected that Shoebox is capable of providing very good early lateral reflections for most of the audience due to its parallel side walls and therefore, it shows the best performance, as far as Binaural Benefit is concerned. Considering the slope of the curves and the amount of the change in Binaural Benefit as a function of ${RT^*}_{250}$, all the shapes show an almost similar behaviour (nearly 0.5 dB drop) except for Horseshoe. In this case the behaviour of the curve is not monotonic and the decrease of Binaural Benefit is almost a half of that of the other shapes. However, when it is combined with STI (resulting in Binaural STI values, see figure 4.6), the behaviour of the curve and also its slope gets almost similar to other shapes. However, it should be mentioned that in this case, Fan shaped shows a slightly higher decrease in Binaural STI with the increase in ${RT^*}_{250}$. In other words, Binaural STI seems to be slightly more sensitive to the change of ${RT^*}_{250}$ for Fan shaped rooms compared to the other three room shapes. Moreover, Shoebox shoes still the best performance regarding the overall Binaural STI value.
Considering ${RT^*}_{125}$, the only significant change is seen on Binaural STI as discussed in section 4.2 and shown in figure 4.5. In this case, the overall slopes of the curves look almost the same where Horseshoe shows a little bit less decrease and Vineyard shows a slightly higher decrease than the others. However, considering the fact that almost all of the curves show a total drop of almost 0.01, the mentioned differences seem to be negligible. Therefore, it can be concluded that shape has no moderating effect on the decrease in Binaural STI with the increasing of the ${RT^*}_{125}$.

5.2 Conclusion

Four room shapes, volumes and reverberation times were investigated where the RT at 125 Hz and 250 Hz octave bands where changed in 7 steps. The speech intelligibility was measured using STI and Binaural Benefit together with a value in which the effect of both is incorporated, denoted by Binaural STI.
The results show that the change of reverberation time at octave band 125 Hz shows no significant effect on STI or Binaural Benefit while this change reveals a significant effect on Binaural STI in this study.
As far as Binaural STI is concerned, Shoebox appears to be the best room shape showing 0.04 better Binaural STI value than the worst case Fan shaped. In this regard, Horseshoe and Vineyard stand on the second and third places, respectively. The decrease of Binaural STI as a result of increasing the reverberation time at 125 Hz octave band is more uniform in case of Shoebox and Fan shaped compared to the other two room shapes.
Change of the reverberation time at 250-Hz octave band shows a significant effect on all three speech intelligibility measures investigated in this study (STI, Binaural Benefit and Binaural STI) and all of these intelligibility measures were in a statistically significant way decreased by the increase of 250-Hz octave band RT.
Additionally, considering the Binaural STI as the dependent variable in the statistical analysis, a higher variance of data could be explained by the independent variables, compared to the cases where STI or Binaural Benefit were considered as dependent variables. It could suggest that incorporating the Binaural Benefit in the STI, has successfully given a better assessment of intelligibility as a function of the independent variables. (Adjusted R squared = 0.86, 0.80 and 0.94, for the cases of STI, Binaural Benefit and Binaural STI, respectively.)
In case of STI, the room shape Vineyard shows the best performance where Fan shaped is still on the last place and Horseshoe and Shoebox come in between at the second and third places, respectively. On the other hand, in case of Binaural Benefit, Shoebox shows the best performance. The reason why Shoebox is at the first place when the binaural hearing comes into play, could be that this room shape can provide better early lateral reflections which support the speech intelligibility. It can be seen that this effect is so strong that it remains even when Binaural Benefit is combined with STI and makes Shoebox the best even when Binaural STI is concerned.
Furthermore, the parameters Overall RT, Volume and Shape show no significant moderating effect on the change of the discussed intelligibility measures as a function of 125-Hz or 250-Hz octave band RT.
The results of this contribution suggest that besides taking measures to reduce the reverberation time in frequency bands, typically between 500 Hz and 4 kHz, low-frequency reverberation time (specifically, at 125-Hz and 250-Hz octave bands) should also be considered as a critical factor in room acoustical designs where speech intelligibility is an important design parameter.
In further studies, the presented results could be validated by means of performing auralizations and listening tests to compare the outcomes, especially in case of $RT = 0.5\ sec$ and $Volume = 16000\ m^3$, where the binaural benefit model could not deliver realistic results to be incorporated in STI.

Bibliography

[1] C. V. Pavlovic. Classroom acoustics for the 21st century. *Proceedings of the 19th International Congress on Acoustics, Madrid, Spain*, 2007.

[2] H. V. Fuchs. *«Raumakustische Grundlagen für größere Räume» in: Raum-Akustik und Lärm-Minderung*. Springer, 4th edition, 2017; doi: 10.1007/978-3-662-53163-1.

[3] A. C. Gade. *«Acoustics in Halls for Speech and Music» in: T. D. Rossing (Ed.), Handbook of Acoustics*. Springer, 2007.

[4] W. C. Sabine. Collected papers on acoustics. *Harvard University Press, Cambridge*, 1923.

[5] W. Ahnert und H. P. Tennhardt. *«Raumakustik» in: S. Weinzierl (Hrsg.), Handbuch der Audiotechnik*. Springer, 2008.

[6] H. Kuttruff. *«The sound field in a closed space» in: Room Acoustics*. Spon Press, 4th edition, 2001.

[7] M. Hodgson. Experimental evaluation of the accuracy of the sabine and eyring theories in the case of non-low surface absorption. *The Journal of the Acoustical Society of America*, 94:835–840, 1993.

[8] L. L. Beranek. Analysis of Sabine and Eyring equations and their application to concert hall audience and chair absorption. *The Journal of the Acoustical Society of America*, 120:1399–1410, 2006.

[9] V. L. Jordan. Acoustical criteria for auditoriums and their relation to model techniques. *The Journal of the Acoustical Society of America*, 47:408–412, 1970.

[10] H. Kuttruff. *«Characterisation of subjective effects» in: Room Acoustics*. Spon Press, 4th edition, 2001.

[11] H. Kuttruff. *«Some facts on sound waves, sources and hearing» in: Room Acoustics*. Spon Press, 4th edition, 2001.

[12] S. A. Fulop. *«Introduction» in: Speech Spectrum Analysis*. Springer, 2011.

[13] S. A. Fulop. *«Phonetics and signal processing» in: Speech Spectrum Analysis*. Springer, 2011.

[14] C. V. Pavlovic. Derivation of primary parameters and procedures for use in speech intelligibility predictions. *The Journal of the Acoustical Society of America*, 82:413–422, 1987.

[15] R. W. Young. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2:567–577, 1994.

[16] L. Nijs and M. Rychtáriková. Calculating the optimum reverberation time and absorption coefficient for good speech intelligibility in classroom design using U50. *Acta Acustica united with Acustica*, 97:93–102, 2011.

[17] B. Hagerman. Reliability in the determination of speech reception threshold (SRT). *Scandinavian Audiology*, 8:4, 195-202, 1979; doi: 10.3109/01050397909076321.

[18] K. S. Rhebergen and N. J. Versfeld. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117:2181–2192, 2005.

[19] W. Reichardt; O. A. Alim and W. Schmidt. Definition and basis of making an objective evaluation to distinguish between useful and useless clarity defining musical performances. *Acta Acustica united with Acustica,*, 32:126–137, 1975.

[20] R. Höhne and G. Schroth. Zur Wahrnehmbarkeit von Deutlichkeits- und Durchsichtigkeitsunterschieden in Zuhörersälen. *Acustica*, 81:309–319, 1995.

[21] J. S. Bradley. Predictors of speech intelligibility in rooms. *The Journal of the Acoustical Society of America*, 80:837–845, 1986.

[22] R. Thiele. Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen. *Acta Acustica united with Acustica*, 3:291–302, 1953.

[23] G. Boré. Kurzton-Meßverfahren zur punktweisen Ermittlung der Sprachverständlichkeit in Lautsprecherbeschallten Räumen. *Dissertation, Technische Hochschule, Aachen*, 1956.

[24] V. Peutz. Articulation loss of consonants as a criterion for speech transmission in a room. *The Journal of the Audio Engineering Society*, 19:915–919, 1971.

[25] T. Houtgast and H. J. M. Steeneken. Evaluation of speech transmission channels by using artificial signals. *Acta Acustica united with Acustica*, 25:355–367, 1971.

[26] T. Houtgast and H. J. M. Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *The Journal of the Acoustical Society of America*, 28:66–73, 1973.

[27] T. Houtgast; H. J. M Steeneken ; R. Plomp. Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics. *Acta Acustica united with Acustica*, 46:60–72, 1980.

[28] R. Plomp; H. J. M. Steeneken and T. Houtgast. Predicting speech intelligibility in rooms from the modulation transfer function. II: Mirror image computer model applied to rectangular rooms. *Acta Acustica united with Acustica*, 46:73–81, 1980.

[29] J. Cowan. *«Building Acoustics» in: T. D. Rossing (Ed.), Handbook of Acoustics*. Springer, 2007.

[30] H. Kuttruff. *«Geometrical room acoustics» in: Room Acoustics*. Spon Press, 4th edition, 2001.

[31] D. Schröder and T. Lentz. Real-time processing of image sources using binary space partitioning. *The Journal of the Audio Engineering Society*, 54:604–619, 2006.

[32] J. B. Allen and D. A. Berkly. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943–950, 1979.

[33] J. Borish. Extension of the image source model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75:1827–1836, 1984.

[34] A. Kulowski. Algorithmic represenation of the ray tracing technique. *Applied Acoustics*, 18:449–469, 1985.

[35] H. Kuttruff. *«Design considerations and design procedures» in: Room Acoustics*. Spon Press, 4th edition, 2001.

[36] American tentative standards for sound level meters Z24.3-1936 for measurement of noise and other sounds. *The Journal of the Acoustical Society of America*, 147-152, 1936.

[37] B. C. J. Moore. *«The Perception of Loudness» in: An Introduction to the Psychology of Hearing.* Brill, 6th edition, 2013.

[38] B. C. J. Moore. *«The nature of sound and the structure and function of the auditory system» in: An Introduction to the Psychology of Hearing.* Brill, 6th edition, 2013.

[39] R. L. Wegel and C. E. Lane. The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review Journals Archive*, 23:266–285, 1924.

[40] N. Durlach. Auditory masking: Need for improved conceptual structure. *The Journal of the Acoustical Society of America*, 120:1787–1790, 2006.

[41] J. M. Pickett. Low-frequency noise and methods for calculating speech intelligibility. *The Journal of the Acoustical Society of America*, 31:1259–1263, 1959.

[42] B. A. Edmonds and J. F. Culling. The spatial unmasking of speech: Evidence for better-ear listening. *The Journal of the Acoustical Society of America*, 120:1539–1545, 2006.

[43] E. C. Cherry. Some experiments on the recognition of speech with one and with two ears. *The Journal of the Acoustical Society of America*, 25:975–979, 1953.

[44] I. J. Hirsh. The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America*, 20:9536–544, 1948.

[45] J. C. R. Licklider. The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20:150–159, 1948.

[46] M. Lavandier and J. F. Culling. Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, 127:387–399, 2010.

[47] S. Jelfs; J. F. Culling and M. Lavandier. Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, 275:96–104, 2011.

[48] R. Patterson; I. Nimmo-Smith; J. Holdsworth and P. Rice. An efficient auditory filterbank based on the gammatone function. *Institute of Acoustics Speech Group on Auditory Modelling.*, 1987.

[49] J. F. Culling. Signal-processing software for teaching and research in psychoacoustics under unix and x-windows. *Behavior Research Methods, Instruments, Computers.*, 28:376–382, 1996.

[50] J. F. Culling; M. L. Hawley; and R. Y. Litovsky. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *The Journal of the Acoustical Society of America*, 116:1057–1065, 2004.

[51] J. F. Culling; M. L. Hawley; and R. Y. Litovsky. Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [j. acoust. soc. am. 116, 1057 (2004)]. *The Journal of the Acoustical Society of America*, 118:552, 2005.

[52] ANSI. Methods for the calculation of the speech intelligibility index. *American National Standards Institute, New York.*, 1997.

[53] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74:750–753, 1983.

[54] S. J. van Wijngaarden and R. Drullman. Binaural Intelligibility Prediction Based on the Speech Transmission Index. *The Journal of the Acoustical Society of America*, 123:4514–4523, 2008.

[55] N. I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, 35:1206–1218, 1963.

[56] J. S. Bradley; H. Sato and M. Picard. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, 113:3233–3244, 2003.

[57] S. Wu; J. Peng and Z. Bi. Chinese speech intelligibility in low frequency reverberation and noise in a simulated classroom. *Acta Acustica united with Acustica*, 100:1067–1072, 2014.

[58] L. Polich and R. S. Segovia. Acoustic conditions in classrooms for hearing impaired in nicaragua. *Journal of the Academy of Rehabilitative Audiology*, 31:29–43, 1999.

[59] GB 4959-85. Methods of measurement for the characteristics of sound reinforcement in auditoria. *Standard of P.R. China*, 1985.

[60] ANSI S3.2-1989 (R 1999): Measuring the intelligibility of speech over communication systems. *American National Standards Institute, ANSI, Washington, DC*, 1999.

[61] E. Mommertz; P. Reents and G. Graber. Bedeutung kurzer Nachhallzeiten bei tiefen Frequenzen für die raumakustische Qualität in Unterrichtsräumen. *DAGA*, 2006.

[62] X. Q. Cha; H. V. Fuchs. Some views on reverberation time characteristics. *Proceedings of the 10th architectural physics, South China University of Technology Press*, 37-40, 2008.

[63] H. V. Fuchs. Din 18041:2016 - Eine Norm im raumakustischen " Abseits". *Lärmbekämpfung, VDI Fachmedien GmbH Co. KG*, 14:123–132, 2019.

[64] J. M. Pickett. Low frequency noise and methods for calculating speech intelligibility. *The Journal of the Acoustical Society of America*, 31:1269–127, 1951.

[65] R. Beutelmann; T. Brand and B. Kollmeier. Revision, extension and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, 127:2479–2497, 2010.

[66] A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple talker conditions. *Acta Acustica united with Acustica*, 86:117–128, 2000.

[67] V. Hohmann. Frequency analysis using a gammatone filterbank. *Acta Acustica united with Acustica*, 88:433–442, 2002.

[68] D. Schröder and M. Vorländer. Raven: A real-time framework for the auralization of interactive virtual environments. *European Acoustics Association*, 1541-1546, 2011.

[69] U. M. Stephenson. An energetic approach for the simulation of diffraction within ray tracing based on the uncertainty relation. *Acta Acustica united with Acustica*, 96:516–535, 2010.

[70] D. Schröder. Physically-based real-time auralization of interactive virtual environments. *PhD thesis, RWTH Aachen university*, 2011.

[71] DIN EN ISO 3382-1: 2009, Akustik–Messung von Parametern der Raumakustik – Teil 1: Aufführungsräume (ISO 3382-1:2009); Deutsche Fassung EN ISO 3382-1:2009.

[72] M. Holden. *«Creating the Building» in: Acoustics of Multi-Use Performing Art Centers*. CRC Press, 2016.

[73] J. Greif; D. Ackermann; O. Kokabi and S. Weinzierl. Kann man die Form eines Konzertsaales hören? Ein audiovisueller Test in simulierten 3D-Umgebungen. *DAGA 2020 Hannover*, 1153-1156, 2020.

[74] F. Brinkmann; A. Lindau; Stefan Weinzierl; G. Geissler; S. van de Par; M. Müller-Trapet; R. Opdam and M. Vorländer. The FABIAN head-related transfer function data base, 2017; doi:10.14279/depositonce-5718.

[75] P. Søndergaard and P. Majdak. *«The Auditory Modeling Toolbox» in: J. Blauert (Ed.), The Technology of Binaural Listening.* Springer, 2013.

[76] O. Kokabi; F. Brinkmann and S. Weinzierl. Segmentation of binaural room impulse responses for speech intelligibility prediction. *The Journal of the Acoustical Society of America*, 144:2793–2800, 2018; doi: 10.14279/depositonce-6725.5.

[77] J. Rennies; T. Brand and B. Kollmeier. Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *The Journal of the Acoustical Society of America*, 130(5):2999–3012, 2011.

[78] P. Larm and V. Hongisto. Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index. *The Journal of the Acoustical Society of America*, 119:1106–1117, 2006.

[79] R. H. Myers and D. C. Montgomery. A tutorial on generalized linear models. *Journal of Quality Technology*, 29:3:274–291, 1997; doi: 10.1080/00224065.1997.11979769.

Appendix

Source and receiver coordinates, as explained in section 3.1.3.

	x (m)	y (m)	z (m)
R1	2.10	-0.39	2.31
R2	3.07	-0.27	0.90
R3	3.49	-0.17	-1.59
R4	4.87	0.10	0.91
R5	4.51	0.04	2.76
R6	1.69	-0.58	-0.11

Table 5.1: Receiver positions, Volume = 250 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
S1	-1.52	-0.22	1.81
S2	-2.12	-0.24	0.03

Table 5.2: Source positions, Volume = 250 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
R1	3.96	0.87	-0.11
R2	4.24	0.90	1.36
R3	6.36	0.87	1.21
R4	8.05	0.81	1.03
R5	9.44	0.89	1.16
R6	9.07	0.89	-0.49

Table 5.3: Receiver positions, Volume = 250 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
S1	-1.52	-0.22	1.81
S2	-2.12	-0.24	0.03

Table 5.4: Source positions, Volume = 250 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
R1	9.56	1.24	5.23
R2	9.31	0.86	1.80
R3	9.59	1.23	-1.66
R4	7.61	0.66	-5.61
R5	5.84	-0.11	-1.43
R6	7.64	0.42	0.42

Table 5.5: Receiver positions, Volume = 1000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
S1	-2.39	-0.24	-1.26
S2	-1.05	-0.24	-0.79
S3	-0.05	-0.22	0.56

Table 5.6: Source positions, Volume = 1000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
R1	15.14	2.67	0.55
R2	6.96	0.87	-1.69
R3	12.95	0.88	-0.27
R4	12.99	0.86	2.01
R5	8.74	0.85	0.48
R6	10.03	0.86	2.21

Table 5.7: Receiver positions, Volume = 1000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
S1	1.86	1.40	-2.53
S2	2.88	1.42	-0.40
S3	2.83	1.40	1.91

Table 5.8: Source positions, Volume = 1000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
R1	17.70	2.35	11.57
R2	13.22	0.75	4.22
R3	19.25	2.37	6.39
R4	15.93	1.49	1.99
R5	19.60	2.67	-3.52
R6	10.93	0.47	8.47

Table 5.9: Receiver positions, Volume = 4000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
S1	1.17	0.00	6.16
S2	-0.84	0.01	1.74
S3	0.22	0.05	3.67
S4	2.82	0.06	2.38

Table 5.10: Source positions, Volume = 4000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
R1	8.39	0.88	3.87
R2	21.23	6.24	1.29
R3	15.46	0.86	2.11
R4	8.12	6.17	7.04
R5	16.27	3.76	-3.87
R6	7.47	0.88	-1.83

Table 5.11: Receiver positions, Volume = 4000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
S1	-0.89	1.47	1.86
S2	1.10	1.47	0.77
S3	1.13	1.47	-2.52
S4	-0.74	1.48	5.21

Table 5.12: Source positions, Volume = 4000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
R1	13.16	1.15	4.24
R2	11.44	1.16	-3.05
R3	7.24	1.02	4.08
R4	17.39	8.61	-0.91
R5	5.68	0.91	6.65
R6	9.60	1.07	-0.77

Table 5.13: Receiver positions, Volume = 4000 m^3, Shape: Horseshoe.

	x (m)	y (m)	z (m)
S1	0.82	1.45	6.83
S2	-0.66	1.46	4.45
S3	-1.77	1.46	0.53
S4	-1.79	1.46	-2.61

Table 5.14: Source positions, Volume = 4000 m^3, Shape: Horseshoe.

	x (m)	y (m)	z (m)
R1	13.02	2.72	-0.31
R2	12.02	2.76	8.36
R3	5.36	5.79	12.29
R4	17.30	4.35	-1.48
R5	9.35	1.70	6.32
R6	16.88	5.51	-5.59

Table 5.15: Receiver positions, Volume = 4000 m^3, Shape: Vineyard.

	x (m)	y (m)	z (m)
S1	3.54	1.47	-0.28
S2	2.31	1.46	2.28
S3	1.56	1.46	4.90
S4	-0.18	1.47	1.05

Table 5.16: Source positions, Volume = 4000 m^3, Shape: Vineyard.

	x (m)	y (m)	z (m)
R1	26.03	3.10	2.54
R2	28.31	4.12	-7.06
R3	14.93	1.29	-9.51
R4	14.22	0.18	-1.06
R5	13.34	-0.15	4.99
R6	6.44	-0.94	13.45
R7	31.48	4.61	-0.58
R8	22.21	2.51	12.08
R9	16.36	2.21	20.02
R10	26.01	5.00	22.65

Table 5.17: Receiver positions, Volume = 16000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
S1	2.49	-0.08	-2.98
S2	-4.40	-0.08	3.33
S3	-2.18	-0.09	-0.67
S4	0.33	-0.08	6.58
S5	4.51	-0.09	2.20

Table 5.18: Source positions, Volume = 16000 m^3, Shape: Fan shaped.

	x (m)	y (m)	z (m)
R1	16.63	5.39	-7.25
R2	27.74	5.42	11.54
R3	16.87	0.59	-0.33
R4	27.24	9.18	-7.35
R5	29.81	0.59	4.37
R6	12.74	0.60	6.59
R7	35.49	9.24	0.73
R8	16.14	9.08	11.03
R9	24.44	0.60	-2.59
R10	35.23	5.40	7.07

Table 5.19: Receiver positions, Volume = 16000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
S1	3.86	1.35	-5.30
S2	0.82	1.35	6.16
S3	-0.43	1.34	-4.89
S4	0.98	1.35	0.96
S5	-1.88	1.34	2.20

Table 5.20: Source positions, Volume = 16000 m^3, Shape: Shoebox.

	x (m)	y (m)	z (m)
R1	8.55	0.73	10.52
R2	18.66	3.56	14.41
R3	12.73	0.88	5.68
R4	22.76	1.21	3.75
R5	28.82	8.21	0.84
R6	25.98	5.89	10.06
R7	16.73	1.05	-4.27
R8	19.53	12.85	-9.40
R9	23.41	10.58	-7.47
R10	9.75	0.76	0.30

Table 5.21: Receiver positions, Volume = 16000 m^3, Shape: Horseshoe.

	x (m)	y (m)	z (m)
S1	-1.32	1.44	10.19
S2	-1.87	1.43	4.70
S3	1.48	1.45	-7.28
S4	-3.33	1.44	-6.72
S5	0.48	1.43	-0.07

Table 5.22: Source positions, Volume = 16000 m^3, Shape: Horseshoe.

	x (m)	y (m)	z (m)
R1	-4.43	4.87	9.40
R2	-0.11	7.18	13.65
R3	18.16	3.99	9.22
R4	20.84	4.08	0.94
R5	18.79	3.65	-4.52
R6	26.62	6.76	-7.01
R7	28.31	6.92	0.38
R8	27.88	7.35	5.67
R9	11.78	1.25	-0.60
R10	10.60	1.69	-9.79

Table 5.23: Receiver positions, Volume = 16000 m^3, Shape: Vineyard.

	x (m)	y (m)	z (m)
S1	-10.05	4.54	-6.95
S2	-12.54	5.93	-8.91
S3	-14.35	6.92	-2.18
S4	-11.52	5.33	-1.37
S5	-10.68	4.94	6.36

Table 5.24: Source positions, Volume = 16000 m^3, Shape: Vineyard.