Kai Jussila

The effects of background noise and test subject on the perceived amount of bass in phase-modified harmonic complex tones

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 25.11.2013

Thesis supervisor:

Prof. Ville Pulkki

Thesis instructor:

M.Sc. (Tech.) Mikko-Ville Laitinen

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Kai Jussila

Title: The effects of background noise and test subject on the perceived amount of bass in phase-modified harmonic complex tones

Date: 25.11.2013 Language: English Number of pages:8+62

Department of Signal Processing and Acoustics

Professorship: Acoustics and Audio Signal Processing Code: S-89

Supervisor: Prof. Ville Pulkki

Instructor: M.Sc. (Tech.) Mikko-Ville Laitinen

The perception of timbre is closely related to the relative levels produced by a sound in each frequency band, called 'critical band', in the cochlea. The magnitude spectrum defines the relative levels and phase spectrum the relative phases of the frequency components in a complex sound. Thus, the timbre of sound depends often only on the magnitude spectrum. However, several studies have shown that the timbre of certain complex sounds can be affected by modifying only the phase spectrum.
Moreover, a recent study has shown that with certain modifications of only the phase spectrum of a 'phase-sensitive' harmonic complex tone, the perceived level of bass changes. That experiment was conducted using two synthetic harmonic complex tones in which adjacent frequency components have a phase-shift of -90° and 90°, respectively, and the fundamental component is in cosine-phase. The greatest difference in perceived level of bass was found at the fundamental frequency of 50 Hz and it corresponds to a 2 – 4-dB amplification of the magnitude spectrum at low frequencies. However, this effect was reported to vary substantially between individuals. Moreover, the differences were found to be easier to detect in the presence of background noise.
The aim of this thesis was to investigate further the roles of background noise and the individual in the perceived level of bass in the phase-sensitive tones. Two formal listening tests were conducted accordingly using headphones. Firstly, the effect of background noise on the discrimination of the phase-sensitive tones based on the perceived level of bass was studied. The effect of increasing background noise level on the perceived loudness difference was found not to be statistically significant, but a trend could be seen towards increasing loudness difference. Additionally, the results indicate that the overall perceived loudness of the test tones decreases with increasing level of background noise. Secondly, an experiment was conducted to find the preferred value of the constant phase shift between adjacent components that produces a tone with the perceptually loudest bass for different individuals. The results show that individuals hear the phase spectrum required to produce the perception of the loudest bass statistically significantly differently from each other.

Keywords: psychoacoustics, timbre, perception, phase spectrum

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Kai Jussila

Työn nimi: Taustakohinan vaikutus ja yksilölliset erot vaihemuokattujen harmonisten äänien bassokkuuden havaitsemisessa

Päivämäärä: 25.11.2013 Kieli: Englanti Sivumäärä:8+62

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittelytekniikka Koodi: S-89

Valvoja: Prof. Ville Pulkki

Ohjaaja: DI Mikko-Ville Laitinen

Äänenvärin havaitseminen liittyy läheisesti äänen tuottamiin suhteellisiin tasoihin simpukassa eri taajuuskaistoilla, joita kutsutaan kriittisiksi kaistoiksi. Äänen magnitudispektri määrittää sen taajuuskomponenttien suhteelliset voimakkuudet ja vaihespektri niiden suhteelliset vaiheet. Äänenväri siis riippuu usein pelkästään magnitudispektristä. Tutkimustulokset ovat kuitenkin osoittaneet, että tietyn tyyppisten äänien äänenväriä voidaan muuttaa myös pelkästään vaihespektriä muuttamalla.
Tämän lisäksi aiempi tutkimus on osoittanut, että muuttamalla harmonisen äänen vaihespektriä tietyllä tavalla havaittu bassokkuus muuttuu. Tällaiset äänet ovat siis 'vaiheherkkiä'. Kyseisessä tutkimuksessa käytettiin kahta tällaista vaihemuokattua ääntä, joista toisessa taajuuskomponenttien välillä oli -90 asteen ja toisessa 90 asteen vaihe-ero, ja perustaajuuskomponentti oli molemmissa kosinivaiheessa. Tutkimus osoitti, että suurin bassokkuusero havaitaan matalilla perustaajuuksilla ja se vastaa keskimäärin 2 – 4 dB:n vahvistusta magnitudispektrissä matalilla taajuuksilla. Tämä ilmiön suuruus riippui kuitenkin huomattavasti testihenkilöstä. Lisäksi huomattiin, että bassokkuuserot ovat helpompia kuulla taustakohinan kanssa.
Tämän työn tavoitteena oli tutkia edelleen taustakohinan merkitystä ja yksilöllisiä eroja tällaisten vaiheherkkien äänien bassokkuuden havaitsemisessa. Kaksi formaalia kuuntelukoetta järjestettiin käyttäen kuulokkeita. Ensiksi tutkittiin taustakohinan vaikutusta kyseisten äänien bassokkuuserojen kuulemiseen olettaen, että nämä erot ovat kuultavissa äänekkyyseroina. Tulokset viittaavat, että taustakohinan tason nousun vaikutus testiäänien äänekkyyseroon ei ole tilastollisesti merkittävä, mutta on lähellä merkittävyyden rajaa ja trendi on nähtävissä äänekkyyseron kasvulle. Lisäksi nähdään, että kyseisten vaiheherkkien äänien yleinen äänekkyys laskee kun taustakohinan tasoa voimistetaan. Toiseksi tutkittiin sitä, minkä vaihespektrin omaavan äänen eri ihmiset kuulevat bassokkaimpana. Tulokset osoittavat, että testihenkilöt eroavat siinä, minkä vaihespektrin omaavan äänen he kuulevat bassokkaimpana, ja että tämä ero on tilastollisesti merkittävä.

Avainsanat: psykoakustiikka, äänenväri, havaitseminen, vaihespektri

Acknowledgements

I would like to thank my instructor Mikko-Ville for encouraging guidance and feedback during the work process. I would also like to thank my supervisor Ville for giving me the opportunity to work on this thesis.

Vantaa, 20.11.2013

Kai Jussila

Contents

Abbreviations and symbols

ANOVA	analysis of variance
BM	basilar membrane (of the cochlea)
CB	critical bandwidth
CF	characteristic frequency
CPH	cosine-phase wave
DTFT	discrete-time Fourier transform
ERB	Effective Rectangular Bandwidth
RPH	random-phase wave
ω	angular frequency

List of Figures

Chapter 1

Introduction

A sound can be represented in the time domain as a waveform and in the frequency domain as a spectrum. The most simple form of sound is the sine wave, which is seen in the frequency domain as a peak corresponding to its frequency. If a sound is harmonic, the waveform is periodic, i.e., the waveform has a cycle that repeats every period of the sound signal. The period refers to a time sequence between two sequential points that are in the same phase. In the frequency domain, a harmonic sound is represented as spectral peaks with certain magnitudes for each of its sinusoidal components. This representation is called the magnitude spectrum. A sound that consists of three or more sinusoidal components is referred to as a complex tone. By changing the relative phases of the frequency components, the waveform of the sound can be changed without affecting the magnitudes of the spectral peaks corresponding to those components. This means that the starting phase of one or more of the individual sine waves comprising the harmonic sound is made to differ from the others. By doing this, the phase spectrum of the sound is altered without affecting the magnitude spectrum.

Sounds can be captured and represented as either analog or digital audio signals, which can then be reproduced as sounds. With the help of such representation sounds can be also synthesized and processed in many ways for different applications. Sounds can be characterized with duration, pitch, loudness and timbre. Timbre is defined as that attribute of sound, in terms of which two sounds that have the same pitch and loudness can be judged to sound dissimilar (ASA, 1960). The timbre of a complex tone is mainly defined by the relative levels of its spectral components (Moore, 1997). If the lowest components of a wide-band complex tone are relatively loud, the tone sounds as containing more bass. On the other hand, if the high frequency components are emphasized, the sound becomes brighter or sharper.

Already since the 19th century, the perception of the phase spectrum of audio signals has been a topic of research. This matter was first studied by Ohm (1843) and von Helmholtz (1863) who both suggested that people cannot hear changes in the phase spectrum. Therefore, for a long period of time, it was thought that the timbre of a sound is determined only by its magnitude spectrum.

However, several studies thereafter (Mathes and Miller, 1947; Plomp and Steeneken, 1969; Patterson, 1987; Moore and Glasberg, 1989; Moore, 2002; Laitinen et al., 2013) have proven with synthetic signals that human hearing is phase-sensitive. Phase-sensitivity has been

studied under many research topics such as hall acoustics (Griesinger, 2010) and timbre perception of, e.g., vowels (Plomp and Steeneken, 1969). For example, a harmonic signal with all frequency components starting at their maximum amplitudes, i.e., in cosine-phase, is perceived as different from a signal with the same magnitude spectrum but random starting phases (Patterson, 1987). Such randomization of phase occurs in nature due to reflections when the cosine-phase wave, or any sound, propagates in a room. Similar phase-randomization occurs also in decorrelation techniques in spatial audio coding (Pulkki and Merimaa, 2006).

Human perception of sound can be considered to occur in frequency bands called auditory filters or critical bands (Fletcher, 1940; Scharf, 1961). Even small changes in the phase within the outputs of these auditory filter channels yield perceivable changes, whereas large changes between auditory channels are needed for them to be perceived (Patterson, 1987). Moreover, recent research suggests that the sensitivity to phase holds also for certain natural sounds such as transient-like signals (Laitinen et al., 2011), and anechoic speech (Laitinen and Pulkki, 2012), trumpet and trombone signals (Laitinen et al., 2013).

In addition to the process of decorrelation, phase distortion is caused in audio signal processing also by other processes such as quantization (Rossing et al., 2002). Despite the research results described above, many audio coding schemes (Pulkki, 2007; Herre et al., 2008; Sergi, 2013) employ the assumption of insensitivity to phase for all types of signals. This assumption is based on the fact that although there are signals for which phase-sensitivity holds, listening tests measuring the performance of audio coding techniques (Herre et al., 2008; Vilkamo et al., 2009) show that for the majority of signals we are not sensitive to changes in the phase spectrum.

The motivation for this thesis originates from recent research about the perceptual differences due to phase spectrum changes in harmonic complex tones, which can be considered as the synthetic correspondent to the natural 'phase-sensitive' signals described earlier. In a study by Laitinen et al. (2013) the perceived amount of bass in phase-modified harmonic complex tones was investigated with a formal listening test. It was found that a certain phase modification of the harmonic tone led to a bass boost that was quantitatively larger than that of a magnitude spectrum modification in which the fundamental component was amplified by 1 dB. This was reported to be true at fundamental frequencies of 50 and 100 Hz. They performed the test with noise in the background, because in informal listening it had been noticed to make the discrimination easier. Additionally, Laitinen et al. (2013) reported large individual differences in the perceived level of bass. These findings led to the experiments performed for this thesis. The role of different factors were investigated considering the effect of phase spectrum changes on the perception of bass. Moreover, two formal listening tests were performed to study further the roles of background noise and the individual in this effect. More specifically, the aim of this thesis is first to define how much background noise affects the perceived differences in bass between phase-modified harmonic complex tones. Secondly, it is aimed to investigate how the modification of the phase spectrum required for the tone to be perceived as the loudest in bass, varies between individuals.

This thesis is divided into background, listening test, and conclusion parts. First, basic audio signal processing theory is presented in chapter 2. Human auditory system, including the structure and function of the ear, and basic concepts of auditory perception, is discussed in chapter 3. Moreover, the perception of phase spectrum changes is discussed in chapter 4.

Fundamentals of audio evaluation and statistics are addressed in chapter 5. Thereafter, the motivation, research methods and results of two listening tests are discussed in chapter 6. Finally, in chapter 7, conclusions are made from the results of the thesis.

Chapter 2

Audio signal processing

In this chapter, the basics of audio signal processing are discussed. As a basis for the discussion in this thesis on phase spectrum modifications the following concepts are introduced. First, the discrete-time Fourier transform and the frequency response of a linear time-invariant system are described. Second, the phase and group delays associated with the frequency response of a system are then discussed. Finally, three classes of phase response characteristics are addressed: zero-phase, linear-phase, and minimum-phase and maximum-phase responses.

2.1 The discrete-time Fourier transform

First, the definition of the discrete-time Fourier transform is presented and its basic properties are introduced. The proofs can be found in the literature, e.g., Mitra (2006). As digital signal processing is based on sampling (Mitra, 2006, p. 1–117), we consider only discrete-time signals in this section for practicality. First of all, a discrete-time sequence $x[n]$ is represented in terms of the complex exponential sequence $e^{j\omega n}$ by the discrete-time Fourier transform (DTFT) of the signal. The discrete-time Fourier transform, or simply Fourier spectrum, of a sequence $x[n]$ is

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}, \tag{2.1}$$

where ω is the real (angular) frequency variable. $X(e^{j\omega})$ is a continuous complex function of ω with a period 2π. If the Fourier transform representation of the sequence exists, it is unique, and the original sequence can be computed from the transform-domain representation by an inverse Fourier transform, or Fourier integral. I.e., the original sequence $x[n]$ can be computed from $X(e^{j\omega})$ as

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n} \, \mathrm{d}\omega. \tag{2.2}$$

The integral is commonly computed over the interval $[-\pi, \pi]$, as in Equation 2.2, even though any interval of the angle 2π could be chosen. Equation 2.1 is referred to as the analysis equation, and, on the other hand, Equation 2.2 is called the synthesis equation. Together Equations 2.1 and 2.2 form a discrete-time Fourier transform pair for the sequence $x[n]$. (Mitra, 2006, p. 117–136)

The Fourier spectrum $X(e^{j\omega})$ can be expressed in rectangular form as

$$X(e^{j\omega}) = X_{re}(e^{j\omega}) + jX_{im}(e^{j\omega}), \tag{2.3}$$

where $X_{re}(e^{j\omega})$ and $X_{im}(e^{j\omega})$ are the real and imaginary parts, respectively, which are real functions of ω. It can alternately be expressed in polar form as

$$X(e^{j\omega}) = |X(e^{j\omega})|e^{j\theta(\omega)}, \tag{2.4}$$

where $|X(e^{j\omega})|$ is called the *magnitude spectrum*, and the argument $\theta(\omega)$ is called the *phase spectrum*, which are both again real functions of ω. Note from the periodicity of the Fourier spectrum that the phase spectrum is not uniquely specified for the DTFT for all values of ω. (Mitra, 2006, p. 117–136)

When the computed phase spectrum is outside the range $[-\pi, \pi]$, the phases are computed modulo 2π to bring them back to the desired range. That is why the phase spectrum plots of some sequences have discontinuities of 2π radians. Sometimes it is useful to consider an alternative representation of the phase spectrum that is a continuous function of ω, and which is derived from the original phase spectrum by removing the discontinuities. The phase spectrum $\theta(\omega)$ can be defined unequivocally by its derivative and can be expressed as

$$\theta(\omega) = \int_0^\omega \left[\frac{\mathrm{d}\theta(\eta)}{\mathrm{d}\eta}\right] \mathrm{d}\eta, \tag{2.5}$$

with the constraint

$$\theta(0) = 0. \tag{2.6}$$

This alternative representation of the phase spectrum is called the *unwrapped phase spectrum* of $X(e^{j\omega})$. (Mitra, 2006, p. 145–171)

2.2 The frequency response: magnitude and phase responses

In order to understand the response of a system with respect to the phase of an audio signal, the concept of frequency response of a system is discussed next. We take a look particularly at the frequency response of a linear time-invariant (LTI) discrete-time system. Note that if there are any sinusoidal components at new frequencies in the output, the system is either nonlinear or time-variant or both. For linear systems the superposition principle always holds. Most discrete-time signals can be presented as a linear combination

of a (possibly infinite) number of sinusoidal signals. Therefore, knowing the response of an LTI system to a single sinusoidal signal and using the superposition property, we can define the response of that system to more complex signals. The output $y[n]$ of an LTI system is calculated from a convolution sum of the impulse response $h[n]$ and a complex exponential input sequence $x[n] = e^{j\omega n}$. It results, that the output is also a complex exponential sequence multiplied by a complex function $H(e^{j\omega})$, which is the frequency response of the system.

Essentially, frequency response $H(e^{j\omega})$ is, similarly to Equation 2.4, the DTFT of the impulse response $h[n]$ of a system, with the quantity $|H(e^{j\omega})|$ being the *magnitude response*, and the quantity $\theta(\omega)$ being the *phase response* of that system. Whenever there is a zero in the magnitude response, a jump by an amount of π can be seen in the phase response. Note however, that these jumps should not be confused with the 2π discontinuities discussed above. (Mitra, 2006, p. 145–171) The transfer function $H(z)$ of the system is expressed simply as a z-transform of the frequency response. The attenuation ratio can be defined from the transfer function $H(z)$ as (Blauert and Laws, 1978)

$$a(z) = -\ln|H(z)|. \tag{2.7}$$

2.2.1 Phase delay and group delay

Two more parameters of an LTI system, that are of special interest considering this thesis, are discussed next. These parameters are called *phase delay* and *group delay*, which are associated with the frequency response of a system. If the input $x[n]$ is a sinusoidal signal of frequency ω_0, the output is of the form

$$\begin{aligned} y[n] &= A|H(e^{j\omega_0})|\cos\big(\omega_0 n + \theta(\omega_0) + \phi\big) \\ &= A|H(e^{j\omega_0})|\cos\big(\omega_0\big(n - \tau_p(\omega_0)\big) + \phi\big) \end{aligned} \tag{2.8}$$

where A is real, and $\tau_p(\omega_0) = -\frac{\theta(\omega_0)}{\omega_0}$ is the phase delay. As the input, the output is also a sinusoidal signal of the same frequency ω_0 but with a phase lag of $\theta(\omega_0)$ radians. As can be seen from Equation 2.8, the output $y[n]$ is a time-delayed version of the input $x[n]$. However, in the discrete-time case considered here, the output will be a delayed replica of the input only if the phase delay is an integer in samples. Therefore, the phase delay is better understood with the underlying (physical) continuous-time waveforms associated with the input and output sequences. In the continuous-time case of a narrowband system with an arbitrary phase delay, the output is a delayed replica of the input waveform. However, in LTI systems with a wide-band frequency response the phase and group delays do not have any physical meanings. (Mitra, 2006)

In practical applications, a more interesting parameter of an audio system is the group delay. If the input sequence contains many frequency components that are not harmonically related, each component will have a different amount of phase delay when going through an LTI discrete-time system. The signal delay is then expressed with the group delay defined

as

$$\tau_g(\omega) = -\frac{\mathrm{d}\theta(\omega)}{\mathrm{d}\omega}. \tag{2.9}$$

It is assumed here that the derivative of the phase function $\theta(\omega)$ exists, i.e., that the phase spectrum is unwrapped. As in the case of the phase delay, also the group delay is better understood when considering the underlying (physical) continuous-time functions associated with the input and output sequences. From trigonometry we get that the phase delay $\tau_p(\omega_0)$ is the negative slope of the straight line from the origin to the point $[\omega_0, \theta(\omega_0)]$ on the phase function plot. The group delay $\tau_g(\omega_0)$, on the other hand, is the negative slope of the phase function $\theta(\omega)$ at the frequency ω_0. The group delay functions are always nonnegative functions of ω. (Mitra, 2006)

Generally, electroacoustic audio systems consist of LTI subsystems (Blauert and Laws, 1978), such as amplifiers, loudspeakers, earphones, etc. It is described by Equations 2.7 and 2.9 that in these systems, the frequency components of the input signal arrive at the output of the system with different attenuations, i.e., distorted. Also the input waveform becomes distorted. Moreover, the group delay response determines the starting location of each frequency bin in time. A typical group delay response of, e.g., headphones is one which increases towards low frequencies, i.e., low frequencies are delayed more than high frequencies.

2.2.2 Phase response characteristics

Different phase response characteristics are discussed next. Transfer functions can be classified based on their phase response characteristics to three classes: *zero-phase*, *linear-phase*, and *minimum-phase* or *maximum-phase* transfer functions. A transfer function has a zero-phase characteristic when its frequency response is real and nonnegative. If $H(z)$ is a real-coefficient rational z-transform with no poles on the unit circle, then the function

$$F(z) = H(z)H(z^{-1}) \tag{2.10}$$

has a zero phase on the unit circle, since we have

$$F(e^{j\omega}) = H(e^{j\omega})H(e^{-j\omega}) = |H(e^{j\omega})|^2. \tag{2.11}$$

(Mitra, 2006)

Considering a causal LTI system with a nonzero phase response, phase distortion can be avoided by allowing the output to be a delayed version of the input, i.e.

$$y[n] = x[n - D]. \tag{2.12}$$

By transforming both sides of the above equation to the frequency domain with a Fourier transform, and using the time-shifting property of Fourier transform, we have

$$Y(e^{j\omega}) = e^{-j\omega D}X(e^{j\omega}). \tag{2.13}$$

Then the frequency response of the system is

$$H(e^{j\omega}) = \frac{Y(e^{j\omega})}{X(e^{j\omega})} = e^{-j\omega D}. \tag{2.14}$$

This frequency response has a unity magnitude response and a linear phase response with a group delay of D samples at all frequencies, i.e.,

$$|H(e^{j\omega})| = 1, \qquad \tau_g(\omega) = D. \tag{2.15}$$

If an LTI system fulfills the above constraints in the frequency band of interest, it passes input signal components undistorted in both magnitude and phase in that frequency range. In practice, it is always possible to design an FIR filter with an exact linear-phase response, whereas it is not possible to design an IIR filter transfer function with an exact linear-phase response. (Mitra, 2006)

Another classification of a transfer function can be made based on its location of zeros with respect to the unit circle. Whether the zeros of a transfer function are inside or outside the unit circle, affects its phase response. A causal stable transfer function with all zeros inside the unit circle is called a minimum-phase transfer function. On the other hand, a causal stable transfer function with all zeros outside the unit circle is called a maximum-phase transfer function. Furthermore, when a transfer function has zeros inside and outside the unit circle, it is called a mixed-phase transfer function. It can be shown that the group delay $\tau_g^{H_m}(\omega)$ of a minimum-phase causal stable transfer function $H_m(z)$ is smaller than the group delay $\tau_g^{H}(\omega)$ of a non-minimum-phase causal transfer function $H(z)$ which has the same magnitude response function than that of $H_m(z)$, i.e., (Mitra, 2006)

$$\tau_g^{(H_m)}(\omega) < \tau_g^{(H)}(\omega). \tag{2.16}$$

According to Blauert and Laws (1978), group delay distortions occur in electroacoustic audio systems because common transducers, such as earphones, 'are not necessarily minimum-phase systems but show additional all-pass characteristics'. Hence, a general LTI audio system can be split into minimum-phase and all-pass subsystems, i.e., its transfer function can be written as $H(z) = H_m(z)H_a(z)$. Blauert and Laws (1978) pointed out that the additional group delay distortions caused by the all-pass characteristics cannot be corrected with usual minimum-phase audio equalizers alone. All-pass networks are then needed to perform group delay correction, which is explained further in Blauert and Laws (1978).

Chapter 3

Human hearing

In this chapter, the human auditory system is discussed. In order to understand the mechanisms underlying the perception of changes in phase spectrum, the function of the ear as well as principles of auditory perception are concerned. In section 3.1, the structure and function of the auditory system is presented, and in section 3.2, neural responses to sound are discussed. In the sections thereafter, the production of auditory sensations from the mechanisms of hearing is discussed. In section 3.3, general perceptual concepts are introduced. Thereafter, the mechanisms involved in loudness perception and pitch perception are discussed in sections 3.4 and 3.5, respectively. The mechanisms that are involved in defining timbre perception of sound are described in section 3.6.

3.1 Structure and function of the ear

The structure of the peripheral part of the auditory system of most mammals is rather similar (Moore, 1997). The structure of the human peripheral auditory system is presented in Figure 3.1. The outer ear consists of the pinna and the meatus or the auditory canal. Particularly at high frequencies, the pinna significantly modifies the sound before it enters the ear canal. The eardrum, or the tympanic membrane, vibrates and converts the air pressure variations into mechanical vibration in the middle ear. The mechanical vibration in the middle ear is transmitted through three small bones, the ossicles, called the malleus, the incus and the stapes. More popular names for these bones are the hammer, the anvil and the stirrup, respectively. The stapes is the smallest bone and makes contact with the oval window of the spiral-shaped organ in the inner ear, the cochlea. The oval window is an opening in the bony wall of the cochlea, and it is covered by a membrane. (Moore, 1997)

The role of the middle ear is to transfer the sound in air to the fluids in the cochlea efficiently (Moore, 1997). The middle ear acts as a transformer matching the acoustical impedances of air and the oval window. This occurs mainly due to the difference in the effective areas of the eardrum and the oval window, and partly due to the lever action of the ossicles. The middle ear also reduces the amount of reflected sound from the oval window and enhances sound transmission. The ossicles have small muscles attached to them. These muscles contract when we are exposed to intense sounds reducing the transmission of sound through the middle ear at low frequencies. This is called the middle ear reflex, which may

help in preventing damage to the fragile structures of the cochlea. In addition, this reflex has been shown to be activated just before speaking and thus it reduces the audibility of self-generated sounds. Furthermore, according to Moore (1997), it has been suggested that, due to the reflex, low frequencies mask less middle and high frequencies.

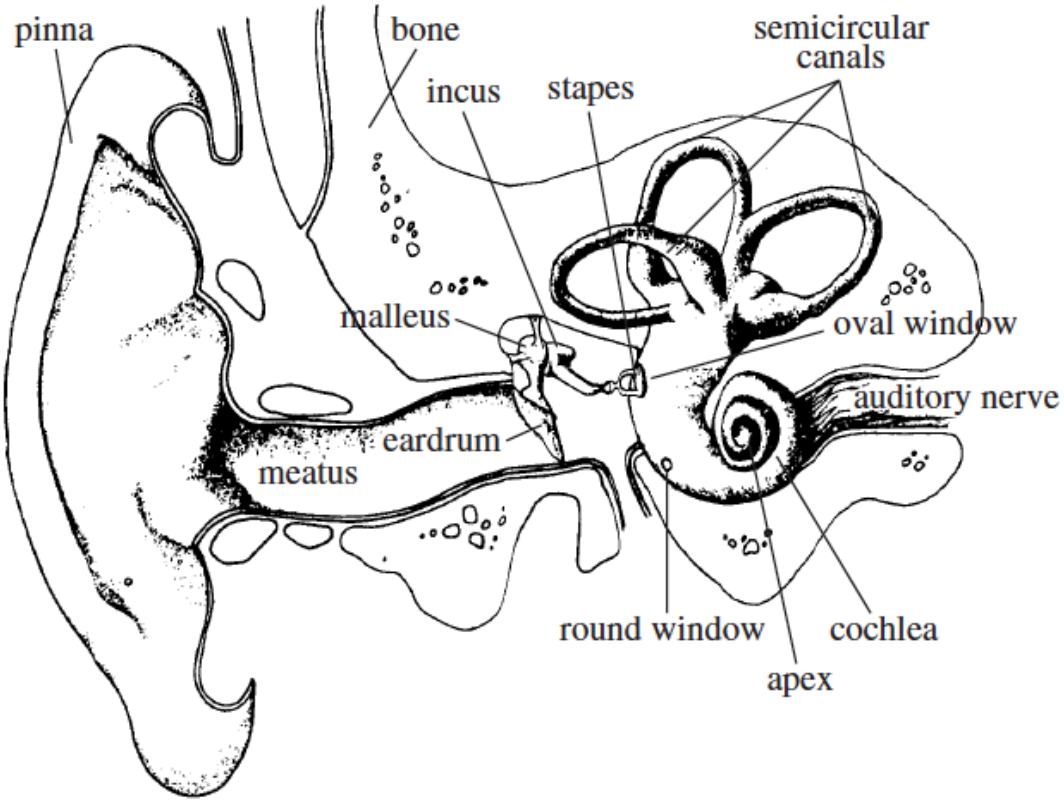


Figure 3.1: The structure of the peripheral auditory system including the outer, middle and inner ear. Redrawn by Moore (2002) from Lindsay and Norman (1972).

The cochlea is the most important part of hearing and provides understanding of many phenomena in auditory perception (Moore, 1995, 1997). A cross-section of the second turn of the cochlea of a guinea pig is presented in Figure 3.2. The cochlea has a spiral shape reminiscent of the snail, and it is divided along its length by Reissner's membrane and the basilar membrane (BM). The starting end of the cochlea where the oval window resides is called the base, and the inner end tip is called the apex. There is a small opening known as the helicotrema between the basilar membrane and the rigid walls of the cochlea. The helicotrema connects the two chambers of the cochlea, the scala vestibuli and scala tympani, shown in Figure 3.2. A second opening in the cochlea, the round window, moves outwards as the oval window moves inwards. (Moore, 1997)

The movement of the basilar membrane in response to sound is of particular interest to us. When an incoming sound moves the oval window, a pressure difference is applied across the basilar membrane in a direction perpendicular to it (Moore, 2002), and it starts then to move. It takes some time for the pressure difference and the pattern of motion to develop, and they vary depending on the location on the BM. The pattern of motion occurring on the BM does not actually depend on which end of it is stimulated; sounds coming to the

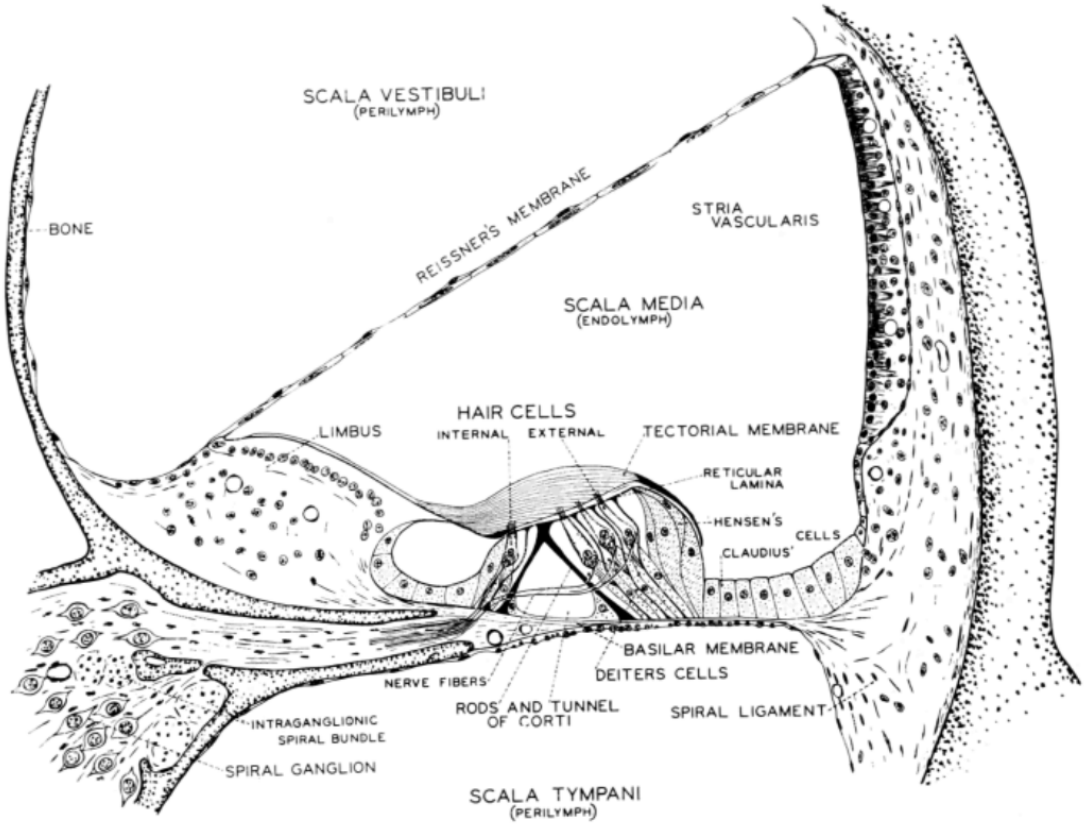


Figure 3.2: A drawing of a cross-section of the second turn of the cochlea of a guinea pig. Adopted from Davis (1962).

cochlea via bone conduction, like our own voices, do not have abnormal responses. (Moore, 1997)

Sinusoidal stimulation results in a response of the form of a traveling wave which moves along the BM from the base towards the apex (Moore, 1997). The form of the traveling wave in response to a 200-Hz sinusoid is presented at two successive time instants in Figure 3.3. First, the amplitude of the wave or the instantaneous displacement of the BM increases slowly, and then after the position of maximum displacement it decreases abruptly. The outer dotted line represents the envelope of the wave in motion. The response of the BM to different frequencies is defined mostly by its mechanical properties (Moore, 1997) which vary depending on the location (Békésy, 1947). At the base it is relatively narrow and stiff, and at the apex it is wider and much less stiff (Moore, 1997). The peak in the pattern of vibration differs in position depending on the frequency of stimulation. Sounds at high frequencies produce a maximum displacement near the stapes, while there is little movement on the remainder of the membrane. Low frequencies produce vibration that extends along the BM and reaches a maximum closer to the apex. The behavior of the cochlea is, however, not explained only by mechanical processes (Békésy, 1947; Moore, 1997, 2002) as will be discussed later in subsection 3.2.3.

The frequency at which there is a maximum response at a particular point on the BM is

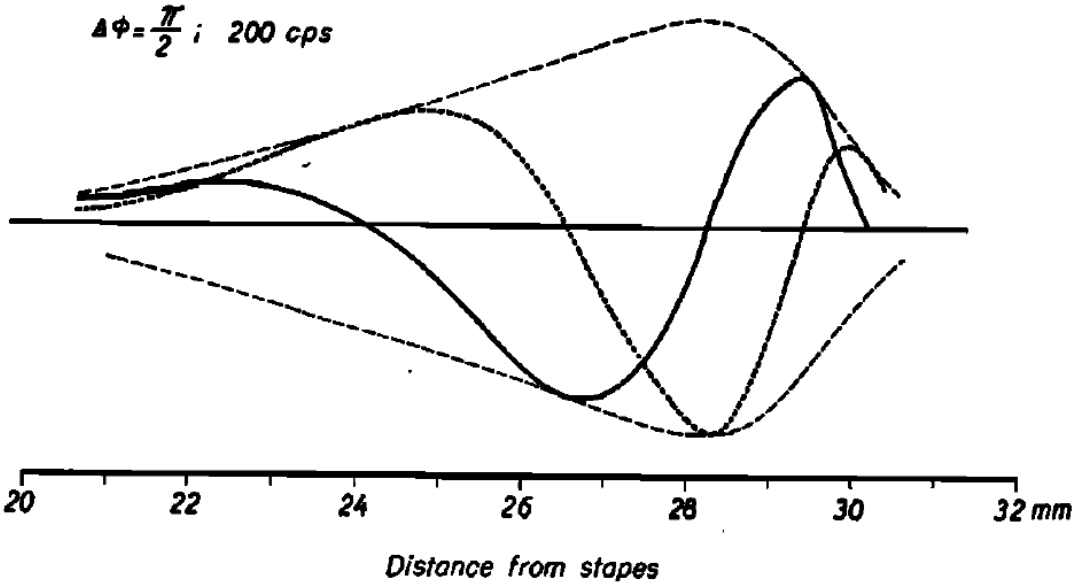


Figure 3.3: The longitudinal instantaneous displacement of the basilar membrane in response to a 200-Hz sinusoid at two instants of time separated by quarter of a period. The ordinate indicates the displacement of the basilar membrane, and the outer dotted line is the envelope of the traveling wave. Adopted from Békésy (1947).

the characteristic frequency (CF) of that point (Moore, 1997). When stimulated with a steady sinusoid the BM responds, at each point where movement can be detected, with vibration at the frequency of that sinusoid. The amplitudes of vibration vary, and the vibrations are in different phases depending on the position on the BM. These rules are true with all sinusoids within the audible range of frequencies. The cochlea acts thus as a spectrum analyzer, or Fourier analyzer, decomposing an incoming sound into its sinusoidal components. Hence, an approximation of the effective stimulating waveform can be derived from the addition of the sinusoidal components. (Moore, 1997)

3.2 Neural responses

In this section the neural impulses originating from the basilar membrane (BM) are discussed in sections 3.2.1 and 3.2.2. In subsection 3.2.3, the active mechanisms influencing the function of the cochlea are discussed.

3.2.1 Neural transduction and neural firings

Let us consider next, how the mechanical movement described earlier is converted into neural signals in the auditory nervous system. In between the Reissner's membrane and the BM is a third membrane called the tectorial membrane, as depicted in Figure 3.2 (Davis, 1962). Between the BM and the tectorial membrane is a structure known as the organ of Corti, which is shown more closer in Figure 3.4. It consist partly of hair cells which

are divided into inner, or internal, and outer, or external, hair cells by the pillars of the tunnel of Corti (Moore, 1997). The outer hair cells get their name from being closer to the outside of the cochlea, and they are arranged in up to five rows. There are about 25 000 outer hair cells, each having about 140 'hairs' known as stereocilia. On the other side of the tunnel of Corti are the inner hair cells arranged in a single row. There are about 3500 inner hair cells, each with about 40 stereocilia. The tectorial membrane lies above the stereocilia, even making contact with the outer hair cells (Moore, 1997).

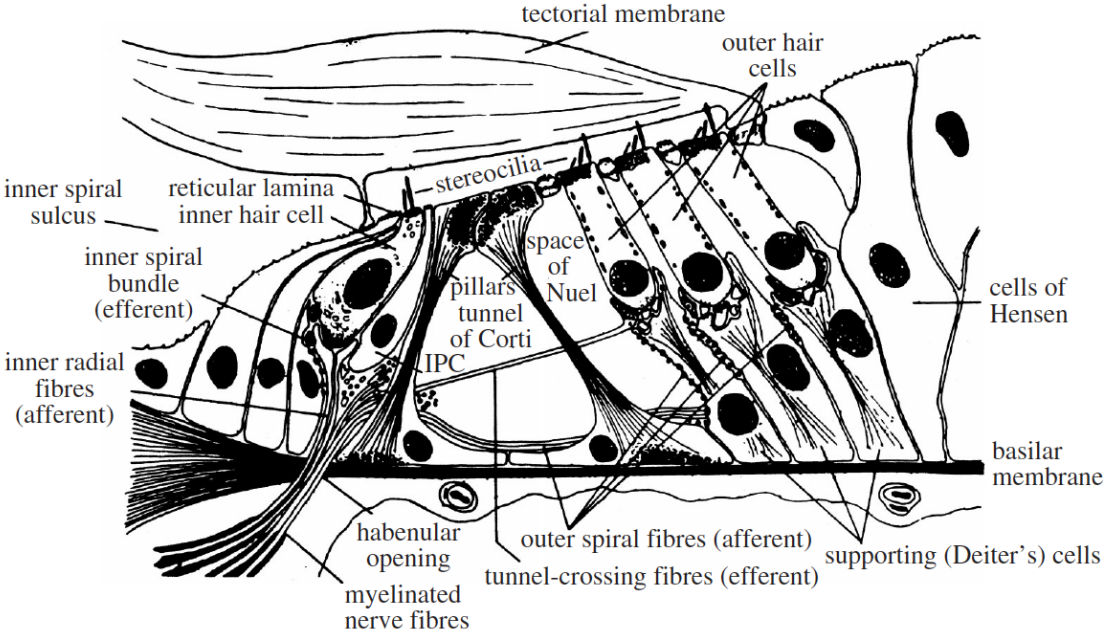


Figure 3.4: A close-up of the organ of Corti as it appears in the basal turn of the cochlea. IPC stands for inner pillar cell. Adopted from Moore (2002).

The tectorial membrane seems to be hinged at one side (the left in Figure 3.2) (Moore, 1997). Thus, when the BM moves up and down, a shearing motion is created between the BM and the tectorial membrane so that the stereocilia at the tops of the hair cells are displaced. It is believed that this leads to excitation of the inner hair cells, which then leads to the creation of action potentials in the neurons of the auditory nerve (Moore, 1997). This neural excitation occurs only at the displacement of the BM towards the tectorial membrane. It is mainly, if not completely, the inner hair cells that transfer the information about sound. (Moore, 1997) In summary, the inner hair cells transduce mechanical movements into neural signals.

There are overall about 30 000 neurons in the auditory nerve (Moore, 1997). Neurons are connected to hair cells with nerve fibers. Afferent nerve fibers carry information from the cochlea towards higher levels of the auditory system, while efferent nerve fibers carry information from the higher levels to the cochlea. The great majority of afferent neurons connect to inner hair cells. Each inner hair cell is contacted by about 20 neurons (Spoendlin, 1970). On the other hand, there are about 1800 efferent nerve fibers, many of which make contact with the outer hair cells (Moore, 1997).

The auditory nerve fiber impulses are often referred to as neural firings. The firing rate

is measured in action potentials (AP), or spikes, per second. There are three properties of these neural responses. The first is that they show spontaneous firing rates in the absence of stimulating sound. Secondly, the nerve fibers are frequency selective, which means that they respond better to certain frequencies than others. Thirdly, the fibers show phase locking to the stimulus waveform, i.e. the nerve fibers show firings at a particular phase of the stimulating waveform. (Moore, 1997) Phase locking is discussed further in subsection 3.2.2. The first two points are addressed next.

The nerve fibers are organized in the auditory nerve in an order with respect to their characteristic frequencies (CFs), so that the place representation of the BM is preserved as a corresponding place representation in the auditory nerve (Moore, 1997). Therefore, it is assumed that the output of a single nerve fiber originates from a particular region on the BM, and that nerve fibers are also highly frequency selective. A nerve fiber has its CF where the threshold of the fiber is the lowest. The threshold of a nerve fiber is the lowest sound level at which a change in the response (or a rise above the spontaneous rate) can be detected. Remote from the CF, the threshold of the nerve fiber increases. A nerve fiber has also a saturation level. The range of levels between threshold and saturation is known as the dynamic range, and it is 20 – 50 dB for most fibers. High spontaneous firing rates are associated with low thresholds, small dynamic ranges and large synapses (Moore, 1997).

3.2.2 Phase locking

There are two important measures of auditory-nerve fiber impulse trains considering the peripheral encoding of sound: the average rate and the instantaneous or phase-locked rate (Sachs and Young, 1980). The latter represents fine temporal details of the response, and the former averages the response. The function of phase locking of the neural responses to the stimulating waveform is discussed next.

When the auditory system is stimulated with a periodic waveform, the nerve firings tend to occur at a particular phase of that waveform (Moore, 1997). This means that the intervals between successive neural impulses are approximately integral multiples of the period of the stimulating waveform. Due to this phenomenon, called phase locking, the firing pattern of a neuron is temporally regular. The precision of phase locking decreases with increasing frequency above 1 – 2 kHz. The upper frequency limit for phase locking in the auditory system is around 4 – 5 kHz. Above that the accuracy, with which the initiation of a nerve impulse gets linked to a specific phase of the stimulus, decreases, and so the distribution of nerve spikes spreads over the whole period of the waveform. In response to a complex tone, a part of the neural activity in the neurons responding to high harmonic frequencies shows phase locking to the overall repetition rate of the stimulus equal to the fundamental frequency, which may be absent as a spectral frequency component. This temporal coding is important in the pitch sensation (see section 3.5) of complex tones as well as in the coding of the relative loudness of each component. (Moore, 1997)

3.2.3 Active mechanisms influencing the cochlea

According to Moore (1995, 1997, 2002), there are active processes that influence the operation of the cochlea. It has been suggested (Moore, 2002) that the outer hair cells

are mainly involved in the active mechanism influencing the mechanics of the cochlea, and thus help to provide sharp tuning and high sensitivity. It is also considered likely that this activity of the outer hair cells is affected by higher levels of the auditory system (Moore, 1997, 2002). It seems that the higher levels of the auditory system control even the earliest stages in the auditory signal analysis. The active mechanisms have an effect also on the interference effects on the BM between closely spaced frequency components of a complex tone (Moore, 2002). Such interference effects influence the perception of changes in phase, the perception of timbre and the perception of pitch.

An evidence of the active mechanisms is the nonlinear behavior of the BM. Recent research has shown that the vibration of the BM is nonlinear (Moore, 2002), i.e., the magnitude of the response does not grow directly in proportion with the magnitude of the input. The input-output function of a point on the BM shows an almost linear response for very low input sound levels (below 20 dB), and the response approaches linearity at high sound levels (above 90 dB) (Moore, 1997). At medium input sound levels, however, the function has a less steeper slope than what would be linear. Thus, this function exhibits a compressive nonlinearity in the processing of the cochlea (Moore, 2002). In other words, a wide range of input sound levels are compressed into a smaller range of responses on the BM. The compression occurs only around the CF, i.e., around the peak of the response pattern on the BM.

Another evidence on the existence of the active biological processes influencing the mechanics of the cochlea are the evoked oto-acoustic emissions (Kemp, 1978), which are reflected sounds from the ear when stimulated with a low-level click. The relative level of this echo is highest at low sound levels and behaves nonlinearly. The middle ear response being of linear nature, the nonlinearity arises from the cochlea. It is sometimes possible that, at a given frequency of the input click, the reflection from the cochlea is amplified. These echoes are individual; each ear gives its characteristic response. (Moore, 1997)

To summarize, the processes influencing the cochlea are strongly nonlinear, they seem to cause the high sensitivity and sharp tuning of the BM. These processes are biologically active and physiologically vulnerable. It seems to be still unclear, which aspects of perception these active mechanisms influence and to what extent such mechanisms are individual. In relation to this, the individuality in the perception of phase modifications will be discussed in section 6.5.

3.3 General concepts of auditory perception

Two general perceptual concepts are discussed in this section. First, the concept of the critical band is introduced. Thereafter, masking is discussed, including simultaneous and non-simultaneous masking as well as release from masking.

3.3.1 The critical band

The auditory system processes sound in frequency bands called critical bands, which are also known as auditory filters. The existence of such bands can be derived from experiments

showing that our perception often differs depending on whether a stimulus is within a critical band or not. A study in which the critical band was discovered is described next.

Fletcher (1940) measured the threshold of a sinusoidal signal as the function of bandwidth of a band-pass noise masker with a constant power spectral density. The results from the experiment show that the signal threshold increases as the noise bandwidth increases until it becomes nearly constant. Further increase in the noise bandwidth does not affect the threshold significantly. The threshold is assumed to depend on the signal-to noise ratio at the output of the auditory filter with a center frequency close to that of the signal. To explain these results, Fletcher (1940) suggested that the peripheral auditory system works as if it contained a set of overlapping bandpass filters, which are now referred to as the auditory filters. The bandpass filtering can be considered to occur on the BM, so that each point on the BM responds to a limited number of frequencies and, thus, corresponds to a filter with a certain center frequency. Fletcher (1940) called the bandwidth, above which the signal threshold no more increased, the 'critical bandwidth' (CB).

The locations on the BM can be considered as a set of bandpass filters with a center frequency corresponding to the characteristic frequency (CF) (Moore, 1997). The bandwidth of the filter, i.e., the CB increases roughly in proportion with CF, as will be seen later. Furthermore, the relative bandwidth is the bandwidth (CB) divided by the CF, and it is often a useful quantity because its reciprocal is a measure of the sharpness of the tuning of an auditory filter, known as the quality factor (Q-factor) (Moore, 1997). In calculations these auditory filters are often approximated as rectangular filters and are often referred to with the term 'critical band'. However, in contrary to this simplification, the auditory filters are assumed to have rounded tops and, outside the passband, sloping skirts (Moore, 1997).

The bandwidth of the auditory filter is often expressed as the equivalent rectangular bandwidth (ERB). The ERB of a given filter is equal to the bandwidth of a perfect rectangular bandpass filter with transmission in its passband equal to the maximum transmission of the specified filter. One way of defining the value of the ERB as a function of center frequency is (Glasberg and Moore, 1990)

$$\text{ERB} = 24.7(4.37f + 1), \tag{3.1}$$

where f is the center frequency in kHz. This function fits well the data from the estimation of the ERB using the notched-noise method (Patterson, 1976) in several research centers (Moore, 1997). In the notched-noise method the threshold of a sinusoidal signal is measured as a function of the width of the stopband of a noise masker. Sometimes ERB is used as the unit of frequency similarly to the Bark scale proposed by Zwicker and Terhardt (1980). A function relating number of ERBs to frequency is (Glasberg and Moore, 1990)

$$\text{Number of ERBs} = 21.4 \log_{10}(4.37f + 1). \tag{3.2}$$

Note, that this scale differs somewhat in numerical values from the Bark scale. For frequencies above 1 kHz the CB is about 10-17% of the center frequency. The asymmetry of the auditory filters have been also measured with an extended version of the notched-noise method of Patterson (1976). The results show that at low and medium sound levels the shape of the auditory filter is about symmetric, but at high sound levels the high-frequency

side becomes somewhat steeper, while the low-frequency slope becomes less steep (Glasberg and Moore, 1990).

Despite of representations of the critical bands such as that of Equation 3.2, there are a continuous set of overlapping critical bands rather than a discrete set of them adjacent to each other (Moore, 1997). Therefore, we may consider an auditory filter centered at any given frequency in the audible range. Furthermore, the CB, or ERB, corresponds to a constant distance of about 0.89 mm on the BM (Moore, 1985).

The processing in the auditory system is not restricted to the critical band, but is capable of integrating over bandwidths much wider than the CB (Moore, 1997). An example of this that the perception of timbre depends partly on the distribution of activity across different critical bands, as will be discussed in section 3.6. Additionally, according to Moore (1997), the detection of phase changes may depend partly on the ability to compare the temporal organization of the outputs of different auditory filters.

The psychoacoustical excitation pattern of a given sound can be derived as the output of the auditory filters to the signal frequency as a function of their center frequency (Moore and Glasberg, 1983). It follows from the shape of the auditory filters that the shape of the excitation pattern of a sinusoidal tone is also reminiscent of a triangle, but less steep at the higher side than the lower side. This happens because the upper side of the excitation pattern is determined from the lower side of the auditory filters. Excitation pattern can also refer to the distribution of neural activity as a function of CF in a single neuron.

In general, in response to a stimulus with a frequency spectrum exceeding the CB, the ear behaves differently from the case of a stimulus not exceeding this band. This difference applies to the absolute and masked threshold, the loudness level, the evaluation of musical intervals, the audibility of partials in complex tones, and the sensitivity to phase (Scharf, 1961). These different aspects of perception are addressed later in sections 3.4 – 4.2.

3.3.2 Masking

Masking is the process in which the threshold of audibility for one sound is raised by the presence of another (masking) sound. The mechanism of masking has been argued to be understood as a combination of swamping and suppression, which are described in more detail in Moore (1997). The problem of signal detection in the presence of a masker can be considered equivalent to the detection of an increment in intensity. In addition to the amount of activity, i.e., firing rates, the temporal patterns of neural firing rates are used in the detection of a signal. The components of a complex stimulus that are most effective in exciting a given neuron define the pattern of phase locking observed in that neuron. The pattern of phase locking is determined by the relative intensities of the components and their frequencies in relation to the response pattern of the neuron. For instance, when evoked with two inharmonically related tones, the neural firings may be phase locked to either of the tones, or both tones simultaneously. (Moore, 1997)

Sometimes the response of a given neuron to a sinusoidal stimulus can be reduced by a second stimulus, even when the second stimulus alone produces no response in the neuron (Moore, 1997). This is called two-tone suppression, and it is a nonlinear process. The mechanical basis of this is discussed in Ruggero et al. (1992). A dominant tone can ‘capture’ the response of a neuron when presented simultaneously with another tone, in which case

the neuron is phase locked only to the dominant tone, when each of the tones in the pair would produce phase locking if presented alone. The temporal pattern of the response may then be indistinguishable from that which occurs when the dominant tone is presented alone. This effect is in analogy with the two-tone suppression, and according to Moore (1997) it may underlie the masking effect between two tones.

The results described above are known to be true only for single auditory nerve fibers, and their application to the response of a group of nerve fibers is somewhat unclear. However, the information about the relative levels of components in a complex sound is contained in the temporal patterns of neural firings, even at sound levels enough to cause saturation in the majority of neurons (Sachs and Young, 1980). The time patterns of response are dominated by the most strongest components in the complex sound, and there may be little or no phase locking to the weak components close in frequency to the strong ones.

Moore (1997) suggests that a tone is masked if a subject cannot detect a change in the temporal structure of the response of the stimulus as a whole. A tone evokes neural discharges with a certain temporal pattern; the nerve spikes are organized in time intervals which are integral multiples of the period of the tone. On the other hand, noise evokes a more irregular pattern of firings in the same neurons. Thus, according to Moore (1997), it holds for the masking of a tone by wide-band noise, that the tone is detected only if there is regularity to a a certain extent in the firing patterns of the nerve fibers responding to that tone. The neurons with CFs close to the tone frequency show the greatest degree of temporal regularity, i.e., phase locking, while the auditory filter performs to reduce the influence of the noise on the neural responses (Moore, 1997).

Previously, simultaneous masking has been discussed. In addition to this, masking can occur between two non-simultaneously presented tones. There are two types of non-simultaneous masking: backward masking, in which the signal is presented before the masker; and forward masking, in which the signal follows the masker. These time effects in masking are studied using short signals. In addition to adaptation and fatigue, forward masking is a third process which may affect the threshold of the short signal after another sound. Forward masking is greatest when the signal is near the masker, i.e., delay D is small, and decreases linearly as a function of $\log(D)$. The rate of recovery from forward-masking increases as the masker level increases. In all cases of forward masking, the masking decays to zero after 100 – 200 ms. Additionally, the amount of forward masking increases with duration up to at least 20 ms. (Moore, 1997) Later in chapter 4, forward-masking will be associated with the detection of within-channel cues in the perception of changes in phase.

Contrary to masking, also release from masking occurs. An example of release from masking is as follows. When a sinusoidal signal is presented in modulated noise which is coherent or correlated between different frequency bands, the signal threshold decreases as the noise bandwidth increases beyond the CB (Hall et al., 1984). This phenomenon, named as 'comodulation masking release' (CMR), is assumed to depend on the comparison of the outputs of different auditory filters.

Although referring much to masking, comparison of temporal envelopes across channels seems likely to be a general feature of auditory pattern analysis (Moore, 1997). According to Moore (1997), detecting signals from noise backgrounds or separating competing sound sources may depend on this ability, as discussed above. Many real-life sounds have temporal intensity peaks and dips, i.e., envelope fluctuations, which are correlated across frequency

bands. One example of such stimuli is speech. Use of these across-channel analysis cues is later discussed to occur also in other perceptual scenarios, especially, in the perception of phase spectrum changes addressed in chapter 4.

3.4 Perception of loudness

Loudness is defined as that attribute of sound in terms of which tones can be ordered on a scale from quiet to loud. Loudness is a subjective measure, and its corresponding physical quantity is intensity. The lowest sound level that can be heard is referred to as the absolute hearing threshold at each frequency. The range of intensities that we can detect is vast; the dynamic range of sounds we can hear can be even 120 dB (Moore, 1997), which corresponds to a ratio of intensities of $10^{12} : 1$.

3.4.1 Absolute thresholds and loudness of sound

The absolute thresholds of hearing can be measured as the minimum audible pressure (MAP), which is usually determined with the use of headphones, or as the minimum audible field (MAF), which is determined using test tones delivered by a loudspeaker usually in a large anechoic chamber (Moore, 1997). The minimum audible sound level is on average 2 dB lower when determined binaurally than when determined monaurally. The absolute threshold data is usually derived from the average of many young listeners. However, it should be noted that a person can have a threshold 20 dB below or above the average at a certain frequency and still be considered as 'normal'. A typical hearing threshold curve is plotted in Figure 3.5. It can be seen from the hearing threshold curve that we are most sensitive to sounds at frequencies of about 1 kHz to 5 kHz. This sensitivity at middle frequencies is at least partly due to the sensitivity of the outer and middle ear to this frequency range. (Moore, 1997)

The absolute thresholds increase rapidly both at the low-frequency and at the high-frequency end. The highest audible frequency is largely dependent on the age of the listener; young children can hear tone frequencies up to 20 kHz, while for most adults the thresholds rise rapidly above 15 kHz. The variability between individuals is also large at high frequencies. According to Moore (1997), it appears that the low-frequency limit of our hearing cannot be determined.

In order to obtain a subjective scale for loudness, loudness level has been defined. The purpose of this measure is to determine the relationship between the physical intensity and judged loudness. The loudness level of a 1000-Hz tone is equal to its sound pressure level in dB SPL (Moore, 1997). Loudness level is defined with the unit 'phon'. Note, that sound pressure level is a physical quantity. The unit sensation level in dB SL is used when the reference intensity is the absolute threshold for a given sound and a given subject.

Furthermore, loudness can be described with the unit 'sone', so that one sone is arbitrarily the loudness of a 1000-Hz tone at 40 dB SPL. If the level of a 1000-Hz tone is fixed and the test sounds at several frequencies are adjusted to sound equally loud, equal-loudness contours are obtained. The equal-loudness contours and the absolute hearing threshold curve are presented in Figure 3.5. It should be noted, however, that equal-loudness contours

from different research instances differ from one another and from the contours in Figure 3.5, which are just one example. With increasing intensity the loudness level of low frequencies (and to some extent that of high frequencies) grows at a higher rate than that of middle frequencies. From this, it follows that equal-loudness contours are significantly different depending on the level of reproduced sound. If sounds are reproduced at a level other than the original (recorded level), the spectral balance is altered. (Moore, 1997)

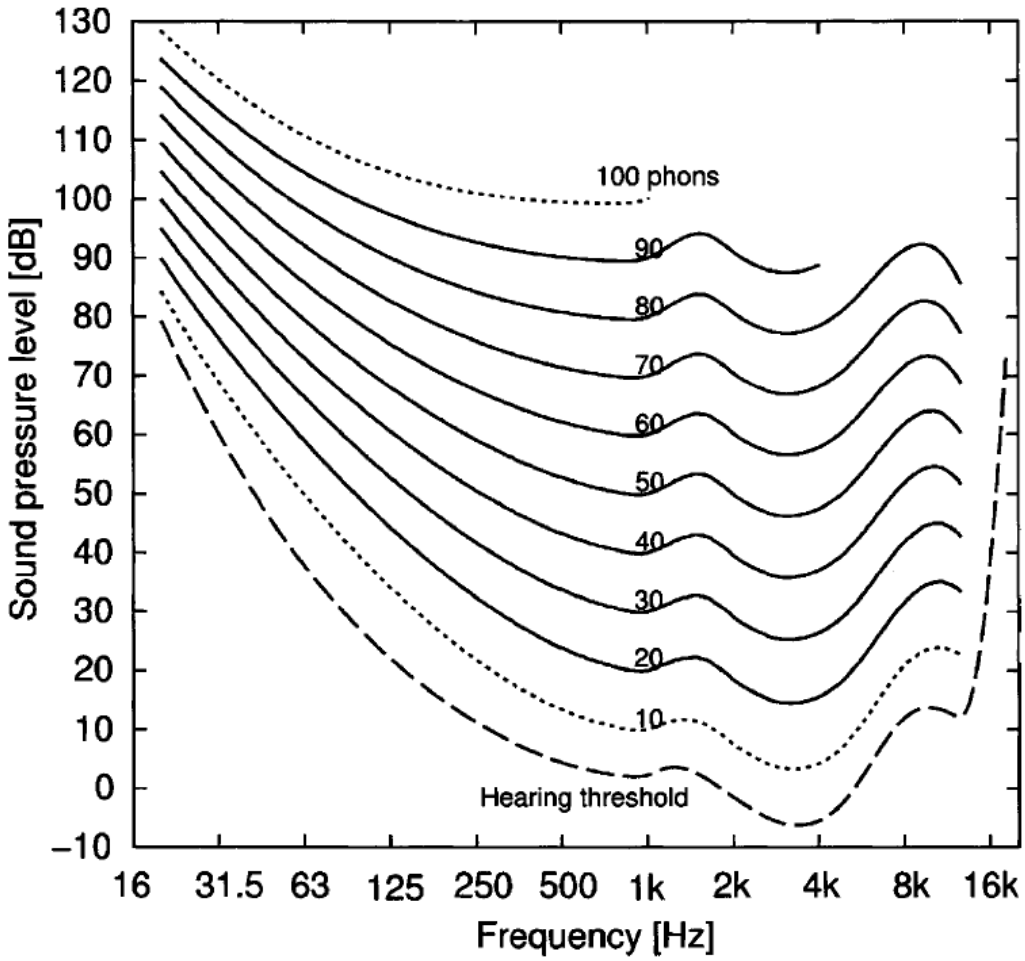


Figure 3.5: The shapes of the equal-loudness contours and the absolute hearing threshold. Adopted from Suzuki and Takeshima (2004).

It is known that both absolute thresholds and the loudness of sound depend upon duration (Moore, 1997). When the duration exceeds about 500 ms, the sound intensity at threshold is independent of duration. However, for durations less than about 200 ms, the sound intensity at threshold increases as the duration decreases. The temporal integration of the auditory system can be described so, that the product of time and the amount by which the intensity exceeds the threshold intensity for a long-duration tone is constant. There exists a limit to the time over which the ear can integrate energy or, in other words, there is a lower limit to the intensity which can be effectively integrated. (Moore, 1997)

3.4.2 Loudness of complex sounds

The perception of loudness is somewhat different for complex sounds than for pure tones. If the bandwidth W of a complex sound of fixed energy (or intensity) is below the critical bandwidth (CB), the loudness of that sound is somewhat independent of W. The loudness would be then judged to be the same as that of a pure tone with same intensity at the center frequency of the critical band (Moore, 1997). However, if W is increased beyond the CB, the loudness of the complex starts to increase. Hence, when a given amount of energy is spread over more than one critical band, the complex sound is louder than in the case when the same energy lies within one critical band. This holds for complex sounds consisting of pure tones whose frequency separation is varied (Scharf, 1961) and for bands of noise (Zwicker et al., 1957). The increase in loudness at bandwidths greater than the CB can be explained with the concept of 'loudness summation' (Zwicker et al., 1957), which is illustrated in Figure 3.6. The dashed line shows the bandwidth, above which the loudness level starts to increase for most of the sound levels depicted. At sensation levels around 10 – 20 dB SL the loudness of a complex sound is approximately independent of bandwidth, even above the CB.

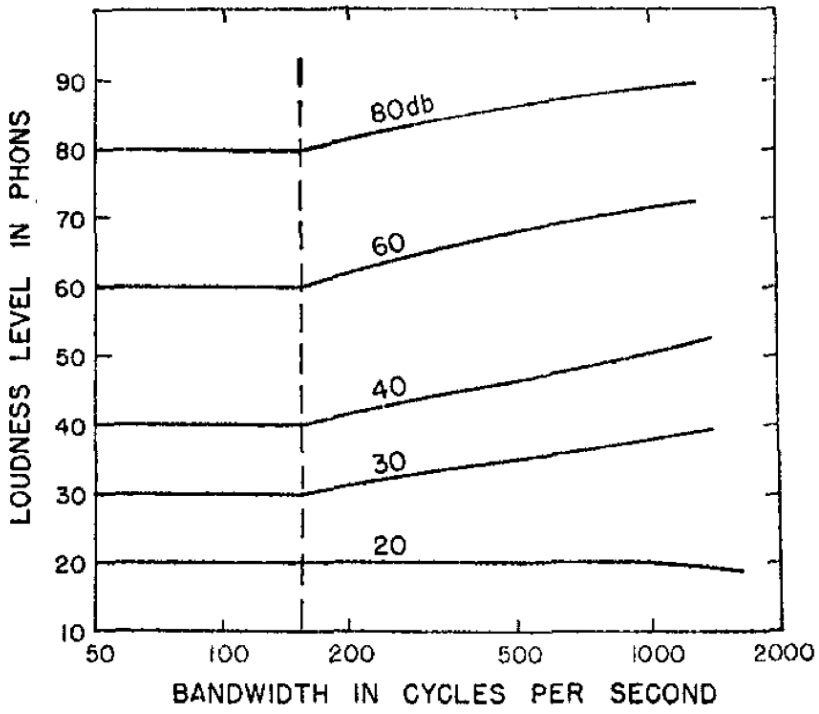


Figure 3.6: Loudness summation. The loudness level in phons of a band of noise centered at 1 kHz measured as a function of bandwidth. The overall sound levels are indicated above each curve in dB SPL. Originally adapted from Feldtkeller and Zwicker (1956) by Scharf (1961).

Furthermore, near threshold at very low sensation levels (below 10 dB), loudness decreases as the bandwidth of a complex sound is increased over the CB. When a fixed amount of energy is situated within a critical band, it is enough for the sound to be audible. But when that energy is spread over more critical bands, the energy in each critical band becomes

insufficient to make the sound audible. (Moore, 1997)

Listening to complex sounds is affected also by the type of headphones used. The physiological noise caused by the vascular system gets 'trapped' inside the ear canal and causes an increase in thresholds when earphones are worn (Moore, 1997). This low-frequency noise varies in level with the leakage of air around the headphone, with the volume of the ear canal and with the subject's heart beat. Due to these reasons, circumaural headphones, which fit around the pinnae, are preferred for listening at low frequencies compared to other types of headphones. On the other hand, supra-aural headphones, which lie on the pinnae, are more reliable and easier to calibrate at high frequencies. Open headphones can markedly reduce the physiological low-frequency noise. (Moore, 1997)

3.4.3 Coding of loudness and intensity discrimination

The underlying mechanisms of the perception of loudness are still not fully understood. However, according to Moore (1997) it is commonly assumed that loudness depends somehow on the total amount of neural activity evoked by a sound. Thus, Moore (1997) suggests that loudness may depend on the summation of neural activity across different auditory channels. Then the loudness of a sinusoidal stimulus would be determined not only by the activity in the neurons with characteristic frequencies (CF) close to the tone frequency but also by the activity in the neurons with adjacent CFs. Furthermore, the determination of the loudness of a sound source is affected by the apparent distance from the source, the context in which it is heard and the nature of the sound. In this way the properties of the sound source itself are estimated, but it may be difficult to estimate the magnitude' of a sensation. (Moore, 1997)

Nevertheless, there is knowledge of intensity discrimination and coding of intensity changes, which are discussed next. Intensity discrimination refers to the detection of a change in intensity or sound level. Thresholds for intensity discrimination are usually defined in decibels as the change in level at threshold, ΔL, which is about 0.5 – 1 dB for wide-band noise. This is true from about 20 dB above threshold to 100 dB above threshold, while ΔL increases for sounds that are close to the absolute threshold. For wide-band or band-pass filtered noise, or bursts of noise, the thresholds for detecting intensity changes follow the Weber's law, which states that the minimum detectable change is proportional to the magnitude of the stimulus. For pure tones Weber's law does not hold. The intensity discrimination of pure tones, and actually also that of narrow-band 'frozen' noise, improves at high sound levels. There are at least two explanations for this. Firstly, the high-frequency side of the excitation pattern grows nonlinearly with increasing sound level, and secondly, we are able to combine information from all the excited frequency channels across the whole excitation pattern. Noise is frozen if it is identical from one presentation to the next and it does not have random fluctuations in energy. (Moore, 1997)

Intensity changes are coded in several different ways in the auditory system. A change in intensity can be detected by changes in the firing rates of neurons with CFs at the center of the neural excitation pattern and also by the spreading of the excitation pattern. The latter means that when the level of a sound increases, more neurons with CFs close to the edges of the excitation pattern respond to the stimulus. This growth in the neural activity is greater at the high-frequency side than at the low-frequency side. However,

the information from the spread of excitation is not essential for intensity discrimination performance at high sound levels. (Moore, 1997)

It has been suggested by Carlyon and Moore (1984) that at low frequencies intensity discrimination depends also on phase locking. However, information from phase locking is not critical since it has been shown, that good intensity discrimination is possible with stimuli restricted to the range above 4 – 5 kHz, where phase locking does not occur. Moreover, according to Moore (1997), the relative levels of the components of a complex tone may be signaled by the degree of phase locking in the neurons with CFs close to the component frequency.

A single neuron is able to code intensity changes depending both upon the shape of the rate versus level function and upon the variability in the firing rate. A given neuron transmits information optimally over a small range of sound levels. At low levels the coding is poor because below or close to the threshold the neuron shows minimal changes in firing rate. On the other hand, at high sound levels the neuron shows poor coding due to neural saturation. If all the information contained in the firing rates of the neurons in the auditory nerve were used in detecting intensity changes, intensity discrimination would be much better than it actually is (Moore, 1997). Thus, again it seems that more central parts of the auditory system are involved in defining its perceptual properties. For most stimuli, it seems that intensity discrimination is not limited by the information carried in the auditory nerve, but by the capacity to make use of that information at more central levels of processing (Carlyon and Moore, 1984).

3.5 Pitch perception

Pitch has been defined as 'that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale' (ASA, 1960). Pitch is related to the repetition rate of the waveform, which for pure tones corresponds to the frequency and for periodic complex tones often to the fundamental frequency. There are, however, exceptions to this rule. Two theories exist for perception of pitch: place theory and temporal theory (Moore, 1997). The place theory accounts for such a spectral analysis in the inner ear, that different frequency components in a complex tone excite different points on the basilar membrane (BM) and, thus, neurons with different CFs. The temporal theory suggests that the pitch of a sound is determined from the temporal pattern of excitations within and across neurons. Thus, information from phase locking, which was explained in subsection 3.2.2, is believed to define the perceived pitch. An auditory model that could always predict the perceived pitch, would involve the use of both place and temporal information depending on the frequency range and the type of sound. Moore (1997) supports the idea that the pitch of pure tones is determined by temporal mechanisms below 5 kHz and by place mechanisms above 5 kHz. Moreover, sequences of pure tones above 5 kHz do not produce a clear sense of musical interval or melody.

The pitch of a complex tone is usually determined by components other than the fundamental (Moore, 1997). An evidence of this is the 'missing fundamental' or residue pitch. A simple example of a residue pitch is, that when stimulated by a group of harmonics lacking the fundamental component, a pitch corresponding to the repetition rate (corresponding to the fundamental frequency) is heard. The residue is the strongest component in a complex

tone, and thus, the pitch of the sound is determined mainly by the residue pitch. Schouten (1940) developed a theory for pitch perception of complex tones, and main points of it are as follows. A number of pitches may be heard when listening to complex tones. For a harmonic complex sound, some of these pitches correspond to the lower harmonics. One or more residue pitches may be heard, and they are produced by high harmonics, which are not well resolved by the ear but interfere on the BM. The residue is determined by the temporal pattern of the waveform at the point on the BM where the partials interfere. More specifically, 'the pitch of the residue is determined by the time interval between peaks in the fine structure of the waveform (on the BM) close to adjacent envelope maxima.' (Moore, 1997) Thus, the overall rate of firing is not the information needed for defining the perceived pitch.

A model of pitch perception proposed by Schouten et al. (1962) and De Boer (1956) accounts for two basic facts. Firstly, the pitch shift of a residue is a linear function of the frequency shift of the components, and secondly, the slope of the line relating pitch and frequency shifts decreases as harmonic number increases (Patterson, 1973). This theory of pitch perception is based on inter-peak durations. Data on this subject suggests that when the components in the wave are all above the fourth harmonic region, it is the second to the lowest component that dominates in the production of a pitch sensation. Patterson (1973) also found that the pitch of the residue is not dependent on the number of components in the complex tone, and that the residue pitch is unchanged when the relative phases of the components are randomized.

In fact, it has been found that only one component is enough to produce a perception of a subharmonic pitch corresponding to the residue, in the presence of background noise (Houtgast, 1976). The background noise makes the sensory information ambiguous and due to this, the perception of these potential pitches is more likely. This potential deteriorates for harmonic numbers of about 7 to 11. This shows that interaction of components is not crucial for the perception of a residue pitch. It has been suggested (Moore, 1997) that a synthetic mode of pitch perception is enhanced in the presence of background noise. Information would then be combined across frequency regions and in time to produce the pitch sensation of a residue.

Related to the hearing of pitches of pure tones and complex tones, is the detection of different partials in a complex tone. A partial can be 'heard out' with about 75% accuracy when the frequency separation from the neighboring partials is about 1.25 times the ERB of the auditory filter (Moore and Ohgushi, 1993; Moore, 1997). Although, Plomp (1964) reported that this limit is in some cases about one CB or the ERB of the auditory filter. For two-tone complexes, however, the frequency separation needed for identifying the partials is less than this, especially at low component frequencies. Moreover, Moore and Ohgushi (1993) found that with frequency spacings more than one ERB, the partials at the low and high 'edges' of an inharmonic complex tone were more easily heard out than the 'inner' partials. This finding suggests that the partials of a two-tone complex are more easily heard out than that of a multi-tone complex. A possible explanation to this phenomenon is that the excitation at CFs corresponding to the inner partials comes from the interaction of several components of similar effectiveness, and thus, there is not a certain CF where the temporal pattern of the response would be determined by one particular partial (Moore, 1997). Whereas, for a two-tone complex, at CFs just below and above the stimulus frequencies, the temporal pattern of the response of the neurons is

dominated by the lower and upper components, respectively (Moore, 1997). Furthermore, it has been found that musicians, i.e., trained or expert listeners, perform better in hearing out partials (Soderquist, 1970).

In general, the lower harmonics of a complex tone are most easily detected (Plomp, 1964), which is explained by the fact that the ERB of the auditory filter is narrower at the low frequency end than towards high frequencies and, therefore, the spacings of the harmonics have a high value in ERBs. Pitches can be observed also when the two ears are stimulated dichotically with two noise stimuli that are time shifted, or phase shifted, in a narrow frequency band with respect to each other (Moore, 1997). This indicates that information about the relative phases of components in the two ears are preserved up to the point in the auditory system where binaural interaction occurs. In summary, it has been suggested (Moore, 1997), that in addition to the frequency resolution of peripheral processing, temporal coding (phase locking) also plays a role in the production of pitch sensations.

3.6 Timbre perception

Next, the perception of timbre of complex tones is discussed. Timbre can be described as the 'color' or 'quality' of sound. Timbre is multidimensional unlike pitch and loudness which can be considered as unidimensional. This means that there is no single scale on which timbres of different sounds can be placed. American Standards Association (ASA, 1960) has defined timbre as 'that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.' The timbre differences between sounds, such as vowels or steady complex tones, can be considered to be closely related to their magnitude spectra presented, for example, as the levels in 18 1/3-octave frequency bands (Plomp, 1970). As the bandwidth of 1/3 octave is slightly broader than the CB over most of the audible frequency range, timbre can be considered to be related to the relative levels produced by a sound in each critical band. In other words, for steady sounds, timbre is related to the excitation pattern of a sound. However, based on the facts about the coding of the perception of loudness and pitch (see sections 3.4 and 3.5), timbre depends also on the temporal pattern of sound.

A well-known example of timbre differences is the sounds of different musical instruments. The identification of musical instruments depends much on onset transients and the time envelopes. Patterson (1994) showed that a sound with an asymmetric time envelope is perceived differently from its time-reversed version even though their long-term magnitude spectra are the same. He used sinusoidal carriers that were amplitude-modulated with a repeating exponential function. This shows that time envelopes play a role in the perception of timbre. Many sounds of real life, such as many musical instruments cannot be recreated simply by the same harmonic structure than the instrument sound has, because those instruments have noise-like qualities which strongly affect the perceived sound quality, i.e., timbre. Such summation of steady component frequencies does not create the dynamic time characteristics of these instruments. When talking about magnitude and phase spectra, it is important to keep in mind that both of them can change due to the transmission path and room reflections.

Chapter 4

On the perception of phase spectrum changes

The perception of changes in the phase spectrum of harmonic complex tones is discussed in this chapter. Phase spectrum is a signal property, as was explained in chapter 2. By modifying this property the waveform of the sound is altered. Furthermore, this can be done without affecting the magnitude spectrum. Several studies about the perception of phase changes have been conducted. Ohm (1843), von Helmholtz (1863) and place theorists after them have believed that there are no temporal mechanisms but only place mechanisms involved in the perception of sound. In other words, they believed that the ear is not sensitive to changes in phase. However, research in the 20th century and after has shown that the phase spectrum can affect the perception of sound in various ways. Many of these studies are referred to in this section. First, general information on the perception of phase changes is presented. Thereafter, the discussion focuses on a classification of perceptual phase changes; local phase changes are discussed in subsections 4.1.1 and 4.1.2, and global phase changes in subsection 4.1.3. These types of phase changes and results from experiments using them are summarized in subsection 4.1.4. Finally, the effect of phase changes on timbre is focused on and summarized in section 4.2.

Local, or within-channel, phase changes refer to changes that can be seen within the outputs of different auditory filters, whereas global, or between-channel, phase changes can be seen as differences between the outputs of auditory filters while the within-channel waveform changes are minimal. Note, that in this chapter and when working with wide-band periodic sounds and place/time models of hearing, spectral frequencies which carry the acoustic energy have to be kept separate from the repetition rate (corresponding to the fundamental frequency) of the signal, which may or may not have energy associated with it. To follow up this distinction in this chapter, the repetition rates are designated with the unit cycles per second (cps).

There are several complex sounds, including many vowels and musical tones, the timbre of which depend on the relative phases of the spectral components as well as their relative amplitudes (Patterson, 1987). In a pilot experiment Patterson (1987) studied the discrimination of cosine-phase (CPH) and random-phase (RPH) waves, which have identical magnitude spectra. A wave is in cosine-phase when all of its components start at their maximum amplitudes, i.e., in cosine-phase. In a random-phase wave the starting-phases

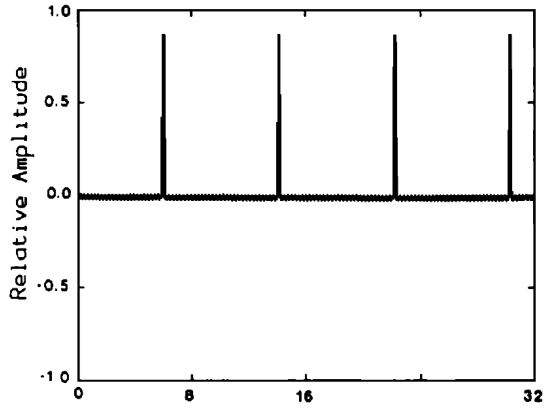
are randomized. It was von Helmholtz (1863) who first claimed that these two types of signals could not be discriminated using 12 harmonics with a repetition rate of 119 cps. This is the B♭2 tone an octave and two semitones lower than the middle C. However, his result was later proven to be false. It is believed that Helmholtz could not do the discrimination because he used only 12 harmonics. However, Plomp and Steeneken (1969) showed that it is possible, though only marginally, to discriminate CPH and RPH stimuli using only ten harmonics. Patterson (1987) showed in his experiment that "the general discrimination between RPH and CPH waves is possible over a wide range of repetition rates and spectral locations." His results showed also that discrimination performance tends to be better at low repetition rates and at higher spectral locations. At 250 cps, for example, the performance fell off when four adjacent harmonics were present and the position of the lowest harmonic was decreased from 8 to 4. The RPH stimulus sounds rougher than the CPH wave, because of within-channel phase changes in high harmonics. In this comparison, between-channel phase shifts are not heard. (Patterson, 1987) The effects of phase spectrum changes on timbre are discussed further in section 4.2.

Patterson (1987) introduced also an auditory model for the multichannel neural firing patterns produced in the auditory nerve. This model predicts whether a phase change will produce a timbre change in monaural perception. The model is called the pulse ribbon model of hearing, and it generates a pulse stream across auditory filters as the output. It is a simplified model of the cochlea with an auditory filter bank of 24 channels with bandwidths of one ERB. The pulse ribbon model has five stages: first two simulate the operation of the cochlea, the next three stages aim to transform the output of the cochlea into something representing an auditory sensation including pitch and timbre perception. The auditory filter bank mentioned above simulates the frequency analysis performed by the cochlea. The output of this frequency analysis is also referred to as driving waves, since these waves drive the primary auditory neurons in a way of controlling the firing pattern observed in these neurons. After that, 24 pulse generators are used to simulate this neural encoding of the frequency analysis, i.e., one pulse generator per auditory filter channel is used. The output of the pulse generators is called the initial pulse ribbon, which is then aligned if needed. The three final stages convert the data into patterns of timbre and pitch perception aiming to produce a change in the pattern only if a change is heard.

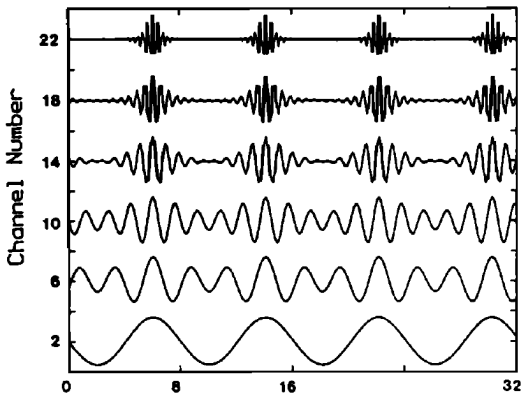
Because the auditory nerve fibers are, at relatively low spectral frequencies, phase locked to the (driving) wave, they generate pulse streams that carry information about the peaks in the wave (Moore, 1997). In the pulse ribbon model (Patterson, 1987), the pulse generators fire precisely on the peaks and their sustained firing rates have fixed upper limits at any given signal level. A CPH wave of 31 equal-amplitude harmonics of 125 Hz and its pulse ribbon are illustrated in Figure 4.1 as an example. As can be seen from this figure, the initial pulse ribbon of the CPH wave is perfectly aligned. A more detailed description of the usage of the model in studying monaural phase perception is presented in Patterson (1987).

4.1 Phase spectrum changes

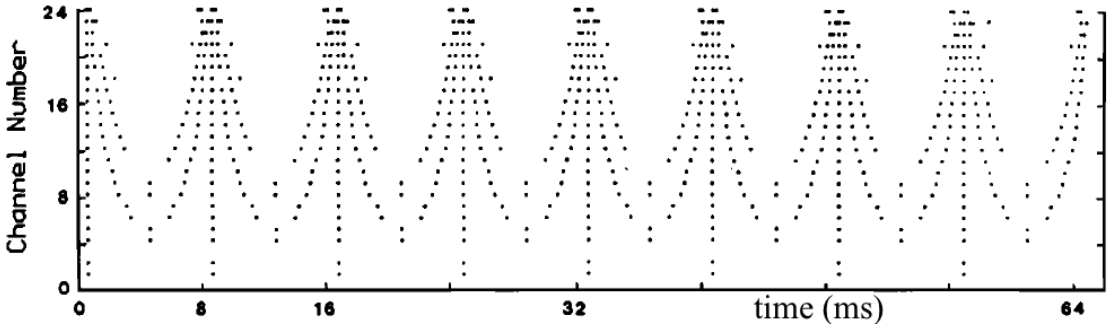
Patterson (1987) studied both local and global phase changes in monaural perception. In that study, the phase-discrimination thresholds were measured subjectively as a function



(a) Waveform



(b) Driving waves



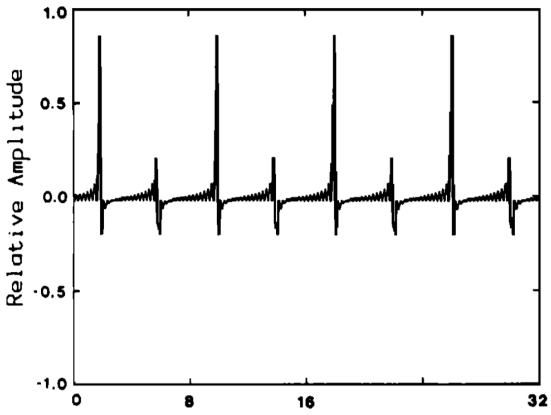
(c) Initial pulse ribbon

Figure 4.1: Outputs of different stages of the pulse ribbon model for a cosine-phase wave (a) with 31 equal-amplitude harmonics of 125 Hz. The driving waves (b) are the outputs of the auditory filter bank driving the corresponding pulse generators, which produce the initial pulse ribbon (c). Adapted from Patterson (1987).

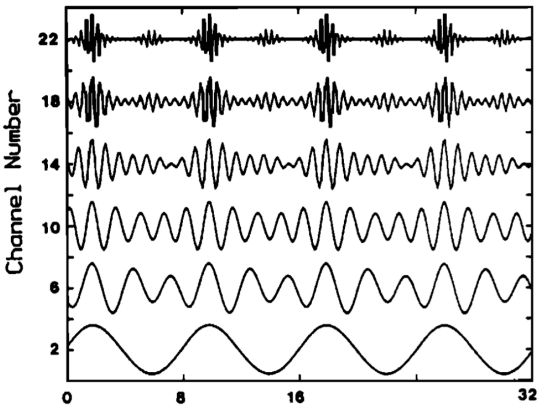
of the repetition rate, level, bandwidth, spectral location and stimulus length. Following this classification by Patterson (1987), local and global phase changes are discussed next.

4.1.1 Local phase changes: alternating-phase wave

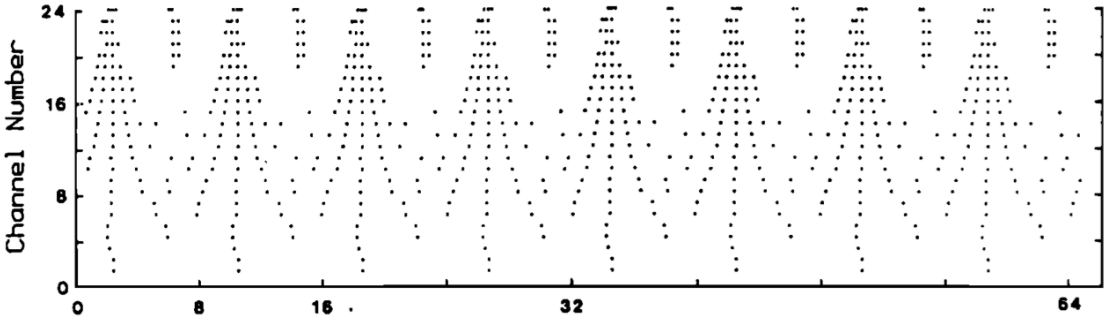
Local, or within-channel, phase changes mean changes in the envelopes of the waves at the outputs of a range of auditory filters. Patterson (1987) studied local phase changes with a stimulus in which every other component of a CPH wave is shifted in phase by a fixed amount D. The phase spectrum of such stimulus alternates between two values, so the stimulus is referred to as the alternating-phase, or APH, wave. As D is increased towards 90°, secondary maxima appear in the driving-wave envelopes of, first, the high-frequency channels, and the main maxima decrease. When D is increased more, secondary pulses appear also in the lower frequency channels. At 90° the primary and secondary pulses are of equal amplitude. An alternating-phase wave, in which the odd harmonics start in cosine-phase and the even harmonics are advanced 40° in phase, is depicted in Figure 4.2. (Patterson, 1987)



(a) Waveform



(b) Driving waves



(c) Initial pulse ribbon

Figure 4.2: Outputs of different stages of the pulse ribbon model for an alternating-phase wave (a) with 31 equal-amplitude harmonics of 125 Hz. The even components are advanced 40° from cosine starting phase. The driving waves (b) are the outputs of the auditory filter bank driving the corresponding pulse generators, which produce the initial pulse ribbon (c). Adapted from Patterson (1987).

The fact that more frequency components are in the passband of high frequency filter channels explains why the secondary pulses appear first in these filter channels. The interaction of the components accounts for the (amplitude) modulation in the envelopes and the variations in the shape of the envelopes produced by shifting the phase. In the lowest channels, the phase change produces an asymmetry in the envelope shape instead of a distinct additional pulse. The secondary pulses in the driving-wave envelopes get encoded in the initial pulse ribbon as a column in between the main columns, as can be seen in Fig. 4.2c in the channel numbers from about 20 upwards. As the phase shift D increases up to 90°, the secondary column in the pulse stream lengthens towards low frequency channels, just as occurs in the driving waves. (Patterson, 1987)

Patterson (1987) showed that the size of the within-channel phase change required for discrimination depends strongly on the repetition rate, intensity, and spectral location of the signal. When the signal level is 40 dB per component and the repetition rate is 125 cps, the D required to distinguish APH and CPH waves is about 40°. As the repetition rate increases, the discrimination threshold increases, as the decrease in the frequency

resolution of the pulse stream predicts. The threshold for discrimination decreases as the spectral location of the stimulus is increased in frequency. A trend was reported that the threshold is lower when the even harmonics are shifted in phase than when the odd harmonics are shifted, but it, however, was not significant. The threshold for discrimination of within-channel changes decreases as stimulus level is increased. The effect of increasing the stimulus bandwidth from 4 to 8 harmonics is small. Signal duration does not affect the discrimination. (Patterson, 1987)

4.1.2 Local phase changes: on the thresholds and sensitivity

In addition to the alternating-phase wave, a change in the relative phase of a single harmonic component can be perceived (Moore and Glasberg, 1989). Moore and Glasberg (1989) showed that a phase shift of a single component produces a within-channel envelope change using complex tones of the first 20 harmonics of fundamental frequencies 50, 100, or 200 Hz. Figure 4.3 shows the output waveforms of the response of a simulated auditory filter centered at 1.5 kHz, to a cosine-phase complex tone (top trace) and to a complex tone with the phase of the 15th harmonic advanced progressively in successive traces. An oscillation at the frequency of the phase-shifted component appears in the previously low-amplitude parts of the envelope. (Moore and Glasberg, 1989)

Furthermore, Moore and Glasberg (1989) found that for normally hearing subjects the thresholds for detecting a phase shift, or the phase difference limens, in a single harmonic were $2° - 4°$ for the fundamental frequency of 50 Hz and for the harmonics above the eighth. At this fundamental frequency, any kind of changes in phase of harmonics below the third or fourth were not detectable. Moreover, at 200-Hz fundamental frequency, phase changes were not detectable below harmonic numbers 5 – 13. The thresholds for a phase shift increased slightly for the highest harmonics for all the studied fundamental frequencies. These results indicate that local phase changes are more easily perceived the lower the fundamental frequency is. (Moore and Glasberg, 1989)

Moore and Glasberg (1989) studied also the effect of level on the threshold of phase shift and found that thresholds increased significantly with decreasing level, except for the highest harmonic. According to the test subjects of their listening tests, the phase shift caused the harmonic component to be 'heard out' as a pure tone. Note, that this shows that the temporal structure affects the detection of partials in a complex tone, which was discussed in section 3.5. When the phases of the components of the reference stimulus, which were all in cosine phase in the other trials, were randomized, the thresholds for a phase shift increased, and were many times impossible to measure (Moore and Glasberg, 1989). Furthermore, the phase shifted harmonic was no more heard out from the complex tone. Subjects with unilateral cochlear hearing impairment had generally poorer phase sensitivity in their impaired than in their normal ears.

The relation between the critical band, discussed earlier in subsection 3.3.1, and the sensitivity of hearing to phase can be demonstrated by comparing our ability to detect modulation in amplitude-modulated (AM) signals to that of frequency-modulated (FM) signals (Zwicker, 1952; Moore, 1997; Feldtkeller and Zwicker, 1956). In amplitude modulation the amplitude of the carrier tone is made to change as the magnitude of the modulating sine wave, while the carrier frequency is unchanged. In frequency modulation the instantaneous frequency of the carrier is varied in proportion to the magnitude of the modulating signal,

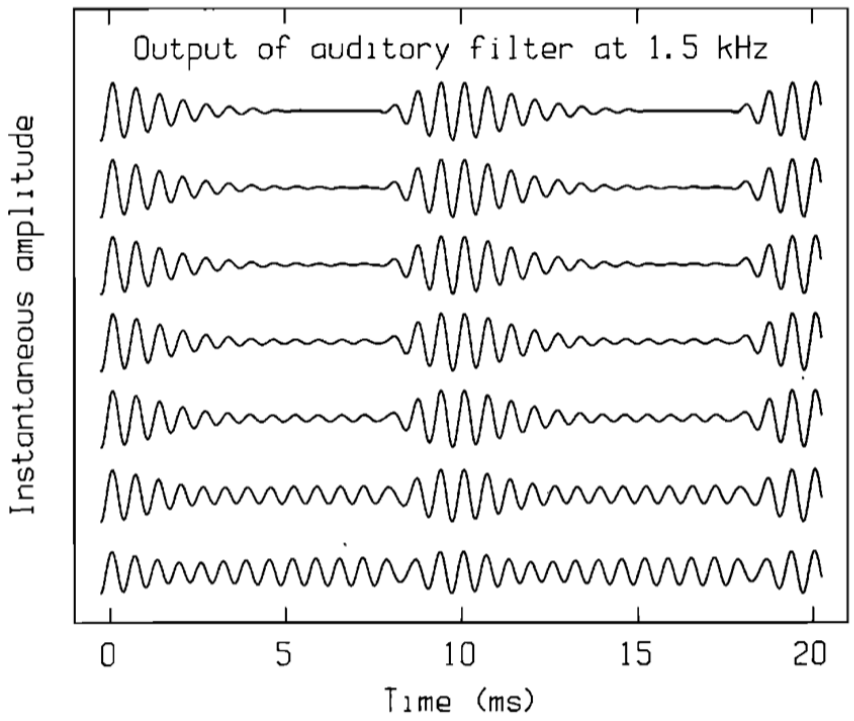


Figure 4.3: The output waveforms of a simulated auditory filter centered at 1.5 kHz in response to complex tones with 20 equal-amplitude harmonics of 100 Hz. For the top waveform, all of the harmonics started in cosine phase. For the successive traces, the 15th harmonic was advanced in phase by 6.0°, 7.1°, 13.4°, 20.1°, 53.6°, and 90°, respectively. In a study by Moore and Glasberg (1989), these were the mean thresholds for detecting a phase shift for stimulus levels of 80, 70, 60, 50, 40, and 30 dB SPL per component, respectively. Adopted from Moore and Glasberg (1989).

while the amplitude is constant. The expression describing the amplitude-modulated signal is

$$\big(1 + m \cdot \sin(2\pi g t)\big)\sin(2\pi f_c t)\,, \tag{4.1}$$

where f_c is the carrier frequency, g is the modulating frequency, t is time and m is the modulation index, which is a constant determining the amount of modulation. The expression for a frequency-modulated signal is

$$\sin\big(2\pi f_c t - \beta\cos(2\pi g t)\big)\,, \tag{4.2}$$

where β is the modulation index. (Moore, 1997)

The complex waveforms of both the AM wave and the FM wave can be analyzed into their sinusoidal components. The AM wave is a complex tone consisting of three tones: the carrier at the center and the other tones, called sidebands, on both sides. The frequency components have frequencies $f_c - g$, f_c and $f_c + g$. The spectrum of an FM wave usually

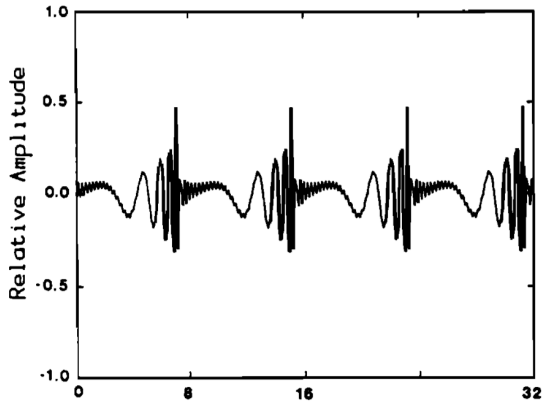
consists of many components, but if β is small, the FM wave can also be considered as containing three components, which have the same frequencies than those of the AM wave. When the carrier frequencies and modulation frequencies are the same and when the modulation indices are equal ($m = \beta$), the components of an AM wave and an FM wave are equal in frequency and amplitude, the only difference being in their relative phases. (Moore, 1997) Moreover, any difference in our sensitivity to AM and FM, is likely due to the auditory system being sensitive to the relative phases of the components (Scharf, 1961).

Furthermore, Zwicker (1952) measured the just-detectable amounts of amplitude and frequency modulation as a function of modulation frequency, using such stimuli. For low modulation frequencies, when the frequency separation Δf between the sidebands is small, AM could be detected when the relative levels of the sidebands were lower than for a wave with just-noticeable amount of FM ($m < \beta$). Whereas, for high modulation frequencies, when Δf is large, the thresholds for detecting AM and FM were equal when the amplitudes of the components of each of the waves were equal ($m = \beta$). Thus, it can be said that subjects appear to be sensitive to the relative phases of the components, when the frequency separation Δf is small, but for wide frequency separations they are not. If the threshold for detecting modulation is expressed as the ratio β/m, it decreases with increasing modulation frequency and approaches an asymptotic value of unity. The modulation frequency at which this ratio first reaches unity is called the critical modulation frequency (CMF). Unlike it may seem and unlike Zwicker (1952) suggested, the CMF does not give a direct measure of the CB, as is explained in the literature (Moore, 1997).

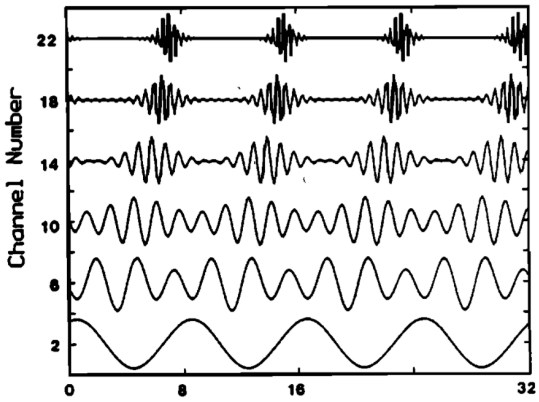
Nevertheless, the study of Zwicker (1952) shows that we are sensitive to relative phases of components, which are within a critical band. It can be also said that the relative phases in the complex tone are not a significant cue in the detection of modulation when Δf is greater than the CB (Scharf, 1961). However, subjects can detect phase changes between components which are spaced considerably wider than the CB, when the components are all well above threshold of audibility (Patterson, 1987; Blauert and Laws, 1978).

4.1.3 Global phase changes

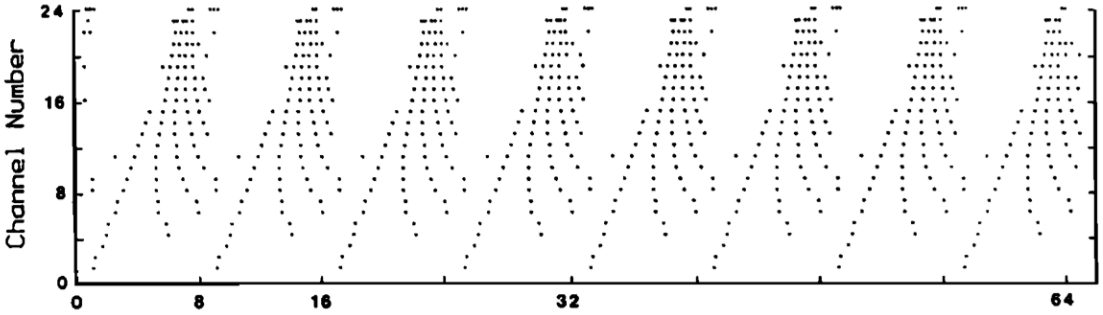
In a stimulus, global, or a between-channel, phase changes are used to produce a progressive phase shift between successive auditory filter channels, without changing the envelopes of the auditory filter signals. Patterson (1987) used a sound wave with such phase modifications to study global phase changes which may affect the timbre. Such a waveform was created using a monotonic phase spectrum with a slow deceleration. This monotonic-phase (MPH) wave consisted of 31 equal-amplitude harmonics of 125 Hz, equally to the CPH wave. The phase spectrum was created calculating the number of ERBs between adjacent harmonics with a function suggested by Moore and Glasberg (1983), and incrementing the phase by 180° per one ERB. This kind of phase function reflects the rate of change of phase in the cochlea and makes between-channel time delays roughly equal, minimizing local phase changes. Furthermore, a scalar was applied to the entire phase spectrum to vary the absolute size of the phase increments from component to component. When the scalar is about 10, the stimulus effectively becomes an RPH wave, because the starting phase of each component is practically arbitrary. The MPH wave and its response to the auditory-filter model of Patterson (1987) are illustrated in Figure 4.4. In this case, a scalar of 1/2 was used.



(a) Waveform



(b) Driving waves



(c) Initial pulse ribbon

Figure 4.4: Outputs of different stages of the pulse ribbon model of Patterson (1987) for a monotonic-phase wave (a) with 31 equal-amplitude harmonics of 125 Hz, used to study global phase changes. The driving waves (b) are the outputs of the auditory filter bank driving the corresponding pulse generators, which produce the initial pulse ribbon (c). Adapted from Patterson (1987).

The purpose of the auditory model by Patterson (1987) is to produce a change in the aligned pulse ribbon only if a change in timbre is perceived, and it has been tuned accordingly. The model involves a mechanism that aligns the initial pulse ribbon produced by the driving waves. Since the envelopes of the driving waves produced by the MPH wave in Fig. 4.4b are practically the same as those of the CPH wave in Fig. 4.1b, the pulse ribbon of the MPH wave can be made the same as that of the CPH wave by just realigning all the channels. However, there is a limit to the size of the between-channel time shifts that the alignment mechanism should correct, so that the aligned pulse ribbon indicates a change in timbre.

Patterson (1987) showed that for broadband signals with a bandwidth of 5 or more auditory channels, the between-channel phase discrimination, or MPH discrimination, occurs when the total time shift across the channels reaches 4 – 5 ms. They investigated the threshold by varying the scalar applied to the phase spectrum so that the slant of the initial pulse ribbon changed. In the case when the slant was positive, the columns of the initial pulse ribbon were tilted to the right, as in Fig. 4.4c. Because the propagation of low frequencies is delayed in the cochlea with respect to high frequencies, the positive slant cancels this

propagation delay behavior. The magnitude of the threshold slant was larger for the negative slant values than those for the positive slant. I.e., for the stimuli that accentuate the propagation delay of the cochlea, the magnitude of the threshold slant is higher than for the stimuli that counteract the propagation delay. (Patterson, 1987)

Patterson (1987) also found that there is much less variation in the thresholds of between-channel phase change discrimination than in those for within-channel phase discrimination. Secondly, repetition rate has only a marginal effect on the results, and the direction of the effect is reverse compared to that of the within-channel experiments. Thirdly, in contrast to the within-channel experiments, level has little effect on MPH threshold and there is no interaction with repetition rate. Similarly to the within-channel experiments, the duration of the stimulus has only a small effect on the results.

4.1.4 Summary on the perception of local and global phase changes

Local phase changes are discriminated depending on the secondary column that appears between the main arches in the response to the pulse ribbon model of Patterson (1987) (see Fig. 4.2c). At threshold, these secondary maxima, seen also in the driving waves (Fig. 4.2b), are just large enough to occupy some of the firing capacity of the pulse generators. Patterson (1987) suggested that this proportional capacity used for the main arches increases in the cases when within-channel discrimination threshold increases, i.e., when repetition rate increases, level decreases or spectral location decreases. As the between-channel cue is resided in the dominant structures of the pulse pattern, it is the last one to disappear when the repetition rate is raised. However, when the repetition rate rises from 250 to 500 cps, the changes in timbre perception associated with phase changes are no longer present. This means that the upper boundary for the residue pitch of similar stimuli is about an octave higher than the upper boundary for phase effects. (Patterson, 1987)

There are different cues to which discrimination of local and global phase changes are based on. Firstly, the cue for the discrimination of local (within-channel) phase changes is in the quiet parts of the envelope; a phase-shifted component appears as a low-level signal in the low-amplitude portions. Secondly, for global phase changes the cue lies in the peaks of the envelopes across the auditory filter channels (Patterson, 1987; Mathes and Miller, 1947; Moore and Glasberg, 1989).

Within-channel phase effects are perceived more easily at high sound levels, which can be explained with the recovery from forward-masking of the main peaks being faster at higher levels (Moore, 1997); as the level is decreased, the secondary low-level maxima decrease below the absolute threshold or below the threshold due to the forward-masking of the preceding main peak (Moore and Glasberg, 1989). As the low-amplitude portion broadens with decreasing repetition rate, the within-channel phase sensitivity increases. More amplitude modulation occurs at high harmonic numbers, because more components interact within the passband of the auditory filter, and therefore phase shifts are more easily detected for high harmonic numbers (Moore and Glasberg, 1989).

In summary, within-channel phase changes are heard between a cosine-phase signal and a signal with every other component advanced in phase by at least 40°, when the level of the 31-component signal is 40 dB per component and the repetition rate is 125 cps. In this case, discrimination depends strongly on the repetition rate, intensity and spectral

location of the signal. The discrimination threshold increases as repetition rate increases. On the other hand, between-channel phase shifts are discriminated once the total time delay across the channels containing the signal reaches 4 – 5 ms. Discrimination is in this case dependent only on signal bandwidth. In other words, a relatively large phase change between auditory filter bands is needed for it to be heard.

4.2 On the effect of phase spectrum changes on timbre in general

The sensitivity to phase has been discussed extensively in the previous section, but only few articles have described how phase changes can affect the timbre of sound (Mathes and Miller, 1947; Plomp and Steeneken, 1969; Ozawa et al., 1993; Galembo et al., 2001; Banno et al., 2002; Laitinen et al., 2013). Patterson (1987) suggested that timbre perception is affected by the relative phases of harmonics below the 12th, the repetition rate being 125 cps or less. He concluded that musical notes below middle C and quality of mens' voices depend on component phase relations. Anechoic recordings of, for example, speech, trumpet, and trombone sounds (Laitinen et al., 2013) as well as applause-type signals (Laitinen et al., 2011) have been reported to be phase-sensitive signals, because randomization of phase of these signals is detectable.

At low repetition rates, say 125 cps or less, the first few auditory filter channels capture the fundamental component of the stimulus. In the low-frequency end also the next frequency components are dominant in more than one channel. The width of the ERB increases with increasing frequency, and thus, at the high-frequency end an auditory filter captures several harmonic components. The activities in different auditory filters contribute to and are a factor of the perception of timbre. As was discussed previously in section 3.4, the relative loudness of a component can increase the more phase locking occurs to that frequency (Moore, 1997). Thus, phase locking is in an important role in the coding of loudness and timbre in the auditory system.

Furthermore, Laitinen et al. (2013) suggest that the auditory system has interaction between adjacent auditory bands. They showed that harmonic signals, which differ in a certain manner only in their phase spectra, are perceived to contain different amounts of bass, i.e., that the perceived loudness of the lowest harmonics depends on the phase spectrum. The timbral effects of the phase spectrum are more prominent the lower the repetition rate, as was mentioned in subsection 4.1.1. In the experiments of Laitinen et al. (2013), the effect of phase spectrum on the perceived amount of bass was clear at repetition rates not higher than 50 and 100 cps. The observations by Laitinen et al. (2013) led to the research done for this thesis and they are discussed further in section 6.1.

Other timbre effects due to phase modification have also been shown (Mathes and Miller, 1947). Rotating the phase of a high harmonic of a narrow-band stimulus leads to a change in the perceived roughness of the sound. The rough-sounding tone had an envelope with intervals of zero or very low amplitude, whereas the envelope of the smooth sound was more nearly uniform in value. The degree of roughness is related to the relative length and depth of the depressions in the envelope wave. The change from roughness to smoothness can be made by changing either the amplitude or phase of a particular component or set of components. Periodicity in the envelope also created a sensation of an apparent pitch,

which seems to be the same as the residue pitch described in, e.g., Moore (1997). Mathes and Miller (1947) studied harmonic tones of fundamental frequencies in the voice range, but found out that the results are identical with inharmonically related but equally spaced frequency components. As was mentioned earlier, the random-phase (RPH) wave was also reported (Patterson, 1987) to sound rougher than the corresponding cosine-phase (CPH) wave.

The maximal difference in timbre obtained with phase-modifications is between a signal with only sine or cosine terms and another signal with alternative sine and cosine terms according to Plomp and Steeneken (1969). Secondly, the maximal difference in timbre due to phase is quantitatively smaller than the effect of changing the slope of the magnitude response by 2 dB/oct. Also this effect is more noticeable at low fundamental frequencies. The effect of phase is larger the higher the number of the harmonic being phase-shifted. Thirdly, the effect of phase on timbre seems to be independent of the effect of the magnitude pattern and of the loudness.

It should be noted, that a signal is phase-modified also due to the group delay characteristics of common transducers. Blauert and Laws (1978) studied the audibility of group delay distortion with two group delay patterns resembling those measured from actual loudspeakers and earphones. They found out that the group delays caused by common loudspeakers and earphones are on the order of 0.4 ms, and that the threshold of perceptibility is on the order of 0.5 – 1.1 ms. Thus, in most cases, it is not necessary to correct those additional group delay distortions.

The fact that phase effects are perceivable suggests that there is a neural mechanism involved (Mathes and Miller, 1947). I.e., only the mechanical resolution of the cochlea does not explain the differences in perception due to modifications of signal phase, because the mechanics of the cochlea are considered to act as a frequency analyzer and phase-modifications occur in time and not in frequency domain. According to Moore (1997), the detection of phase changes between components separated by more than a CB may depend partly on the ability of the auditory system to compare the time patterns of the outputs of different auditory filters. Moreover, Moore and Glasberg (1989) suggested that the sensitivity to a phase shift of a single component depends both upon the frequency selectivity and the temporal resolution of the ear. Plomp and Steeneken (1969) stated that the way phase affects timbre can be considered, up to a certain frequency limit, as originating from the correlation between the vibration patterns at different points on the basilar membrane and the corresponding time patterns of the neural firings.

Chapter 5

Audio evaluation

The purpose of this chapter is to describe the theory and practice behind audio evaluation, and in this way to provide a basis for the listening tests conducted for this thesis. The basics of audio evaluation is discussed first. Thereafter, basic theory of statistical analysis including linear and circular statistics is presented.

5.1 Fundamentals of audio evaluation

The nature of audio and the perception of audio is multidimensional. The physical characteristics of audio signals can be measured with rather high accuracy. What we cannot do yet, is to perform a direct measurement of human perception of an audio signal. Therefore, in order to specify how the human auditory system will interpret that audio signal, we need to ask listeners to quantify their experience or perception. This is commonly done as a formal listening test. However, before the decision of applying a listening test, the researcher should consider the following. One should figure out whether there is a physical measurement of the signal that would give the needed information. Secondly, it should be determined whether there is a direct measurement of the perceived audio quality. Thirdly, the researcher should consider whether there is a predictive model of perceived audio quality applicable for the evaluation of the stimuli in question.

Listening tests are a common practice in the field of, e.g., evaluation of audio codec performance. If a listening test is the only way to measure the perception of the stimuli in question, it should be noted, that listeners quantify their experience in a manner which is on one of two levels, as defined by Bech and Zacharov (2006):

- **Perceptual measurement**, which means 'objective quantification of the sensorial strength of individual auditory attributes of the perceived stimulus'

- **Affective measurement**, which means 'objective quantification of an overall impression of the perceived stimulus.'

In other words, perceptual measurement refers to assessment of individual attributes of sound, and affective measurement refers to the overall impression of the sound. With listening tests it can be identified whether audio stimuli are perceptually identical or not, which is relevant considering also the experimental part of this thesis. Listening test results

can also indicate whether a stimulus is superior to another, and to which degree it is superior in terms of audio quality. (Bech and Zacharov, 2006) The qualities of audio stimuli can be assessed in detail using perceptual attributes, e.g., the amount of bass or low-end. Furthermore, familiarization or training prior to a listening test is recommended (ITU-R, 1997), because it can turn subjects with low assessment ability to expert listeners for the purposes of the test in a short period of time.

As was mentioned already, perceptual evaluation can alternatively be estimated by means of predictive modeling techniques. Usually, this method employs a perceptual model of the human auditory system and cognitive processes (Bech and Zacharov, 2006). Therefore, such predictive models simulate the response of listening test subjects. Predictive models are desired because they are cost-effective and significantly faster than formal listening tests. These predictive models can be divided in to two classes, namely, those that are made to predict a particular perceptual attribute, such as loudness, and those that can quantify the overall performance, such as audio quality. Speech and audio codecs have developed in recent years, and therefore many predictive models for speech and audio quality have been developed and standardized. One problem is that such predictive models usually have a restricted domain of application, and they cannot be used beyond that reliably, because the prediction accuracy becomes unknown. (Bech and Zacharov, 2006) For the purpose of this thesis, we are particularly interested in predictive perceptual models taking into account the effect of phase spectrum on timbre, such as those of Patterson (1987) and Laitinen et al. (2013). Although there are models which predict audio quality and its attributes, listening tests will still be a crucial part of perceptual audio evaluation as well as the development and verification of predictive models.

Scientific argumentation is an important matter in every scientific study regardless of the experimental method. When planning a scientific experiment, a testable statement based on a research hypothesis and initial conditions needs to be formulated (Bech and Zacharov, 2006). There are two main principles of scientific argumentation. One is the principle of empiricism, which claims that only experiments and observations can verify whether theories and hypotheses are true, i.e., whether they are scientific knowledge. The other is the principle of rationalism, which claims that the correct way to obtain scientific knowledge is deduction using common sense. These principles are also known as the inductive and the deductive principles, respectively. The deductive principle states that if the hypothesis and the initial conditions, i.e., the two premisses, are true, the testable statement (the conclusion) must be also true. If the conclusion is false and the initial conditions are true, the hypothesis is false. The testable statement should be formulated in terms of falsification, i.e., so that it is aimed to be proven false. (Bech and Zacharov, 2006)

These two principles are the most commonly used in scientific articles (Bech and Zacharov, 2006). Perceptual experiments usually involve a mixture of these principles as they are within the domains of exact sciences (physics) and social sciences (psychology). The initial conditions include the experimental conditions, and, thus, the truthfulness of the initial conditions depends on the quality of the experimental part of the research. Experimental planning is therefore very important. (Bech and Zacharov, 2006)

There are also other relevant types of scientific argumentation used in practical experiments in engineering, in contrast to theoretical approaches, e.g., mathematics. These are, e.g., probabilistic reasoning, ‘argumentum ad hominem’, and conclusion by analogy. Probabilistic reasoning means that the truthfulness of the hypothesis depends on a limited number

of observations from a given population. It is the typical approach for subjective audio evaluation, as only a limited number of subjects, samples, listening rooms, loudspeaker positions etc. can be tested. 'Argumentum ad hominem' refers to the principle of using personal statements from workers in the field as facts without referencing them, which should be generally avoided. Finally, the principle of conclusion by analogy means that the conclusion is derived to be valid from its similarity to the premisses, but this is rarely used. (Bech and Zacharov, 2006)

There are a number of standards in the field of perceptual audio evaluation. These key standards include ITU-T and ITU-R recommendations for perceptual evaluation of speech and audio quality, respectively. Examples of general guidance for the ITU-R recommendations are ITU-R (2003b) and ITU-R (2003c). ITU-R (1997) and ITU-R (2003a) are recommendations on listening test methods. When a standard exists in the desired domain, it should be considered whether that standard is applicable to the task to be performed. Moreover, if a standardized approach is used outside its original domain, modification of the approach may be needed. (Bech and Zacharov, 2006)

5.2 Statistical analysis

In this section, statistics that is relevant considering the statistical analysis performed to the results of this thesis is presented. Prior to conducting a listening test an experimental plan should be defined. Such plan includes the used methods for statistical analysis of the listening test results. First, basics of statistical analysis are discussed with a focus on linear statistics. Second, essential differences in the statistical analysis of circular data compared to that of linear data is discussed in subsection 5.2.2.

5.2.1 Basics of statistical analysis

An important part of experimental research is the statistical experimental planning. The factors under study are divided into two classes: independent and dependent variables. Independent variables are factors controlled by the experimenter. These can be, for example, the phase spectrum of the stimulus or the background noise with four levels. The response(s) by the test subject is the dependent variable. The purpose of statistical analysis in evaluation of audio is to find out whether the controlled variables of the listening test have an influence on the perceptual response(s) (dependent variable) to the stimulus (Bech and Zacharov, 2006). The research question of a listening test is whether the variation in the subjective response is a result of the intended variation of the experimental independent variables, or is it likely to be random variation. Secondly, by statistical analysis it can be determined how generalizable the conclusions are. Therefore, the main goal of an experiment is to statistically test the independent variables, and to test any possible interactions.

Statistical analysis is performed on three levels: descriptive, inferential and measurement levels. On the descriptive level an overview of the collected data is obtained by verification of the statistical assumptions, detection of outliers and by a summary of the observations as means and associated variances. On the inferential level the experimental hypothesis is rejected or accepted, i.e., the effect of the independent variables on the dependent

variables is estimated. Measurement level includes the estimation of the relationship between the independent and dependent variables. This relationship is often analyzed for each dependent variable separately, and for this univariate analysis methods such as analysis of variance (ANOVA) and regression analysis are used. (Bech and Zacharov, 2006)

Before choosing the statistical analysis method, the type of the resulting data should be considered. The data can be classified as quantitative or categorical data (Bech and Zacharov, 2006). Quantitative data refers to a continuous quantity or the degree of a given attribute. This type of data can be, for example, ratings reported on an interval scale and are usually normally distributed. Such data is analyzed using methods such as the standard t-test, ANOVA, etc., and this type of methods are also referred to as parametric statistics. Categorical data are based on categorical scales such as 'prefer' or 'not prefer', and they are often binomially or chi-squared distributed. The analysis methods used for this type of data are referred to as non-parametric statistics. The distinction between these two data types depends on the measurement scale used and leads to the question, when data from a scale can be considered continuous? (Bech and Zacharov, 2006) Moreover, data can be separated also depending on whether they are on linear or circular scale.

The meaning of statistical analysis at the inferential level is to make conclusions about the relationship between the true mean and the estimated mean values, and to compare two or more estimated mean values (Bech and Zacharov, 2006). The confidence interval of the mean gives the relationship between the true population mean and the estimated mean. The statistical meaning of the confidence interval can be explained as follows. If several identical listening tests were performed using different samples from the population(s) (of, e.g., subjects) each giving a sample mean and a 95% confidence interval, then 95% of these confidence intervals would include the true mean value. (Bech and Zacharov, 2006)

ANOVA is used to identify whether any significant differences exist within a group of means. The purpose of the ANOVA is to 'compare the deviation between the sample means of the treatments to the random deviation within the samples and test whether the magnitude of the ratio is higher than a certain critical value.' (Bech and Zacharov, 2006) Standard ANOVA can be applied only to data on a linear scale, which is usually the case. For the linear data gathered from the first experiment of this thesis, a repeated-measures analysis of variance (RM-ANOVA) was applied (see subsection 6.4.2). If Mauchly's test revealed that sphericity was violated, a correction was used. The correction method was Greenhouse-Geisser when $\epsilon < 0.75$, and Huynh-Feldt correction was used when $\epsilon > 0.75$. The results including all the significant main effects and interactions are presented in subsection 6.4.2.

5.2.2 Circular statistics

Essential differences in the theory of statistical analysis of directional data compared to that of linear data is presented next. First, descriptive measures are discussed, and thereafter, inferential statistics made use of later in this thesis are addressed.

Circular statistics is a subfield of statistics, and it involves statistical techniques for the use with angular data. In addition to variables that are naturally measured in angles, circular statistics also applies to such types of data as time of the day, phase of the moon, or day of the year, that all have a different periodicity. This type of data can be converted to a

common angular scale in radians by

$$\alpha = \frac{2\pi x}{k}, \tag{5.1}$$

where x is the original representation of the data, k is the total number of steps in the original scale that x was measured in, and α is the angular direction. A sample of N directional observations α_i is denoted as $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$.

Two essential descriptive measures of angular data are considered next: the mean and the mean resultant vector. The mean of an angular sample $\boldsymbol{\alpha}$ can not be calculated as the average of the data. Instead, the directions of the observations are first transformed to unit vectors on the unit circle by

$$\bar{r_i} = \begin{pmatrix} \cos \alpha_i \\ \sin \alpha_i \end{pmatrix}. \tag{5.2}$$

Second, the vectors $\bar{r_i}$ are vector averaged according to

$$\bar{r} = \frac{1}{N} \sum_{i=1}^{N} \bar{r_i}\,, \tag{5.3}$$

where $\bar{r}$ is the mean resultant vector for the sample. The sample mean angular direction can be then expressed as (Berens, 2009)

$$\bar{\alpha} = \mathrm{Arg}(\bar{r}). \tag{5.4}$$

The length of the mean resultant vector is an important quantity considering hypothesis testing in directional statistics. The population mean resultant vector length is computed by

$$R = \|\bar{r}\|, \tag{5.5}$$

and it has a value between 0 and 1. The closer it is to zero, the more the data sample is spread around the mean direction.

Next, we will consider inferential statistics for circular data. Let us consider first the calculation of **confidence intervals** with the significance level δ for the mean angular direction. The $100 \cdot (1 - \delta)\%$-confidence intervals for the population mean are computed according to the following formulae. For $R \leq 0.9$ and $R > \sqrt{\chi^2_{\delta,1}/2N}$, the confidence interval is (Berens, 2009)

$$d = \arccos\left[\frac{\sqrt{2N(2R_n^2 - N\chi^2_{\delta,1})}}{R_n}\right], \tag{5.6}$$

where $R_n = R \cdot N$. For $R \geq 0.9$, the confidence interval is (Upton, 1986)

$$d = \arccos\left[\frac{\sqrt{N^2 - (N^2 - R_n^2)\exp\left(\chi^2_{\delta,1}/N\right)}}{R_n}\right]. \tag{5.7}$$

In both cases, the lower confidence limit is $L_l = \bar{\alpha} - d$, and the upper confidence limit is $L_u = \bar{\alpha} + d$.

The Watson-Williams two- or multi-sample test is a circular analogue of the two-sample t-test or the one-factor analysis of variance (ANOVA) in the linear scale. This test defines whether the mean direction of two or more groups are identical or not. Thus, the null hypothesis and the alternative hypothesis are (Berens, 2009):

H_0: All of s groups share a common mean direction, i.e., $\bar{\alpha}_1 = \cdots = \bar{\alpha}_s$.
H_A: Not all s groups have a common mean direction.

The test statistic is calculated by (Watson and Williams, 1956)

$$F = K\frac{(N-s)\left(\sum_{j=1}^{s} R_j - R\right)}{(s-1)\left(N - \sum_{j=1}^{s} R_j\right)}, \tag{5.8}$$

where R is the mean resultant vector length when all samples are combined and R_j the mean resultant vector length of the jth group. The correction factor K is calculated from

$$K = 1 + \frac{3}{8\kappa}, \tag{5.9}$$

where κ is the maximum likelihood estimate of the concentration parameter of a von Mises distribution. The Watson-Williams test assumes underlying von Mises distributions with a common concentration parameter, but has proven to be robust against deviations from these assumptions (Berens, 2009). The von Mises distribution $VM(\mu, \kappa)$ can be considered as a circular analogue of the normal distribution, and it is the most common unimodal circular distribution. Its probability density function can be found in, e.g., Berens (2009). The sample size for the test should be at least 5 for each sample. The use of binned data is advised only if bin widths are not larger than 10°. (Berens, 2009) It should be noted that when the null hypothesis is rejected, it indicates only that not all groups have the same mean, and not how many of the means differ and how much they differ from each other. The application of the statistical tests for the results with circular data are described in the corresponding section (subsection 6.5.2).

Chapter 6

Listening tests

The experimental part of this thesis is described in this chapter. The motivation for the current research is discussed first in section 6.1. The used test setup and the studied stimuli are described in section 6.2 and section 6.3, respectively. Two listening tests were conducted and they are discussed in section 6.4 and section 6.5, respectively. Both of these sections are divided into method, results and discussion parts.

6.1 Motivation

The perception of phase spectrum modifications has been studied in many articles, as was discussed in chapter 4 and section 4.2. Some of those articles describe also how the timbre of a tone can be affected by such modifications, as was discussed in section 4.2. For example, randomizing the starting phases of a harmonic complex tone makes it sound rougher. On the other hand, random-phase stimuli have been described as thinner and colored compared to a signal in which the phase between the harmonics is aligned (Laitinen et al., 2013). However, to the author's knowledge, the changes in perceived bass due to phase-modifications of such stimuli has been studied only by Laitinen et al. (2013).

In Laitinen et al. (2013) a conclusion was made, that with a certain cumulative constant phase shift for successive harmonics a continuous harmonic complex tone becomes perceptually louder in its low end, i.e., a bass boost is heard. They assumed that the bass boost is maximal between the tone with a certain progressive phase shift and its inverse wave. The two studied phase-modification cases were signals with phase shift constants 90° and -90°, of which the former was, in general, perceived with a higher level of bass. In both cases, the fundamental component started in cosine-phase. At fundamental frequencies 50 and 100 Hz, the increase in the perceived amount of bass was reported to correspond to a magnitude spectrum amplification at low frequencies by 2 – 4 dB on average. Interestingly, different subjects perceived the amount of bass very differently (Laitinen et al., 2013). Additionally, Laitinen et al. (2013) suggest a predictive auditory model for indicating whether a phase change produces a change in perceived timbre of the harmonic complex tone. This model has a similar purpose than that of Patterson (1987), which was discussed in chapter 4.

The findings by Laitinen et al. (2013) led to the current research, and several observations

were made by informal listening. By informal listening it was noticed that the inversion of the phase-modified wave with the perceptually loudest bass produces a tone which has less bass than the original cosine-phase tone. Thus, it seemed likely that the largest difference in perceived bass was between the tone with the perceptually loudest bass and its inversion in amplitude. Additionally, Laitinen et al. (2013) noticed that the discrimination between the different phase-modified tones seemed to be easier in a noisy environment, or when background noise was added. This result was confirmed by informal listening. Therefore, background noise was anticipated to affect the results of listening tests in which these type of stimuli are employed. By informal listening to two-tone complexes with the first two harmonics of the fundamental frequency of 50 Hz and with different phase shifts for the second harmonic, it was noticed that at least one person heard the amount of bass contrary to the author's judgement. With such two-tone complexes, the differences in perceived amount of bass were much more vague. It was also observed that when keeping the phase shifts for successive harmonics of the wide-band complex tones unchanged (i.e., 90° and -90°), but changing the starting phase of the fundamental component from cosine-phase to sine-phase, the tone with previously less bass became now the tone with more bass and vice versa. Furthermore, out of the tested headphones, the largest bass effects were perceived with Sennheiser HD 598 headphones. Moreover, it was noticed that the perceived pitch changed slightly depending on the value of the additive phase shift for successive harmonics.

For the current study, two formal listening tests were conducted in order to investigate the roles of background noise and individuality in the perception of bass of the phase-modified harmonic complex tones. In the first listening test, the effect of background noise on the discrimination of the phase-modified tones based on the perceived amount of bass was studied. The purpose of this test, described in section 6.4, was to find a suitable level of background noise for the second experiment, which is described in section 6.5. The second listening test was formulated in a way to determine the preferred value of the constant phase shift that produces a tone with the perceptually loudest bass and how much this varies between individual subjects.

6.2 Test setup

A MOTU UltraLite mk3 Hybrid audio interface was used in the tests, because its magnitude and group delay responses were known to be relatively flat. Sennheiser HD 650 headphones were used in both of the tests. These are open headphones, which are preferred in listening at low frequencies, as was mentioned in section 3.4. The HD 650 headphones are known to have a typical group delay response; the group delay increases when approaching low frequencies. The group delay response is possible to be corrected in theory Blauert and Laws (1978), but for practical reasons the correction was omitted. The sound pressure levels of the test samples were measured with a B&K Artificial ear and a linear-phase microphone B&K 4192 in a quiet listening booth. All audio samples were created with Matlab software, and Max/MSP was used to program the procedures and to create the user interfaces of the listening tests.

6.3 Phase-sensitive stimuli

The stimuli used in the experiments were synthetic harmonic complex tones created according to

$$x(t) = G\sum_{n=1}^{\infty}\frac{1}{n}\cos(2\pi n f_0 t + \phi_{\mathrm{n}}), \tag{6.1}$$

where G is a gain controlling the overall level of the signal, n is the sequential number of the harmonic and ϕ_{n} is a frequency-dependent phase angle for controlling the phase spectrum. All harmonics were created starting from the fundamental component until 20 kHz. The phase spectrum was set according to $\phi_{\mathrm{n}} = (n-1)\cdot\alpha$. The angles α producing eight different signals, depicted in Figure 6.1, were, from 1 to 8: $0°, 45°, 90°, 135°, 180°, -135°, -90°$, and -45°, i.e., one value for each signal. Thus, these signals have a phase shift of α between successive harmonics. Note, that the fundamental component started always in cosine-phase, i.e., $\phi_1 = 0$.

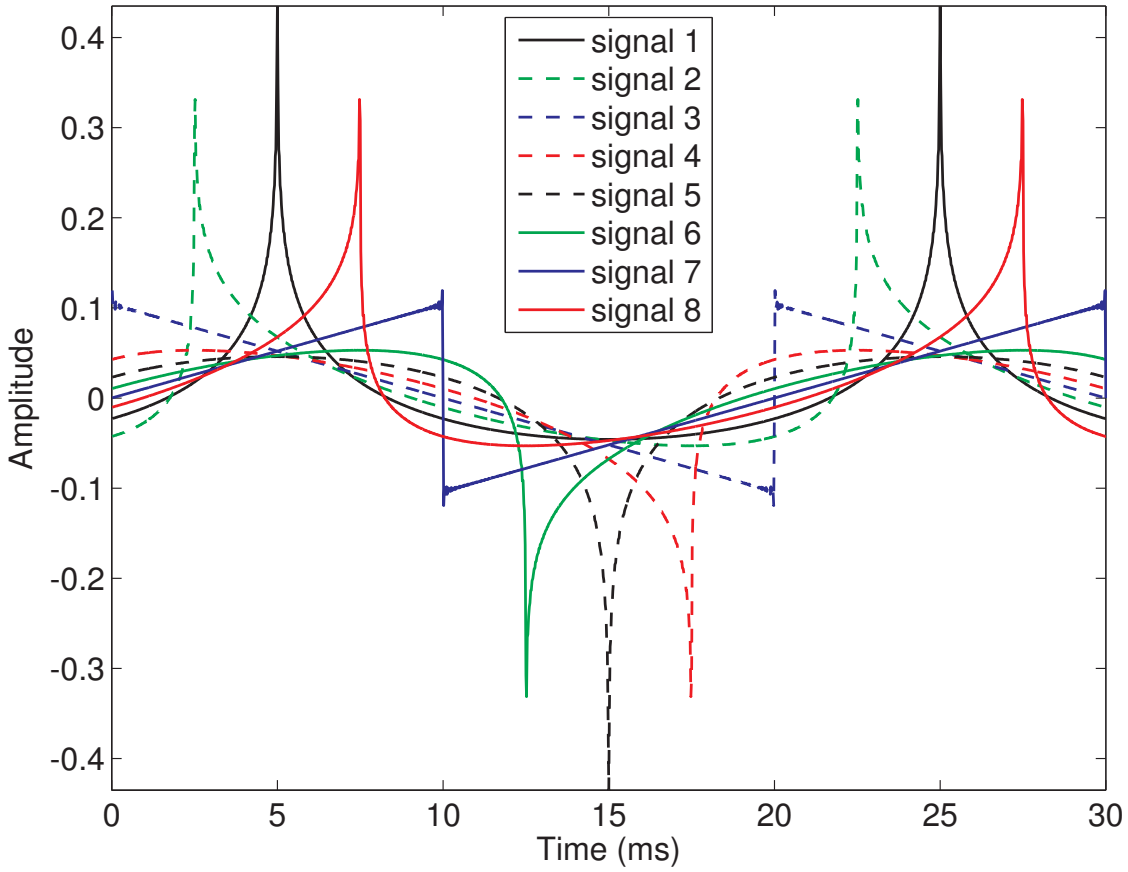


Figure 6.1: The phase-sensitive stimuli used in the listening tests. The signals have a fundamental frequency of 50 Hz and were created with the following additive phase shifts (α): signal 1: 0°; signal 2: 45°; signal 3: 90°; signal 4: 135°; signal 5: 180°; signal 6: −135°; signal 7: −90°; signal 8: −45°.

Note from Figure 6.1 that signals 5 – 8 are the negations of signals 1 – 4, and that signals 3 and 7 (in blue) are essentially sawtooth waves. These stimuli are perceived as sharp, 'buzzy' and with different amounts of bass. Thus, they are perceived to differ in loudness and timbre. These type of phase-modified stimuli can be considered to correspond to, for example, anechoic voiced vowels. These harmonic complex tones have relatively large phase-shifts in degrees between adjacent harmonics, except signal 1, in which all the harmonics start in cosine-phase. On the other hand, the phase-shifts 'wrap around' the period of the waveform (2π). Therefore, a phase-shift essentially moves the corresponding frequency bin to another location in time within the period of the complex waveform.

6.4 Listening test 1: Effect of background noise on the discrimination of loudness differences due to phase spectrum modifications

6.4.1 Method

As informal listening suggested, the discrimination of changes in the perceived amount of bass of the phase-sensitive stimuli seemed to be easier in a regular room with a little background noise from ventilation etc. than in anechoic conditions. In other words, the noise made the increase in perceived bass noticeably larger, which is against intuition. To the author's knowledge similar phenomena have not been studied before. It was noticed by informal listening that the perceived difference in loudness between these stimuli was mainly caused by the changes in perceived level of bass. Thus, a formal listening test was conducted in order to find out how the amount of background noise affects the discrimination of subjective changes in loudness of the phase-sensitive stimuli. The listening test was organized as follows.

Three types of continuous audio signals were used in the listening test: two harmonic complex tones (sawtooth waves), a target noise, and background noise at four different levels including silence. The sawtooth waves are harmonic complex tones at 50-Hz fundamental frequency with identical $1/f$ magnitude spectra as described by Equation 6.1, and they are depicted in Figure 6.1 as signals 3 and 7. Note, that these two signals are each other's negations in amplitude as well as inversions in time. It was assumed that the difference in perceived bass was the largest between these two tones. The target noise has also $1/f$ magnitude spectrum, i.e., it is Brownian noise. The background noise was in all cases a stereo signal with 50% coherence between right and left channels, and it was created from three different white noise sequences with uniform distributions. All the other samples used in the test were presented diotically, i.e., the same signal was played to both ears.

The test was constructed as follows. The graphical user interface of the listening test is presented in Figure 6.2. At each trial, the test sequence consisted of one-second samples of either of the sawtooth waves and the Brownian noise, which were presented in cascade twice, i.e., as 'A, B, A, B'. Additionally, the background noise was ten seconds long; it started one second before and ended one second after the A and B samples. The test subject was asked 'which one is louder?' and was told to choose A or B. Hence, the listening test was a 2-alternative forced choice procedure. The test subject was allowed to repeat each sample sequence once before answering.

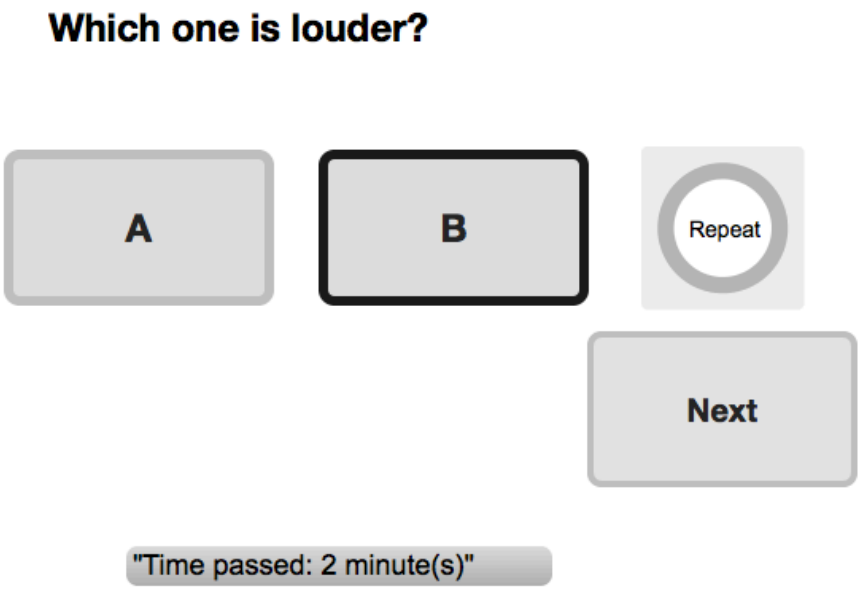


Figure 6.2: The graphical user interface used in test 1. Only one repetition per question was allowed using the 'Repeat' button.

The loudness difference between the sawtooth signals was assessed at four background noise levels, by comparing the loudness of both of them with the target noise. Thus, the test consisted of eight test cases. The level of the target noise was adjusted adaptively during the test starting from -4 dB, because it was assumed to be assessed quieter than the sawtooth signals. At the reference (0 dB) level, the RMS of the target noise equals that of the sawtooth waves. The target noise level was increased by 2 dB after each correct answer. After the first incorrect answer, the test proceeded with a 2-Up/1-Down rule with a step size of 1 dB. This rule means that after two correct responses, the level of the target noise was increased by the step size. Conversely, when the response was incorrect, the level of the target noise was decreased by one step. 2-Up/1-Down adaptive procedure converges to the 29.3% point on the psychometric function (Levitt, 1971). All of the eight adaptive tracks were presented in an interleaved manner and in random order. Each adaptive track ended after ten turns.

As was mentioned in chapter 5 to be advisable, each test subject had a training session prior to the listening test. The practice run was about 8 minutes long on average, and the actual listening test took around 25 to 40 minutes to complete. The practice run was the same as the main test until the first turn in each adaptive track, with the exception that the target noise level started from -8 dB. Because complete silence was desired, the tests were performed in an anechoic chamber using the gear described in section 6.2. The test subjects used an iPad to give their answers. 20 test subjects, excluding the author, participated in the test. All of the test subjects had earlier experience in listening tests.

The signals under test (sawtooth waves), had sound pressure levels of 67 dB. In addition to the case without background noise, the sound pressure levels of the background noise samples were about 45 dB (low), 50 dB (mid), and 55 dB (high). The studied stimuli were reported to be very 'buzzy' and to cause listening fatigue easily, as was anticipated beforehand. Therefore, the test subjects were given the option for a pause during the test. The data from each test subject consisted of the ten turn points as gain values for the target noise, for all of the eight test cases.

6.4.2 Results

One test subject judged the harmonic tones to be of equal loudness when the target (brown) noise was amplified as much as 14 dB. This judgement was at the upper limit of the applied level scale of the test procedure and, therefore, this subject had to be excluded from the results. For one of the test subjects the overall sound level of the test had to be adjusted 6 dB lower than for others because of listening fatigue. Based on informal listening, it was known that the phase effects are more difficult to hear the lower the sound level, similarly to what Patterson (1987) showed. Nevertheless, this subject was included in the results. This is because also the overall level of assessed thresholds varied significantly between test subjects in general, and because this subject was fairly consistent with their judgements. Therefore, data from 19 test subjects was included in the results.

The thresholds in all of the eight test cases were calculated as the mean of the six last turn points of the corresponding adaptive track. First, a two-factor RM-ANOVA was applied for the statistical analysis of the data. The results are presented next, and based on them, further analysis is then performed. The within-subjects factors were *signal*, which refers to the two sawtooth waves, and *noise*, which refers to the four levels of background noise. It was revealed with Mauchly's test that the assumption of sphericity was violated in the cases of factor $noise, \chi^2(5) = 32.235, p < 0.05, \epsilon < 0.75$; and factor $signal, \chi^2(0) = 0, p < 0.05, \epsilon > 0.75$. RM-ANOVA revealed the following significant factors: main effect $signal, F(1, 27.512) = 26.061, p < 0.05$; and main effect $noise, F(1.402, 25.238) = 32.111, p < 0.05$.

The significant main effects are studied further. Pairwise comparison of the noise measure showed that the mean differences between all the pairs are statistically significant. Furthermore, pairwise comparison of the signal measure showed that also the mean difference between the two signals is statistically significant ($p < 0.05$). The used confidence interval adjustment was Bonferroni. The initial results of the listening test are presented in Figure 6.3, which shows the 95% confidence intervals. From this data, it cannot be said whether the background noise makes the difference in perceived bass of the two signals larger, because the interaction is not significant: $noise * signal, F(3, 54) = 2.218, p = 0.097$. However, the significance of the interaction is very close to the confidence limit. The difference between the thresholds of the two sawtooth waves for each test subject are investigated further because the overall level of the assessed thresholds varied substantially depending on the test subject.

The data was studied further by recalculating it for each test subject and background noise combination as the difference in assessed level between signal 3 and signal 7. The second statistical analysis was that a one-factor RM-ANOVA was applied to the recalculated data with the threshold differences of the studied signals. The within-subjects factor

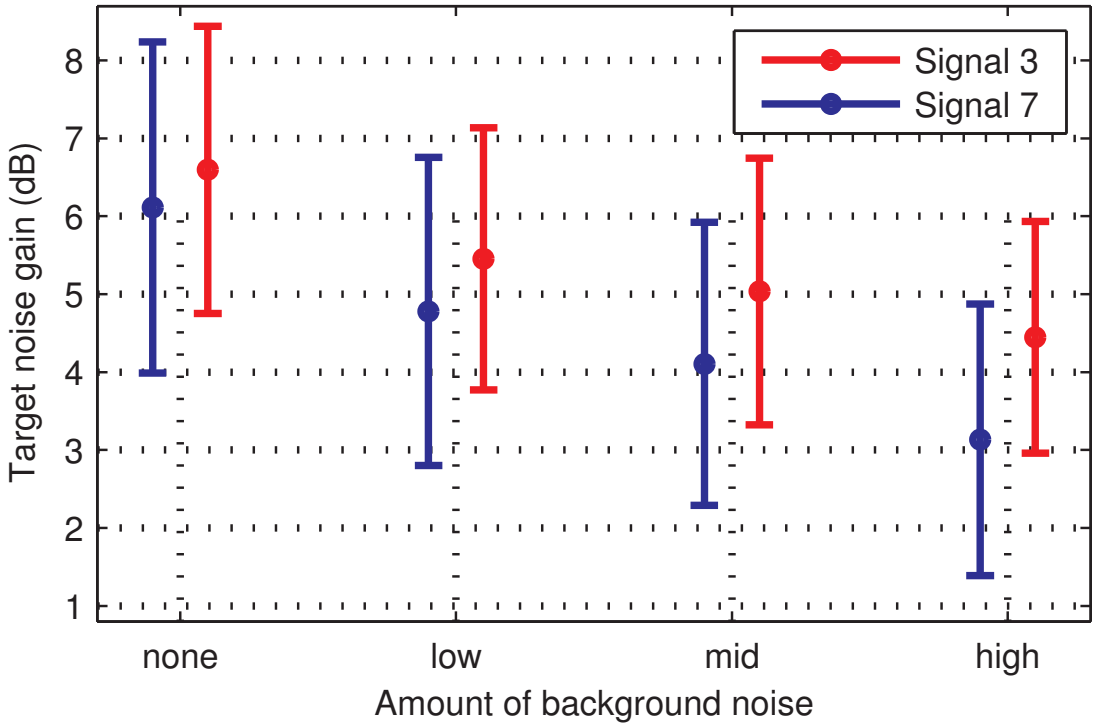


Figure 6.3: Results of listening test 1 plotted as means of 19 test subjects for all of the test cases. 95% confidence intervals are shown. The signals 3 and 7 are those depicted in Figure 6.1.

was (background) *noise*, $F(3, 54) = 2.218, p = 0.097$. Thus, this factor is not statistically significant but it is very close to the confidence limit ($p = 0.05$). In Figure 6.4, the means of the differences between assessed thresholds for the two signals are plotted. It can be seen, that there is a general trend upwards in perceived difference towards higher background noise levels. More specifically, the means and confidence intervals of the cases 'mid' and 'high' are further away from zero than those of the two lowest background noise levels. This and the low p-value suggest that background noise does increase the perceived level difference between the signals under study. However, the growth of the difference is not statistically significant.

Additionally, Figure 6.3 shows a trend that the overall loudness of the studied samples are assessed lower the higher the background noise level is. Therefore, for each test subject and for both signals under study the original data was recalculated so that the threshold for the case without background noise was subtracted from all threshold values. Thus, data indicating the decrease in overall level was obtained. As the third analysis of the results, a one-factor RM-ANOVA was applied separately to such data of both signals under study.

The data for signal 3 is analyzed first. The within-subjects factor was (background) *noise*. Based on Mauchly's test the assumption of sphericity was violated in the case of this factor: *noise*, $\chi^2(5) = 25.248, p < 0.05, \epsilon < 0.75$. RM-ANOVA revealed that the factor was significant: main effect *noise*, $F(1.855, 33.397) = 21.013, p < 0.05$. The data for signal 7 is analyzed next. The within-subjects factor was *noise*. Based on Mauchly's

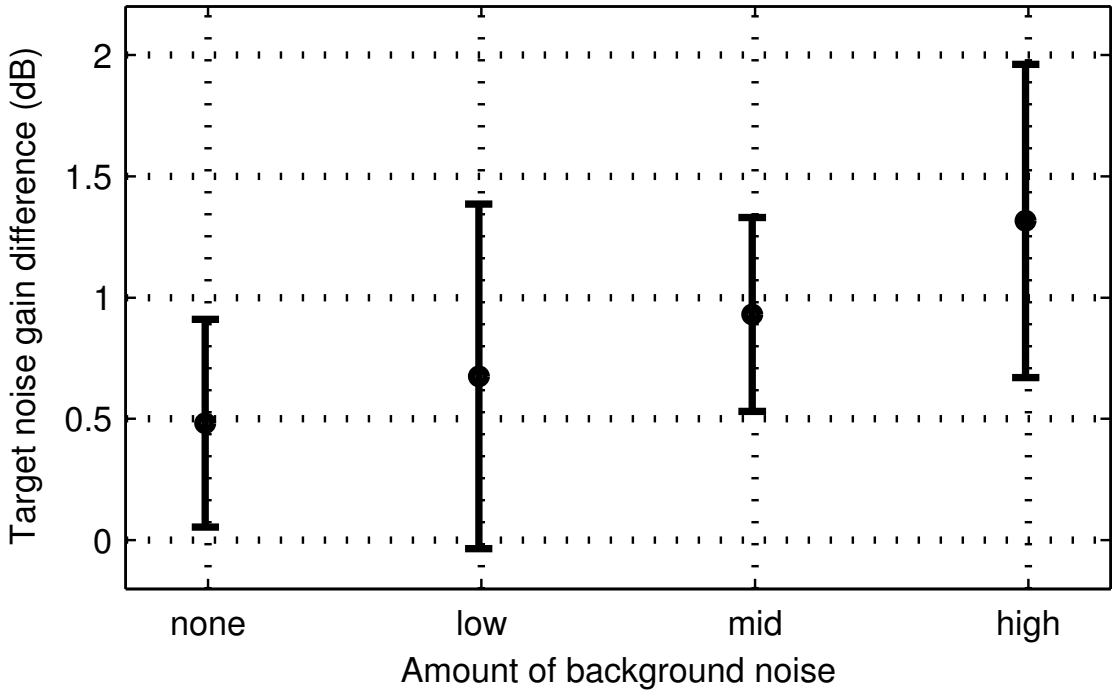


Figure 6.4: Results of listening test 1 plotted as the difference of the assessed thresholds for the signals under study. 95% confidence are shown.

test the assumption of sphericity was violated in the case of this factor: $noise, \chi^2(5) = 22.318, p < 0.05, \epsilon < 0.75$. RM-ANOVA revealed that this factor was significant: main effect $noise, F(1.823, 32.809) = 25.104, p < 0.05$. The means of the data for both signals are plotted in Figure 6.5 with 95% confidence intervals. The confidence intervals for the cases with background noise are apart from those of the cases without background noise ('none'). Thus, it can be seen from this figure, that the overall perceived level of the signals decreases as the level of background noise increases.

6.4.3 Discussion

Because it seems likely that the two signals with 90° and -90° phase shift constants, respectively, do not produce the maximal difference in bass for every subject, the data for the effect of background noise is ambiguous. The results do not prove that the perceived loudness difference between the studied two signals would increase statistically significantly with increasing background noise level. However, the results do neither show that background noise would have no effect on the perceived loudness difference. Therefore, according to the upward trend of the means and confidence intervals towards high background noise levels in Figure 6.4, it is justifiable to conclude that background noise makes the discrimination of loudness differences between the phase-sensitive tones easier. Moreover, based on informal listening the loudness differences between the studied signals can be assumed to be associated mainly with the boost at low frequencies. Thus, the bass boost becomes more notable with increasing background noise level.

As was discussed in section 3.5, a similar phenomenon has been reported by Houtgast

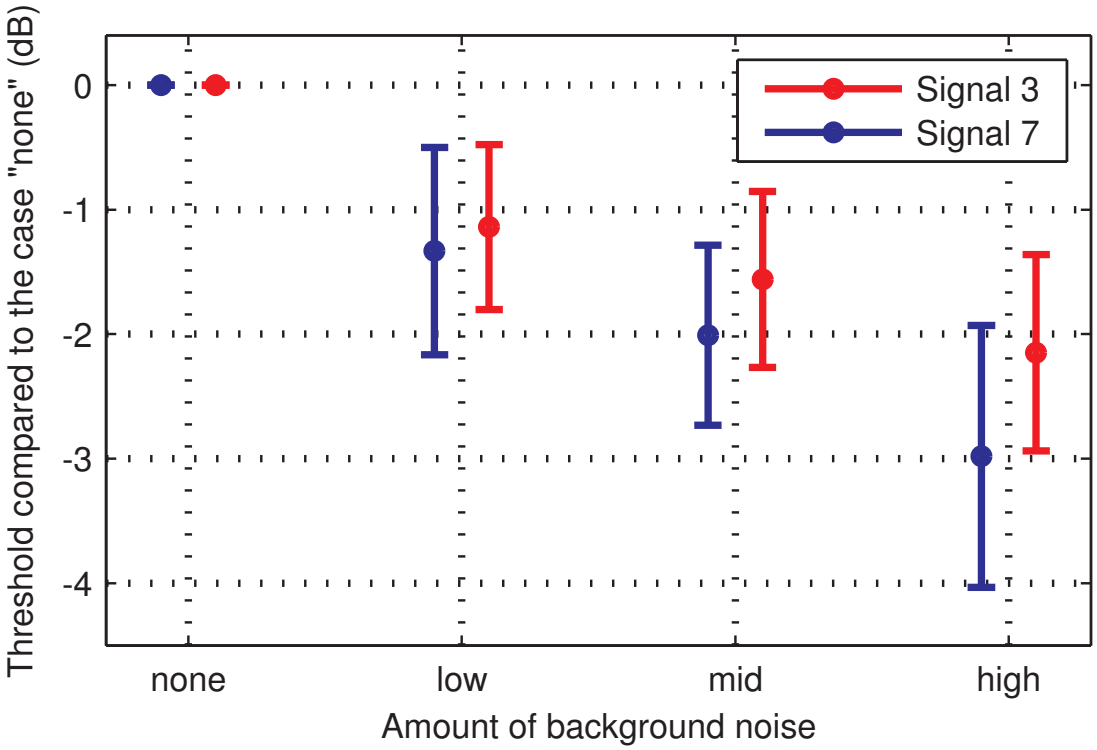


Figure 6.5: Results of listening test 1 plotted for data in which, for both signals under study, the thresholds are calculated relative to the corresponding background case 'none'. The means and 95% confidence intervals are shown.

(1976) who showed that only a single sinusoidal component evoked a subharmonic pitch, namely a residue pitch, in the presence of background noise. If similar auditory mechanisms control the perception of complex tones, increasing background noise would then increase the fundamental and a few other low harmonic components in relative loudness, and thus, alter the timbre. As the harmonic complexes extend to the upper end of the audible range, high frequency components within many auditory filters interact on the basilar membrane, and therefore according to Schouten (1940) many residue pitches may be heard. It was also mentioned in section 3.5 that these residue pitches are determined by the temporal pattern of the waveform at the point on the basilar membrane where the harmonics interfere. Moreover, the most prominent of these pitches corresponds to the perceived pitch of the complex. It should be noted also, that the white background noise gradually starts to mask more high harmonics of the complex tones as its level is increased, which then focuses the attention more to the low frequency components and corresponding (residue) pitches.

As the pairwise comparisons showed, the mean difference between the two signals is statistically significant. This indicates that the studied signals are assessed to differ in loudness. Additionally, as background noise increased, the reference noise needed to be adjusted lower to sound equally loud with the harmonic complexes. Thus, it can be concluded that the overall perceived loudness of these phase-modified signals decreased with increasing background noise level.

6.5 Listening test 2: The preferred additive phase shift constant for successive harmonics for the maximum perceived amount of bass

6.5.1 Method

The results of the first listening suggested that some background noise should be added when listening to the phase-sensitive stimuli, so that the subjective differences in bass are more noticeable. Based on that result, another listening test was conducted in order to find out the value of the phase shift constant that produces a tone with the perceptually loudest bass. Especially, the variation of this phase shift constant between individuals was of interest. The test was organized as follows.

The tested phase-sensitive signals were the harmonic complex tones described by Equation 6.1 with a fundamental frequency (f_0) of 50 or 100 Hz. Formal listening tests (Laitinen et al., 2013) had shown, that already at fundamental frequency of 200 Hz the perceived differences in bass between the phase-sensitive stimuli are marginal. This was confirmed by informal listening and, thus, fundamental frequencies over 100 Hz were not studied in this experiment. Eight different phase-modifications were tested, and the waveforms are plotted in Figure 6.1 for fundamental frequency of 50 Hz. The length of each of the studied tones was four seconds, and the background noise sample was six seconds long. All of the samples were played in a loop. The subjects used a graphical user interface, shown in Figure 6.6, to listen to the eight samples in each test set. The subjects were able to change between different samples freely and were asked to choose the sample with the loudest bass. Based on informal listening, the subjects were advised to compare first the samples opposite to each other in the circle, because the samples with least and most bass can be found this way. There were eight repeats for both fundamental frequencies and, thus, 16 test sets in total. The order of the samples in the circle was kept the same between test sets, while the rotation of the circle was randomized. Additionally, the order of the test sets was randomized.

The listening test was organized in two parts. First, there was a brief training session, where the participants could listen to all of the different samples as much as they wanted and get familiarized with the user interface. The training sessions took about 2-3 minutes at most, depending on the test subject. With the training session the subjects could become familiar with the differences under test. Second, there was the actual listening test, which took about 13 minutes on average. Because the test sounds were noticed in listening test 1 to cause listening fatigue easily, this time the test was designed to be relatively short.

Based on the results of listening test 1, the sound pressure level of the background noise was chosen to be about 49 dB, which is a little lower than the 'mid' noise level in that test. The studied samples had sound pressure levels of about 69 dB. Thus, the relative loudness of the background noise was between the 'low' and 'mid' noise levels of test 1. The listening test was performed in a soundproof listening booth. 20 test subjects took part in the listening test, and 15 of them had participated also in the first listening test. All of the participants had earlier experience in listening tests. The resulting data from each test subject consisted of the number of the sample which was judged to be the loudest in bass for each of the 16 test sets.

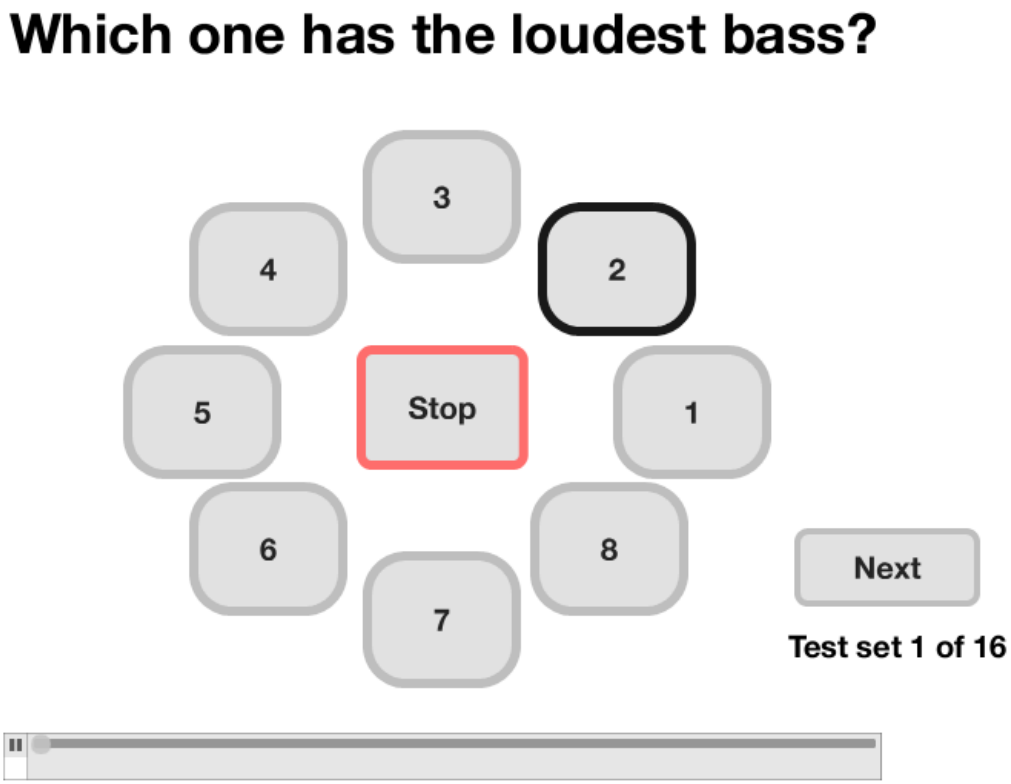


Figure 6.6: The graphical user interface of test 2.

6.5.2 Results

The results from the 20 test subjects are presented in Figure 6.7 for both fundamental frequencies. The data points in cyan are the mean values for data that did not meet the conditions for the calculation of confidence intervals of the circular data. These conditions were presented previously in subsection 5.2.2.

Parametric Watson-Williams multi-sample test for equal means was used for statistical analysis of the results. This test can be considered as a one-way analysis of variance (ANOVA) for circular data. The test was computed for both fundamental frequencies. The within-subjects factor was *participant*. The Watson-Williams test revealed that this factor is significant for the fundamental frequency of 50 Hz: $participant, F(19, 140) = 17.441, p < 0.05$. Similarly for the fundamental frequency of 100 Hz, the test revealed that the factor *participant* is significant: $F(19, 140) = 10.839, p < 0.05$. Thus, in both cases the test shows that the null hypothesis, that the means are equal in the population, can be rejected. From this it follows, that the test subjects hear the phase spectrum required to produce a tone with the perceptually loudest bass, differently from each other, and that this difference is statistically significant.

Indeed, it can be seen from Figure 6.7 that test subjects assessed the samples very differently but consistently. It is also interesting to see, that for each test subject, at the higher

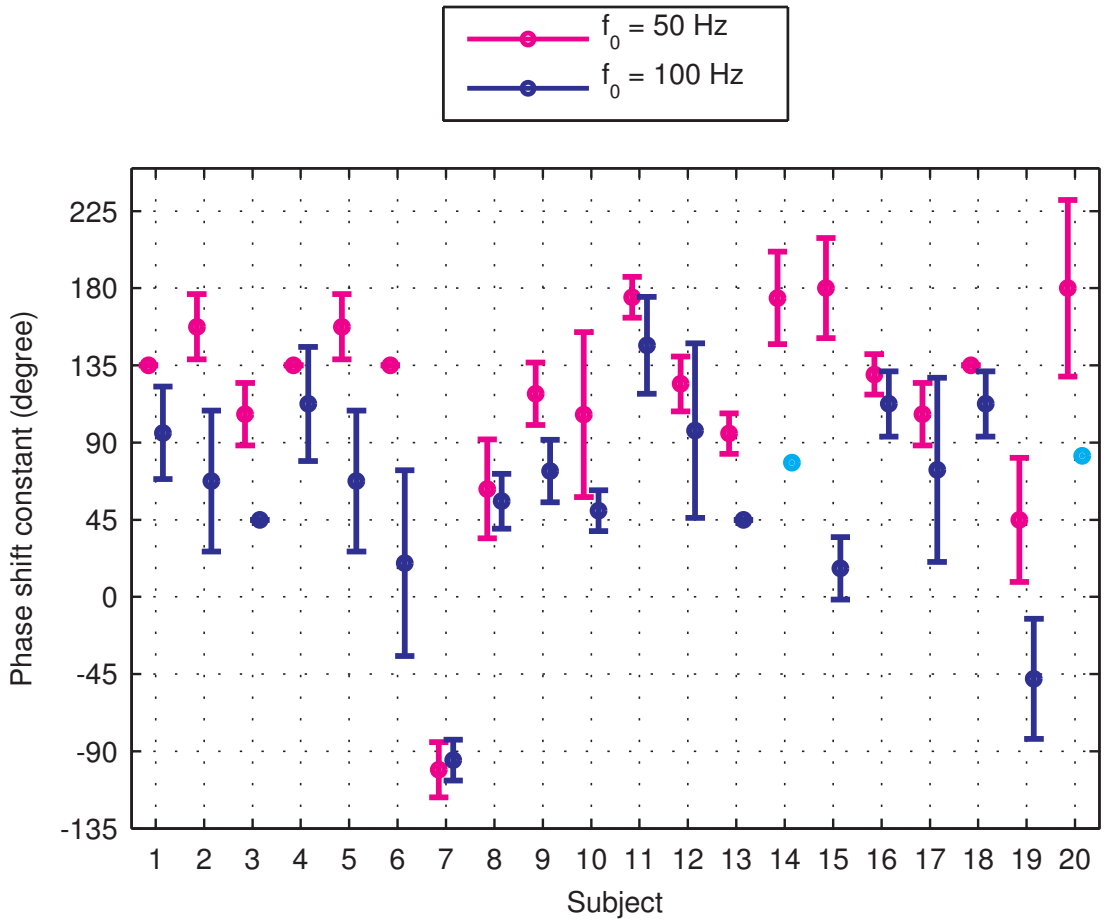


Figure 6.7: Results of listening test 2 plotted as the means and 95% confidence intervals for each test participant. The test subjects were asked to adjust the phase shift constant so that the perceived amount of bass is maximal. The cyan data points represent the mean for subjects whose data did not fulfill the requirements for the resultant vector length in order to calculate the confidence intervals.

fundamental frequency the additive phase shift constant is the same or a value clockwise on the unit circle from the case of the lower fundamental frequency. I.e., the blue/cyan data points are next to or lower than the red data points for each test subject in Figure 6.7. Some subjects reported changes in perceived pitch between the samples, and that these pitch shifts were not correlated with the perceived amount of bass. Moreover, some subjects reported that it was difficult to determine the 'amount of bass' as the samples differed in perceived loudness, timbre and pitch.

6.5.3 Discussion

The results show that in the case of harmonic complex tones at fundamental frequency of 50 Hz, the loudest low end can be obtained by modifying the phase spectrum in a heavy manner, significantly away from the situation where all components start in cosine-phase. However, at the fundamental frequency of 100 Hz, the additive phase shift required to

produce the perception of the loudest bass was judged to be sometimes also 0°, which corresponds to the cosine-phase wave.

It can be noticed from Figure 6.7 and Figure 6.1, that for the majority of the test subjects, the signal in which in each period the amplitude rises rapidly and decreases gradually is perceived as containing the most bass (e.g., constant 135°). Note, that the signal produced by the phase shift constant that is ±180° from its counterpart (on the unit circle) is the negation of the counterpart signal. These are the signal pairs shown with the same color in Figure 6.1, e.g., signals with constants 135° and -45°. Based on informal listening, the signal with the phase shift angle producing maximal bass and its ±180° counterpart signal differ the most in perceived bass, i.e., the counterpart is perceived with the least amount of bass.

As can be seen from Figure 6.7, there is significant variation between individuals in the perceived amount of bass of the studied phase-sensitive stimuli. For the lower repetition rate one individual, and for the higher repetition rate two individuals differ substantially in their judgements from the others. For example, for subject 7 the phase shift constant that produces the maximum amount of perceived bass differs almost 180° from the average. These results are in line with the study by Laitinen et al. (2013) who reported that in their listening tests, individuals had very differing but consistent judgements for the perceived level of bass of similar stimuli.

The results are an evidence that for harmonic complexes timbre is affected also by the temporal structure of the stimulating waveform. The results also indicate that there exists an individual mechanism in the coding of relative loudness of the lowest components of a harmonic complex tone. An explanation of the timbre differences due to phase spectrum modifications could be a change (increase) in the amount of phase locking to the lowest component frequencies (see subsection 3.4.3) or an emphasis of the residue pitch(es) (see section 3.5). However, the coding of the timbre changes in these cases is beyond the scope of this thesis and is left for future studies.

Chapter 7

Conclusions

The perception of bass of phase-sensitive signals was studied in this thesis. First, an introduction on the research field and the topic of the thesis was given. Thereafter, basic signal processing theory was introduced. The meaning of the phase response of audio systems was discussed. In common electroacoustic audio systems the magnitude spectrum of the input signal as well as the input waveform are distorted. Transducers such as earphones can be defined as systems with minimum-phase and all-pass subsystems. Thereafter, human hearing was discussed. The function of the ear and properties of neural firing was dealt with. An acoustic stimuli is transduced into neural firings. In response to a periodic sound, these neural firings tend to occur at particular phase of the stimulating waveform, which is called phase locking. Phase locking occurs at low frequencies and its precision decreases with increasing frequency above $1 - 2$ kHz until at about $4 - 5$ kHz phase locking no more occurs. It was shortly explained that there are active biological processes in the auditory system that affect the mechanics of the cochlea. It seemed that the higher centers of the auditory system can control even the earliest stages of auditory processing.

Auditory perception was discussed starting from the concepts of the critical band and masking. This was continued by the discussion about the perception of loudness, pitch and timbre. It was noted that the relative levels of the partials in a complex tone depend partly on the amount of phase locking to those component frequencies. Thereafter, research on the perception of phase spectrum changes was reviewed. It was concluded that phase spectrum is thought to affect the timbre of certain types of sounds. These sounds include applause-type, speech, trumpet and trombone signals. Similar signals can be synthesized as harmonic complex tones. The perception of phase changes is based on a combination of cues, which includes at least cues within and between auditory filter channels. The auditory system uses comparison of temporal envelopes across auditory filters to perform pattern analysis. Thereafter, experimentation on audio evaluation was discussed followed by a brief description of statistical analysis methods including circular statistics.

The motivation for the listening tests came from recent research which had shown that the phase spectrum of harmonic complex tones seems to have an effect on the perceived amount of bass. Such tones were called phase-sensitive stimuli. It was noticed that already a tone with two harmonics of a low fundamental frequency the perceived amount of bass seemed to change when the second component was shifted different amounts in phase. By informal

listening, it was noticed also that the phase shift that caused the maximal difference in bass varied between listeners. This bass effect was found to be more substantial with wide-band harmonic stimuli with a constant phase shift for successive harmonics. Recent research had shown also that, surprisingly, the perceived difference in bass between two of the phase-sensitive tones seems to increase when background noise is added. This was confirmed by informal listening in an anechoic chamber by adding white noise.

Based on these observations, two listening tests were conducted using similar synthetic harmonic (phase-sensitive) complex tones. The aim of the first listening test was to determine the level of background noise that is optimal for hearing the differences in bass. Two sawtooth signals were studied, and the difference in perceived loudness was measured. According to the results, the loudness difference does not increase statistically significantly with increasing level of background noise. It should be noted that the results do neither prove that background noise would have no effect on the perceived loudness difference. In fact, the low p-value of the interaction suggested that a significant effect possibly exists. The results are ambiguous at least because, as was studied further in the second experiment, the constant phase shifts that the sawtooth signals represent are not the pair with the largest possible perceived difference in bass for all of the subjects. It was shown, however, that there is a significant perceived loudness difference between the studied sawtooth signals, and that the overall perceived level of the stimuli decreases as background noise level increases.

The purpose of the second listening test was to investigate how different listeners perceive the bass of the phase-sensitive signals. In this test, eight different phase spectra created with eight different values for the phase shift for successive harmonics were studied. The results show that the constant phase shift that produces the signal with the loudest perceived bass varies significantly between individuals. The results are an evidence that timbre can be affected by modifying the phase spectrum only. It was debated that the timbre differences in question could be signaled as changes in the phase locking pattern of the neural signals. Whether changing the phase spectrum of a complex stimulus affects the pattern of phase locking (to the low spectral frequencies) is a question for future studies. Additionally, experiments on this phenomenon using loudspeakers is left for future studies.

In conclusion, phase spectrum modifications occur in the signal processing of, for example, audio coding. In the case of certain signals that are similar to harmonic complex tones, it should be taken into account that phase spectrum changes can cause changes in timbre. These timbre changes are associated at least with the changes in perceived bass. These effects of phase on timbre depend also on the group delay spectrum of the acoustic system and, especially, the headphones.

Bibliography

ASA. American Standards Association and others. Acoustical terminology SI, 1–1960. *American Standards Association, New York*, 1960.

H. Banno, K. Takeda, and F. Itakura. The effect of group delay spectrum on timbre. *Acoustical Science and Technology*, 23(1):1–9, 2002. ISSN 13463969. doi: 10.1250/ast.23.1. URL http://joi.jlc.jst.go.jp/JST.JSTAGE/ast/23.1?from=CrossRef.

S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2006.

G. Békésy. The variation of phase along the basilar membrane with sinusoidal vibrations. *The Journal of the Acoustical Society of America*, 1947. URL http://link.aip.org/link/?JASMAN/19/452/1.

P. Berens. CircStat: A MATLAB Toolbox for Circular Statistics. *Journal of Statistical Software*, 31(10), 2009. URL http://kyb.mpg.de/publications/attachments/J-Stat-Softw-2009-Berens_6037[0].pdf.

J. Blauert and P. Laws. Group delay distortions in electroacoustical systems. *The Journal of the Acoustical Society of America*, pages 1478–1483, 1978. URL http://link.aip.org/link/?JASMAN/63/1478/1.

R. Carlyon and B. Moore. Intensity discrimination: A severe departure from Weber's law. *The Journal of the Acoustical Society of America*, 76(November):1369–1376, 1984. URL http://link.aip.org/link/?JASMAN/76/1369/1.

H. Davis. Advances in the Neurophysiology and neuroanatomy of the cochlea. *The Journal of the Acoustical Society of America*, 409(1952), 1962. URL http://link.aip.org/link/?JASMAN/34/1377/1.

E. De Boer. *On the "residue" in hearing*. PhD thesis, Uitgeverij Excelsior, 1956.

R. Feldtkeller and E. Zwicker. Das Ohr als Nachrichtenempfänger, 1956.

H. Fletcher. Auditory patterns. *Reviews of modern physics*, 1940. URL http://rmp.aps.org/abstract/RMP/v12/i1/p47_1.

A. Galembo, A. Askenfelt, L. L. Cuddy, and F. a. Russo. Effects of relative phases on pitch and timbre in the piano bass range. *The Journal of the Acoustical Society of America*, 110(3):1649, 2001. ISSN 00014966. doi: 10.1121/1.1391246. URL http://link.aip.org/link/JASMAN/v110/i3/p1649/s1&Agg=doi.

B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.

D. Griesinger. Phase Coherence as a Measure of Acoustic Quality, part three: Hall Design. In *Proceedings of 20th International Congress on Acoustics, ICA*, number August, 2010. URL http://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ICA2010/papers/p604.pdf.

J. Hall, M. Haggard, and M. Fernandes. Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America*, 76(July):50–56, 1984. URL http://link.aip.org/link/?JASMAN/76/50/1.

J. Herre, K. Kjörling, and J. Breebaart. MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding. *Journal of the Audio Engineering Society*, 56(11), 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14643.

T. Houtgast. Subharmonic pitches of a pure tone at low S/N ratio. *The Journal of the Acoustical Society of America*, pages 405–409, 1976. URL http://link.aip.org/link/?JASMAN/60/405/1.

ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Rec. ITU-R BS.1116-1, 1997.

ITU-R. Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA). Rec. ITU-R BS.1534-1 Annex 1, 2003a.

ITU-R. A guide to ITU-R Recommendations for subjective assessment of sound quality. Rec. ITU-R BS.1283-1, 2003b.

ITU-R. General methods for the subjective assessment of sound quality. Rec. ITU-R BS.1284-1 Annex 1, 2003c.

D. T. Kemp. Stimulated acoustic emissions from within the human auditory system. *The Journal of the Acoustical Society of America*, 64(5):1386–91, Nov. 1978. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/744838.

M.-V. Laitinen and V. Pulkki. Utilizing Instantaneous Direct-to-Reverberant Ratio in Parametric Spatial Audio Coding. *Audio Engineering Society Convention 133*, 2012. URL http://www.aes.org/e-lib/browse.cfm?conv=133&papernum=8804.

M.-V. Laitinen, F. Kuech, S. Disch, and V. Pulkki. Reproducing Applause-Type Signals with Directional Audio Coding. *Journal of the Audio Engineering Society*, 2:29–43, 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=15774.

M.-V. Laitinen, S. Disch, and V. Pulkki. Sensitivity of human hearing to changes in phase spectrum of continuous signals. *Accepted to Journal of the Audio Engineering Society*, 2013.

H. Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, pages 467–477, 1971. URL http://link.aip.org/link/?JASMAN/49/467/1.

P. H. Lindsay and D. A. Norman. *Human information processing*. ACADEMIC PRESS, 1972.

R. Mathes and R. Miller. Phase Effects in Monaural Perception. *The Journal of the Acoustical Society of America*, 1947. URL http://link.aip.org/link/?JASMAN/19/780/1.

S. K. Mitra. *Digital signal processing: a computer-based approach*. McGraw-Hill Higher Education, New York, 3rd edition, 2006.

B. C. Moore. Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. *Scandinavian audiology. Supplementum*, 25:139–152, 1985.

B. C. Moore, editor. *Hearing (Handbook of Perception and Cognition)*. Handbook of Perception and Cognition. Academic Press, 2nd edition, 1995. ISBN 0125056265. URL http://books.google.fi/books?id=31NRAAAAMAAJ.

B. C. Moore. *An introduction to the psychology of hearing*. ACADEMIC PRESS, 4th edition, 1997.

B. C. Moore. Interference effects and phase sensitivity in hearing. *Philosophical Transactions of the Royal Society*, pages 833–858, 2002. URL http://rsta.royalsocietypublishing.org/content/360/1794/833.short.

B. C. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(September):750–753, 1983. URL http://link.aip.org/link/?JASMAN/74/750/1.

B. C. Moore and B. R. Glasberg. Difference limens for phase in normal and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, 86(4):1351–65, Oct. 1989. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/2808909.

B. C. Moore and K. Ohgushi. Audibility of partials in inharmonic complex tones. *The Journal of the Acoustical Society of America*, 93(1):452–61, Jan. 1993. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/8423261.

G. S. Ohm. Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. *Annalen der Physik*, 1843. URL http://onlinelibrary.wiley.com/doi/10.1002/andp.18431350802/abstract.

K. Ozawa, Y. Suzuki, and T. Sone. Monaural phase effects on timbre of two-tone signals. *The Journal of the Acoustical Society of America*, pages 1007–1011, 1993. URL http://link.aip.org/link/jasman/v93/i2/p1007/s1.

R. D. Patterson. Effects of relative phase and the number of componentson residue pitch. *The Journal of the Acoustical Society of America*, pages 1565–1572, 1973. URL http://asadl.org/jasa/resource/1/jasman/v53/i6/p1565_s1.

R. D. Patterson. Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3):640–54, Mar. 1976. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/1254791.

R. D. Patterson. A pulse ribbon model of monaural phase perception. *The Journal of the Acoustical Society of America*, 82(5):1560–86, Nov. 1987. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/3693696.

R. D. Patterson. The sound of a sinusoid: spectral models. *The Journal of the Acoustical Society of America*, 1994. URL http://link.aip.org/link/?JASMAN/96/1409/1.

R. Plomp. The ear as a frequency analyzer. *The Journal of the Acoustical Society of America*, 32(1960):1628–1636, 1964. URL http://link.aip.org/link/?JASMAN/36/1628/1.

R. Plomp. Timbre as a multidimensional attribute of complex tones. *Frequency analysis and periodicity detection in hearing*, 397, 1970.

R. Plomp and H. J. Steeneken. Effect of phase on the timbre of complex tones. *The Journal of the Acoustical Society of America*, 46(2):409–21, Aug. 1969. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/5804112.

V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

V. Pulkki and J. Merimaa. Spatial Impulse Response Rendering II : Reproduction of Diffuse Sound and Listening Tests. *Journal of the Audio Engineering Society*, 54(1), 2006. URL http://www.aes.org/e-lib/browse.cfm?elib=13664.

T. Rossing, R. Moore, and P. Wheeler. The science of sound. 2002.

M. Ruggero, L. Robles, and N. C. Rich. Two-tone suppression in the basilar membrane of the cochlea: mechanical basis of auditory-nerve rate suppression. *Journal of neurophysiology*, 68(4):1087–99, Oct. 1992. ISSN 0022-3077. URL http://www.ncbi.nlm.nih.gov/pubmed/1432070.

M. B. Sachs and E. D. Young. Effects of nonlinearities on speech encoding in the auditory nerve. *The Journal of the Acoustical Society of America*, 68(3):858–75, Sept. 1980. ISSN 0001-4966. URL http://www.ncbi.nlm.nih.gov/pubmed/7419821.

B. Scharf. Complex sounds and critical bands. *Psychological bulletin*, 58(3):205–17, May 1961. ISSN 0033-2909. URL http://www.ncbi.nlm.nih.gov/pubmed/13747286.

J. Schouten, R. Ritsma, and B. Cardozo. Pitch of the residue. *The Journal of the Acoustical Society of America*, 294(1940):1418–1424, 1962. URL http://link.aip.org/link/?JASMAN/34/1418/1.

J. F. Schouten. The residue and the mechanism of hearing. In *Proc. K. Ned. Akad. Wet*, volume 43, pages 991–999, 1940.

G. Sergi. Knocking at the door of cinematic artifice: Dolby Atmos, challenges and opportunities. *The New Soundtrack*, 3(2):107–121, 2013.

D. R. Soderquist. Frequency analysis and the critical band. *Psychonomic Science*, 1970.

H. Spoendlin. Structural basis of peripheral frequency analysis. *Frequency analysis and periodicity detection in hearing*, pages 2–36, 1970.

Y. Suzuki and H. Takeshima. Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, 116(2):918, 2004. ISSN 00014966. doi: 10.1121/1.1763601. URL http://link.aip.org/link/JASMAN/v116/i2/p918/s1&Agg=doi.

G. Upton. Approximate Confidence Intervals for the Mean Direction of a von Mises Distribution. *Biometrika*, 1986. URL http://biomet.oxfordjournals.org/content/73/2/525.short.

J. Vilkamo, T. Lokki, and V. Pulkki. Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation. *Journal of the Audio Engineering Society*, pages 709–724, 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=14838.

H. L. F. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. F. Vieweg & Sohn, 1863.

G. Watson and E. Williams. On the Construction of Significance Tests on the Circle and the Sphere. *Biometrika*, 43(3):344–352, 1956. URL http://www.jstor.org/stable/10.2307/2332913.

E. Zwicker. Die grenzen der hörbarkeit der amplitudenmodulation und der frequenzmodulation eines tones. *Acustica*, 2(Beih. 3):125–133, 1952.

E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5):111523–111525, 1980. URL http://link.aip.org/link/?JASMAN/68/1523/1.

E. Zwicker, G. Flottorp, and S. Stevens. Critical Band Width in Loudness Summation. *The Journal of the Acoustical Society of America*, 29(1937):113–115, 1957. URL http://link.aip.org/link/?JASMAN/29/548/1.