AALTO UNIVERSITY SCHOOL OF ELECTRICAL ENGINEERING
Department of Signal Processing and Acoustics

Lari Paunonen

Measurements in Perceptual Annoyance of Audio Coding Artifacts

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, February 3, 2013

Supervisor: Prof. Paavo Alku

Instructor: Tom Bäckström, D.Sc. (Tech)

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author:	Lari Paunonen	
Name of the Thesis:	Measurements in Perceptual Annoyance of Audio Coding Artifacts	
Date:	February 3, 2013	**Number of pages:** 73 + 13
Department:	Department of Signal Processing and Acoustics	
Professorship:	Acoustics and Audio Signal Processing (S-89)	
Supervisor:	Prof. Paavo Alku	
Instructor:	Tom Bäckström, D.Sc. (Tech)	

This thesis discusses the perceptual annoyance of several audio coding artifacts that have become of interest during the development of USAC, a new low-bitrate speech and audio coder. A total of four different coding-related phenomena, all of which are explained below, were investigated in this study. All artifacts were artificially generated using MATLAB(R) and evaluated in listening tests with approximately ten participants in each. This work was commissioned by Fraunhofer IIS, Germany – a leader in audio coding technology and the home of MP3.

In audio coding, signals are usually processed in frames with a length of 20 to 50 milliseconds, which may cause rapid variations in artifacts. In our tests, the level of critical-bandwidth noise or single harmonics was altered with various speeds. The results suggest that moderate-speed variations are considered the most annoying.

Harmonic bandwidth extension is a method that generates artificial harmonics by stretching spectra in frequency. It is useful in audio compression because upper harmonics need not be encoded explicitly, but can be approximately reconstructed in the decoding phase. However, the generated harmonic patch will inevitably be incomplete, which may cause a false additional pitch sensation. The perceived strength of this ghost pitch was examined with synthetic tones as a function of fundamental and crossover frequencies.

The masking curve of a signal frame can be efficiently modelled with a spectral envelope. It can then be used for transferring the frame to the perceptual domain for quantization. The resulting quantization noise will be less audible if the smoothness of the envelope is properly adjusted in the first place by modifying the transfer function with a constant. A proposal for the optimal constant value is provided in this study.

Strong parts of a signal spectrum can be boosted and weak parts diminished by multiplying the spectrum with its modified envelope. This technique, known as formant enhancement, enables a better masking of quantization noise, but tends to render the overall tone unnatural. A model for selecting the optimal spectrum modification parameter values as a function of perceptual signal-to-noise ratio is proposed.

Keywords: Annoyance, audio coding, listening test, noise, psychoacoustic measurements, psychoacoustics, speech coding.

AALTO-YLIOPISTON
SÄHKÖTEKNIIKAN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Lari Paunonen
Työn nimi:	Audionkoodausartifaktien ärsyttävyyden mittauksia
Päivämäärä:	3.2.2013 **Sivuja:** 73 + 13
Laitos:	Signaalinkäsittelyn ja akustiikan laitos
Professuuri:	Akustiikka ja äänenkäsittely (S-89)
Työn valvoja:	Prof. Paavo Alku
Työn ohjaaja:	TkT Tom Bäckström

Tässä diplomityössä tutkitaan matalan bittinopeuden puhe- ja audiokooderin USACin kehityksessä merkittäväksi koettujen koodausartifaktien psykoakustista ärsyttävyyttä. Tutkielmassa käsitellään neljää ilmiötä, jotka on eritelty alempana. Artifaktit mallinnettiin MATLAB(R)-ohjelmistolla ja niiden ärsyttävyyttä arvioitiin kuuntelukokein. Työn toimeksiantaja on saksalainen Fraunhofer-instituutti, joka tunnetaan muun muassa MP3-koodekin kehittäjänä.

Audionkoodauksessa signaaleja käsitellään yleensä noin 20–50 millisekunnin pituisina kehyksinä, jolloin koodausartifaktit voivat vaihdella nopeastikin. Tämän ilmiön ärsyttävyyttä tutkittiin varioimalla kapeakaistaisen kohinan sekä yksittäisten harmonisten voimakkuutta eri nopeuksilla. Koetulosten perusteella keskinopea vaihtelu koetaan ärsyttävimmäksi.

Harmoninen kaistanleveyden laajennus (harmonic bandwidth extension) on menetelmä, jolla voidaan luoda harmonisia komponentteja rajataajuuden yläpuolelle alkuperäistä spektriä venyttämällä. Näin audiosignaalin bittinopeutta voidaan laskea, kun ylimpiä harmonisia ei tarvitse koodata eksplisiittisesti, vaan ne voidaan generoida dekoodauksessa. Koska luotujen harmonisisten joukko on kuitenkin aina puutteellinen, saattaa syntyä vaikutelma ylimääräisestä sävelkorkeudesta (ghost pitch). Kuuntelukokeessa tutkittiin synteettisillä äänillä, miten tämän ilmiön voimakkuus riippuu äänen perustaajuudesta ja valitusta rajataajuudesta.

Kuulon peittokäyrää voidaan approksimoida tehokkaasti spektrin verhokäyrällä, jota käyttäen itse signaalikehys voidaan siirtää perkeptuaaliseen alueeseen kvantisoitavaksi. Kvantisointikohinan peittymistä voidaan tehostaa säätämällä verhokäyrän pehmeyttä sen siirtofunktioon sijoitetulla vakiolla. Työssä esitetään ehdotus tämän parametrin arvoksi.

Sopivasti muokattua verhokäyrää voidaan käyttää myös spektrin voimakkaiden osien vahvistamiseen ja heikkojen osien vaimentamiseen. Puhesignaaleilla huomattiin, että tällä formanttien korostamisella voidaan peittää kvantisointikohinaa, mutta samalla sointiväri muuttuu epäluonnollisemmaksi. Tekstissä esitetään malli optimaalisten muokkausvakioiden valitsemiseksi perkeptuaalisen signaali-kohinasuhteen funktiona.

Avainsanat: Ärsyttävyys, audionkoodaus, kohina, kuuntelukoe, psykoakustiikka, psykoakustiset mittaukset, puheenkoodaus.

Acknowledgements

First and foremost, I am indebted to my instructor Dr. Tom Bäckström who taught me a million things I wish I knew before starting this project.

The thesis would have remained a dream without Fraunhofer IIS and the USAC team who provided me with the unique opportunity to work with and learn from the best.

Back at the university, my supervisor Prof. Paavo Alku and the always helpful staff definitely deserve to be credited.

As always, I owe my deepest gratitude to my family and friends.

Finally, special thanks to Heidi for waiting for me while I was in Erlangen. You are a big deal.

Espoo, February 3, 2013

Lari Paunonen

Contents

Symbols and Abbreviations

E	Signal Energy
F	Formant Enhanced Spectral Envelope
f_0	Fundamental Frequency
f_c	Center Frequency
f_s	Sampling Frequency
f_x	Crossover Frequency
f_Δ	Frequency Range
R^2	Goodness of Fit
SNR	Signal-to-Noise Ratio
SNR_P	Perceptual Signal-to-Noise Ratio
V	White Noise Spectrum
W	Spectral Envelope
x, y	Input/Output Signal in Time Domain
X, Y	Input/Output Signal in Frequency Domain
$\hat{X}, \hat{Y}$	Input/Output Signal in Perceptual Domain
α	SNR Normalization Coefficient
γ_0	Spectral Envelope Smoothing Constant
γ_z	Numerator Coefficient in Formant Enhancement
γ_p	Denominator Coefficient in Formant Enhancement
σ	Loudness Normalization Coefficient

BTL	Bradley-Terry-Luce
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
GUI	Graphical User Interface
HBE	Harmonic Bandwidth Extension
HF	High Frequency
IFFT	Inverse Fast Fourier Transform
IMDCT	Inverse Modified Discrete Cosine Transform
IQR	InterQuartile Range
ITU	International Telecommunication Union
LF	Low Frequency
LPC	Linear Predictive Coding
MDCT	Modified Discrete Cosine Tranform
MOA	Method Of Adjustment
MPEG	Moving Pictures Experts Group
MUSHRA	MUltiple Stimuli with Hidden Reference Anchor
RWR	Rating Without Reference
SNR	Signal-to-Noise Ratio
TDAC	Time-Domain Aliasing Cancellation
USAC	Unified Speech and Audio Coder

Chapter 1

Introduction

Audio coding refers to methods for reducing the bitrate of an audio signal. This is of interest in applications in which transmission or storage capacity is limited. Raw audio data requires so much space and bandwidth that without effective compression methods, many everyday applications – mobile phones, internet radios, or portable music players, to name a few – would not be feasible.

To reach high compression ratios of 1:10 or more, audio coders use lossy compression schemes, meaning that the reconstructed signal is not bit-identical to the original. The fundamental idea behind a common paradigm known as perceptual audio coding is that if the coding errors caused by bit reduction are not audible, they can be considered harmless. Following that principle, sophisticated psychoacoustic models are used to determine how signals could be compressed as efficiently as possible, yet achieving a nearly transparent perceptual quality. [1]

In this thesis, the perceptual annoyance of several important audio coding artifacts is investigated to help improve the new state of the art Unified Speech and Audio Coder (USAC) [2]. Traditionally, there have been separate coders optimized either for speech or generic content while USAC is an endeavor to combine the best speech and audio compression techniques into a single coder. Because of the low bitrates it provides, artifacts cannot be totally avoided and hence the focus is on minimizing their annoyance.

1.1 Milestones in Audio Coding

The first significant breakthroughs of digital audio date back to the 1910's–1930's when such innovations as pulse code modulation (PCM) and the Nyquist-Shannon sampling theorem were introduced [3]. However, digital audio did not find its way into widespread use until the T1 telephony in the early 1960's [4]. The era of commercial digital recording began in the 1970's and digital audio finally entered the consumer market with the introduction of the compact disc (CD) in 1982 [3]. Since then, analog devices have been disappearing, both in professional and consumer use.

Digital speech coding has attracted substantial interest since the advent of digital telephony [5]. The first inventions adopted into widespread use were waveform coders that enhanced the efficiency of PCM. These inventions include companding

techniques and differential PCM [6]. Around the 1970's, increasing computing power enabled the use of methods that are based on modelling the sound source and estimating its parameters from the signal [5]. Since the 1980's, much effort has been placed on developing hybrid coders that combine both parametric and waveform coding concepts, one notable example being code excited linear prediction (CELP) presented in 1985 [7]. As the amount of speech and data transferred in mobile networks is constantly increasing, research on speech coding remains active. Today, the 3rd Generation Partnership Project (3GPP) has a major role in organizing the development and standardization of speech codecs. The 3GPP standardized AMR-line codecs are often considered state of the art for compressing narrowband (AMR-NB, formerly simply known as AMR), wideband (AMR-WB), or superwideband (AMR-WB+) speech [8].

The huge capacity requirements of high-quality raw audio[1] led to a need for effective – and therefore inevitably lossy – compression also for music and generic content. In the late 1980's, the Moving Pictures Experts Group (MPEG) started an initiative to develop a standard that would include suitable coding methods for many application domains [10]. Three audio layers with different complexities were introduced and the most effective one, layer III, was a significant success in the consumer market. This layer, commonly known as MP3, allowed compression ratios of approximately 1:10 without significant loss in perceived quality [11]. Five years later, a new non-backwards compatible codec referred to as Advanced Audio Coding (AAC) was introduced as a part of the MPEG-2 standard [12]. Other technologies with a notable market share today include AC-3 by Dolby Laboratories (used in DVD players and digital television) [13], Windows Media Audio (WMA) by Microsoft [14], and the open source project Vorbis [15].

In 2007, MPEG initiated a process aiming towards the standardization of a new universal codec for both music and speech. It was required to be of superior quality to any existing speech or audio codec at low bitrates. A framework developed jointly by Fraunhofer IIS and VoiceAge Corporation was selected as the basis of the new MPEG Unified Speech and Audio Codec, abbreviated as USAC. [2]

1.2 Technical Overview of Unified Speech and Audio Coder

Figure 1.1 illustrates the basic framework of audio coding. First, the raw audio data is fed to an encoder that analyzes it in many ways and attempts to compress the signal as efficiently as possible. An essential part of this process is quantization in which the bitrate is decreased by reducing the precision of sample values in the frequency domain. In the receiver end, a decoder is responsible for transforming the compressed data back to the listenable form. The fundamental objective of all audio coders is to preserve the perceived signal quality as well as possible. [1]

[1]For example, one second of CD audio requires 1.41 million bits [9].

Figure 1.1: An overview of a typical audio coding system.

Figure 1.2 shows the basic functionalities of the USAC coder (explained thoroughly in [2]). It attempts to combine the best parts of the current state of the art music and speech coders, High-Efficiency Advanced Audio Coding (HE-AAC) and Extended Adaptive Multi-Rate – Wideband (AMR-WB+). USAC is based on the switched core principle, which means that it selects the coding scheme depending on the signal content. The signal classifier module is responsible for identifying whether the content is pure speech or something else. However, transitions between the vastly different coding schemes remain a challenge and might result in audible artifacts – an issue studied in Chapter 5.

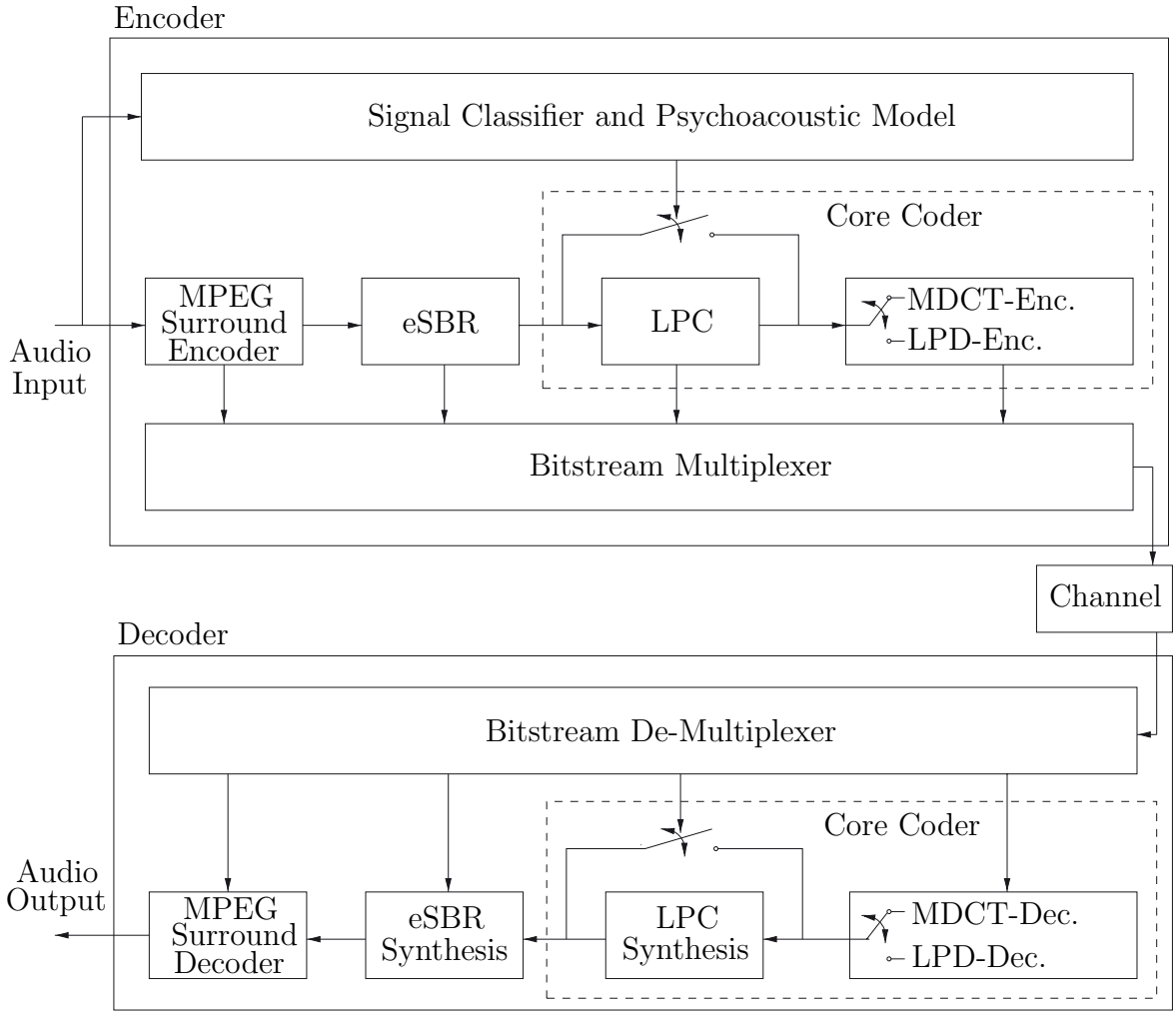

Figure 1.2: The building blocks of USAC. Figure used with permission from the authors. [2]

The available coding schemes share two common processing modules. The tailored MPEG Surround 2-1-2 mode is used for downmixing stereo signals to mono, as well as, for extracting parametric information about their spatial properties. Compared with coding both channels separately, this technique enables significant bit savings as spatial parameter data is of low bitrate.

Another mutual module, enhanced spectral band replication (eSBR), is an extended version of the standard spectral band replication (SBR) module used in AAC. Bitrate can be decreased significantly with the help of the eSBR module as spectra do not have to be stored as a whole: data above the crossover frequency is omitted in the encoding process and only the basic information, the spectral envelope for example, is saved. The traditional way of regenerating high frequencies in the decoder is to simply copy or mirror the lower part of the spectrum, but the eSBR module also supports a method known as harmonic bandwidth extension (HBE). It has an advantage of preserving the harmonic relations of tonal components, but as the high frequency patch has some harmonics missing, audible virtual pitches, often called ghost tones, might emerge. This problem is investigated in Chapter 6.

For non-speech material, the AAC-based frequency domain coding scheme is selected. Briefly stated, this coding scheme consists of the following stages: first, the time-domain signal is transformed to the frequency domain with the modified discrete cosine transform (MDCT); second, the resulting frequency bins are carefully quantized with the help of psychoacoustic models; and third, the resulting data as well as essential parametric side-information are encoded and saved. In order to spread the quantization noise in frequency so that it would be masked as efficiently as possible, quantization is habitually done in the so called perceptual domain. This domain is discussed in Section 7.1.

Speech signals are directed to the linear prediction domain (LPD) coding scheme based on AMR-WB+. First, a short-term linear prediction filter is used for extracting the spectral envelope. After that, the LPD coder makes a rapid selection between two modes: if the segment seems to fit the speech production model, algebraic code excitation linear prediction coding (ACELP) is used; otherwise, the weighted linear predictive coding (LPC) residual is coded in the frequency domain following the principles of the transform coded excitation (TCX) technology. To further improve the perceived coding quality, the ACELP module incorporates a technique known as formant enhancement that is discussed in Section 7.2.

1.3 Research in This Work

During the development of USAC at Fraunhofer IIS, the perceptual annoyance of several related coding artifacts has raised great interest. Two of the questions investigated in this thesis are closely related to audio coding while the rest are of more general interest as well. All of the phenomena were known and in many ways well understood already, but measurements were needed to deepen the knowledge and to enable the use of related techniques in practice. The research is based on

well-known psychoacoustic concepts and models, but thoughts and experiences of the development team also were essential sources of information.

Psychoacoustics is about personal experience and perception, which calls for organizing systematic listening tests. In each test, the number of participants was approximately ten, all of them being Fraunhofer employees having at least some earlier test experience. The phenomena were modelled and the test audio files created with Matlab. Finally, the results were analyzed with the help of statistical methods and are presented in this text in a graphical form along with models and conclusions derived of them. The research topics are briefly presented in the following paragraphs.

Quantization noise caused by bit reduction is an inherent feature of all coders. Signals are normally processed in frames of a length ranging from a few up to a hundred or more milliseconds. Consequently, quantization noise level may vary between frames because of core switching, for instance. Similarly, the level of a single harmonic may vary because of changes in quantization aggressiveness. The annoyance caused by this time-variance was examined, leading to results that can be used to optimize bit allocation as well as timing of core switching.

As explained in the previous section, approximating the upper part of a spectrum with the harmonic bandwidth extension is an efficient way of decreasing the bitrate, but it has a curious side-effect of producing new pitch sensations in some cases. This phenomenon was explored by creating synthetic tones with several fundamental and crossover frequencies and organizing a listening test to evaluate the strength of the possibly perceived ghost pitch. In reality, the crossover frequency is usually fixed, but together with an estimated fundamental frequency of the harmonic signal the strength of the ghost pitch can be estimated and missing harmonics manually reinserted to weaken the phenomenon.

A spectral envelope can be efficiently extracted from an input frame with linear predictive coding. The envelope acts as a simple estimate for the psychoacoustic masking curve and can therefore be used to transfer signal frames to and back from the perceptual domain. The masking curve approximation and hence the transformation can be easily improved by smoothening the LPC spectrum by plugging a single constant into its transfer function. A suggestion for the optimal value of that constant is provided in the text.

An appropriately modified spectral envelope can also be used to boost strong peaks and diminish weak valleys of the spectrum of a signal frame. To investigate how this formant enhancement technique can be efficiently used to hide quantization noise, we organized a listening test in which the said processing was applied to signals with different perceptual SNRs. A model for estimating the optimal processing parameter values as a function of perceptual signal-to-noise ratio is proposed.

1.4 Thesis Structure

First, the essential mathematical tools and techniques used in this thesis are presented in Chapter 2. After that, in Chapter 3, a brief summary of the psychoacoustic

models and concepts on which our research is based is provided. The background part is concluded in Chapter 4 with facts and experiences about organizing and analyzing listening tests.

The next three chapters are devoted to the empirical research in this work. Chapter 5 discusses time-variance of artifacts, the ghost tone phenomenon is investigated in Chapter 6, and methods utilizing spectral envelopes are studied in Chapter 7.

The thesis is concluded with a brief summary in Chapter 8. All individual answers in the listening tests are included in a graphical form in Appendix A and the written instructions given to the participants can be found in Appendix B.

Chapter 2

Mathematical Background

In this chapter, the most important mathematical tools and techniques used in this thesis for designing and analyzing listening tests are presented. The aim is to provide the reader with a compact overview of the essential concepts.

2.1 Signal Transforms

In audio coding applications, time domain signals are often transferred to the frequency domain in which frequency bins are quantized. The modified discrete cosine transform (MDCT) is usually favored because it allows 50% overlapping of consecutive frames without increasing the data rate. Overlapping, in turn, is desired as it helps to avoid audible artifacts at frame boundaries. [16]

2.1.1 Discrete Fourier Transform

The discrete Fourier transform (DFT) is among the widely used tools in audio signal processing. The DFT is used for transforming finite-length, discrete-time signals into finite-length, discrete-frequency signals (forward transform), and vice versa (inverse transform). The forward transform is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\mathrm{j}2\pi kn/N}, \quad k = 0, \ldots, N-1,$$

and the inverse transform as

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{\mathrm{j}2\pi kn/N}, \quad n = 0, \ldots, N-1.$$

The DFT of a time signal of the length N is thus an equally long sequence of evenly spaced frequency bins. [17]

It is widely recognized that the above stated equations are computationally heavy. Therefore, an efficient algorithm called the fast Fourier transform (FFT) is typically used for computing the DFT. [17]

2.1.2 Windowing

Signals are usually processed as separate frames with a length typically in the range of $20 - 50$ ms. A trade-off between the time and frequency resolutions is always present: the longer the frame, the better the frequency resolution, and vice versa. Frame length is usually chosen to be a power of two as that is required by many FFT implementations. There are, however, efficient algorithms designed for transforming frames of any length. [16, 18]

Frames are "cut" by multiplying the time signal with a window function that is defined to be zero outside the desired interval. The simplest function is the rectangular window

$$w_{\mathrm{r}}[n] = \begin{cases} 1, & \text{for } n = 0, \ldots, N-1 \\ 0, & \text{otherwise.} \end{cases}$$

Unfortunately, using the rectangular window results in pronounced discontinuities at the frame borders. That, in turn, may lead to severe aliasing problems because of the artificial high frequency components that are produced. By choosing a window that diminishes smoothly to zero at the edges, the risk of aliasing is greatly reduced, but the frequency resolution is deteriorated. This trade-off should always be taken into account when selecting a window function. [16]

The two windows used in this work are the sine window

$$w_{\mathrm{s}}[n] = \begin{cases} \sin(\frac{\pi n}{N-1}), & \text{for } n = 0, \ldots, N-1 \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

which is particularly suitable for the MDCT and the FFT, and the Hamming window

$$w_{\mathrm{h}}[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), & \text{for } n = 0, \ldots, N-1 \\ 0, & \text{otherwise,} \end{cases}$$

which is well-suited for linear predictive coding [16]. Figure 2.1 illustrates all three mentioned window functions.

2.1.3 Overlap-Add Technique

When reconstructing the original time domain signal from frequency domain frames, the initial windowing has to be cancelled. An obvious solution would be to divide the time domain output frame with the same window function that was used in the first place. This approach is, however, usually not acceptable because quantization errors near the edges are magnified greatly. A solution to this problem is the overlap-add technique in which consecutive frames are partly on top of each other. Usually, windowing is done both before the forward transform and after the inverse transform with the functions referred to as the analysis and synthesis windows, respectively.

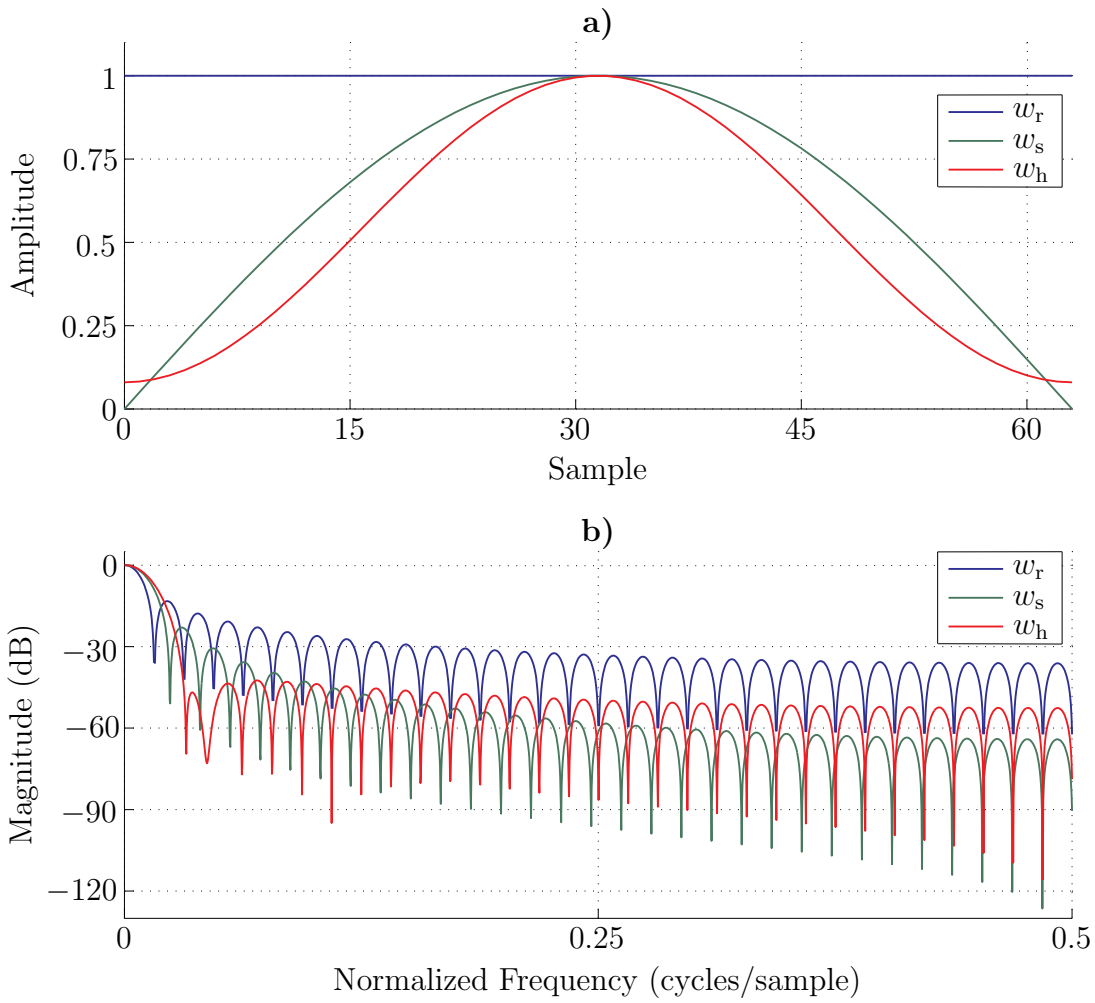

Figure 2.1: The rectangular, sine, and Hamming windows in the (a) time and (b) frequency domains.

The synthesis window helps to keep the quantization errors small near the edges of the inverse-transformed frames. [16]

To ensure that the output signal is identical to the input signal, we require that the sum of the overlapping parts of the windows in the frame i and the previous frame $i-1$ equals one. In the usual case with a 50% overlap, the necessary condition can be expressed as

$$w^i_{\text{an}}[n] \cdot w^i_{\text{syn}}[n] + w^{i-1}_{\text{an}}[N/2+n] \cdot w^{i-1}_{\text{syn}}[N/2+n] = 1, \quad n = 0, \ldots, N/2-1, \quad (2.2)$$

where $w_{\text{an}}[n]$ and $w_{\text{syn}}[n]$ are the analysis and synthesis windows, respectively. If they are identical and do not change over time, the condition can be simplified to

$$w^2[n] + w^2[N/2+n] = 1, \quad n = 0, \ldots, N/2-1,$$

where $w[n]$ is the window function. The sine window defined in Equation (2.1) is an example of a function satisfying the above conditions. [16]

2.1.4 Modified Discrete Cosine Transform

Unlike the DFT, the MDCT is not invertible as single frames, but relies on a phenomenon called time domain aliasing cancellation (TDAC). In other words, the inverse-transformed frames have to be added together with an ovelap of exactly 50% in order to achieve a perfect reconstruction of the original signal. Notably, the number of frequency domain samples in an MDCT frame is only a half of the number of time domain samples. Hence, the number of input samples in a larger dataset equals the total number of MDCT output samples and the process can be said to be critically sampled. To achieve perfect results, the analysis and synthesis windows must fulfill Condition (2.2). [16]

The forward MDCT transform is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] \cos\left[\frac{2\pi}{N}(n + \frac{N}{4} + \frac{1}{2})(k + \frac{1}{2})\right], \quad k = 0, \ldots, \frac{N}{2} - 1,$$

while the inverse transform (IMDCT) can be written as

$$x'[n] = \frac{4}{N} \sum_{k=0}^{N/2-1} X[k] \cos\left[\frac{2\pi}{N}(n + \frac{N}{4} + \frac{1}{2})(k + \frac{1}{2})\right], \quad n = 0, \ldots, N - 1.$$

A single time domain frame of N samples is thus converted to $N/2$ equally spaced frequency bins. [16]

The TDAC concept is best explained graphically. In Figure 2.2a, an artificial signal of 48 points is plotted in the time domain. The signal is then divided into two frames of 32 samples with an overlap of 16 samples. The subfigures b and d show the MDCT coefficients of the frames while the subfigures c and e are the inverse transformations of the corresponding frames. When the frames are added together, we arrive in a perfect reconstruction of the original signal as in the subfigure f.

2.2 Signal Energy

In many cases, such as when normalizing loudness, it is necessary to determine the energy of a signal frame. The energy of a discrete-time signal $x[n]$ during an interval $0 \leq n \leq N - 1$ can be computed as [17]

$$E = \sum_{n=0}^{N-1} |\, x[n] \,|^2.$$

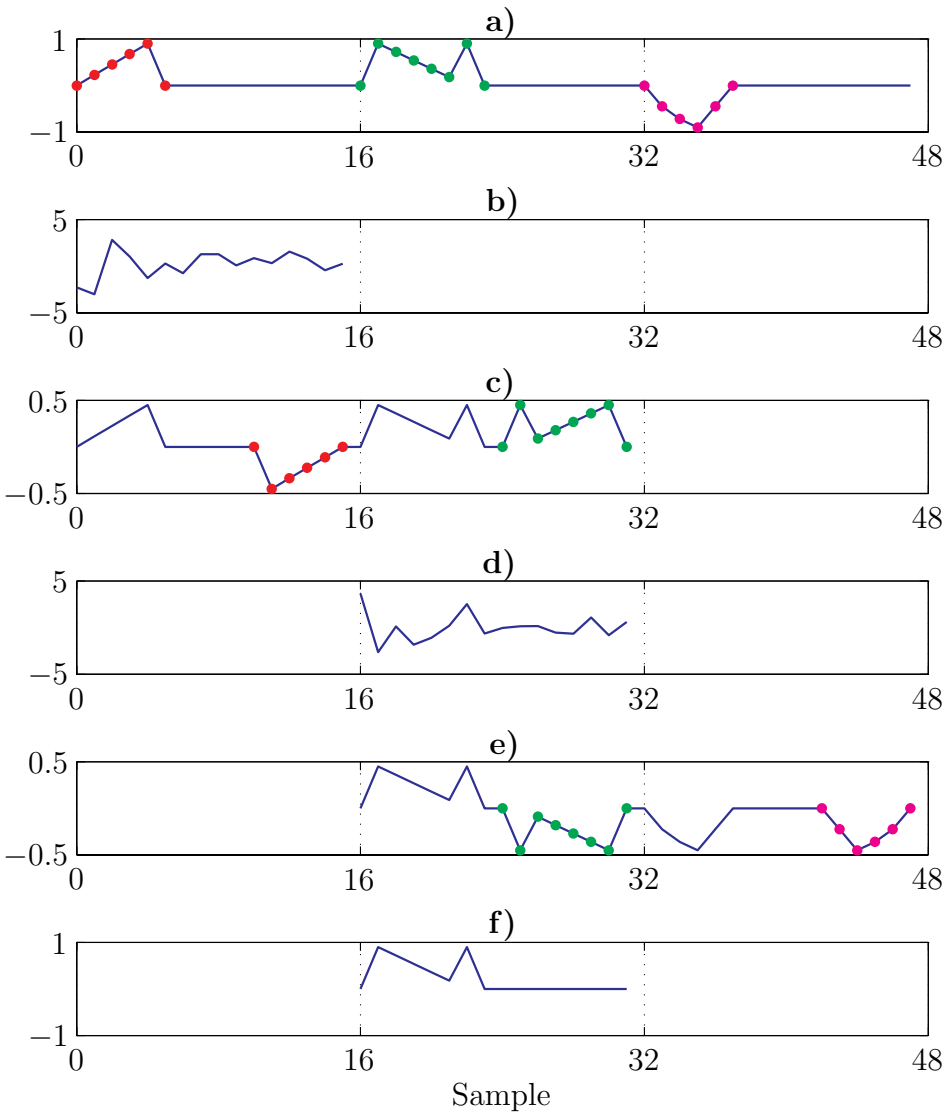

Figure 2.2: An illustration of the MDCT and TDAC: a) is a 48-point signal in the time domain, b) and d) show the MDCT coefficients of the frames $[0, 31]$ and $[16, 47]$, c) and e) are the corresponding inverse transformed frames, and f) is the resulting signal when c) and e) are added together.

According to Parseval's theorem, the above equation can also be expressed in the frequency domain as [17]

$$E = \frac{1}{N} \sum_{k=0}^{N-1} | X[k] |^2. \tag{2.3}$$

2.3 Linear Predictive Coding

Linear predictive coding (LPC) is a widely used method for storing and transmitting speech in a compressed form. In this work, however, LPC is only used for extracting spectral envelopes of signal frames.

As explained in [19], if the transfer function is relatively stationary during the selected period, we can predict the nth sample of a sequence $x[n]$ by forming a weighted sum of the p previous samples as

$$\dot{x}[n] = -\sum_{k=1}^{p} a_k \, x[n-k],$$

where a_k are real constants. The number of samples p defines the order of the resulting transfer function. We want to find the optimal values for a_k so that the error $e[n]$ is minimized:

$$e[n] = x[n] - \dot{x}[n] = x[n] + \sum_{k=1}^{p} a_k \, x[n-k].$$

The total energy of the error terms is

$$E = \sum_{n=0}^{N-1} e^2[n]$$

for a frame of length N. The minimum value of E can be found by setting the gradient to zero:

$$\frac{\partial E}{\partial a_k} = 0, \quad k = 1, \ldots, p$$

and solving for a_k. When $x[n]$ is windowed, it has to be zero outside the interval $0 \leq n \leq N-1$. Hence, we arrive in a form

$$\sum_{k=1}^{p} a_k \sum_{n=0}^{N-1-(i-k)} x[n] \, x[n+i-k] = -\sum_{n=0}^{N-1+p} x[n] \, x[n-i], \quad i = 1, \ldots, p. \tag{2.4}$$

By using the definition of autocorrelation

$$R(l) = \sum_{m} x[m] x[m-l]$$

and noticing the symmetry $R(l) = R(-l)$, Equation (2.4) can be written as

$$\sum_{k=1}^{p} R(|i-k|) \, a_k = -R(i).$$

This results in k linear equations with k unknowns and can be conveniently written in a matrix form as

$$\begin{pmatrix} R(0) & R(1) & R(2) & R(3) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & R(2) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & R(1) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & R(p-4) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = - \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix}.$$

This system of equations can be solved with any standard method, such as Gaussian elimination. As the autocorrelation matrix is of Toeplitz form, the computationally effective Levinson-Durbin recursion is often used.

Finally, the transfer function of the LPC envelope spectrum can be written as an all-pole filter

$$W(z) = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}}.$$

2.4 Statistical Methods

The statistical methods explained in the following were essential in analyzing the results of the listening tests. The methods can be used for extracting patterns and dependencies from data and for estimating the reliability of conclusions.

2.4.1 Median, Interquartile Range, and Boxplot

The median is one of the two most common measures for describing the center of a sample[1]. It is defined to be the observation separating the higher half of an ordered sample from the lower half. If the number of observations is even, the median is the arithmetic mean of two middle values. The first and third quartiles are defined similarly: the first quartile cuts off the lowest 25% of a sample while the third quartile cuts off the lowest 75%. The second quartile is the same as the median. The interquartile range (IQR), on the other hand, is a measure for describing the spread of a sample. It is defined as the range of the middle 50% of the sample:

$$I = Q_3 - Q_1,$$

where Q_1 and Q_3 denote the first and third quartiles. [20]

In this text, the above measures are presented with convenient boxplot figures (see Figure 7.3 for an example). The upper and lower limits of a box indicate the first and third quartiles and therefore the length of the box equals the IQR. The middle line inside the box is the median, while whiskers visualize the range of the

[1]In the statistical terminology, a sample refers to a set of observations which in this work are listening test answers and should not be confused with samples in audio signals.

whole sample, excluding the extreme values. The observations being $1.5I$ or more but less than $3I$ away from the box are called outliers and are denoted by a black dot. Similarly, those observations being $3I$ or more away from the box are called extreme outliers and are denoted by a red circle. [20]

2.4.2 Mean and Confidence Interval

The arithmetic mean is another widely used measure for the center of a sample. It is defined as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

where x_i are the observations and N is the sample size. The sample mean is also the best unbiased estimator for the mean of the underlying population from which the observations are randomly picked. [21].

The 95% confidence interval is usually included with the mean. The statistical intepretation is that the true mean of the population lies inside the interval with the probability of 0.95. If the random variable is approximately normally distributed, the Student's t distribution can be used for computing the confidence interval:

$$\left[\bar{X} - t_{(1-\frac{\alpha}{2};N-1)}\frac{s_{\mathrm{x}}}{\sqrt{N}};\ \bar{X} + t_{(1-\frac{\alpha}{2};N-1)}\frac{s_{\mathrm{x}}}{\sqrt{N}}\right],$$

where s_{x} is the standard deviation of the sample, $\alpha = 0.05$ is the level of significance corresponding to the confidence level of 95%, and $t_{(1-\frac{\alpha}{2};N-1)}$ is the $1-\frac{\alpha}{2}$ quantile of the one-tail t distribution with $N-1$ degrees of freedom. [22]

In this thesis, means and confidence intervals are presented graphically as in Figure 5.7. Wide horizontal lines represent means while whiskers denote confidence intervals.

2.4.3 Linear Regression

Linear regression can be used for modeling the relationship between a dependent (response) variable y and one or more independent (explanatory) variables x. The relationship is assumed to be linear, hence the name.

As explained in [21], a linear model with i independent variables and n equations is of form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1i} \\ x_{21} & \cdots & x_{2i} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{ni} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or simpler as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{\epsilon}$ is a vector of random error terms. The unknown coefficients $\boldsymbol{\beta}$ are commonly estimated with the method of ordinary least squares which minimizes the sum of squared residuals:

$$\min \parallel \boldsymbol{e} \parallel^2 = \min \parallel \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \parallel^2,$$

where $\hat{\boldsymbol{\beta}}$ are the estimated coefficients. The best estimator for them can be shown to be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y}.$$

The explanatory power of the model can be expressed with the goodness of fit measure that is computed as

$$R^2 = 1 - \frac{\parallel \boldsymbol{\epsilon} \parallel^2}{\parallel \boldsymbol{y} \parallel^2}.$$

The R-squared always lies in the range $0 \leq R^2 \leq 1$ where higher values indicate that the model explains the variations in the dependent variable better. In addition, it is customary to test for the statistical significance of the acquired coefficients in order to estimate whether the linear relationships really exist.

2.4.4 Bradley-Terry-Luce Pairwise Comparison

Bradley-Terry-Luce (BTL) is an analysis method used for scaling preferences from research data. In principle, it should be applied to data collected with pairwise comparisons (i.e. when the listeners are given two choices and are asked to select one of them). However, the method is also feasible when more than two conditions are rated at once, assuming that they can be freely compared against each other. BTL attempts to derive a ratio scale expressing the relative preferences between them. [23]

First, for a dataset consisting of N conditions, an $N \times N$ probability matrix $\boldsymbol{M}$ is constructed. The value in each cell (i, j) represents how many times the listeners have altogether preferred – that is, given a better rating – the stimulus i over the stimulus j. If both stimuli are given the same rating, the number of occurrences is increased by 0.5 in both cells (i, j) and (j, i). [24]

To ensure that the test data is suitable for the BTL method, three transitivity prerequisites are tried. The three stochastic transitivies (weak, moderate, and strong stochastic transitivities abbreviated as WST, MST, and SST, respectively) imply that if the probability of choosing x over y is $P_{xy} \geq 0.5$ and the probability of choosing y over z is $P_{yz} \geq 0.5$, then

$$P_{xz} \geq \begin{cases} 0.5 & \text{(WST)} \\ \min\{P_{xy}, P_{yz}\} & \text{(MST)} \\ \max\{P_{xy}, P_{yz}\} & \text{(SST)} \end{cases} \tag{2.5}$$

should hold for all conditions. It is unlikely that all these prerequisites hold for the whole dataset, but the BTL model can still be used if the number of violations is moderate. [25]

Finally, BTL attempts to link the probability P_{xy} and a ratio scale $u(\cdot)$ with maximum likelihood estimation so that [25]

$$P_{xy} = \frac{u(x)}{u(x) + u(y)}.$$

In this thesis, a Matlab function provided by Wickelmaier and Schmid was used for extracting ratio scales as well as for testing the validity of models [24].

Chapter 3

Psychoacoustic Concepts

Psychoacoustics is a well-established scientific branch studying the human perception of sound. Perceptual audio coding fundamentally relies on psychoacoustic models and theories as they are needed to estimate the audibility and annoyance of coding artfacts. Likewise, designing listening tests and interpreting the results correctly requires a vast amount of psychoacoustic knowledge. The research in this thesis rests on the concepts introduced in the following.

3.1 Critical Bands

Critical bands were first introduced by Harvey Fletcher in 1940 [26]. The concept is derived from the physiological phenomenon of tones close to each other in frequency creating overlapping physical responses on the basilar membrane in the inner ear. Critical bands have a few important implications for audio coding. First, if the bandwidth of noise with a constant intensity varies but does not exceed the critical bandwidth, the perceived loudness is also constant. The loudness starts to increase only after the noise covers more than one critical band, as described in Section 3.2. Second, if two tones fall into the same critical band, it is often difficult to distinguish them because the stronger tone tends to mask the softer one, as explained in Section 3.3. [27]

Critical bandwidth depends on the center frequency of the band. Mathematically, the bandwidth in hertz can be estimated with the popular equation proposed by Zwicker:

$$f_\Delta = 25 + 75\,(1 + 1.4\,f_c^2)^{0.69}, \tag{3.1}$$

where f_c is the center frequency in kilohertz [27]. In Figure 3.1, critical bandwidth is plotted as a function of center frequency.

3.2 Loudness

Loudness refers to the perceived "strength" of a sound. It is a subjective measure that is primarily dependent of intensity, but the frequency, bandwidth, and duration have a slight effect as well [27]. Loudness is an important factor to be controlled

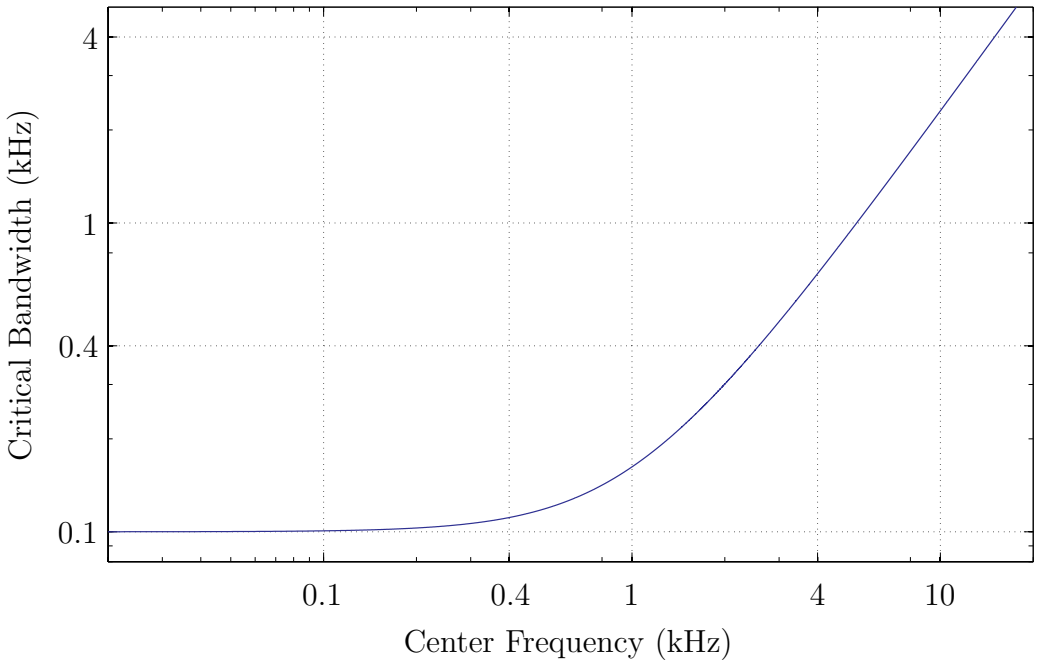

Figure 3.1: Critical bandwidth as a function of center frequency according to Equation (3.1).

in listening tests because louder sounds are regularly judged to be of better quality than softer sounds, which might distort the results [28].

Loudness is measured with the phon scale on which the sound pressure level of 1 dB at the frequency of 1 kHz equals 1 phon [27]. Equal-loudness contours in Figure 3.2 illustrate how the sound pressure level of a pure tone has to be varied with frequency to make the sound equally loud in the whole audible range.

As mentioned in the previous section, critical bands have an important role in loudness sensation. Figure 3.3 shows the effect of bandwidth when the sound pressure level of spectrally flat noise centered at 1 kHz is kept constant. Clearly, loudness starts to increase only after the bandwidth has exceeded one critical band.

Numerous models for estimating the loudness of a complex wideband signal have been proposed in the literature (see e.g. [30]). They are, among other things, useful for normalizing the loudness of conditions in listening tests. In this research, however, there were only minor differences in the conditions to be compared and hence no loudness normalization was seen necessary.

3.3 Masking

Auditory masking is a pivotal concept in perceptual audio coding. It refers to the everyday phenomenon in which a sound (maskee) is rendered partially or completely inaudible because of a presence of a louder sound (masker). The concept explains, for example, why it is difficult to follow a conversation next to a road with heavy traffic. [31]

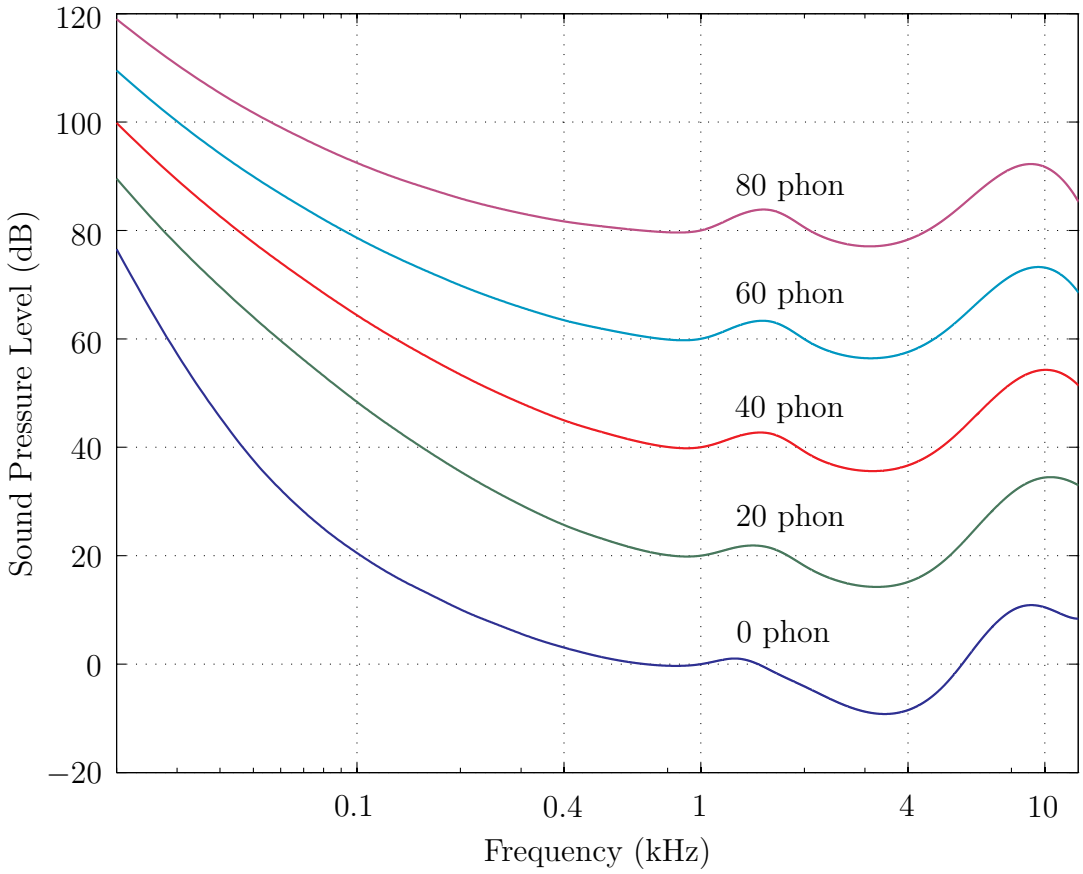

Figure 3.2: Equal-loudness curves for pure tones (adapted from [29]).

In this work, we concentrate mostly on the effects of simultaneous masking, meaning that the masker and the maskee are present at once. Besides that, perception of a maskee might be affected up to 200 ms after the masker is switched off (forward masking) or even 20 ms before the masker is turned on (backward masking). The latter phenomena are known as temporal masking. [27]

In audio coding, it is of interest how spectral peaks are able to mask nearby frequency components because bitrate reduction is achieved by tolerating artifacts that will be masked by the original signal. The level below which tones are masked is called masking threshold [27]. As can be seen from Figure 3.4, the effect is the strongest near the masker, especially below it. The LPC based weighting filter, as applied in AMR-WB, is an example of a fairly accurate and computationally effective method for estimating the masking threshold of an arbitrary spectrum [32].

Masking does not necessarily lead to a complete vanishing of the maskee but it might be only partly hidden instead. That particular phenomenon is accordingly referred to as partial masking [27]. Figure 3.5 illustrates partial masking by showing how the loudness of a tone is affected by critical bandwidth noise present at the same time.

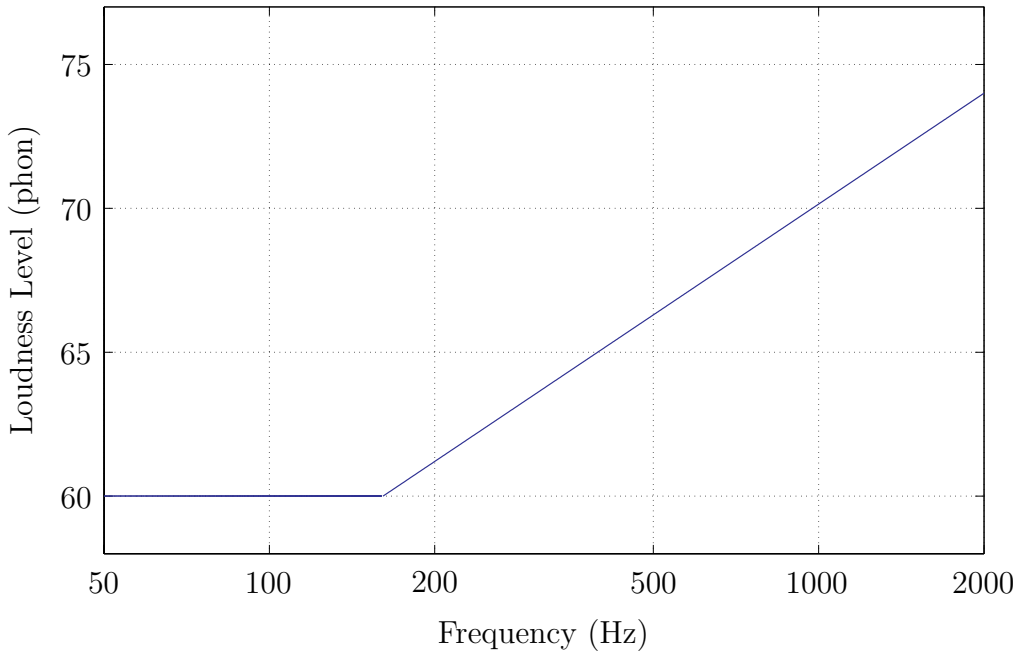

Figure 3.3: Loudness level of noise with the center frequency of 1 kHz and the sound pressure level of 60 dB as a function of the bandwidth of the noise (adapted from [31]).

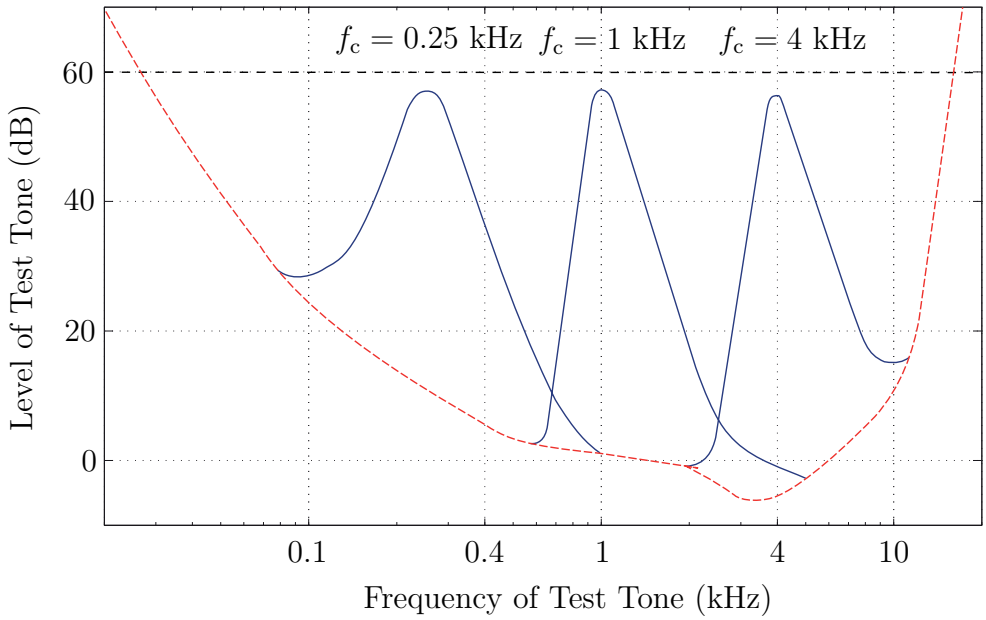

Figure 3.4: Masking curves for critical-band-wide noise maskers with the loudness level of 60 dB centered at 0.25, 1, and 4 kHz (adapted from [27]).

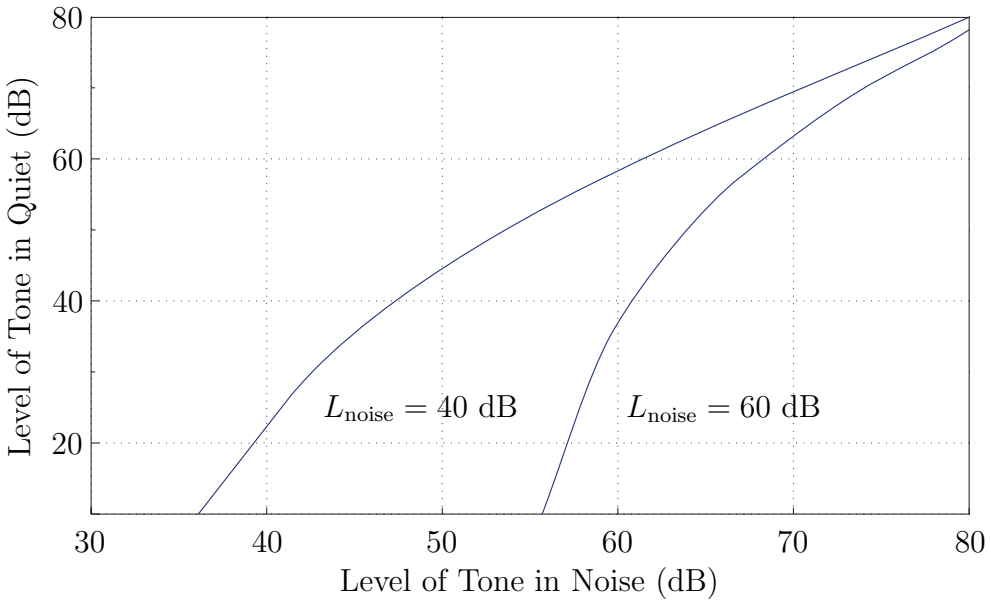

Figure 3.5: A 1 kHz pure tone adjusted to be equally loud with and without the presence of a critical-band-wide noise masker (adapted from [33]).

3.4 Perceptual Domain

If a signal frame is simply transferred to the frequency domain and quantized there uniformly (i.e. all frequency bins are rounded or truncated equally), the resulting quantization noise will be flat. It would, however, be beneficial if the noise was shaped according to the masking function of the frame so that the energy would be directed to the bins with a higher masking capacity. The perceptual domain transformation provides a solution to this question by treating the spectral envelope of the frame as a model for the masking curve and using it for normalizing the spectral coefficients [34]. In Figure 3.6, a spectrum $X(z)$ of a 32 ms speech frame along with its spectral envelope $W(z)$ attained with LPC are shown. The spectrum $X(z)$ is transformed into its perceptual domain counterpart $\hat{X}(z)$ as

$$\hat{X}(z) = X(z)W^{-1}(z).$$

Accordingly, the transformation back to the frequency domain is achieved simply by

$$X(z) = \hat{X}(z)W(z)$$

From the above equations it can be seen that if white (quantization) noise is added in the perceptual domain, it will be favorably shaped once the signal is transferred back to the original domain. This is because quantization noise is heard easier and is thus more harmful when added on soft regions of a signal spectrum than on loud regions.

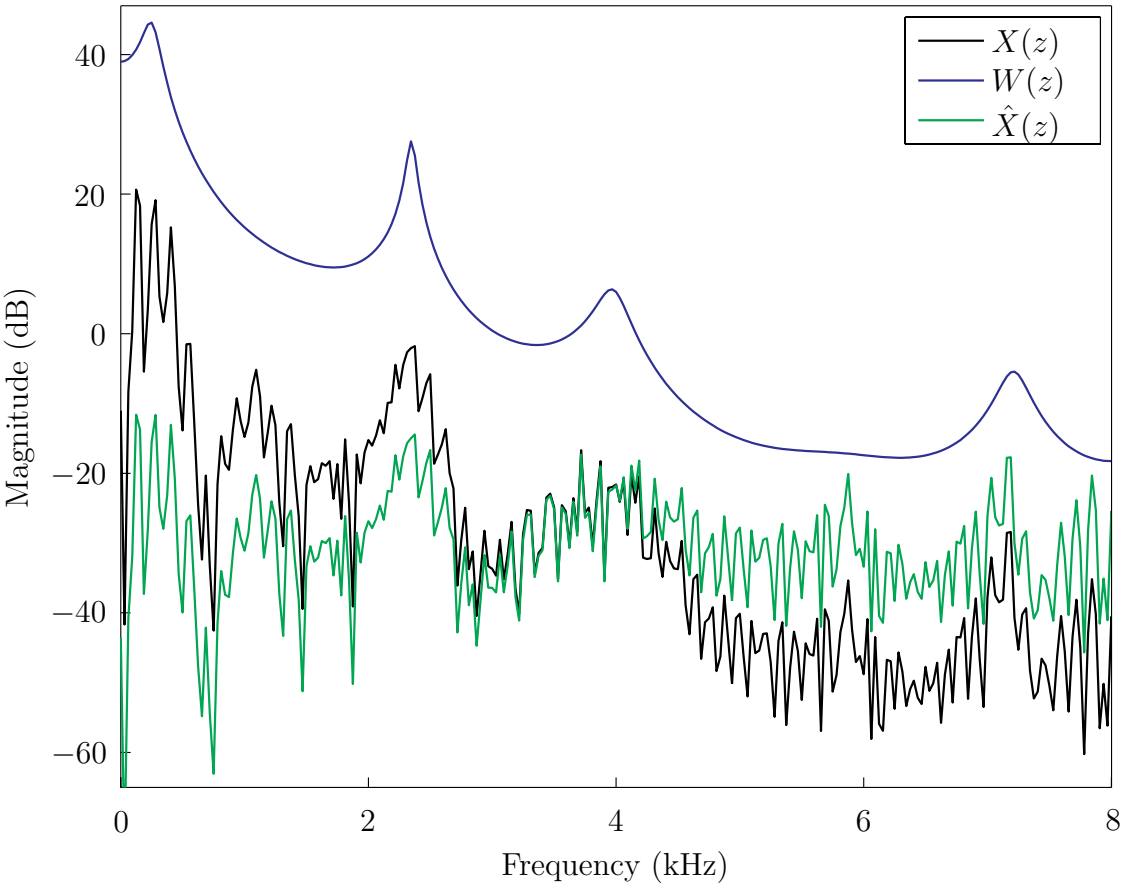

Figure 3.6: A 32 ms frame $X(z)$ of a speech signal transformed into the perceptual domain frame $\hat{X}(z)$ by using the spectral envelope $W(z)$.

3.5 Perceptual Signal-to-Noise Ratio

Signal-to-noise ratio (SNR) is a widely used measure for quantifying the relative amount of noise mixed in a signal frame [35]. It is usually defined in a logarithmic scale as an energy ratio

$$SNR = 10 \log_{10} \frac{E_\mathrm{s}}{E_\mathrm{n}},$$

where E_s and E_n are the signal and noise energies, respectively. According to Equation (2.3), SNR can also be computed in the frequency domain as

$$SNR = 20 \log_{10} \frac{\| X[k] \|}{\| N[k] \|}, \tag{3.2}$$

where $N[k]$ is the noise.

In audio coding, it is generally desired that the output and input signals would be as close to (perceptually) identical as possible. Therefore, it is sensible to insert

their difference $Y[k] - X[k]$ in the previous equation for $N[k]$:

$$SNR_{\mathrm{P}} = 20 \log_{10} \frac{\| Y[k] \|}{\| Y[k] - X[k] \|}.$$

If both the nominator and the denominator are transferred to the perceptual domain, we arrive in the equation

$$SNR_{\mathrm{P}} = 20 \log_{10} \frac{||Y[k]W^{-1}[k]||}{|| (Y[k] - X[k]) W^{-1}[k]||}, \tag{3.3}$$

which defines the perceptual SNR of a coding process.

Chapter 4

Listening Test Methods and Procedures

The listening test methods and procedures used in this thesis as well as the ways of analyzing the results are explained in this chapter. In addition, some test designing and organizing experiences of the author are shared.

4.1 Methods

As different questions call for different test methods, the research problem should first be formulated carefully. Some typical cases are assessing the amount of signal degradation, finding the optimal level for a parameter, or examining whether an artifact is audible. The test should always be simple enough so that the results would be reliable considering the often limited number of listeners. The recommended procedure is to start with piloting the test with a few listeners and once it seems to be able to provide valid answers to the research question, the whole test can be executed. Naturally, it is neither advisable nor ethical to deliberately tune the experiment so that it would support the researcher's hypothesis.

4.1.1 Multiple Stimuli with Hidden Reference Anchor

The popular multiple stimuli with hidden reference anchor (MUSHRA) listening test method was developed especially for evaluating intermediate audio quality, which makes it particularly suitable for testing low-bitrate coders. All investigated coders are present at the same time in each item, thus they are compared not just with the reference, but with each other too. Besides guaranteeing more reliable results, this setting makes the execution faster. [36]

In a MUSHRA test, each item includes the original signal as a reference and a set of modified signals called conditions. Along with the coded excerpts, a few special conditions known as anchors are included. The hidden reference anchor is just a duplicate of the reference signal, but there are also one or more low-quality anchors, at least one of them being of clearly inferior quality to the other conditions. The ITU-R Recommendation requires that at least the hidden reference and the original

signal lowpass filtered at 3.5 kHz always have to be present. The purpose of the anchors is to create reference points for ratings and to make the results of different tests comparable. The listener is implicitly expected to give at least one condition a full rating of 100 because the hidden reference is invariably present while the other anchors may be rated freely. The rating scale is specified as shown in Table 4.1. Because of the high resolution of the scale, the results may be treated in the analysis phase as if they were on a continuous scale. [36]

Table 4.1: The MUSHRA grading scale.

80–100	Excellent
60–80	Good
40–60	Fair
20–40	Poor
0–20	Bad

Wavswitch, a program developed earlier at Fraunhofer IIS, was used for executing the MUSHRA tests (see Figure 4.1 for a screenshot). Wavswitch supports switching between the reference signal and the conditions in an item seamlessly while keeping them time-aligned. This is not mandatory according to the ITU-R Recommendation, but it is sometimes seen advisable [22]. The listeners were also allowed to listen to and set loops on audio freely.

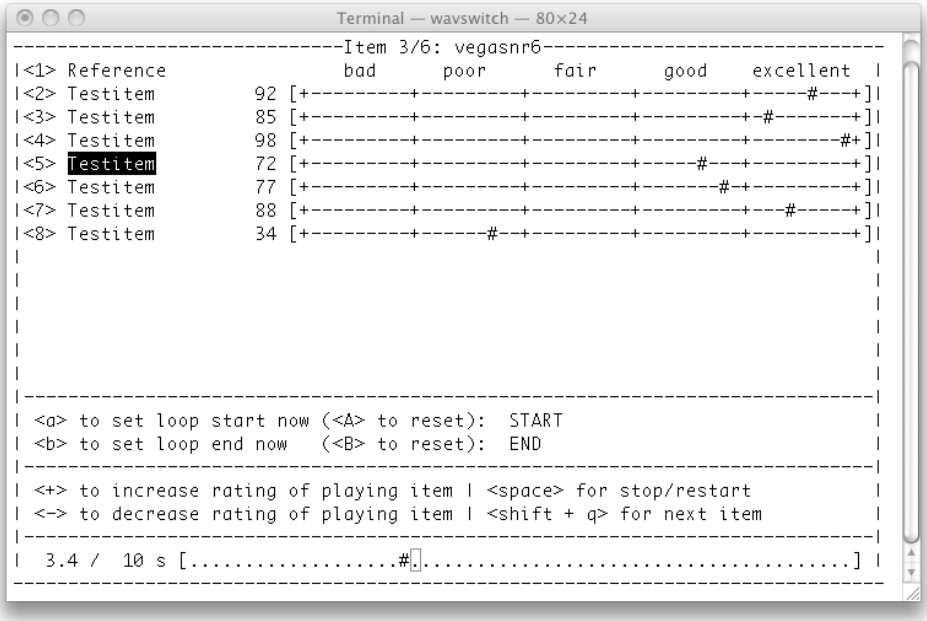

Figure 4.1: A screenshot of Wavswitch that was used in the MUSHRA tests.

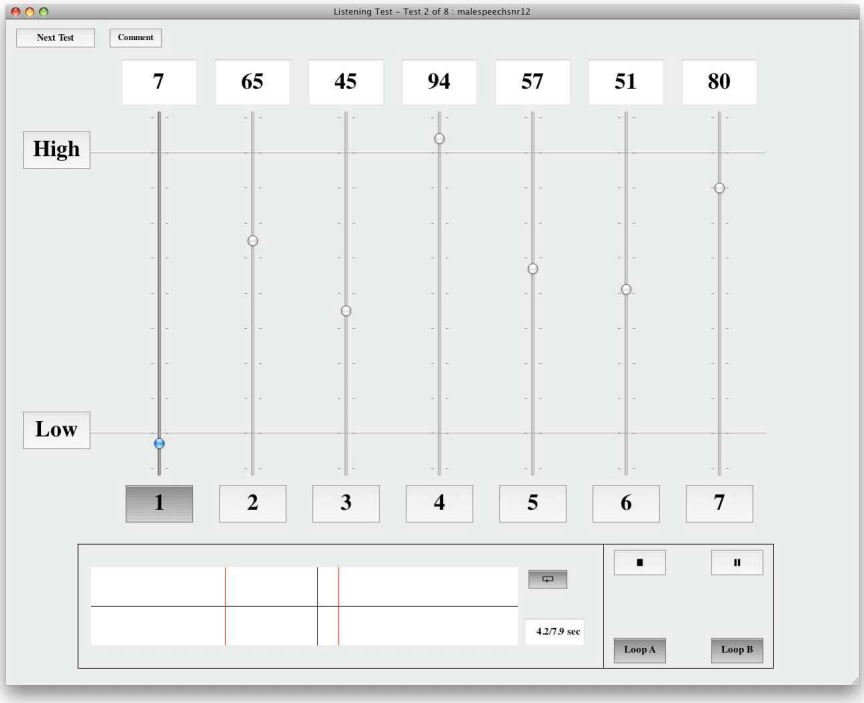

Figure 4.2: A screenshot of ListeningTestGUI used in the modified MUSHRA and rating without reference tests.

4.1.2 Modified MUSHRA

For a few test, we used a slightly modified version of MUSHRA as the test procedure. It is a non-standard modification that is used internally at Fraunhofer. In standard MUSHRA tests, the reference signal corresponds to the maximum value of 100 of the rating scale and the listener is asked to evaluate the magnitude of artifacts in conditions. However, since in some of our measurements the goal was to investigate subjective quality or pleasantness instead, it was desirable to allow the use of higher ratings than what the reference is said to represent because conditions might have been perceived sonically superior to the reference.

In our modified MUSHRA setting, the rating scale is nailed with two reference signals at fixed values. The original signal (called a high reference) corresponds to the reference signal in MUSHRA and is said to represent the rating of 90. The low reference is a low-quality version of the original signal and is fixed at the rating of 10. To retain some compatibility with the original MUSHRA, the condition set of each item includes the hidden reference (advised to be given the rating of 90) and the 3.5 kHz lowpass anchor (rated freely).

The Modified MUSHRA tests were executed with the ListeningGUI software developed earlier at Fraunhofer IIS. A screenshot of the GUI is shown in Figure 4.2. The high and low references could be played by clicking the corresponding buttons on the left while the conditions were played with the numbered buttons and rated with the sliders above them.

4.1.3 Method of Adjustment

In method of adjustment (MOA) tests, listeners are asked to set the level of the parameter in question, such as volume, according to some guidelines. MOA is convenient if the researcher wants to know, for example, when an artifact is barely detectable or – as in this thesis – when it is as annoying as a reference artifact. Some or all items might be evaluated multiple times and the answers finally averaged to get more reliable results. In addition, the reliability of single listeners may be estimated by inspecting the coherence of her answers related to the same item. Despite the repetition, MOA tests are relatively fast to execute. [37]

A simple Matlab program was developed for conducting the MOA tests in this thesis. A screenshot of the graphical user interface (GUI) is shown in Figure 4.3. The listener could play the reference signal (containing the investigated artifact at a fixed level) and the condition signal (the level of the investigated artifact adjusted with the slider) freely.

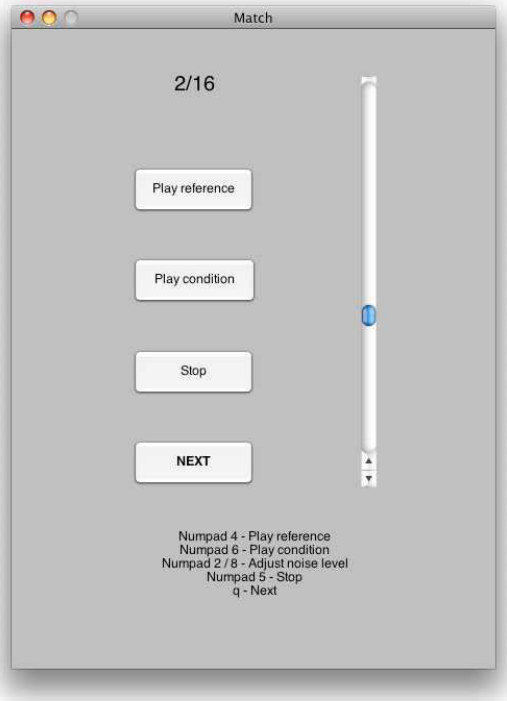

Figure 4.3: A screenshot of the custom Matlab GUI used in the MOA tests.

4.1.4 Rating Without Reference

Rating without reference (RWR) is a non-standard listening test method that we developed for a special case when conditions could be rated without a separate reference signal. This is possible e.g. when the listener is asked to rate the proportion of one artifact to another, and they both exist at the same time, in the same au-

dio signal. Self-explanatory anchors, preferably with audible examples, should be provided to help with using the rating scale.

The main advantage of RWR is that it is almost as fast to execute as the on-off procedure in which the listener just answers whether some feature seems to exist or not. Thus, it is possible to evaluate up to 100 items without the test being overly time-consuming. However, the listeners may find rating difficult, which may lead to unreliable results with a high variance.

The RWR method was realized with ListeningTestGUI presented above. In contrast to the screenshot in Figure 4.2, only one slider existed at a time and there was no reference signal. The listeners were again permitted to listen to and set loops on items freely. The rating scale was defined from 0 to 100.

4.1.5 Test Equipment

All tests were conducted with Apple Mac Mini computers running either OSX 10.4 or 10.5. The computers were equipped with an Edirol audio interface (either UA-1000 or UA-25) and high quality Stax headphones with a dedicated amplifier. Background noise in the listening rooms was low enough to be considered negligible. Listeners were free to adjust the volume in every test.

4.2 Procedures for Analyzing Test Results

In a nutshell, there are three major steps in handling the acquired listening test data. First, the clear outliers should be excluded in the procedure often referred to as post-screening [22]. The second step is to determine and use the analysis methods that fit the situation best and are the most likely to reveal useful information. Finally, the results should be presented as informatively as possible.

4.2.1 Post-Screening

To ensure a high quality of the final results, unreliable listeners are rejected from further analyses. There are many possible causes for the unreliability: perhaps the listener did not understand the test procedure correctly, she did not really hear the artifacts and was just guessing, or she could not provide consistent answers for some other reason. However, listeners should be rejected very cautiously and with larger sample sizes it might not be necessary at all [36]. Depending on the test method, the post-screening phase might consist of some or all of the following phases [22, 23]:

> Verify that the listener understood the test procedure correctly. For example, at least one condition must always be given a rating of 100 in MUSHRA tests.
>
> Check that the listener detected the possible hidden reference correctly. Minor and not repeated errors may be regarded as acceptable.

Check that the ratings of other anchors make sense. For example, the anchor of the lowest quality should not be given a high rating compared with other conditions.

Check the consistency of the answers if there were repetitions.

Reject the listener if a too small part of the rating scale was used.

If the listener gave very different answers from the others, figure out whether she is too inexperienced, found artifacts not detected by the other participants, or if she just has a different preference.

4.2.2 Analysis Methods

It is convenient to begin with briefly illustrating the distribution of the answers with a boxplot. The average ratings can, in turn, be illustrated with a plot showing the means and 95% confidence intervals. From such figure, it can be concluded that if the confidence intervals of two conditions are not overlapping, there probably was some real consensus about their relative quality.

Despite the anchors, listeners tend to use the rating scale differently. Especially if the conditions in items are very similar to each other, it might be beneficial to extract the differences between the ratings. The individual bias is then removed and the statistical differences might be more pronounced, but without a cost of reducing the degrees of freedom [22]. This technique is, however, only feasible if the conditions were present at once, as in MUSHRA tests.

If the above assumption holds, the ratings can also be treated as if they were acquired with a paired comparison test and be analyzed with BTL. Each possible pair of conditions in an item is examined and the number of times one condition was preferred over any another is counted. If two conditions were given equal ratings, the interpretation is that the listener was not able to hear a difference or could not decide which one was superior. In that case, the preference count for both conditions is increased by 0.5 because the probability of selecting either of them would have been 0.5 in case the listener was forced to make a choice. [24]

In addition, linear regression can be used for extracting the linear dependency between variables. Even though causality is not definitely implied, a regression model is often useful for understanding how the investigated parameters affect the perceived outcome. The model can also be used for selecting the optimal values for the parameters in an encoding process.

4.3 Guidelines for Organizing Listening Tests

In this section, some of our informal experiences about designing and conducting listening tests are shared. To begin with, the overall execution time of a test should not exceed approximately 20 minutes in order to avoid listening fatigue and losing concentration. Most tests in our research were considered particularly tiring as they

either consisted of the same sound played repeatedly with only slight variations in it or because the artifacts were hard to hear.

The question of whether naive or experienced listeners should be invited to take the test has to be carefully assessed. If the test is subjective and straightly related to consumer applications, a natural choice would be to invite naive listeners as they supposedly represent the everyman perspective. However, it turned out that most participants not used to analytic listening were relatively inconsistent with their answers and struggled with hearing even relatively clear differences in conditions. In other words, obtaining statistically significant results would require an impractical number of participants. Therefore, it is often advisable to work with experienced listeners because it leads to more meaningful results with fewer participants. The BS.1534 Recommendation also states that experts should be favored for that very reason [36].

It is particularly important to confirm that the listeners have understood the test method and the instructions correctly. It is advisable to provide all the necessary information also in a written form and to make the test interface as simple and intuitive as possible. In addition, the pitfall of making the artifacts too subtle should be avoided: while the test organizer himself is usually thoroughly familiar with the artifacts and hence recognizes them easily, it should not be forgotten that the listeners probably hear them for the first time. The participants may be assisted in this regard by providing them with some clear examples before the actual test, but sometimes it is justifiable not to reveal the artifacts beforehand.

As always, all variables other than those to be examined should be kept constant as precisely as possible. Loudness, in particular, is a factor that deserves special attention because even simple modifications, such as adding noise to the original signal, have an effect on it. It is widely known that even experienced listeners have a tendency to favor louder sounds in terms of quality [28].

One factor that is often overlooked is collecting qualitative insights from the listeners. Especially in such tests that compare unusual attributes or are otherwise atypical, it is beneficial to receive verbal feedback on what sonic properties the listener paid attention to. Qualitative opinions can be easily collected by having a short conversation with the listener right after the test and it is often useful to also include a chance to write short comments during the test.

Chapter 5

Time-Varying Artifacts

The perceived effect of amplitude modulation has been of some interest in psychoacoustic research. A measure called fluctuation strength is often used to describe slow or medium speed amplitude variation (up to approximately 20 Hz) while faster variation is measured with roughness [27, 38]. The sensitivity of human ear to amplitude modulation is found to follow a reverse-U shaped curve and peak at the variation frequency of approximately 4 Hz (Figure 5.1) [27, 39]. The fluctuation strength is zero until a modulation depth of about 3 dB is reached. After that point, the effect increases linearly with the logarithm of the modulation depth until a saturation point at approximately 30 dB is reached. [27] The model of fluctuation strength has been widely used for describing the annoyance of noise caused e.g. by wind turbines [40, 41, 42] and engines [43, 44].

In this chapter, the effect of time-variance on the annoyance of audio coding artifacts is investigated. First, the perceived effect of alternating the level of critical-bandwidth noise is discussed, after which the study is extended to time-variance of single harmonics of a periodic signal.

5.1 Band-Limited Noise

As explained in Chapter 1.2, USAC is based on the switched core principle meaning that the basic functionality is changed according to the type of the input signal. Figure 5.2 illustrates a case in which the coder erroneously switches between two modes even though the input signal is a stationary chord sung by a male choir. This behavior leads to clearly audible, and possibly annoying, fluctuations in quantization noise.

In this section, it is examined with a listening test if time-variance in the level of critical-bandwidth noise affects the annoyance caused by the noise. The results can be used to tackle an important trade-off in audio coding: switching the coder might help push the noise floor down momentarily, but the annoyance of abrupt changes might cancel the benefits.

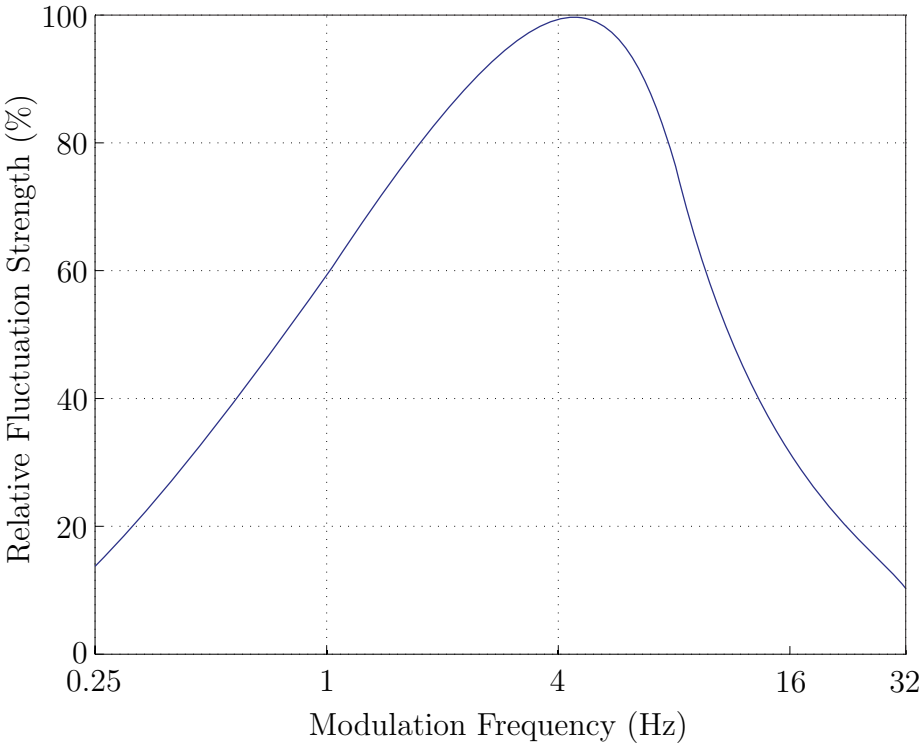

Figure 5.1: Fluctuation strength of amplitude-modulated broad-band noise of 40 dB modulation depth (adapted from [27]).

5.1.1 Methods

As illustrated in Figure 5.3, the listening test was built around a principle of adjusting the level of critical-bandwidth stationary noise so that it would be equally annoying to fluctuating but otherwise similar reference noise. Fluctuation was mostly periodic, but also one item with a varying, randomized period was included for completeness. To make the test more realistic, the noise was always added on the pitchpipe signal depicted in Figure 5.4.

The noise level was always adjusted frame by frame in order to maintain the SNR constant. The SNR was computed as in Equation (3.2) over the frequency band the noise covered. Computing the SNR in perceptual domain instead would not have made a significant difference as the frequency range was such narrow. In the actual listening test, the SNR difference (in this case, the opposite of the power level difference between the two noise types) related to the selected and the reference noise was recorded for each item.

All essential information related to the test arrangements is shown in Table 5.1. As our initial informal tests indicated that the frequency range would not have had a significant impact on the results, the noise was centered around the fifth harmonic that lay in the middle area of the human hearing range [31]. The lower frequency limit was $f_\mathrm{l} = 1.37$ kHz and the upper $f_\mathrm{u} = 1.57$ kHz, which according to Equation (3.1) correspond to approximately one critical band.

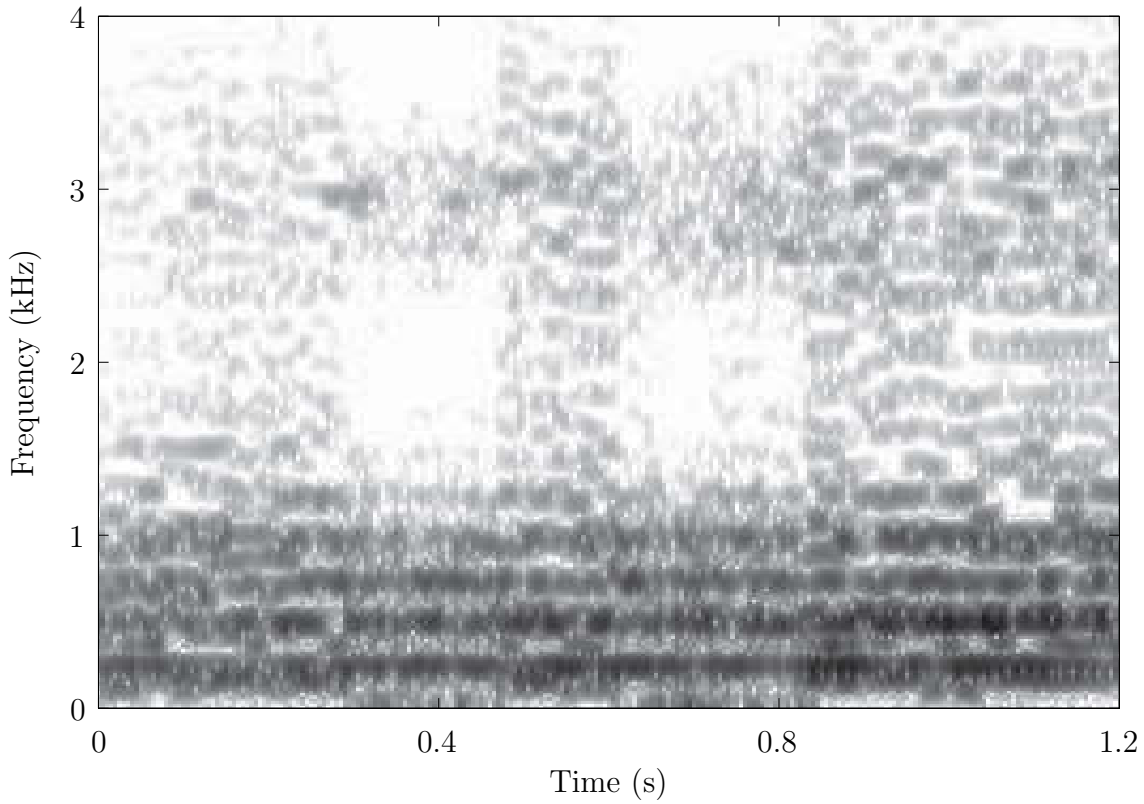

Figure 5.2: A coded excerpt of a stationary chord sung by a male choir. Note the abrupt changes in the quantization noise.

Table 5.1: The essential information about the test comparing the annoyance of static and fluctuating noise.

On/off time (ms)	32	64	128	256	512	$64-256$
Frequency range	$1350-1600$ Hz					
SNR (dB)	10 (for 128 ms only)		15		20 (for 128 ms only)	
Signal	Pitchpipe					
Frame length (ms)	64					
Sampling frequency (kHz)	16					
Bit depth	16					
Number of listeners	10					
Test method	MOA					

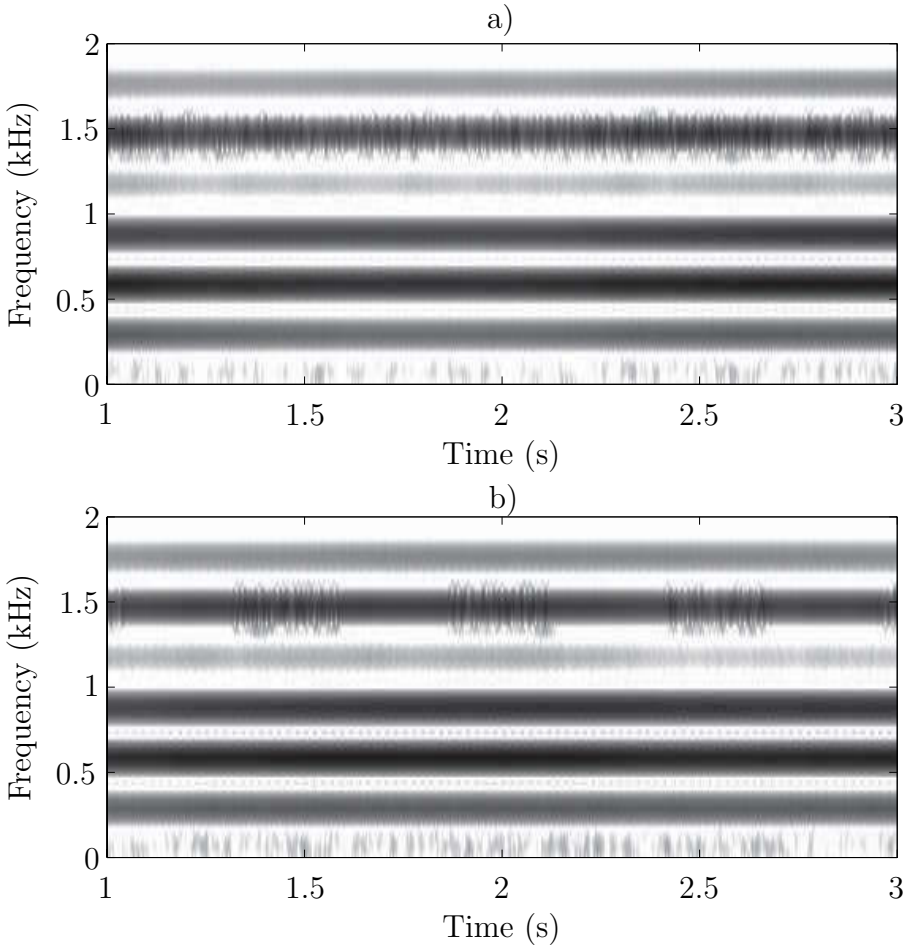

Figure 5.3: The listeners adjusted the level of (a) stationary noise while time-varying noise in the (b) reference signal was fixed. In this example, the modulation frequency is 0.98 Hz.

In addition, it was briefly investigated if the critical-band SNR in the reference signal would affect the results. For the majority of the items, the reference noise level was selected so that the noise would not be particularly prominent but still clearly audible (SNR = 15 dB). For comparison, the test included one item with just hearable (SNR = 20 dB) and another with very prominent (SNR = 10 dB) noise.

A block diagram of the Matlab script used to generate the items is shown in Figure 5.5. The signals were processed in 50% overlapping frames of 1024 samples and the noise was added in the MDCT domain. Because of the windowing used in the process, the noise blocks did not have sharp edges but were turned on and off smoothly. The α parameter was used to scale the noise to keep the critical-band SNR constant.

The method of adjustment test was conducted with the custom Matlab interface presented in Chapter 4.1.3. The level of time-varying noise could be controlled in steps of 0.5 dB. As the step size is less than the just-noticeable level change of

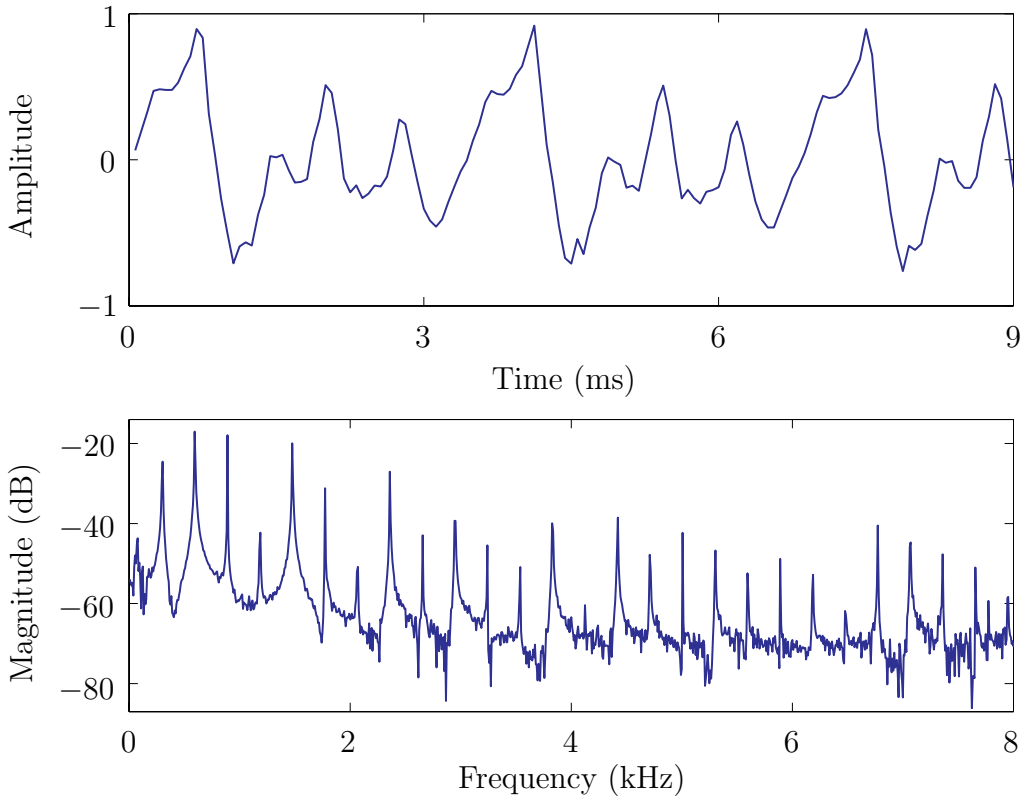

Figure 5.4: A frame of the pitchpipe signal that was used in the time-variance tests. The fundamental frequency is approximately $f_0 = 300$ Hz.

approximately 1 dB [27], it was possible to analyze the results as if the slider values had been continuous.

There were 10 participants in the listening test. Six of them could be considered as experienced listeners while four had little prior experience. Each item was evaluated twice and in random order and the test began with two training items with unique parameters. The participants were specifically asked to avoid matching the physical levels of the two noise types and to concentrate on the annoyance instead. The complete written test instructions can be found in Appendix B.

5.1.2 Results

The distribution of the test answers before post-screening is illustrated in a boxplot in Figure 5.6. The variance of the answers was relatively high, which was expected as the test setting was rather unusual. All individual answers are included in Appendix A, Figure A.1. Compared with the naive listeners (anon07–anon10), the experienced listeners (anon01–anon06) were clearly more consistent in their repeated answers and seemed to be following a common preference pattern. In post-screening, all items with a difference greater than 5 dB between the answers were excluded from further analysis. In addition, anon09 was considered too unreliable and was rejected entirely.

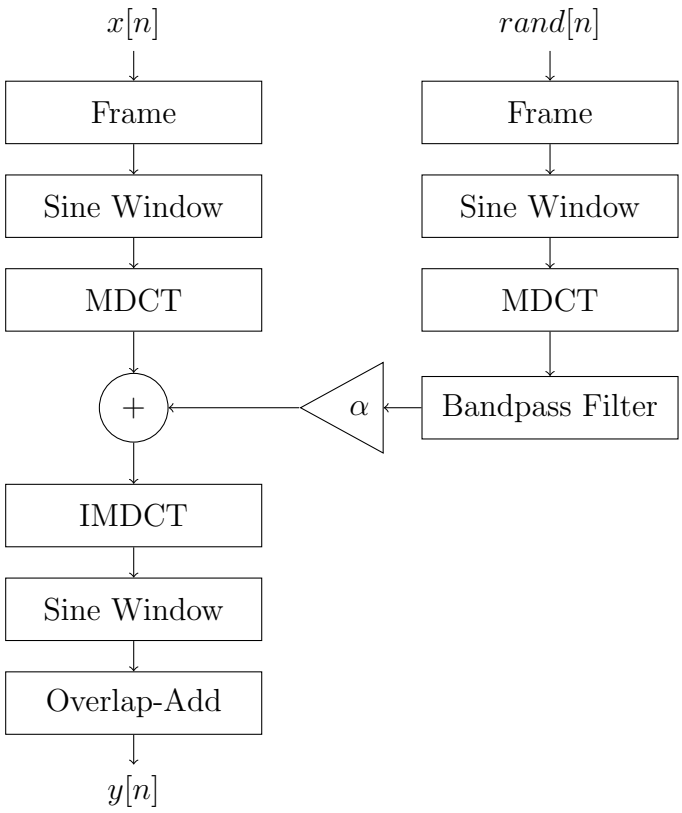

Figure 5.5: A block diagram of the signal generator script in the test comparing the annoyance of stationary and time-varying noise.

The means and confidence intervals after post-screening are shown in Figure 5.7. As there is clear overlapping in the confidence intervals of items, conclusions about annoyance differences should be drawn with caution. However, the two items with the fastest fluctuations and the SNR of 15 dB were seen significantly worse than the others with the same SNR, excluding the slowest variation. In addition, the results suggest that SNR had some effect on the perception of annoyance.

5.2 Amplitude Variation in Harmonics

A common artifact in audio coding is erroneous variation in the harmonic content of the coded signal. Weak harmonics, usually in the upper frequency range, might totally vanish in the quantization process and even stronger ones might have their level altered. To illustrate this, Figure 5.8 shows an excerpt of a coded quasiperiodic pitchpipe signal. Especially the fourth harmonic vanishes from time to time.

In this section, the annoyance of variation in the level of harmonics is discussed. The two scenarios explained in the previous paragraph were simulated and their annoyance evaluated in a listening test. It was designed to be a continuum of the test examining time-variance of noise, explained in the previous section.

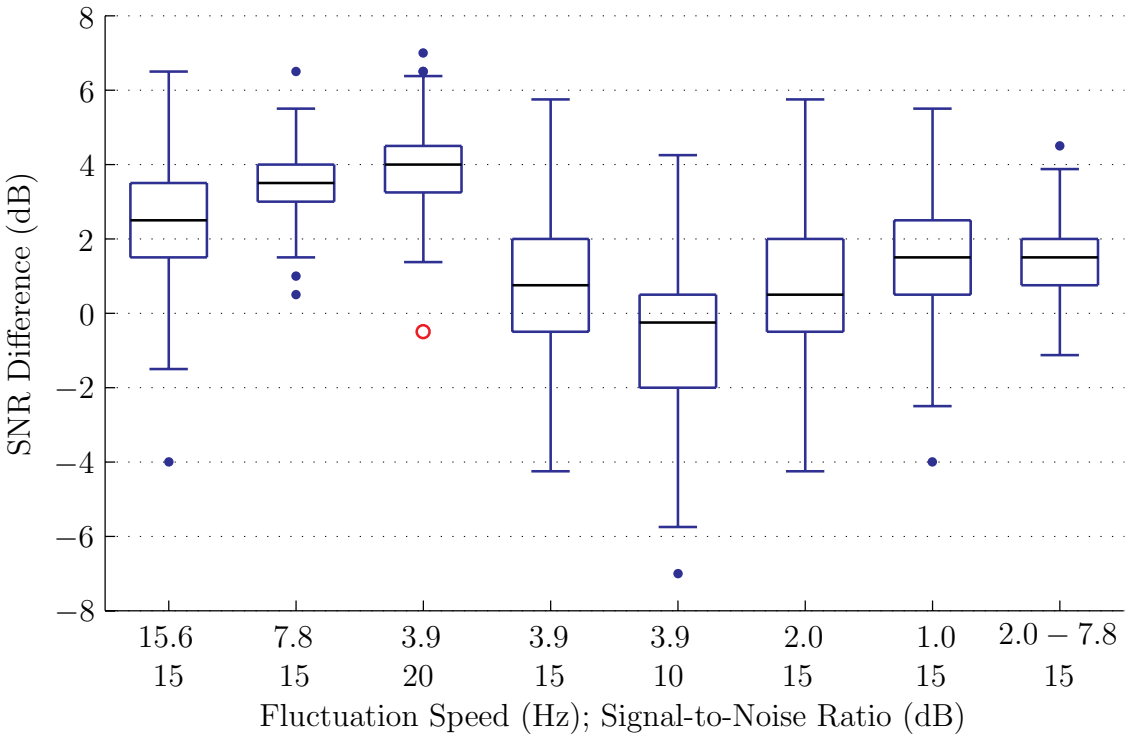

Figure 5.6: A boxplot of the answers of all listeners in the test comparing the annoyance of stationary and time-varying noise (before post-screening, $n = 10$).

5.2.1 Methods

In order to acquire comparable results, the listening test was designed and conducted principally the same way as the previous one described in Section 5.1. The main difference is that no noise was added to the reference signal, but harmonics were varied in one of the two ways instead. As before, the listeners were asked to set the level of stationary noise bandlimited to a frequency range comprising one critical band and the difference between the condition and reference signal SNRs was recorded.

In the first scenario (see Figure 5.9a for a graphical explanation), one of the middle harmonics, located at $f = 1.47$ kHz, was "blinking", i.e. the critical band centered on the harmonic was attenuated by 0 dB and 4 dB, in turns. The amount of attenuation was selected so that the phenomenon would be clearly audible but still realistic. The SNR of the conditions was computed frame by frame over a critical band $f_\Delta = [1.35, 1.57]$ kHz as in Equation 3.3. In the frames with zero harmonic attenuation, the SNR computed from the last attenuated frame was used to set the noise level for the reference signal.

In the second scenario (see Figure 5.9b), two of the upper harmonics were totally attenuated in turns. They were located at $f_1 = 4.7$ kHz and $f_2 = 5.0$ kHz, and the SNR was computed over a frequency range $f_\Delta = [4.35, 5.35]$ kHz that encompassed both harmonics. To keep the noise level in the reference signal reasonably stable,

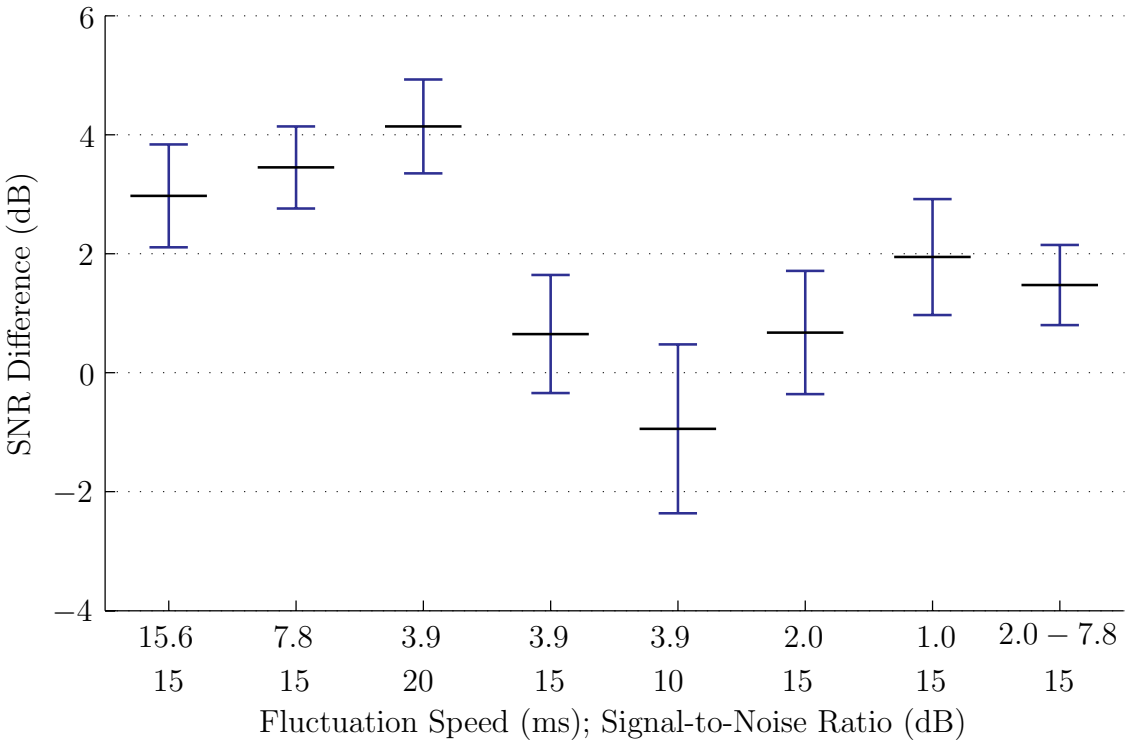

Figure 5.7: The means and 95% confidence intervals of the answers of all listeners in the test comparing the annoyance of stationary and time-varying noise (after post-screening, $n = 9$).

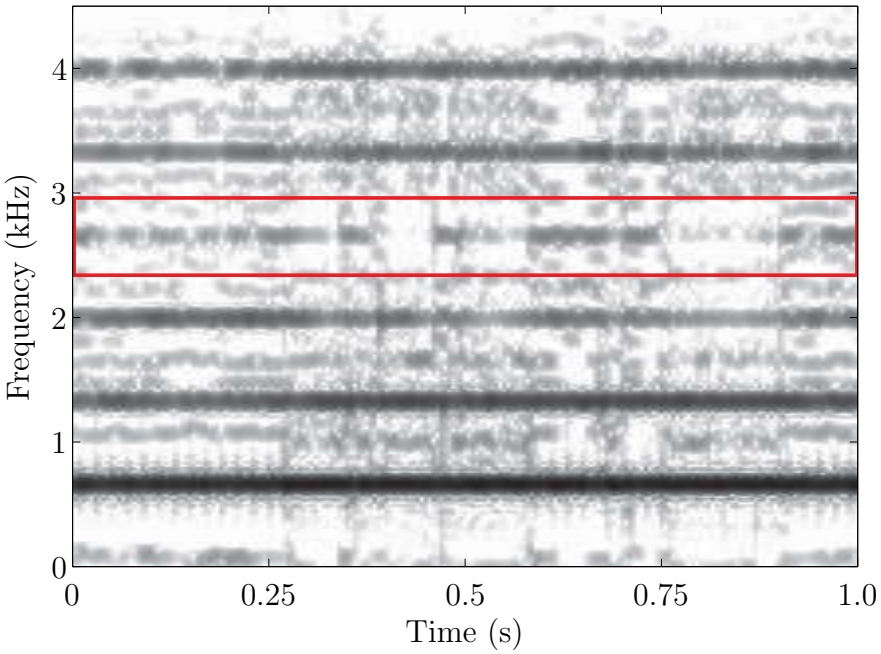

Figure 5.8: A coded excerpt of a pitchpipe signal. Note the fluctuations in the fourth harmonic.

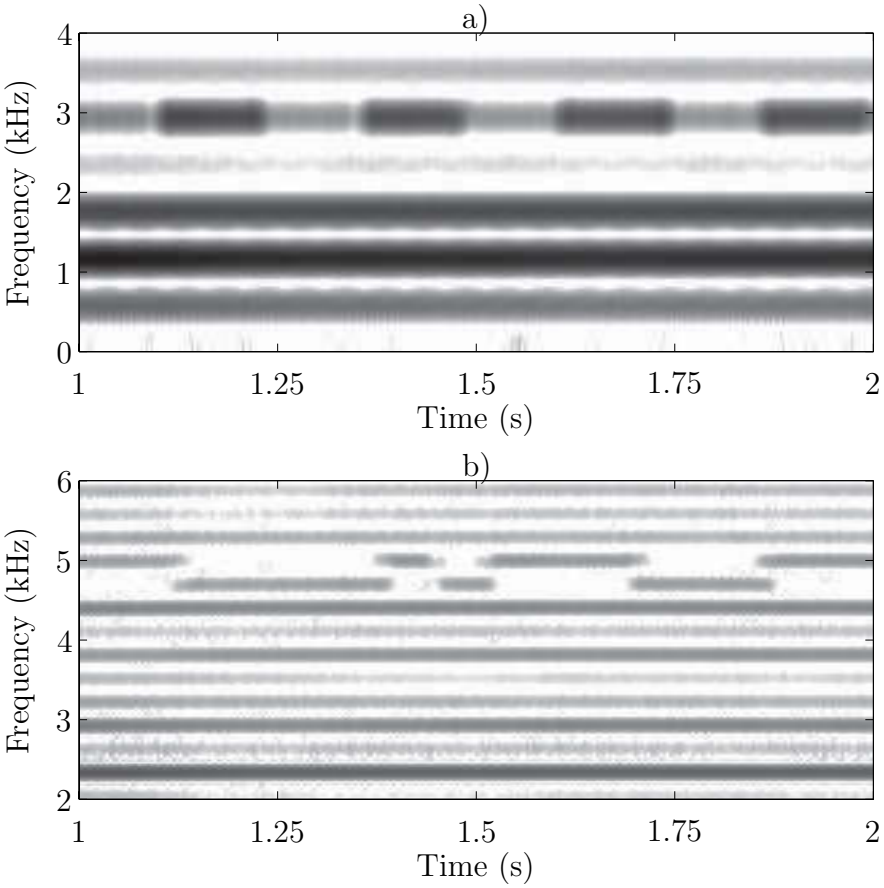

Figure 5.9: The two scenarios in the listening test examining time-variance of harmonics: a) the level of one harmonic fluctuating and b) two harmonics alternating.

the SNR in Equation 3.3 was computed in longer frames of 16,384 samples or 1,024 ms with a 50% overlap.

The Matlab script used for generating the files was principally similar to that explained in the previous section (see Figure 5.5) and the general test parameters, such as the sampling rate, were also the same (see Table 5.1). There were nine participants in this test, all being experienced listeners. Each item was evaluated only once to keep the test duration reasonable and the items were in a random order. The test began with two practice items (one for both scenarios) with unique parameters. The complete written test instructions given to the participants are included in Appendix B.

5.2.2 Results

The distribution of the test results before post-screening in the case of one harmonic is shown in Figure 5.10. Despite the experienced listeners, the variance of the answers was again quite high. All individual answers are shown in Appendix A,

Figure A.2. The performance of participants cannot be assessed based on repetitions as there were not any. Hence, no post-screening actions were taken.

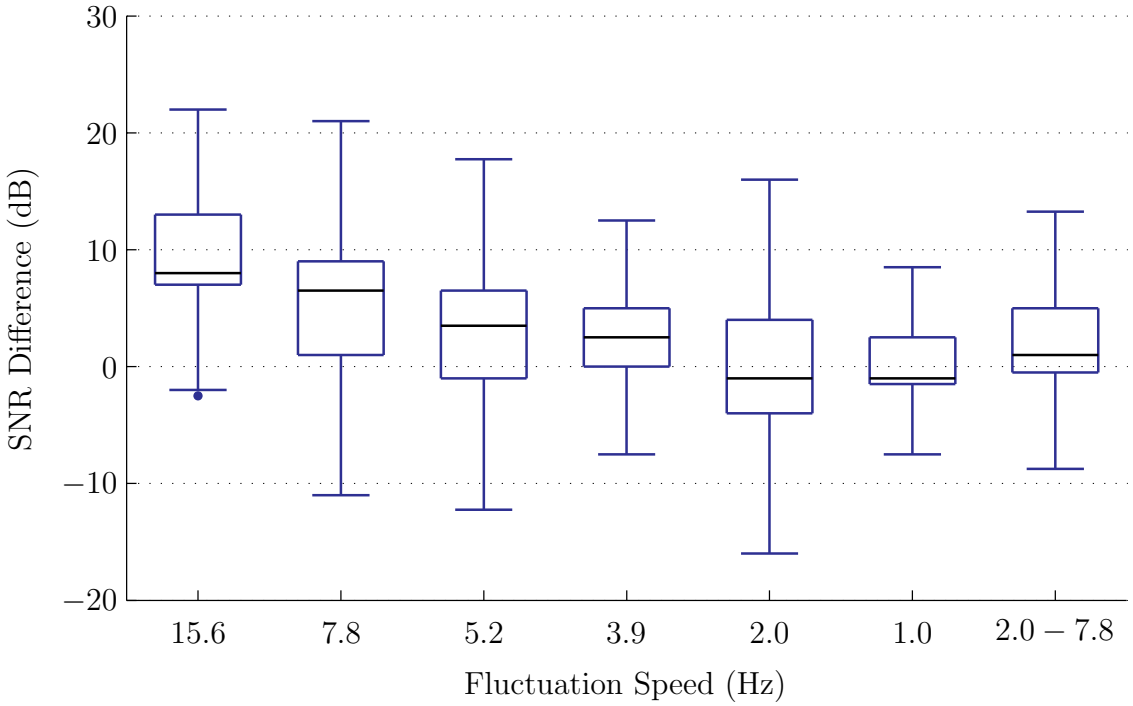

Figure 5.10: A boxplot of the answers of all listeners in the test investigating the annoyance of time-variance in the level of one harmonic (before post-screening, $n = 9$).

The means and confidence intervals after post-screening are shown in Figure 5.11. As the confidence intervals are wide and overlapping, statistically significant conclusions cannot be drawn from this dataset alone. However, there is a slight hint of the familiar U-shaped pattern.

A boxplot of all the answers in the case of two harmonics alternating is shown in Figure 5.12. Again and for the same reasons as before, the variance was relatively high. All individual answers are included in Appendix A, Figure A.3. There were no repetitions or an apparent common pattern that most listeners would follow. Hence, no listeners were rejected, but extreme outliers were discarded from single answers.

The means and confidence intervals of the answers are shown in Figure 5.13. There is again a slight hint of a U-shaped curve, but due to the limited number of participants, the results are not statistically significant.

5.3 Conclusions

The listening tests investigating the annoyance of time-variance were considered difficult among the participants, especially as the test procedure was somewhat unusual. Because of the time limitations of this project, the number of participants was slightly insufficient for getting thoroughly satisfactory results having enough

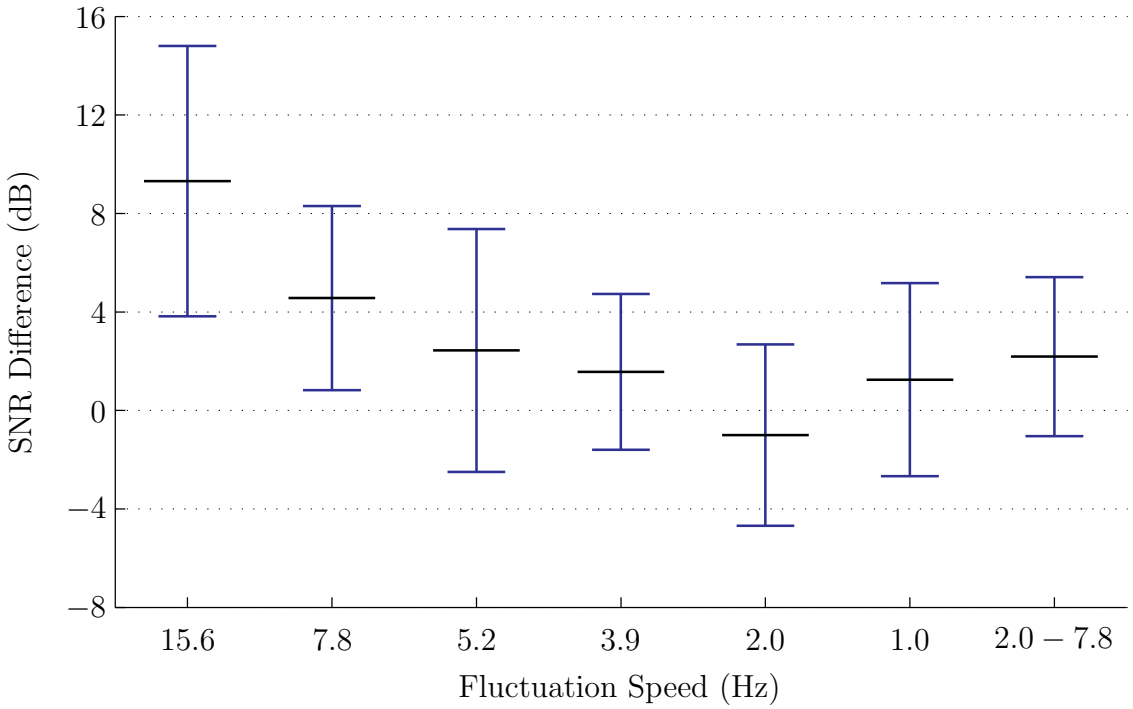

Figure 5.11: The means and 95% confidence intervals of the answers of all listeners in the test investigating the annoyance of time-variance in the level of one harmonic (no post-screening, $n = 9$).

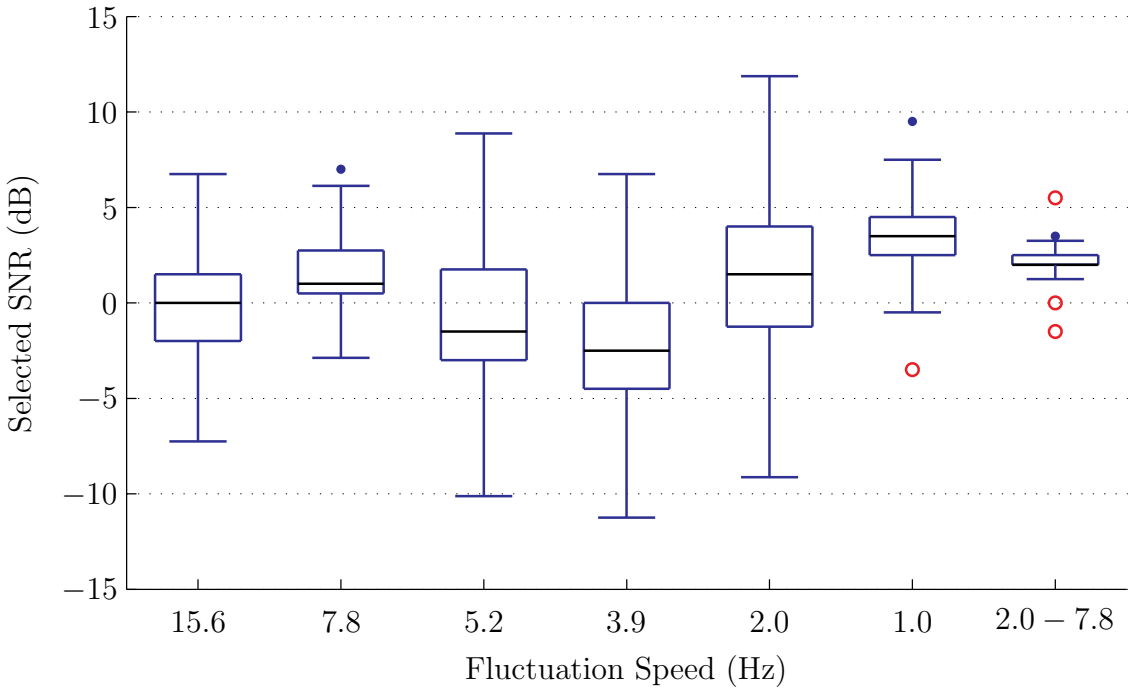

Figure 5.12: A boxplot of the answers of all listeners in the test with two harmonics alternating (before post-screening, $n = 9$).

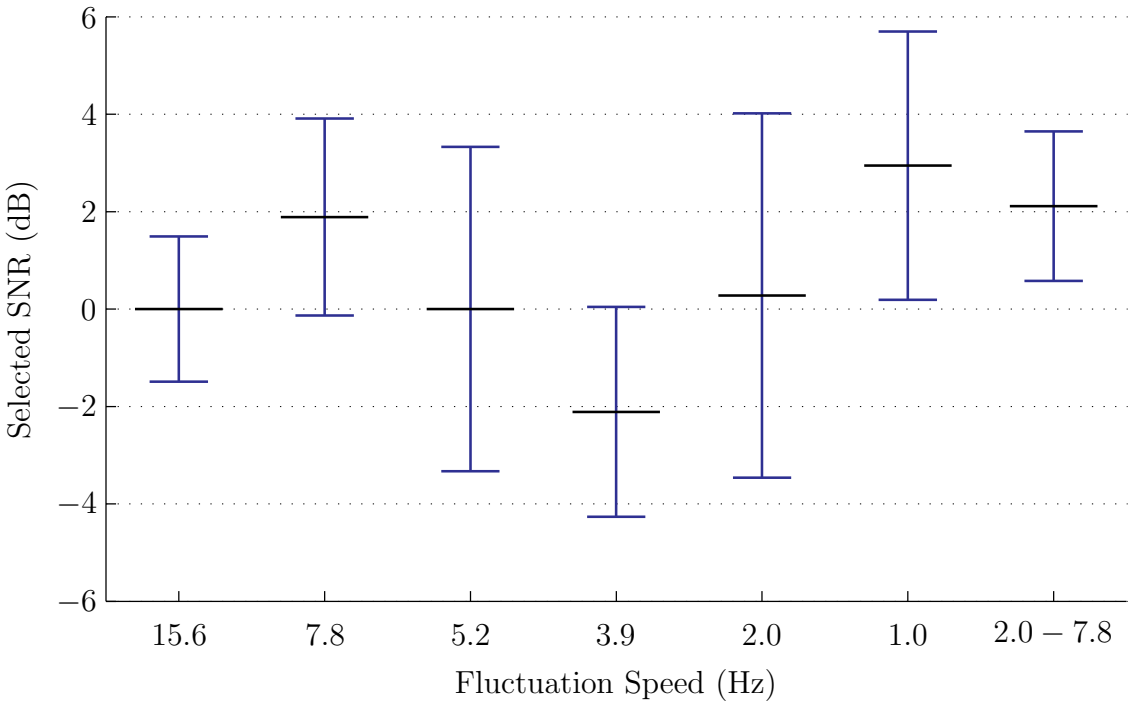

Figure 5.13: The means and 95% confidence intervals of the answers of all listeners in the test with two harmonics alternating (no post-screening, $n = 9$).

statistical significance. However, the emerged preference patterns were similar in all tested scenarios, which increases the reliability of our conclusions.

In accordance with the literature concerning fluctuation strength, the U-shaped annoyance patterns emerged from each test suggest that moderate-speed amplitude modulation at around 2-5 Hz is considered the most annoying, independent of the artifact type. Comments of the listeners also indicate that fast fluctuation is not heard per se, but such signals are perceived as stationary, albeit modulated.

The results of the noise scenario also suggest that the SNR might be a significant factor. When time-varying noise in the reference signal became louder, the listeners had a tendency to push the stationary noise even louder. This indicates that the annoyance of time-variance might be of more significance when the artifact is easily heard. However, the results might be somewhat distorted as some listeners reported that the lowest noise level was virtually inaudible in the test.

The research could be continued with verifying our results with more listeners and perhaps with other types of artifacts too. More modulation speed values should be included in further tests to enhance the accuracy of the findings.

Chapter 6

Ghost Pitch

Bandwidth extension refers to a family of various audio coding methods in which only the lower part of a spectrum is explicitly encoded while the upper part is artificially reconstructed on the decoder side. As the amount of information that has to be saved is much lower than in explicit encoding, significant bit savings can be achieved.

Perhaps the best-known bandwidth extension method is spectral band replication (SBR) in which the high frequencies (HF) are generated by copying or mirroring the low frequency (LF) band [45]. USAC also supports a novel method known as harmonic bandwidth extension (HBE) that is based on stretching the LF band. It has an advantage of preserving the harmonic relations of the tonal components, but unfortunately the HF patch will necessarily be incomplete [46]. As the pitch of a tone is determined by the whole set of harmonics (being multiples of the fundamental frequency by definition) and not just by the fundamental frequency [47], a possibility of an additional pitch sensation, called ghost pitch, emerges.

It has been long known that even if the fundamental component of a tone is missing, a pitch perception at that frequency might still occur [48]. This virtual pitch sensation is thought to be possible because the listener mentally resolves which fundamental frequencies could be related to the set of harmonics in question. For example, the only common submultiple of harmonics at 600 Hz and 750 Hz is 150 Hz which must thus be the fundamental frequency [49].

It is widely accepted that harmonics in the so called frequency dominance region are particularly important in the pitch perception [50, 51]. For typical fundamental frequencies below approximately 1000 Hz, the dominant harmonics are the few first ones [51, 50, 52], but there seems to be individual differences in their relative importance [52]. For higher fundamental frequencies, the role of upper harmonics decreases and the pitch is mainly determined by the fundamental frequency [51]. It has also been shown that the importance of a single low harmonic typically increases with its level as related to adjacent harmonics [52]. On the other hand, the phase of harmonics does not seem to have a significant effect [53].

In this chapter, the ghost pitch phenomenon is examined specifically in relation to the use of HBE in coding applications. The goal is to understand how the fundamental frequency of a harmonic tone and the crossover frequency separating

the LF and HF bands are related to the perceived strength of the possible ghost pitch. It is useful to be able to estimate the likelihood of a ghost pitch sensation in encoding as it helps to decide whether the problem should be tackled by manually reinserting some of the missing harmonics, for instance.

6.1 Background

Figure 6.1 illustrates how an HF band can be generated from an LF band with HBE. In that example, the LF spectrum is first stretched by an integer factor of 2 with a phase vocoder, which effectively stretches the signal in time to twice the original length while keeping frequencies intact. Decimating the signal by the same factor of 2 (i.e. dropping every second sample) results in returning to the original length whereupon the frequency of each harmonic is doubled. Finally, the generated HF band is exctracted with a bandpass filter and the output signal is constructed by combining the original LF and the new HF bands. Figure 6.2 illustrates how the resulting spectrum is formed. Note that all odd harmonics are missing in the HF spectrum. [46]

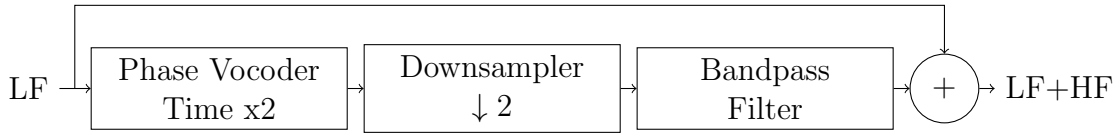

Figure 6.1: A block diagram of HBE with a factor of 2.

The LF band can also be stretched with larger factors to extend the generated frequency range, but in that case the harmonic patch will be even sparser. One useful possibility is to mix HF parts produced by different stretch factors to get better harmonic coverage. To ensure a high perceived quality of the outcome, the envelope of the synthetic HF part should always be shaped to match the envelope of the original spectrum.

6.2 Methods

In the listening test explained in this chapter, the strength of the ghost pitch sensation in HBE was evaluated with periodic signals as a function of fundamental frequency f_0 and crossover frequency f_x. Six values were selected for both of those parameters and all possible combinations of them were evaluated with two different spectral envelopes, totalling 72 items to be rated in terms of perceived ghost pitch loudness. All signals were synthetic as full control over harmonics was desired. The test arrangements are summarized in Table 6.1.

In real applications, crossover frequencies are likely to be set to integer divisions of the Nyquist frequency. For example, if the sampling rate was $f_s = 16$ kHz and the HBE factor $h = 2$, an obvious choice for the crossover frequency would be $f_\mathrm{x} = 0.5 f_\mathrm{s}/h = 4$ kHz. Following that logic, the typical sampling rates of 16, 22.05,

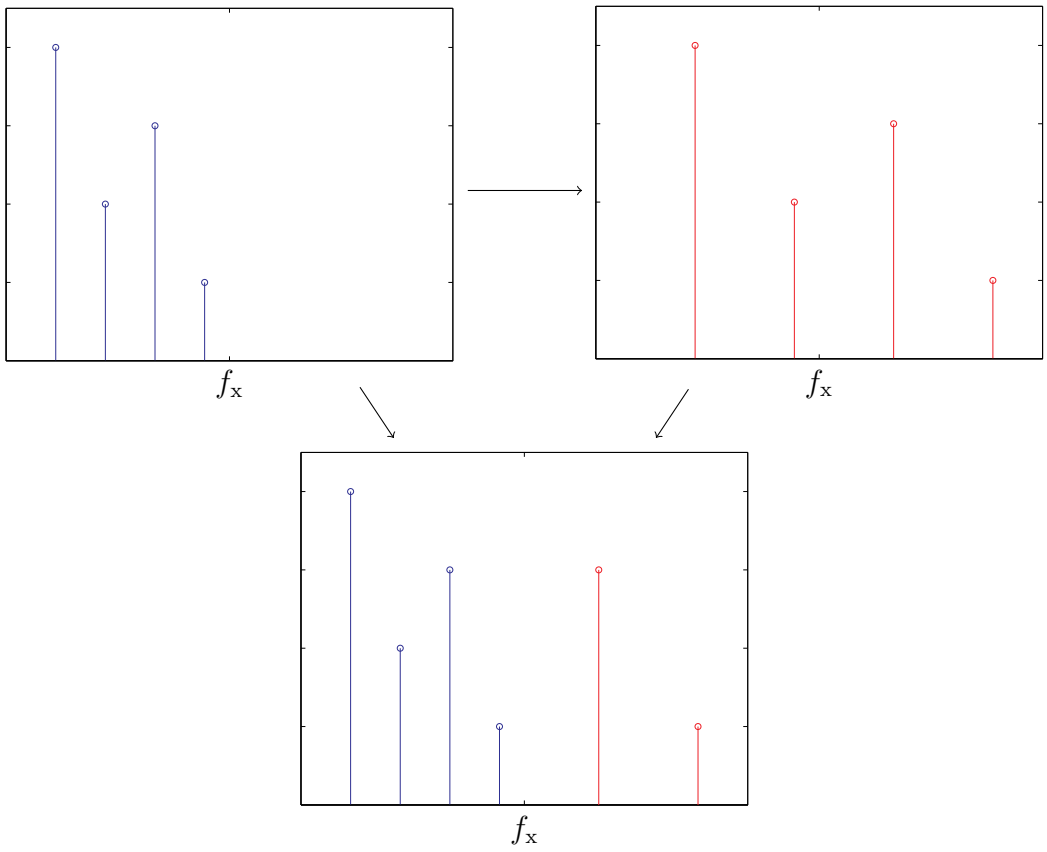

Figure 6.2: In HBE, the HF spectrum (red) is generated by stretching the original LF spectrum (blue) after which the two bands are combined. The crossover frequency is denoted with f_x.

Table 6.1: A summary of the ghost pitch test arrangements.

$f_0\,(Hz)$	100	170	289	491	835	1420
f_x (kHz)	4	5.513	6	8	11.025	12
Spectral envelope	Loudness			−3 dB/octave		
Signal generation	Additive synthesis; spectra extended to Nyquist frequency					
Fade in/out	Sine/cosine window, 150 ms					
Sampling frequency	48 kHz					
Bit depth	16					
Test method	Rating without reference					

24, 32, 44.1, and 48 kHz would produce the crossover frequencies evaluated in this test. Fundamental frequencies f_0, on the other hand, were selected to cover the normal range of typical musical instruments [31]. The frequencies were distributed

logarithmically so that in terms of musical intervals they would be approximately a major 6th apart.

Two different spectral envelopes were applied in the items: one following the equal-loudness contour at 60 phon (Figure 3.2) and another with a −3 dB per octave tilt. The former can be argued to be the most neutral choice since all harmonics are equally loud while the latter was included to represent a more natural tone.

The timbre and pitch of a tone are mostly independent of the phase of harmonics [31, 53]. In our test, the phase was simply randomized with two functions, one for each spectral envelope. On the one hand, it is known that the attack of a tone plays a crucial role in a timbre sensation [31]. The possible impact of attack was excluded by fading the signals in and out smoothly using the first quarters of the sine and cosine windows as the amplitude functions.

The items were generated offline with Matlab in which HBE was simulated with additive synthesis. In reality, HBE with the factor of 2 would produce harmonics up to twice the crossover frequency, but we chose to ignore this limitation by extending all signal spectra up to the Nyquist frequency.

The total number of listeners was eight with varying amounts of listening test experience. Five of them could be considered experienced while three were relatively inexperienced. Because of the large number of items, the time-efficient rating without reference method was selected for the listening tests. The listeners were asked to rate the ghost pitch loudness using a scale with a step size of 1 by considering the following guidelines:

> If you hear only one pitch, set the slider to 0.
>
> If you hear two equally loud pitches, set the slider to 100.
>
> If you hear two pitches, but one is stronger than the other, set the slider in between 0 and 100 according to the perceived loudness ratio of the pitches (e.g. set the slider to 50 if the loudness of the softer pitch is 50% of the louder one).

The complete written test instructions are included in Appendix B.

6.3 Results

All individual answers are included in Appendix A, Figures A.4 and A.5 (among the listeners, anon01-anon02 and anon04-anon07 can be considered experienced). First, it can be clearly seen that the bias in the ratings is vastly different between listeners (compare anon02 and anon03, for example). Some participants commented afterwards that it was hard to adjust to the rating scale as there were no examples at the beginning of the test. Second, even the relative ratings are not nearly consistent among listeners. For instance, anon06 and anon07 seemed to perceive the effect of fundamental frequency almost oppositely. While it is possible that the participants perceived the ghost pitches differently, it is more likely that some of them did not understand the objective, the test method itself turned out to be unreliable, or that

the synthesized items were not suitable and should have been substituted with real sounds.

To summarize the distribution of the answers of the listeners, the interquartile ranges are plotted in Figure 6.3. There seems to be a consensus on some items while the IQR is very high in others, without any apparent logic. This cannot be caused by the time needed for adjusting to the scale as the items were in random order.

The means of the answers are shown in Figure 6.4. As there were no repetitions, no clear common rating patterns, or anything revealing whether a single listener had understood the test correctly, nothing could be rejected in post-screening. It seems that, at least in the items with the tilted envelope, the listeners on average perceived a stronger ghost pitch sensation with lower fundamental and crossover frequencies.

Our limited data suggests that the nature of the ghost pitch phenomenon cannot necessarily be fully captured with a simple linear model. However, to quantify our findings, a two-variable linear regression model describing the perceived ghost pitch strength with a scale from 0 to 100 was derived from the results (Figure 6.5):

$$g(f_0, f_\mathrm{x}) = 52.4 - 0.0141 f_0 - 0.0009 f_\mathrm{x}. \tag{6.1}$$

As expected, the model does not fit the data that well and the goodness of fit is hence low, $R^2 = 0.051$. Nevertheless, some linear dependence on both regressors seems to exist as the significance level for the model as a whole is $P = 0.0000$ ($F = 15.4$), for the constant coefficient $P = 0.0000$ ($t = 13.6$), for the f_0 coefficient $P = 0.0426$ ($t = -2.03$), and for the f_x coefficient $P = 0.0000$ ($t = -5.17$).

6.4 Conclusions

While our test results are relatively noisy and the dependence might be more complicated than what can be uncovered from our limited data, the results suggest that ghost pitch sensation is louder for lower fundamental and crossover frequencies. This is in line with what is known about virtual pitch. First, it is suggested in the literature that lower harmonics play the most important role in pitch sensation. Increasing the crossover frequency in HBE makes the sparse harmonic patch begin higher in frequency and hence provides less indication of the virtual pitch suggested by the harmonic composition of the HF band. Second, it is proposed in the literature that harmonics of a tone become less important in pitch sensation when the fundamental frequency gets higher. Increasing the fundamental frequency of a tone in HBE also increases the frequency of the suggested virtual pitch and therefore the incompleteness of the HF patch becomes less significant.

To investigate ghost pitch further, more participants would be needed for a test. It is not clear if the rating without reference method itself was too unreliable in our case, but at least there should be some example items and answers in the beginning. In addition, the test could be tried with real tones as the sensation seems to be stronger and clearer in that case. In the meanwhile, the two-variable linear regression model in Equation (6.1) can be used as a starting point for further research or encoder designs.

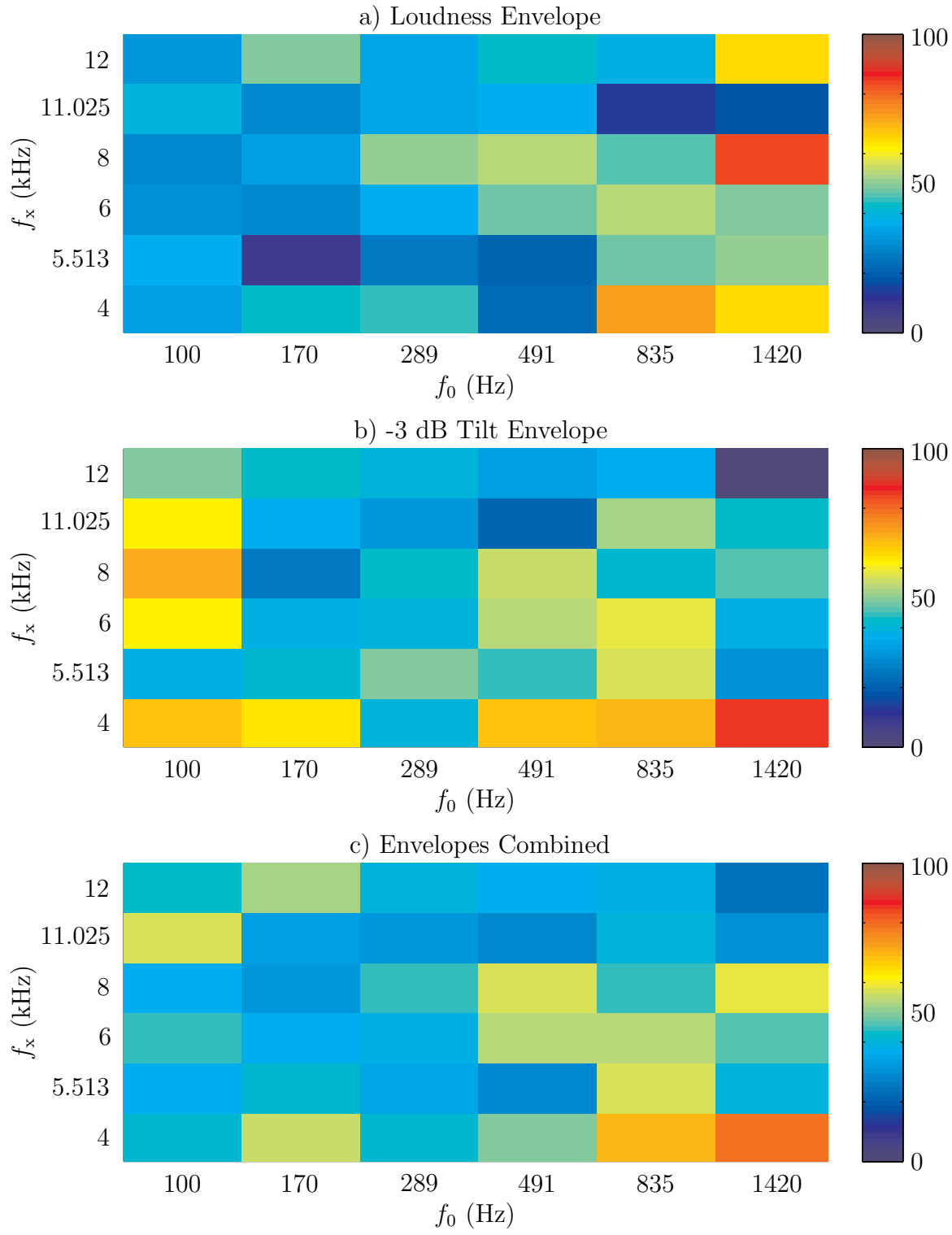

Figure 6.3: The interquartile ranges of the answers of all listeners for a) the loudness envelope, b) the tilted envelope, and c) both envelopes combined ($n = 8$, no post-screening).

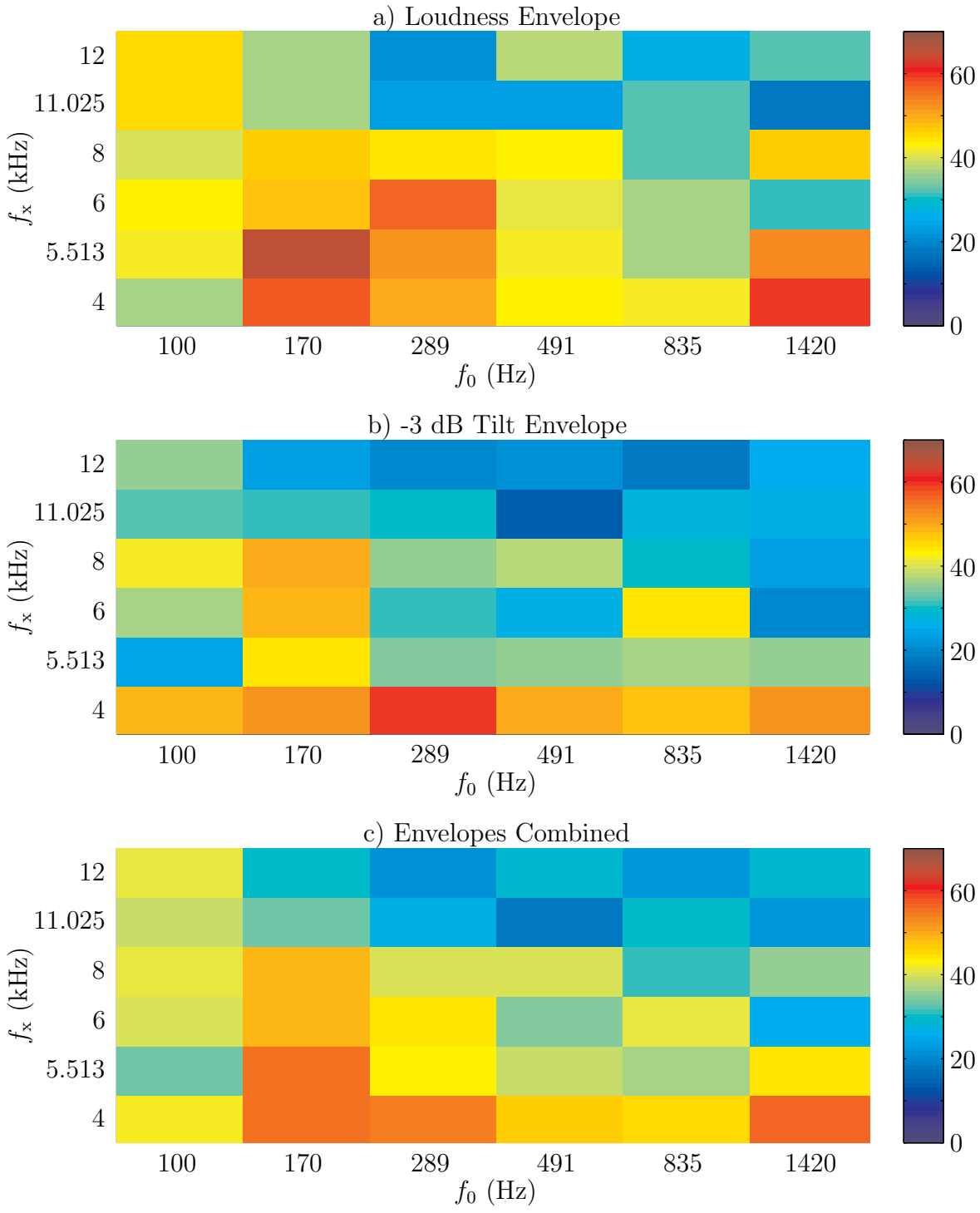

Figure 6.4: The means of the answers of all listeners for a) the loudness envelope, b) the tilted envelope, and c) both envelopes combined ($n = 8$, no post-screening).

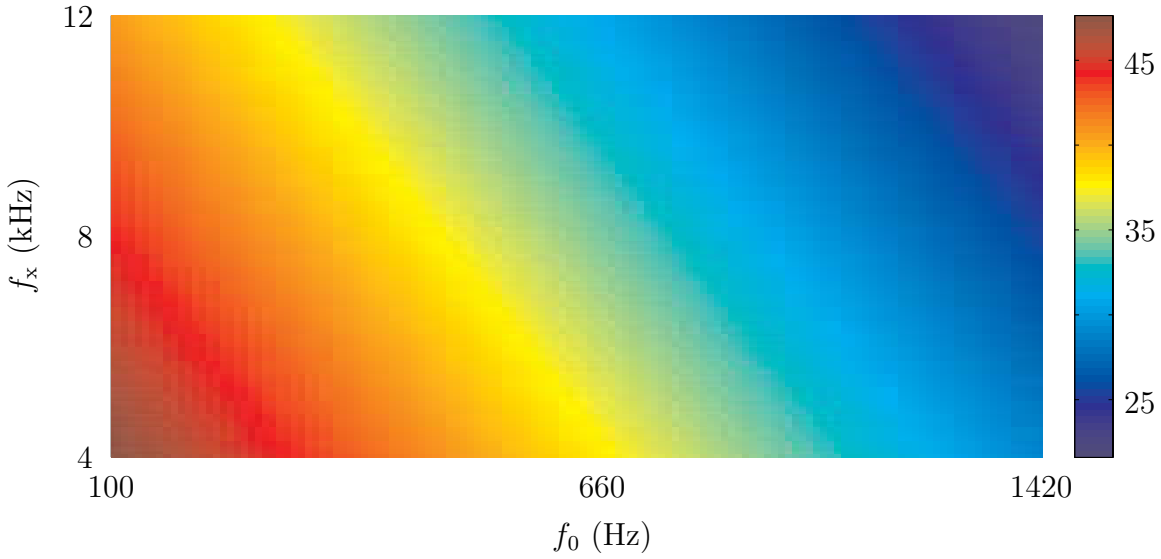

Figure 6.5: A two-variable linear regression model (Equation (6.1)) derived from the answers of all items in the ghost pitch test.

Chapter 7

Spectral Envelopes in Audio Coding

As explained in 3.4, signals are customarily quantized in the perceptual domain in order to minimize the annoyance of the resulting noise. The first part of this chapter discusses an improvement for that process by modifying the transfer function of the spectral envelope used in the transformation.

Formant enhancement is another technique for shaping and hiding quantization noise. It refers to boosting strong and diminish weak parts of the spectra of the frames in encoding. Using spectral envelopes as the basis for that procedure is examined in the second part of this chapter.

7.1 Optimizing Perceptual Domain Transformation

In a perceptual domain transformation, an estimate for the masking curve is required for each frame, for which a spectral envelope is often used. To improve the masking curve approximation and thus enhance the masking of quantization noise, the envelope can be modified with a method belonging to the class of bandwidth expansion: the spectral peaks of the envelope are widened and the spectrum thus smoothed by plugging a constant into its transfer function [54]. The goal of this listening test was to find the optimal value for the smoothing constant. In this section, the technique is introduced in more detail and the results and methods of a listening test aimed toward finding the optimal parameter for the bandwidth expansion are presented.

7.1.1 Background

In an example shown in Figure 7.1, an envelope $W(z)$ of the spectrum $X(z)$ of a signal frame is extracted with LPC. The poles of this all-pole transfer function $W(z)$ can be moved inwards by choosing a constant $\gamma_0 < 1$ and evaluating the transfer

function as

$$W(z/\gamma_0) = \frac{1}{1 + a_1\gamma_0 z^{-1} + a_2\gamma_0^2 z^{-2} + ... + a_k\gamma_0^k z^{-k}},$$

where a_k are the constant coefficients of the transfer function [54]. Moving the poles towards the origin has an effect of flattening the envelope as can be seen in Figure 7.1.

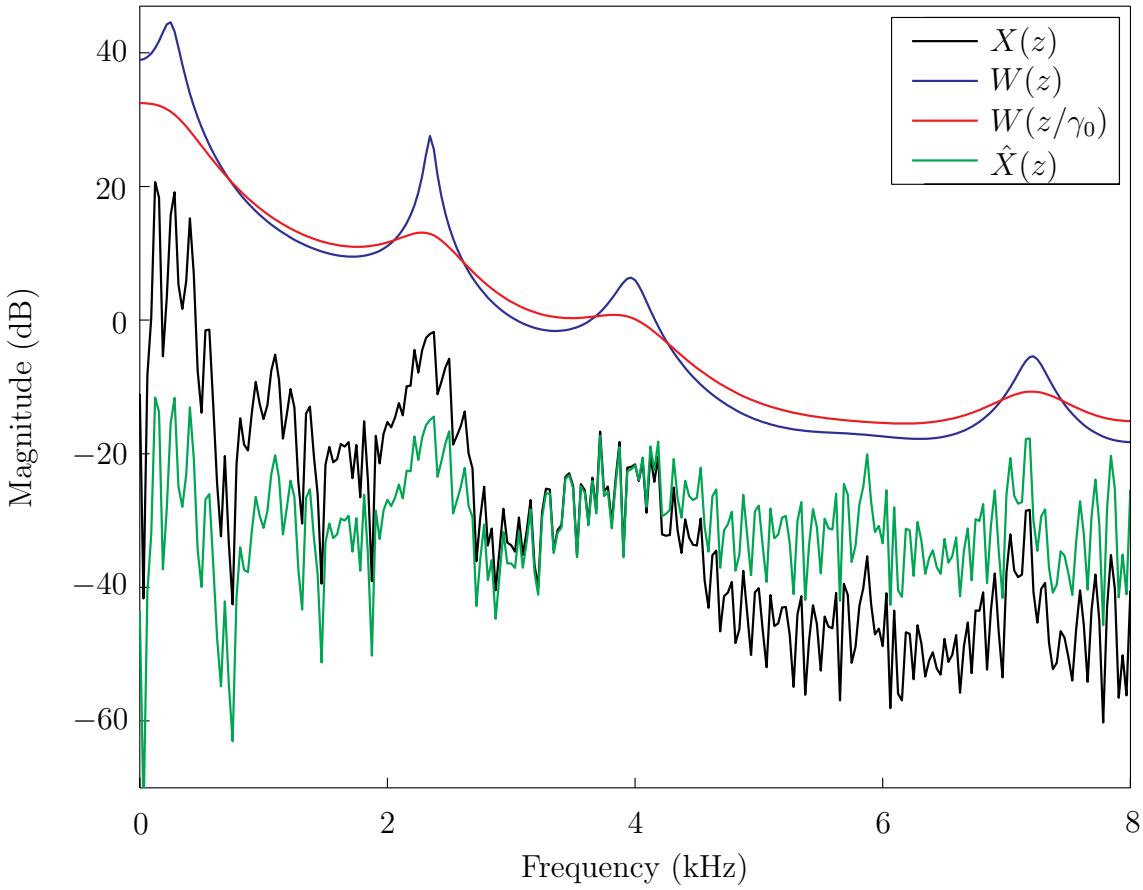

Figure 7.1: A 32 ms frame of a speech signal transformed from a frequency domain representation $X(z)$ into a perceptual domain counterpart $\hat{X}(z)$. The spectral envelope $W(z)$ used for the transformation is smoothed with the constant $\gamma_0 = 0.91$.

If all frequency bins of a signal frame are quantized equally, the resulting noise can be modelled as white noise $V(z)$ (having a flat spectrum by definition) added to the frame. In the frequency domain, the outcome is simply

$$Y(z) = X(z) + V(z).$$

However, if the quantization is done in the perceptual domain instead, back in the original domain the signal becomes

$$Y(z) = X(z) + V(z)W(z/\gamma_0)$$

The perceptual SNR of the outcome can be computed as

$$SNR_{\mathrm{P}} = 10 \log_{10} \frac{\| X[k] W_{\gamma_0}^{-1}[k] \|}{\| V[z] \|}.$$

The smoothing parameter γ_0 should be chosen so that the noise is concentrated under strong parts of the input spectrum, which would lead to a higher perceived SNR and, supposedly, a decreased annoyance.

7.1.2 Methods

Our methods can be described as follows. First, the input signals were divided into overlapping frames. Second, the envelope spectra of the frames were extracted with LPC and smoothened with different γ_0 values. Third, for each input signal frame, a frame of white noise was created and shaped with the corresponding modified LPC envelope. Finally, the output signals were constructed by adding the input signal frames and the corresponding noise frames. In the listening tests, the participants were then asked to compare and rate the perceived quality of the conditions and hence evaluate the effect of γ_0.

As listed in Table 7.1, five γ_0 values that clearly covered the practically usable range were selected for the test. Two pure vocal excerpts were chosen for the test signals: one from an English-speaking male newsreader and another from the infamous track "Tom's Diner" by Suzanne Vega. Music and mixed content were ignored because our informal tests suggested that the effect of adjusting γ_0 would be almost inaudible in those cases. This is probably because spectral envelopes in musical content tend to be quite flat and thus smoothing would not make a significant difference.

Table 7.1: The parameters of the listening test investigating perceptual domain transformation.

γ_0	0.84	0.88	0.92	0.96	1.00
SNR_{P} (dB)	6		12		18
Signal	Male newsreader (English)		Suzanne Vega: Tom's Diner		
Frame length (ms)	32				
Sampling frequency (kHz)	16				
Bit depth	16				
Number of listeners	10				
Test method	MUSHRA				

Perceptual SNR was also varied to investigate whether it has an impact on the listener preferences. It was kept constant in conditions by multiplying each generated noise frame with a scalar α before adding to the signal frame:

$$\alpha = \sqrt{\frac{1}{SNR_{\mathrm{P}}}} \frac{\parallel X[k]W_{\gamma_0}^{-1}[k] \parallel}{\parallel V[k] \parallel}.$$

Furthermore, to keep the loudness of conditions roughly equal to that of the reference signal, each output frame was multiplied with a scalar σ given by

$$\sigma = \frac{\parallel X[k]W_{\gamma_0}^{-1}[k] \parallel}{\parallel X[k]W^{-1}[k] + V[k] \parallel}.$$

This equalized the energy of the input and output frames in the perceptual domain and seemed to be reasonably accurate for loudness normalizing. The test signals were generated with a Matlab script illustrated as a block diagram in Figure 7.2.

MUSHRA was chosen as the test method as it enabled comparing all γ_0 values related to a single item at once. In addition, the total number of conditions was quite large which made MUSHRA a reasonable choice as it is fast to execute. The test was anticipated to be challenging and therefore only experienced listeners were invited. There were 10 participants in total. The complete written test instructions can be found in Appendix B.

7.1.3 Results

The distribution of the listening test answers before post-screening is summarized in a boxplot in Figure 7.3. All individual answers are included in Appendix A, Figure A.6. All listeners were experienced and very familiar with the MUSHRA method, hence no post-screening actions other than discarding extreme outliers were seen necessary. Altogether, the hidden reference anchor (HidRef) was detected correctly in all but one items and the 3.5 kHz lowpass anchor (LPAnchor) was consistently rated reasonably. The means and confidence intervals of the ratings after post-screening are shown in Figure 7.4 for each item.

As expected, there was clear positive correlation between perceptual SNR and the given ratings. However, the differences of the ratings of conditions were generally small and the confidence intervals are clearly overlapping, even though there seems to be a reverse U-shaped pattern of preferences shared by most items. From Figure A.6 can be seen that the listeners used the rating scale very differently, especially in terms of bias. To overcome this problem, the means and 95% confidence intervals are calculated for the differences between the ratings of adjacent γ_0 values in Figure 7.5. The individual bias in using the scale is thus disregarded, but the magnitude of the perceived differences is preserved. Apparently, the differences were clear in the items with a low perceptual SNR, but especially the low-noise items with $SNR_{\mathrm{P}} = 18$ dB were difficult.

As the patterns of answers are relatively similar in each item, we can reasonably judge that perceptual SNR did not affect the preferences significantly. Therefore,

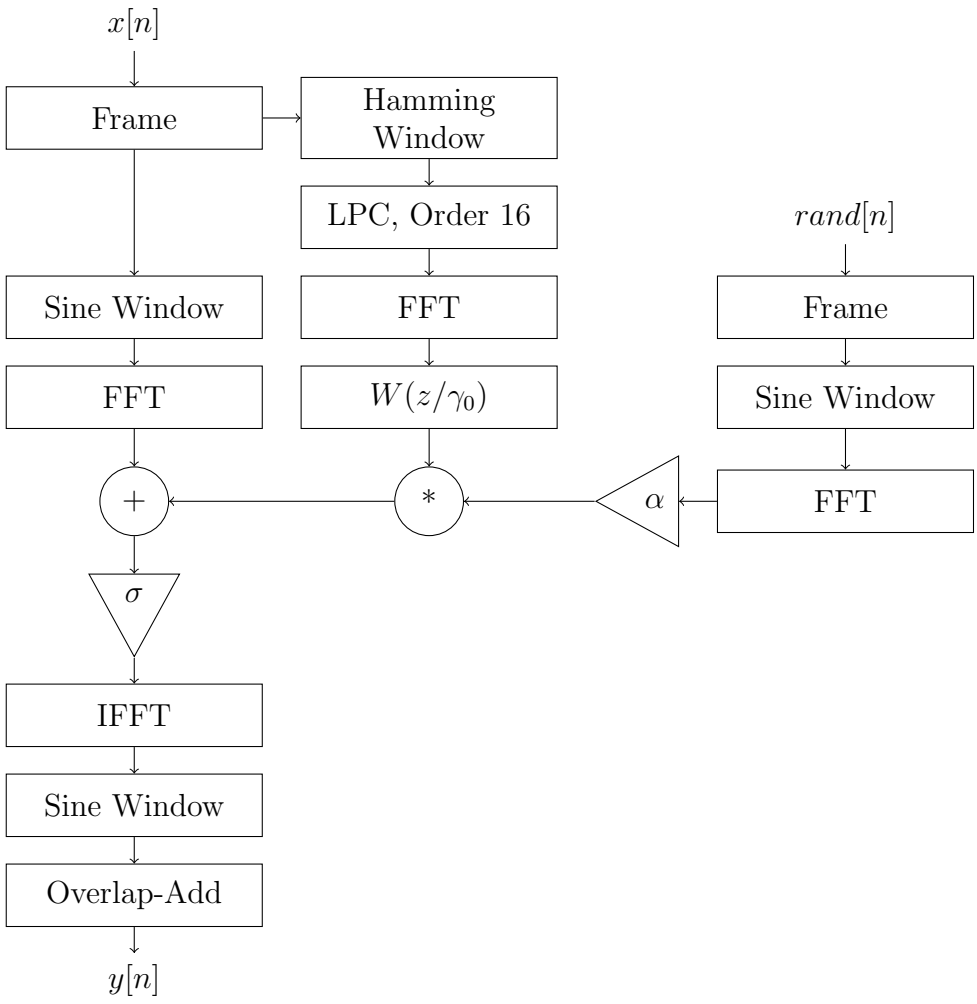

Figure 7.2: A block diagram of the condition generator in the test investigating perceptual domain transformation.

all items could be combined to get the curve shown in Figure 7.6, in which the lowest rated value $\gamma_0 = 1.0$ is (arbitrarily) set to zero and the points are connected with cubic spline interpolation. The overshoot in the red curve is a result of the interpolation and is not related to the preferences of the listeners. The maximum of the curve, and hence the candidate for the optimal value, is at $\gamma_0 = 0.913$.

Still another way to interpret the results is to analyze them as if they were acquired with pairwise comparisons, as explained in Section 4.2.2. In this method, even the magnitudes of the rating differences between conditions are disregarded. The data satisfied the stochastic transitivity requirements in Equation 2.5 reasonably well and hence the ratio scale could be derived with the BTL method from the combination of all items. For easier comparison, the BTL curve, shown in Figure 7.6, was scaled vertically so that it would roughly match the curve found earlier with the difference method. The maximum value of the BTL curve is at $\gamma_0 = 0.915$ which is very close to that found with the difference method.

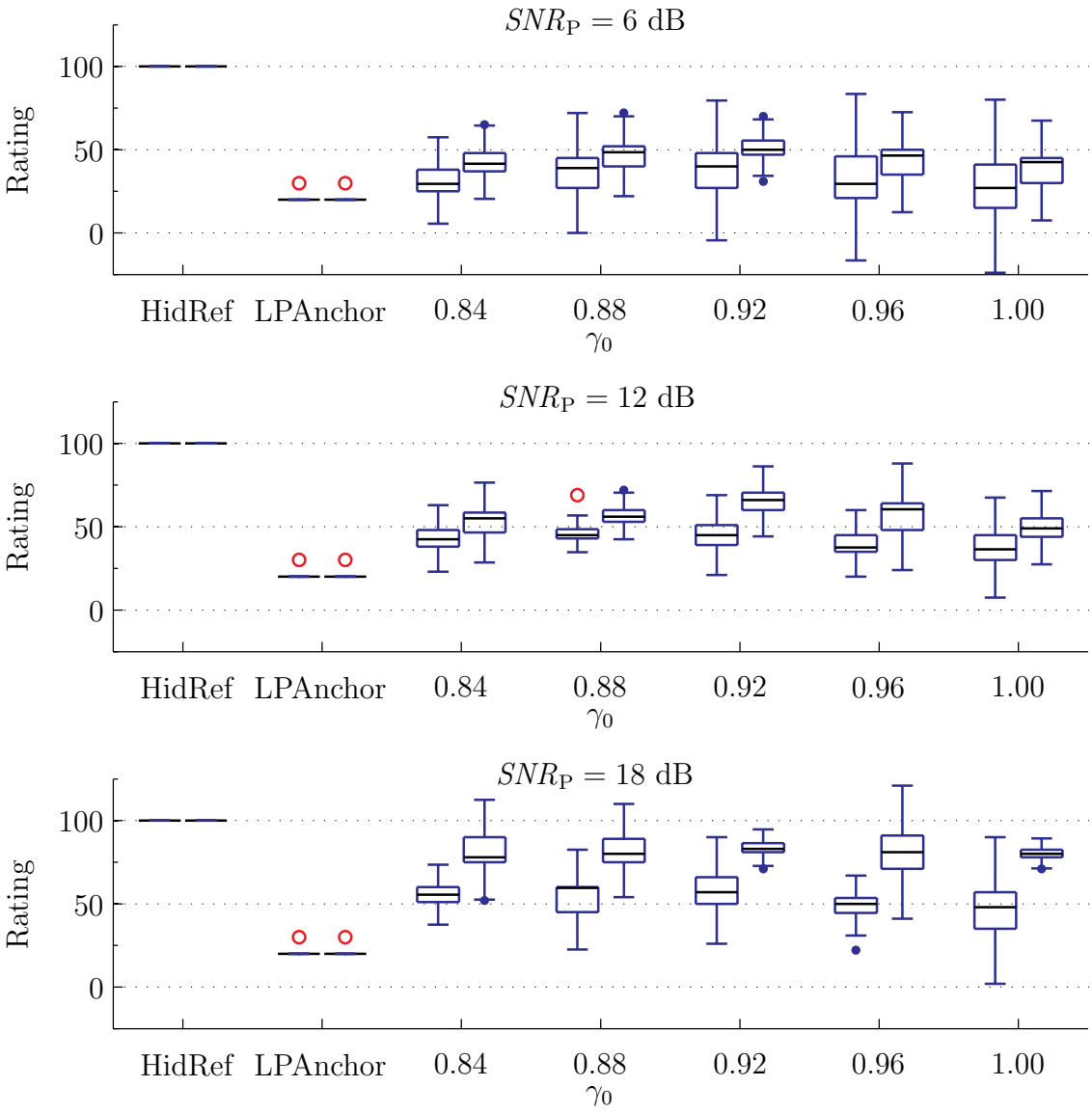

Figure 7.3: A boxplot of the answers of all listeners in the test investigating perceptual domain transformation (before post-screening, $n = 10$, left = male speech, right = Suzanne Vega).

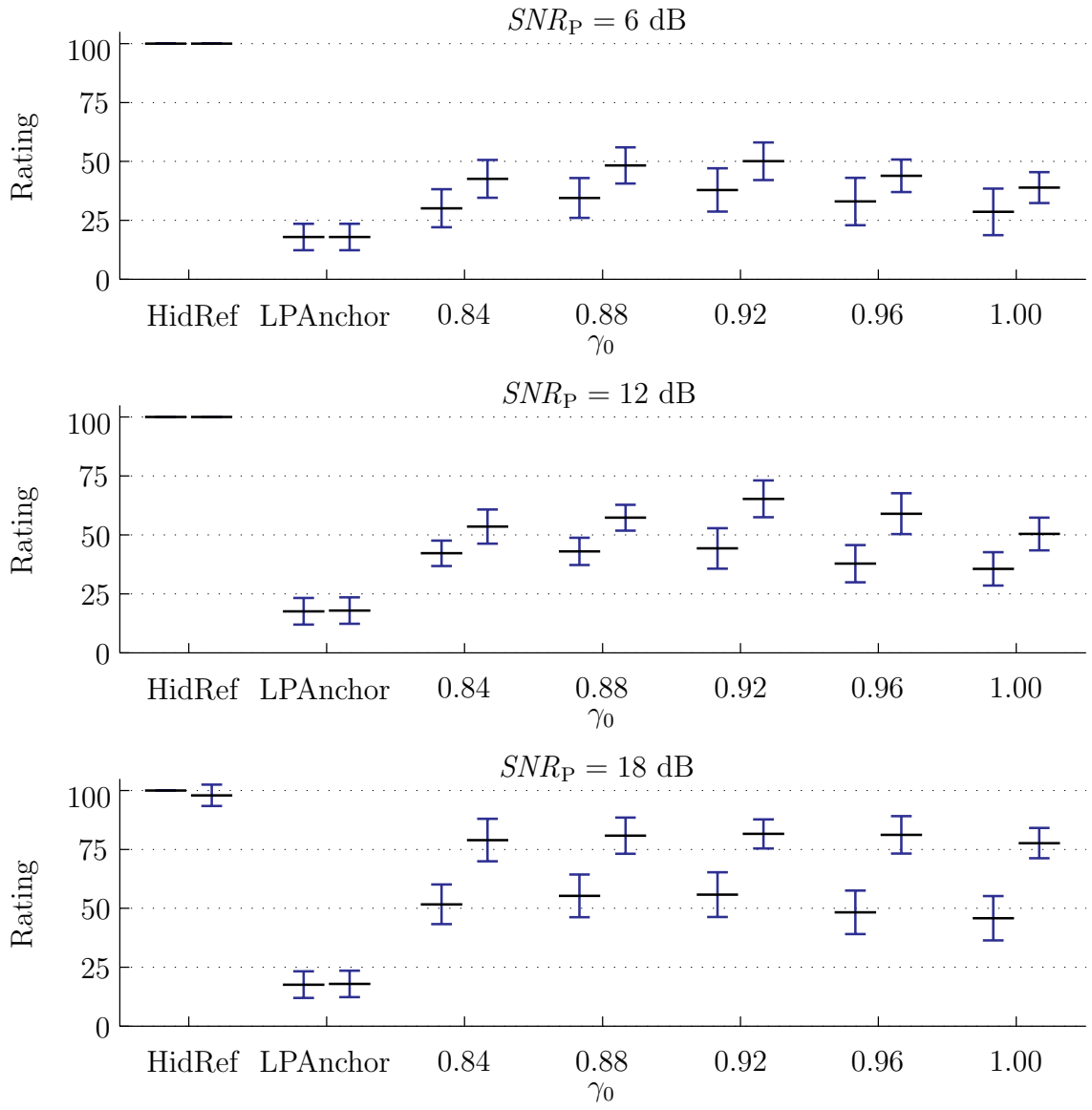

Figure 7.4: The means and 95% confidence intervals of the answers of all listeners in the test investigating perceptual domain transformation (after post-screening, $n = 10$, left = male speech, right = Suzanne Vega).

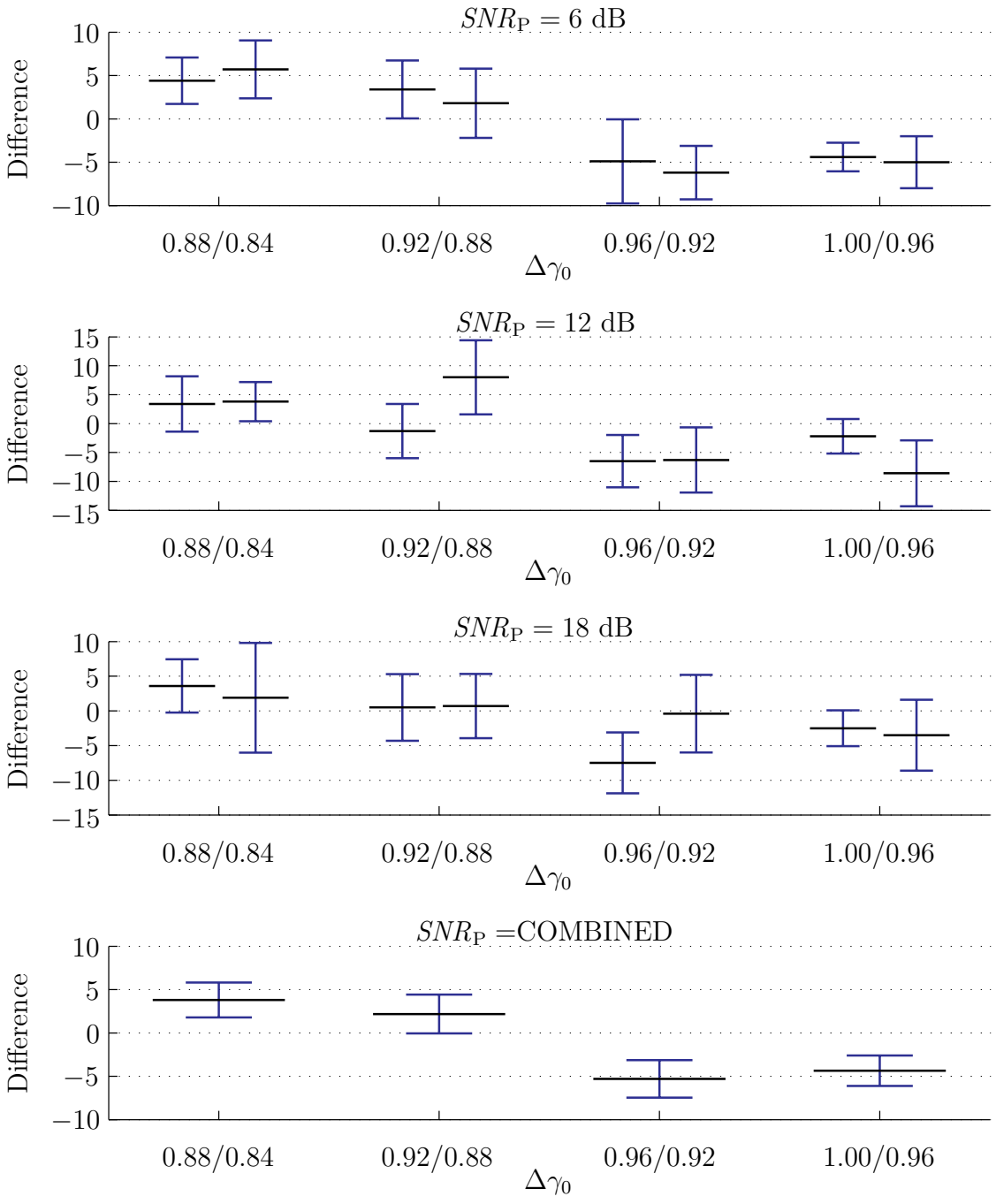

Figure 7.5: The mean values and 95% confidence intervals of the differences of the ratings given to adjacent γ_0 values in the test investigating envelope smoothing (after post-screening, $n = 10$, left = male speech, right = Suzanne Vega).

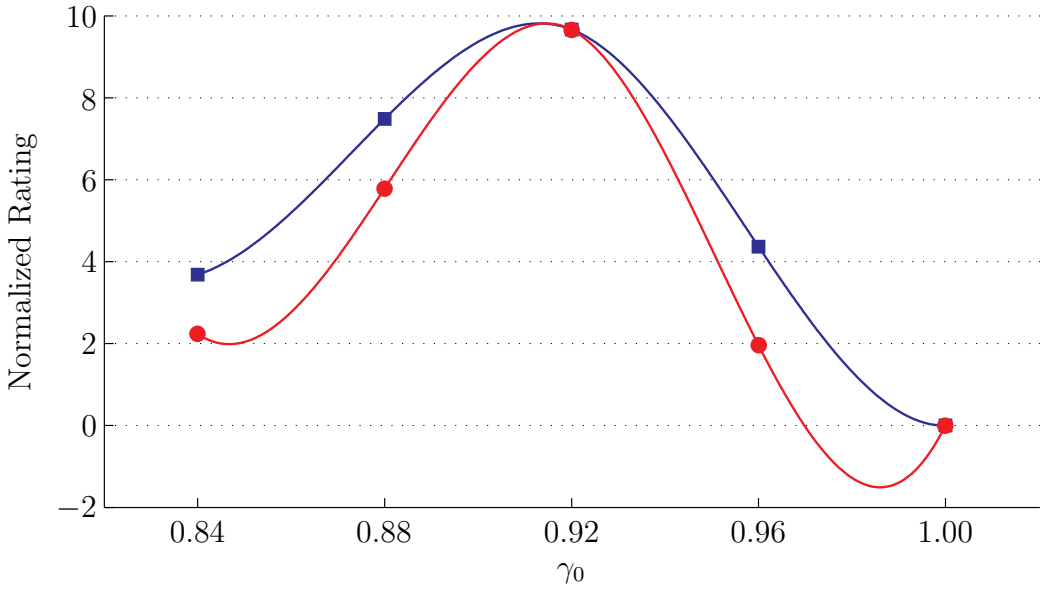

Figure 7.6: A γ_0 rating scale constructed from the difference values (squares) as well as using BTL (circles) in the test investigating envelope smoothing (after post-screening, $n = 10$, all items combined). The points are connected with cubic spline interpolation.

7.2 Formant Enhancement

In formant enhancement, quantization noise is hidden by boosting strong and diminishing weak parts of the spectrum of a signal frame in encoding. This can be easily done by multiplying the frame with its own spectral envelope which is modified similarly than is often done in the perceptual weighting filters in audio coders belonging to the code excited linear prediction (CELP) family (see e.g. [55] and [56] for examples). In this second part of the chapter, the objective was to find optimal values for the formant enhancement parameters in terms of perceived quality of quantized signals.

7.2.1 Background

Let us begin with extracting a spectral envelope $W(z)$ of an input frame $X(z)$ as in Section 7.1.1. However, this time the all-pole transfer function is converted to the form

$$F(z) = \frac{W(z/\gamma_\mathrm{z})}{W(z/\gamma_\mathrm{p})} = \frac{1 + a_1\gamma_\mathrm{z}z^{-1} + a_2\gamma_\mathrm{z}^2z^{-2} + \ldots + a_k\gamma_\mathrm{z}^kz^{-k}}{1 + a_1\gamma z_p^{-1} + a_2\gamma^2 z_p^{-2} + \ldots + a_k\gamma^k z_p^{-k}},$$

where $\gamma_\mathrm{z} \leq 1$ and $\gamma_\mathrm{p} \leq 1$ are constants. By adjusting the relation of γ_z and γ_p, the positions of the poles and zeros, and hence the shape of the spectrum, can be controlled. Figure 7.7 shows an example of a modified spectral envelope $F(z)$ with $\gamma_\mathrm{z} = 0.7$ and $\gamma_\mathrm{p} = 0.9$. The modified envelope inherits some of the waving of the original envelope, but, unlike in bandwidth expansion, is is not tilted.

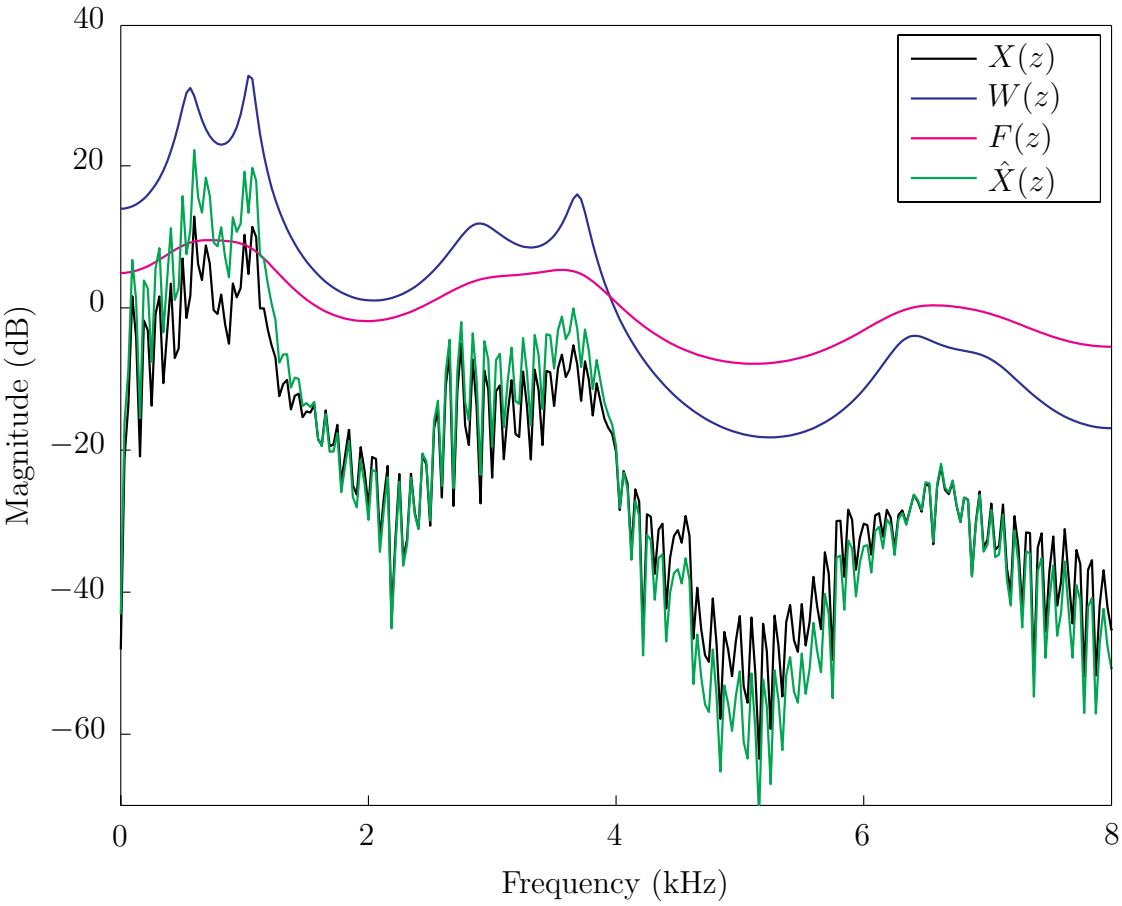

Figure 7.7: A 32 ms frame of a speech signal transformed from a frequency domain representation $X(z)$ into a perceptual domain counterpart $\hat{X}(z)$ with the spectral envelope $W(z)$. The frame is pre-processed by multiplying it with its modified spectral envelope $F(z)$.

As shown in Figure 7.7, an input frame $X(z)$ can be pre-processed by multiplying it with its modified spectral envelope $F(z)$ before a perceptual domain transformation. After quantization and back in the frequency domain, the output frame thus becomes

$$Y(z) = X(z)F(z) + V(z)W(z/\gamma_0).$$

7.2.2 Methods

In the preparation of this listening test, the audio signals were first processed with formant enhancement using different combinations of γ_z and γ_p. Quantization noise, modelled again with white noise in the perceptual domain, was then added. Finally, the participants of the listening test were asked to compare and rate the perceived quality of the conditions and hence evaluate the impact of γ_z and γ_p.

As summarized in Table 7.2, a wide range of five γ_z values was selected for the test while γ_p was kept constant. The selection of the latter was not critical per se

because the outcome specifically depends on the relation of these to parameters. The source signals were selected for the same reasons as in the test investigating perceptual domain transformation, the only difference being that the male speech excerpt was changed to one of a slightly higher technical quality.

Table 7.2: The parameters of the test investigating the perceived effect of formant enhancement.

γ_z	0.55	0.70	0.80	0.85	1.00
γ_p	0.90				
γ_0	0.91				
SNR_P (dB)	6		12		18
Signal	Male speech (German)			Suzanne Vega: Tom's Diner	
Frame length (ms)	32				
Sampling frequency (kHz)	16				
Bit depth	16				
Number of listeners	8				
Test method	Modified MUSHRA				

Perceptual SNR was also varied in the conditions because our informal tests as well as general reasoning suggested that it probably has a significant effect on γ_z preferences. Perceptual SNR was kept constant and loudness normalized as in the perceptual domain transformation test. The signals were generated with the Matlab script illustrated in Figure 7.8.

Modified MUSHRA was a natural choice for the test method as we expected that listeners might perceive some conditions superior to the reference signal and the method had to support that occasion. As the conditions were created to be rather easy to distinguish, four experienced and four naive listeners were invited to take the test. The complete written test instructions are included in Appendix B.

7.2.3 Results

The distribution of the listening test results before post-screening is shown in a boxplot in Figure 7.9 and all individual answers can be found in Appendix A, Figure A.7. As the hidden reference was missed several times, anon06 was discarded in post-screening as too unreliable. In addition, the Suzanne Vega item with $SNR_P = \infty$ of anon08 was rejected because the hidden reference was severely off. The extreme outliers were also removed from any item as usual.

The means and 95% confidence intervals of the answers in Figure 7.10 indicate that the γ_z preference is dependent of perceptual SNR. To see this more clearly,

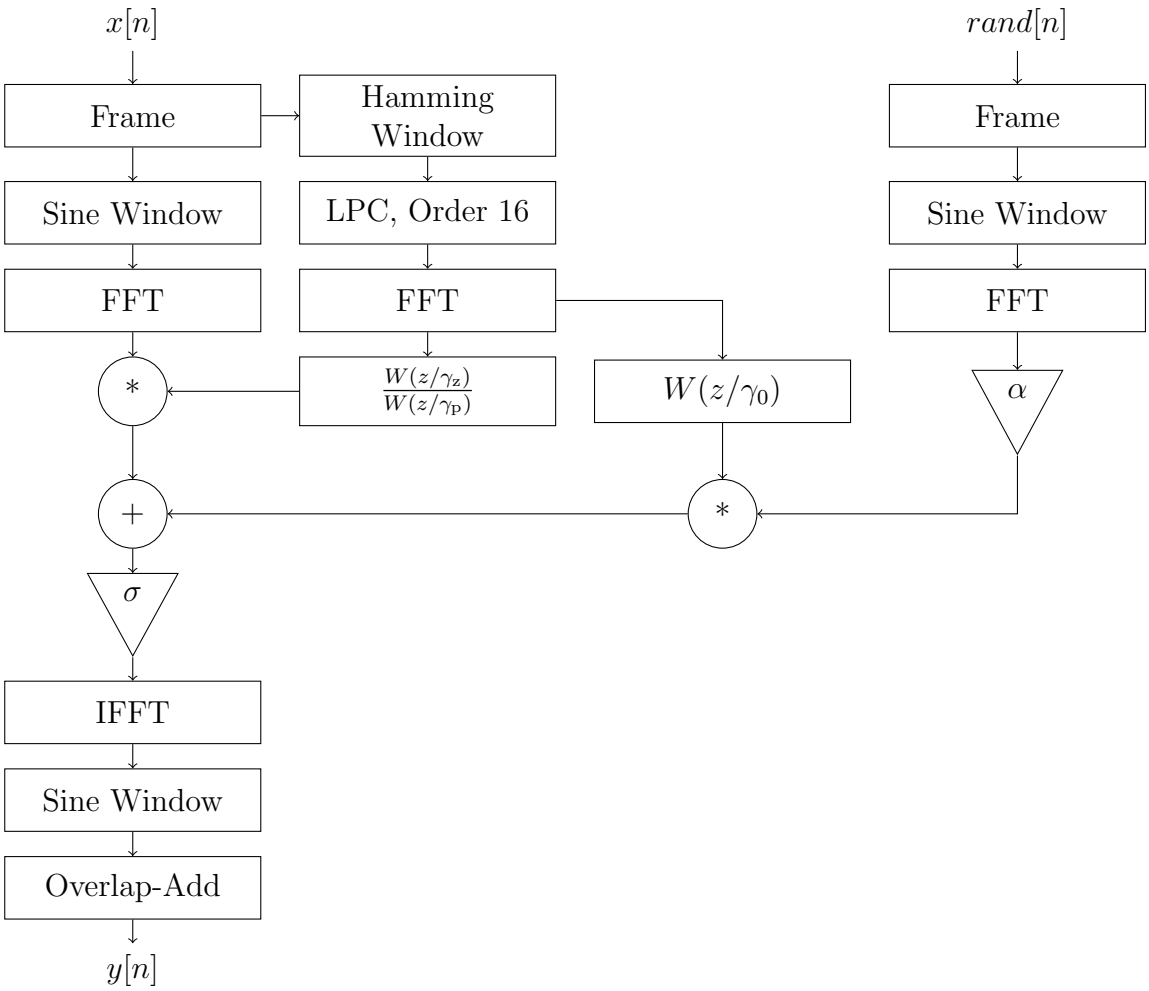

Figure 7.8: A block diagram of the Matlab script for generating the conditions in the formant enhancement test.

a rating scale constructed with BTL is plotted for each perceptual SNR in Figure 7.11. The points are connected with cubic spline interpolation that has a tendency to overshooting, which produces artificial waving in the preference function. The maxima of the red (Suzanne Vega) and blue (male speech) curves are located at $\{0.590, 0.643, 0.859, 0.861\}$ and $\{0.550, 0.634, 0.856, 0.900\}$, respectively. Hence, a linear regression model (plotted in Figure 7.12) for estimating the optimal γ_z as a function of SNR_P could be expressed as

$$\gamma_z(SNR_P) = 0.019 SNR_P + 0.45, \tag{7.1}$$

when $\gamma_p = 0.9$. The model fits the data reasonably well since $R^2 = 0.90$. In addition, both regression coefficients are significant with the level $\alpha = 0.01$.

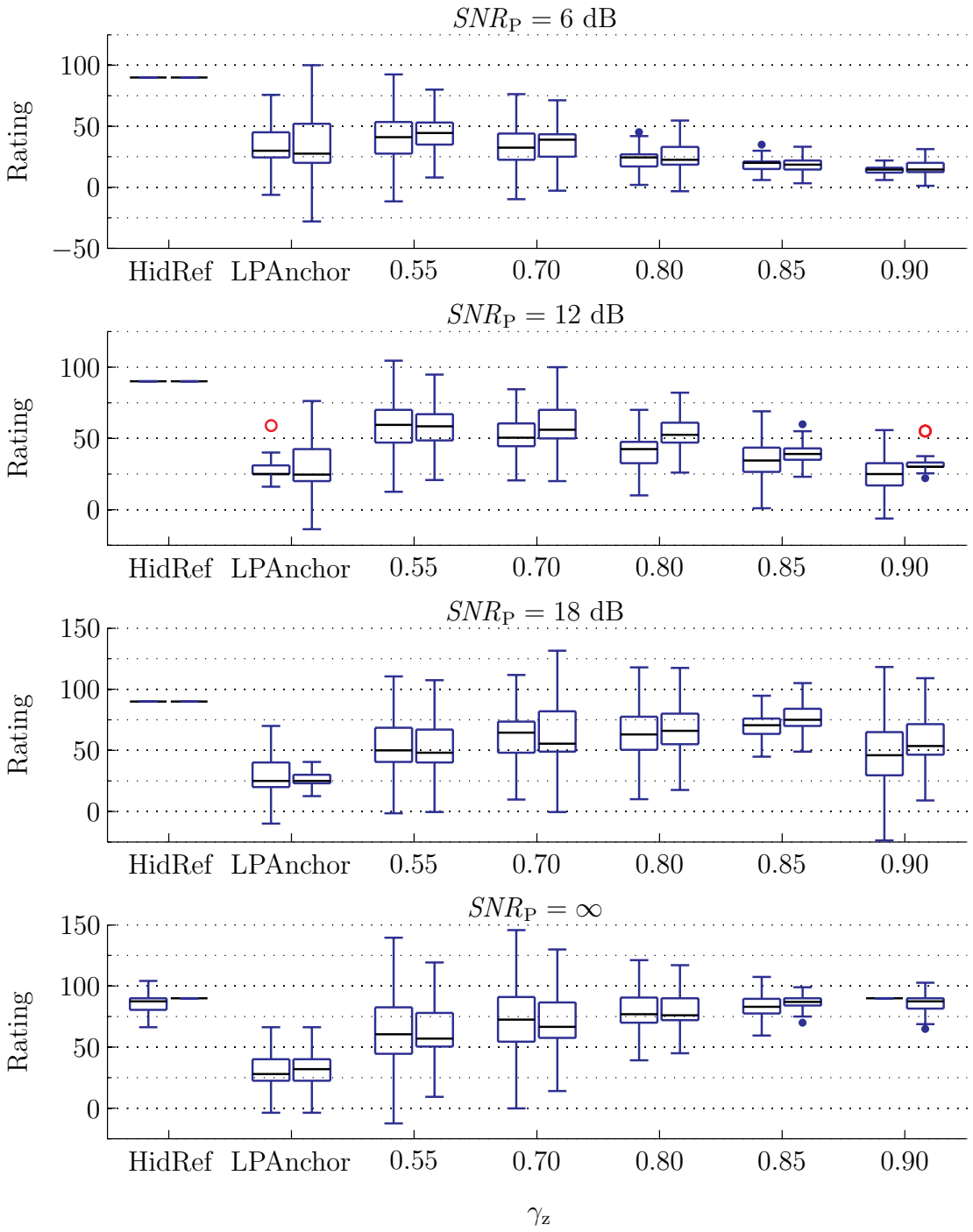

Figure 7.9: A Boxplot of the answers of all listeners in the formant enhancement test (before post-screening, $n = 8$, left = Suzanne Vega, right = male speech). Note the varying y-axis limits.

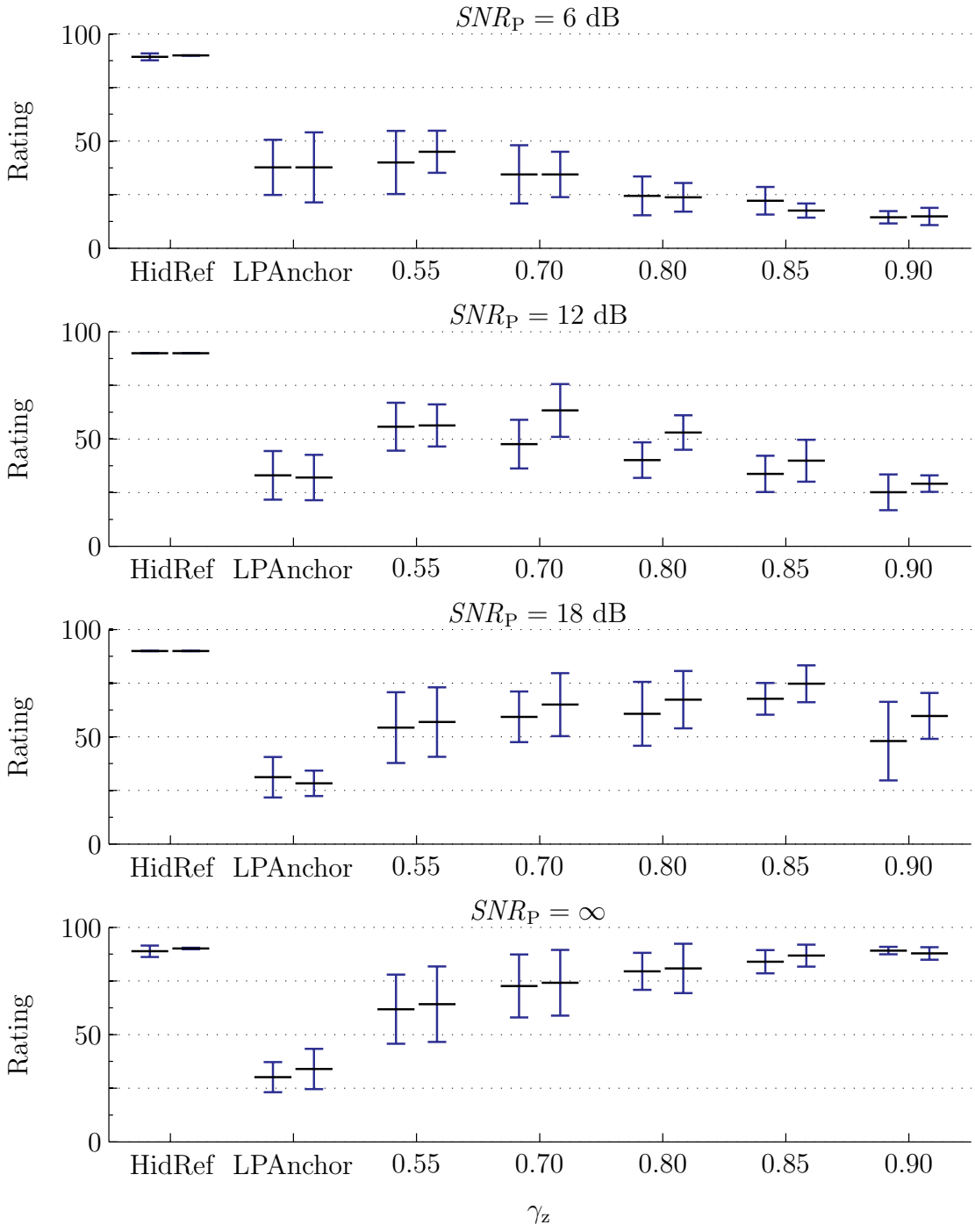

Figure 7.10: The means and 95% confidence intervals of the answers of all listeners in the formant enhancement test (after post-screening, $n = 7$, left = Suzanne Vega, right = male speech).

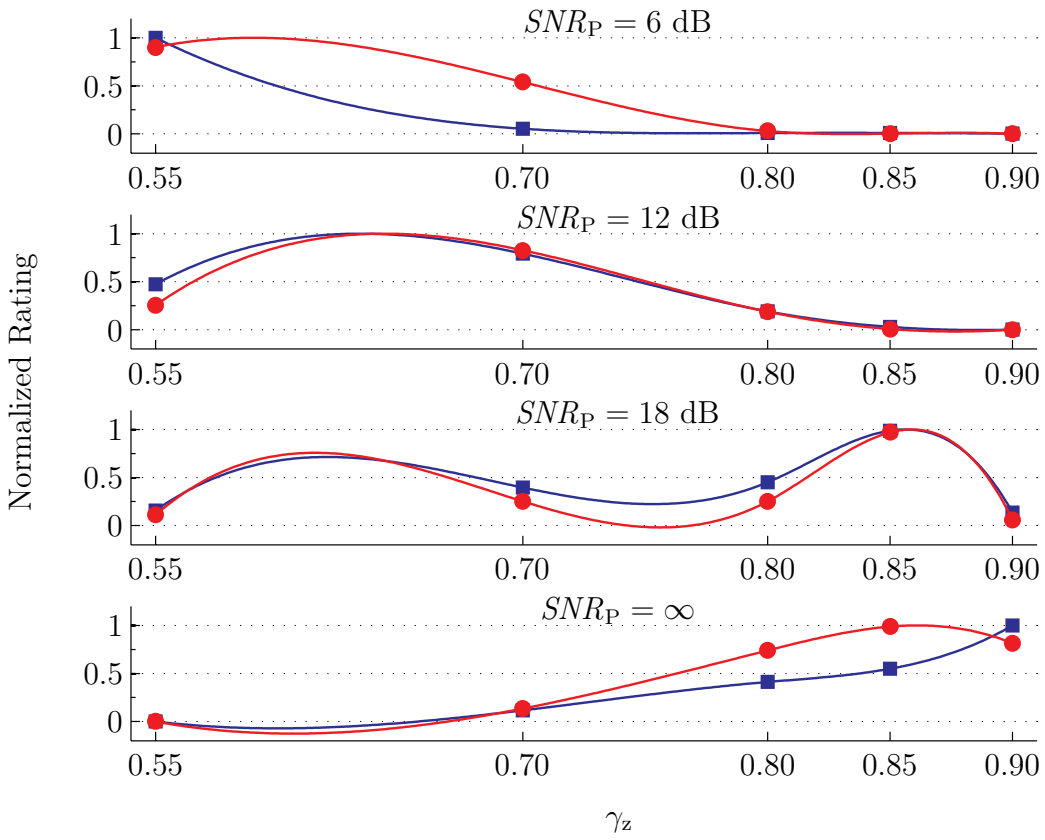

Figure 7.11: A rating scale derived with BTL for each perceptual SNR in the formant enhancement test (after post-screening, $n = 7$, blue squares = Suzanne Vega, red circles = male speech). The points are connected with cubic spline interpolation.

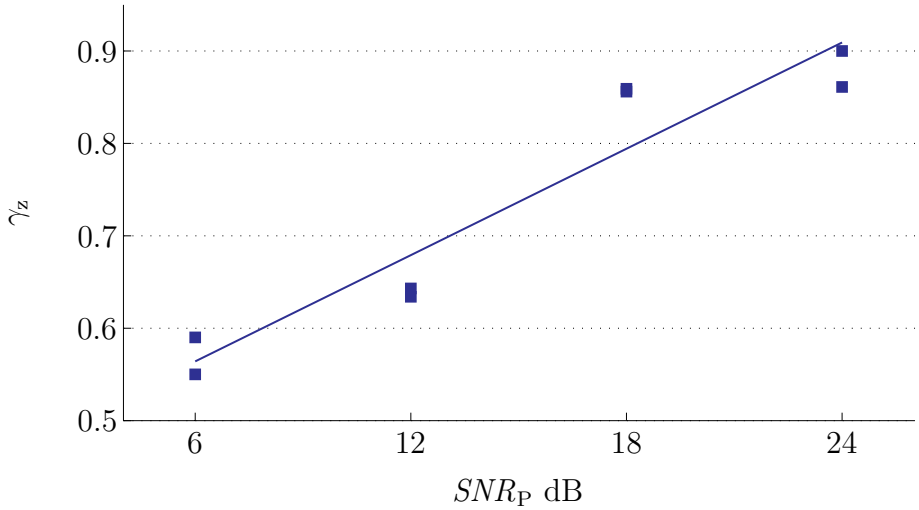

Figure 7.12: A linear regression model for estimating the optimal γ_z as a function of perceptual signal-to-noise ratio in the formant enhancement technique. The perceptual SNR $SNR_P = \infty$ included in the listening test was substituted for $SNR_P = 24$ dB as it is roughly the limit where noise becomes inaudible.

7.3 Conclusions

Our results indicate that the differences in the perceived impact of different γ_0 values in the perceptual domain transformation are small with speech and in case of music with flatter spectra differences would probably be mostly nonexistent. Hence we conclude that this parameter is not particularly critical, but it is, however, recommended to use the value $\gamma_0 = 0.91$ for optimal results, at least in frequency domain speech coders. Voice type or perceptual SNR do not seem to have a significant effect on the optimal value.

According to our test results and listener feedback, formant enhancement is an efficient way to hide quantization noise, but boosting formants tends to render the overall tone unnatural and boomy. This trade-off explains why heavy formant boosting is apparently preferred when the noise is prominent, and vice versa. Voice type does not seem to affect the parameter selection. Furthermore, the method is likely to be much less effective with musical content. If perceptual SNR can be estimated during the encoding process, Equation (7.1) can be used for controlling the parameters of formant enhancement. Regarding real world applications, it would be of interest to study further whether the resulting boominess could be tempered without sacrificing the improvement in the perceived annoyance of the noise.

Chapter 8

Summary

The purpose of this work was to investigate the annoyance of a few audio coding artifacts and some methods for minimizing it. The studied artifacts and methods were already known at Fraunhofer IIS, the commissioner of the thesis, but listening tests were needed to acquire quantitative knowledge to help in tuning USAC compliant encoders.

The effect of time-variance in the perceived annoyance of coding artifacts was investigated in Chapter 5. In the listening tests, the level of critical-bandwidth noise, single harmonics, and pairs of harmonics was varied at different speeds and the resulting annoyance was compared with that caused by stationary noise. The results suggest that moderate-speed variation at approximately 2 to 4 Hz is perceived as the most annoying. It seems that very slow variation does not draw the listener's attention as strongly and, on the other hand, faster variation is seemingly heard more as a continuous, though modulated, signal. Further tests should be organized to verify these results and it would probably be fruitful to extend the focus to other types of artifacts too.

Ghost pitch refers to an extraneous pitch sensation that might emerge when harmonic bandwidth extension is used to generate harmonics above the selected crossover frequency. It is the incompleteness of the generated harmonic patch which might manifest itself as a perceived virtual pitch. The strength of that pitch as a function of fundamental and crossover frequencies was examined in Chapter 6 with synthetic tones. The listening tests provided vague indications that lower fundamental and crossover frequencies might generally lead to stronger ghost pitch sensations. However, further research with more participants, more parameter values, and possibly more natural tones is clearly needed.

Performing quantization in the perceptual domain instead of the ordinary frequency domain is advantageous as the resulting noise becomes favorably shaped and its annoyance is consequently decreased. The transformation can be further enhanced in terms of noise hiding by smoothing the spectral envelope used in the process. In the first part of Chapter 7, the objective was to find the optimal value for the parameter controlling the intensity of the smoothing. According to the listening tests, the optimal value is approximately 0.91 for speech signals. Music or mixed

signals were not included in the tests as the effect of smoothing was found to be too subtle to get reliable results with the limited number of listeners available.

In order to further hide quantization noise in speech signals, strong parts of a spectrum can be boosted and weak parts diminished in encoding by multiplying the signal frame with its modified spectral envelope. The results of the listening test presented in the second part of Chapter 7 suggest that this formant enhancement technique is efficient in hiding quantization noise, but overuse tends to render the overall tone of the audio boomy. A model for selecting the spectral modification parameters as a function of perceived signal-to-noise ratio is presented in the text. An interesting topic for further research would be to study whether the boominess could be decreased without sacrificing the ability to hide noise.

Many of our tests were somewhat unusual in a sense that the listeners were asked to make atypical comparisons and highly subjective judgments. As a result, most participants found it difficult to rate the conditions and thus the answers varied greatly. However, our experiences and results suggest that even in more subjective tests opinions begin to converge when the number of listeners becomes large enough.

References

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[2] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *ICASSP 2009*, 4 2009.

[3] K. A. S. Immink, "Any song, anytime, anywhere," *J. Audio Eng. Soc*, vol. 58, no. 1/2, pp. 73–79, 2010.

[4] A. Z. Dodd, *The Essential Guide to Telecommunications*. Upper Saddle River, NJ, USA: Pearson Education, 4th ed., 2005.

[5] A. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[6] B. Juang and T. Chen, "The past, present, and future of speech processing," *Signal Processing Magazine, IEEE*, vol. 15, no. 3, pp. 24–48, 1998.

[7] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, vol. 10, pp. 937–940, IEEE, 1985.

[8] A. Ramo, "Voice quality evaluation of various codecs," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4662 –4665, 2010.

[9] IEC 60908:1999-02, *Audio Recording – Compact Disc Digital Audio System*. International Electrotechnical Commission, Geneva, Switzerland, 1999.

[10] ISO/IEC 11172-3:1993, *Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s – Part 3: Audio*. International Organization for Standardization, 1993.

[11] A. Pras, R. Zimmerman, D. Levitin, and C. Guastavino, "Subjective evaluation of MP3 compression for different musical genres," in *Audio Engineering Society Convention 127*, 10 2009.

[12] ISO/IEC 13818-7:1997, *Information Technology – Generic Coding of Moving Pictures and Associated Audio Information – Part 7: Advanced Audio Coding (AAC)*. International Organization for Standardization, 1997.

[13] ETSI: TS 102 366 V1.2.1, *Digital Audio Compression (AC-3, Enhanced AC-3) Standard*. European Broadcasting Union, France, 2008.

[14] J. Ribas-Cordeba, "Windows Media 9 series – a platform to deliver compressed audio and video for internet and broadcast applications," *EBU Technical Review*, 2003.

[15] Xiph.Org Foundation, *Vorbis I specification*. August 11, 2011. Retrieved January 14, 2012.

[16] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[17] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill, 2nd ed., 2001.

[18] M. Frigo and S. Johnson, "The design and implementation of FFTW3," *Proceedings of the IEEE*, vol. 93, no. 2, pp. 216–231, 2005.

[19] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[20] D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers*. Wiley, 2010.

[21] R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*. McGraw-Hill, 4th ed., 1998.

[22] F. Nagel, T. Sporer, and P. Sedlmeier, "Toward a statistically well-grounded evaluation of listening tests - avoiding pitfalls, misuse, and misconceptions," in *Audio Engineering Society Convention 128*, 5 2010.

[23] T. Sporer, J. Liebetrau, and S. Schneider, "Statistics of MUSHRA revisited," in *Audio Engineering Society Convention 127*, 10 2009.

[24] F. Wickelmaier and C. Schmid, "A Matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods, Instruments, & Computers*, vol. 36, pp. 29–40, 2004.

[25] F. Wickelmaier, N. Umbach, K. Sering, and S. Choisel, "Comparing three methods for sound quality evaluation with respect to speed and accuracy," in *126th Convention of the Audio Engineering Society, Munich,Germany, May 7–10*, 2009.

[26] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, no. 1, pp. 47–65, 1940.

[27] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer-Verlag, 2nd ed., 1999.

[28] T. A. Nousaine, "Can you trust your ears?," in *Audio Engineering Society Convention 91*, 10 1991.

[29] ISO 226:2003, *Acoustics – Normal Equal-Loudness Level Contours*. International Organization for Standardization, 2003.

[30] S. H. Nielsen and E. Skovenborg, "Evaluation of different loudness models with music and speech material," in *Audio Engineering Society Convention 117*, 10 2004.

[31] T. D. Rossing, *Science of Sound*. Addison-Wesley, 2nd ed., 1990.

[32] ITU-T G.722.2:2003, *Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB)*. International Telecommunication Union, 2003.

[33] E. Zwicker, "Ueber die Lautheit von ungedrosselten und gedrosselten Schallen," *Acustica*, vol. 13, pp. 194–211, 1963.

[34] N. Iwakami, T. Moriya, and S. Miki, "High-quality audio-coding at less than 64 kbit/s by using transform-domain weighted interleave vector quantization (TwinVQ)," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 5, pp. 3095 –3098 vol.5, 5 1995.

[35] E. Ifeachor and B. Jervis, *Digital Signal Processing*. Essex, England: Pearson Education, 2nd ed., 2002.

[36] Recommendation ITU-R BS.1534-1, *Methods for the subjective assessment of intermediate quality level of coding systems*. International Telecommunication Union, 2003.

[37] G. A. Gescheider, *Psychophysics: the fundamentals*. Routledge, 3rd ed., 1997.

[38] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acustica*, vol. 30, pp. 201–213, 1974.

[39] J. Bradley, "Annoyance caused by constant-amplitude and amplitude-modulated sounds containing rumble," *Noise Control Engineering Journal*, vol. 42, no. 6, pp. 203–208, 1994.

[40] K. Waye and E. Öhrström, "Psycho-acoustic characters of relevance for annoyance of wind turbine noise," *Journal of sound and vibration*, vol. 250, no. 1, pp. 65–73, 2002.

[41] A. Bockstael, L. Dekoninck, B. De Coensel, D. Oldoni, A. Can, and D. Botteldooren, "Wind turbine noise: annoyance and alternative exposure indicators," in *Proceedings of Forum Acusticum 2011*, 2011.

[42] D. Siponen, "Noise annoyance of wind turbines," *VTT research report VTTR-00951-11*, 2011.

[43] G. Pinero, A. Gonzalez, and M. De Diego, "Time-frequency analysis applied to psychoacoustic evaluation of car engine noise quality," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 4, pp. IV–IV, IEEE, 1993.

[44] A. L. Hastings, *Sound quality of diesel engines*. PhD thesis, Purdue University, 2004.

[45] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*, 4 2002.

[46] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 145–148, IEEE, 2009.

[47] D. M. Howard and J. Angus, *Acoustics and Psychoacoustics*. Ofxord, UK: Focal Press, 2nd ed., 2000.

[48] A. Seebeck, "Ueber die Sirene," *Annalen der Physik*, vol. 136, no. 12, pp. 449–481, 1843.

[49] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1496–1516, 1973.

[50] R. J. Ritsma, "Frequencies dominant in the perception of the pitch of complex sounds," *The Journal of the Acoustical Society of America*, vol. 42, no. 1, pp. 191–198, 1967.

[51] R. Plomp, "Pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 41, no. 6, pp. 1526–1533, 1967.

[52] B. C. J. Moore, B. R. Glasberg, and R. W. Peters, "Relative dominance of individual partials in determining the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 77, no. 5, pp. 1853–1860, 1985.

[53] R. D. Patterson and F. L. Wightman, "Residue pitch as a function of component spacing," *The Journal of the Acoustical Society of America*, vol. 59, no. 6, pp. 1450–1459, 1976.

[54] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, pp. 309 – 321, 1 1975.

[55] Z. Wen, Z. Tao, Z. Liang, and Z. Hai, "Performance analysis and evaluation of AVS-M audio coding," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, pp. 31 –36, 11 2010.

[56] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, pp. 620 – 636, 11 2002.

Appendix A

Complete Listening Test Results

This appendix includes all individual responses in each listening test discussed in the main matter. To ensure good readability, axes or ticks are not labelled in every case.

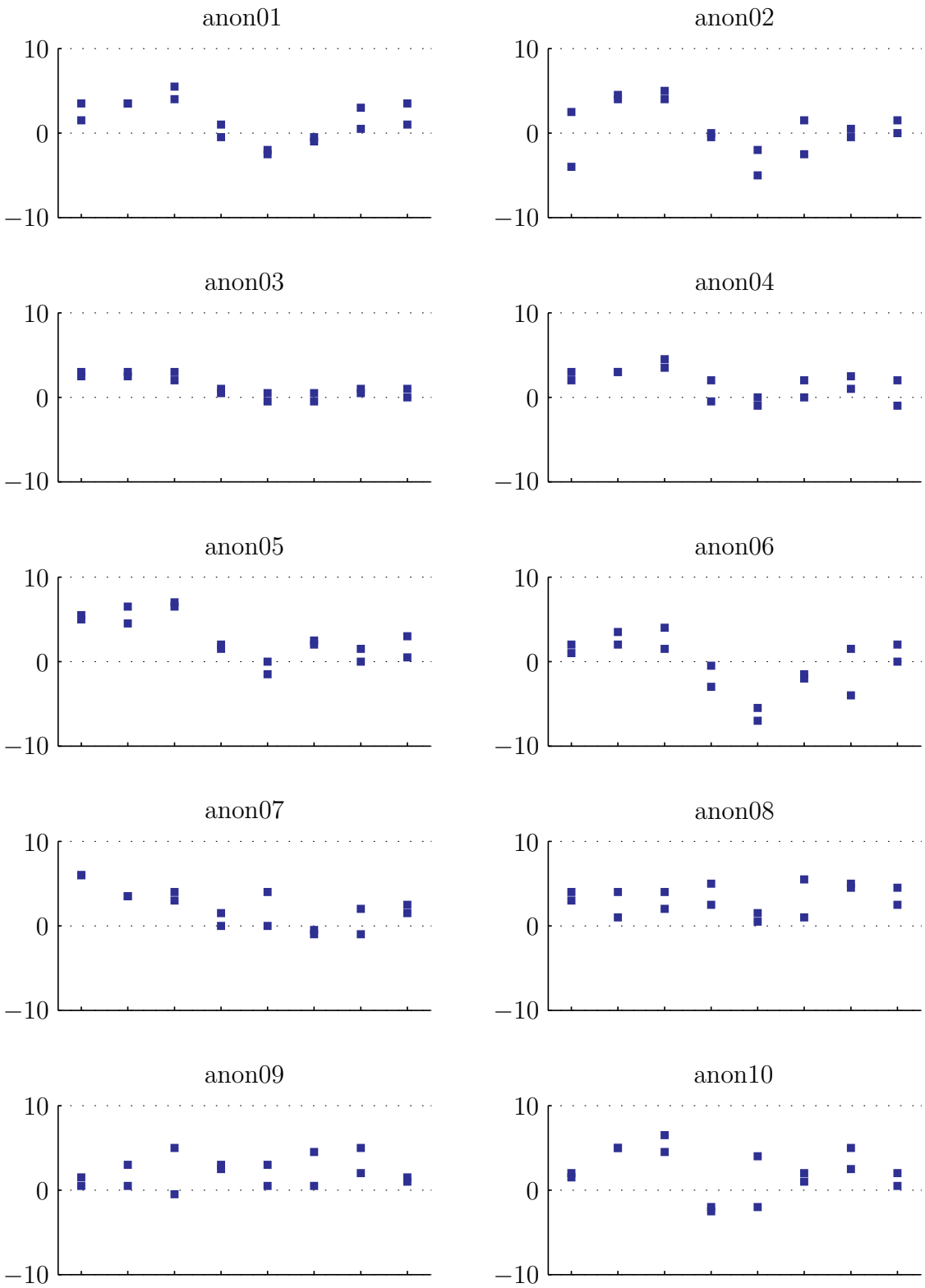

Figure A.1: Individual answers of all listeners in the test examining fluctuation of noise (Section 5.1). Refer to Figure 5.6 for x-axis tick labels. Each item was evaluated twice.

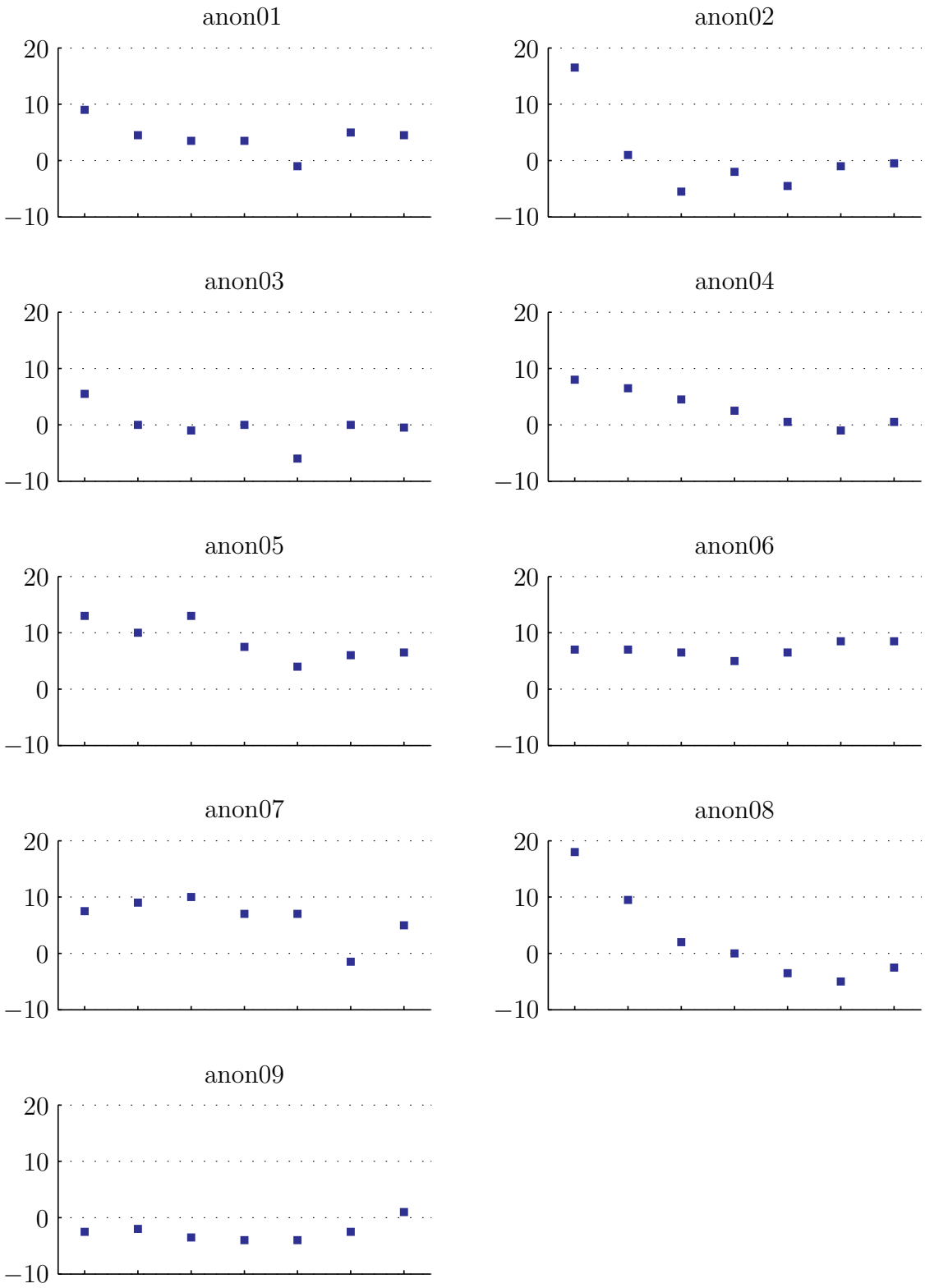

Figure A.2: Individual answers of all listeners in the test examining fluctuation of one harmonic (Section 5.2). Refer to Figure 5.10 for x-axis tick labels.

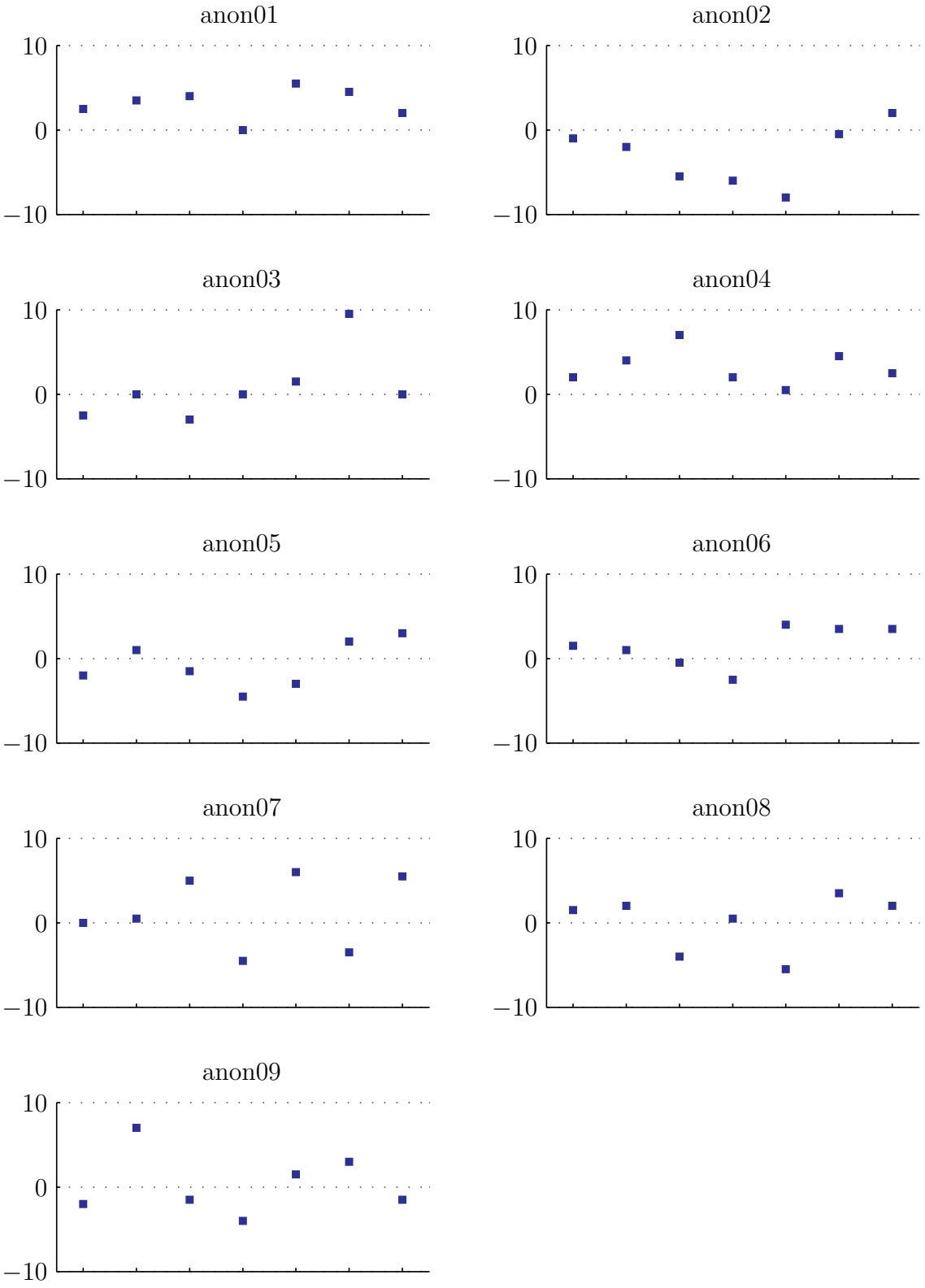

Figure A.3: Individual answers of all listeners in the test examining fluctuation of two harmonics (Section 5.2). Refer to Figure 5.12 for x-axis tick labels.

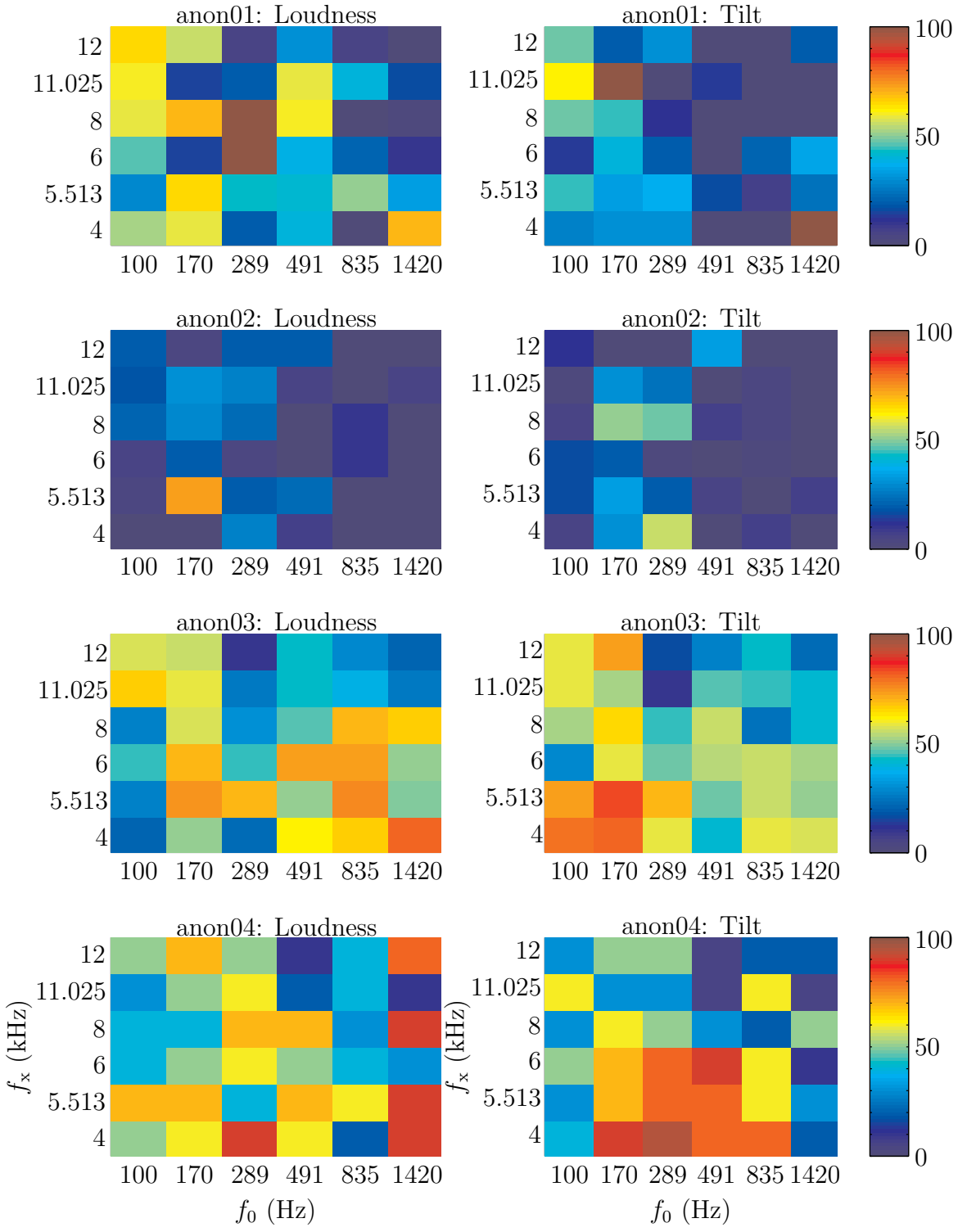

Figure A.4: Individual answers of listeners 1–4 in the ghost pitch test (Chapter 6).

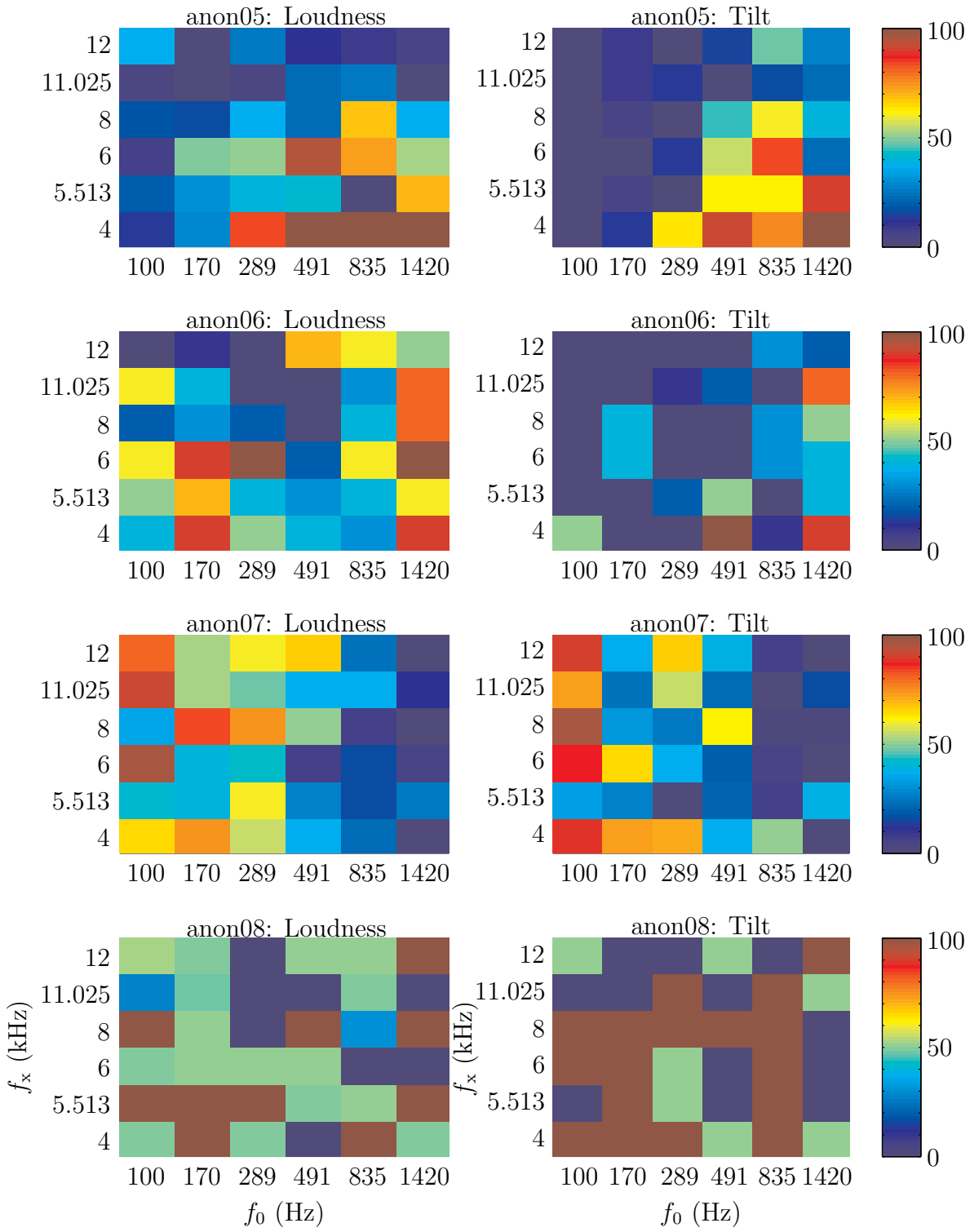

Figure A.5: Individual answers of listeners 5–8 in the ghost pitch test (Chapter 6).

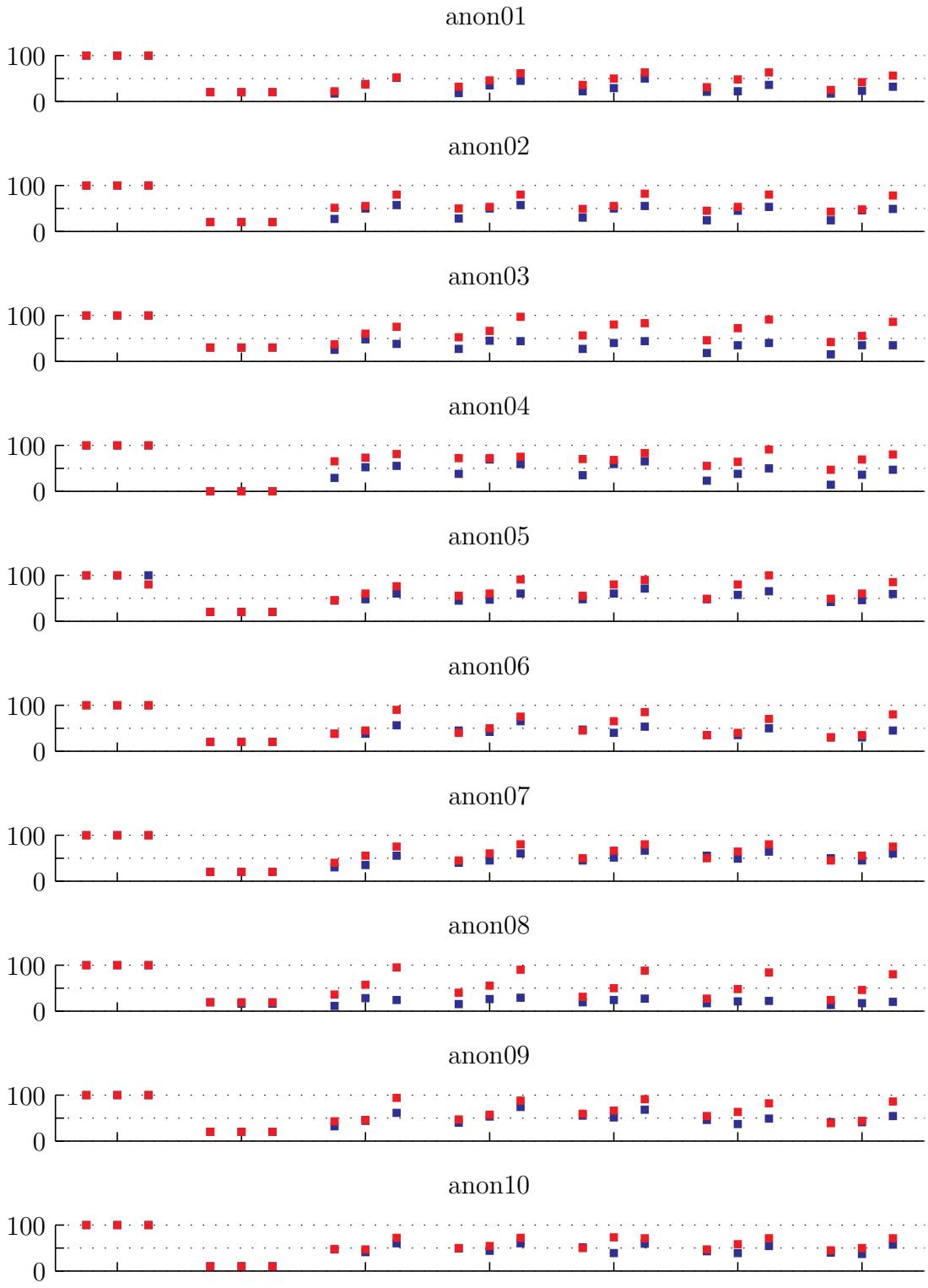

Figure A.6: Individual answers of all listeners in the perceptual domain transformation test (Section 7.1). Refer to Figure 7.3 for x-axis tick labels. Each triplet represents perceptual SNR values of 6, 12, and 18 dB, respectively. Male speech items are marked with blue and Suzanne Vega items with red.

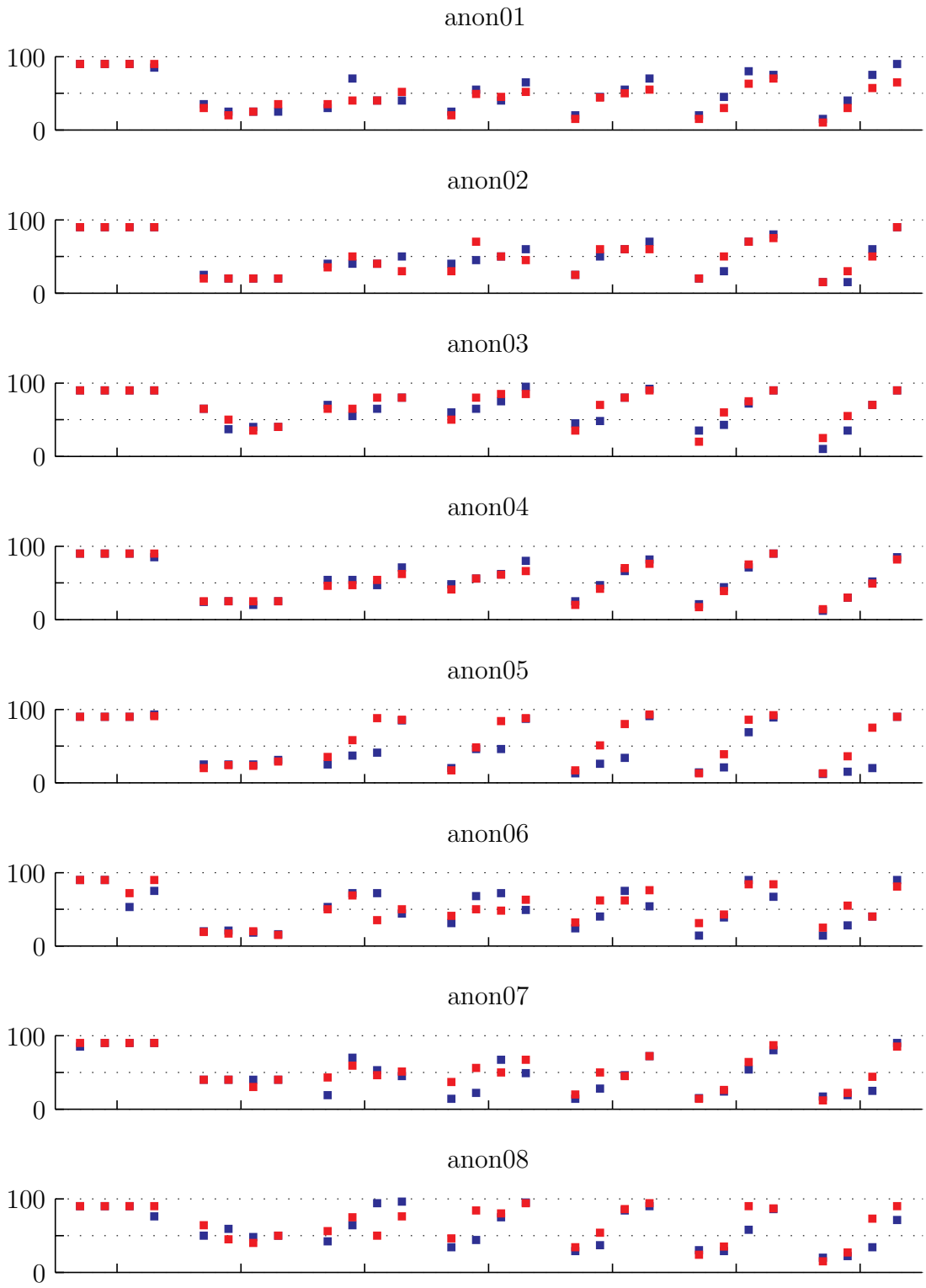

Figure A.7: Individual answers of all listeners in the formant enhancement test (Section 7.2). Refer to Figure 7.9 for x-axis tick labels. Each quartet represents perceptual SNR values of 6, 12, 18, and ∞ dB, respectively. Suzanne Vega items are marked with blue and male speech items with red.

Appendix B

Listening Test Instructions

OVERVIEW

The purpose of this test is to investigate how annoying it is if critical-band noise in a signal is constantly switching on and off instead of being stationary. The signals in this test are:

- Reference = pitchpipe with **fluctuating noise** (fixed)
- Condition = pitchpipe + **stationary noise** (noise level adjustable)

Your task is to **adjust the level of the stationary noise** (condition) **so that you find it equally annoying as the fluctuating noise** (reference).

SOFTWARE USAGE

The test software is to be used with **keyboard only** (an exception being the user name window where the mouse must be used to click OK...). All keyboard commands are shown in the UI window.

STARTING THE TEST

1) Open Matlab
2) Navigate to /Users/wavswitch/paunonli/var_noise
   (e.g. by typing cd /Users/wavswitch/paunonli/var_noise on the command window…)
3) Type Match(' var_noise ') to begin

Thanks for your time,

Lari (6781, paunonli@iis...)

Figure B.1: Written instructions for the test examining fluctuation of noise (Section 5.1).

OVERVIEW

The purpose of this test is to investigate how annoying it is if harmonics are 'blinking', that is, constantly switching on and off. The signals in this test are:

- Reference = pitchpipe with **one harmonic pulsating** (fixed)

  OR
- Reference = pitchpipe with **two harmonics switching on and off in turns** (fixed)

- Condition = pitchpipe + **stationary noise** (noise level adjustable)

Your task is to **adjust the level of the stationary noise** (condition) **so that you find it equally annoying as the blinking phenomenon** (reference). Yep, it truly is like comparing apples (strange behavior in harmonics) and oranges (stationary noise)... good luck!

SOFTWARE USAGE

The test software is to be used with **keyboard only** (an exception being the user name window where the mouse must be used to click OK...). All keyboard commands are shown in the UI window.

STARTING THE TEST

1) Open Matlab
2) Navigate to /Users/wavswitch/paunonli/varharmtest
   (e.g. by typing cd /Users/wavswitch/paunonli/varharmtest on the command window…)
3) Type Match('varharmtest') to begin

WHAT IS THE "RESET AUDIO" BUTTON USED FOR?

Well, earlier there have been some issues with the audio library used in this GUI. If you experience problems with audio, this button *might* work... Otherwise you can just ignore it.

Thanks for your time,

Lari (6781, paunonli@iis...)

Figure B.2: Written instructions for the test examining fluctuation of harmonics (Section 5.2).

INTRODUCTION

Harmonic Bandwidth Extension is a technology in which only lower frequencies are encoded explicitely and higher harmonics are later generated from the LF part in the decoding process. It has a curious side effect of sometimes creating pitch sensations that were not present in the original signal.

This test consists of synthetic waveforms that simulate the HBE technique. Your task is to evaluate the presence of these new pitch sensations called ghost pitches.

TEST PROCEDURES

The test procedure is simple: each item consists of one signal, and you are asked to **rate the loudness of the ghost pitch as compared to the most prominent pitch** (which is in most cases the fundamental pitch…). You can use the **whole scale continuously from 0 to 100** by keeping these "anchors" in mind:

- **If you hear only one pitch -> set the slider to 0**
- **If you hear two equally loud pitches -> set the slider to 100**
- **If you hear two pitches but one is stronger than another -> set the slider in between 0 and 100 accordingly** (i.e. according to how you perceive the ratio of their loudnesses… rating of 50 would mean that the loudness of the softer tone is 50% of the louder one etc.)

There are quite a few items, but they should be fast to rate because this test does not require highly analytical listening. Just listen to the condition for a while, set the slider and move on – takes perhaps about 15 seconds or so.

SOFTWARE USAGE

The test uses the MUSHRA mode of ListeningTestGUI. However, because the test procedure is not really MUSHRA, **there is no reference signal** (nor hidden reference or lowpass anchor). You are kindly asked to just listen to the condition 1 and set the slider from 0 to 100 as described above. Please **ignore the written scale** ("excellent, good, fair…") next to the slider!

WHERE AND HOW CAN I TAKE THE TEST?

It is available in two listening rooms: **Hendrix** and **Beethoven**.

To start the test:

1) Start ListeningTestGUI (available on the desktop or on the dock)
2) Click "open config -file"
3) Navigate to /wavswitch/paunonli/ghosttest/ghost1.ltg (Hendrix) or /wavswitch/users/paunonli/ghosttest/ghost1.ltg (Beethoven)

Thanks for your time,

Lari (6781, paunonli@iis...)

Figure B.3: Written instructions for the ghost pitch test (Chapter 6).

INTRODUCTION

Signals are often quantized in the perceptual domain instead of the normal MDCT domain in order to minimize the annoyance of the resulting noise. In the transformation, an estimate of the masking curve is required for each frame, for which a spectral envelope might be used. To enhance the masking of quantization noise, the envelope can be smoothed by modifying its transfer function slightly. The goal of this listening test is to find the optimal parameters for the smoothing.

TEST PROCEDURES

The test method is MUSHRA with normal hidden reference and lowpass filtered anchors. You are kindly asked to rate the conditions in terms of the annoyance of quantization noise. The reference signal is noiseless.

WHERE AND HOW CAN I TAKE THE TEST?

It is available in **Hendrix** and uses Wavswitch.

To start the test:

1 Go to folder /wavswitch/paunonli/envelope/
2 Run prac-envelope.sh for warmup rounds
3 Run envelope.sh for the actual test

Thank you for your time,

Lari (6781, paunonli@iis...)

Figure B.4: Written instructions for the perceptual domain transformation test (Section 7.1).

INTRODUCTION

The purpose of this listening test is to study the effects of a pre-processing method in which strong parts of the spectrum are boosted and valleys are diminished before quantization. Each item consists of a signal excerpt with a certain predetermined perceptual SNR and the processing is applied to conditions with different parameters.

TEST PROCEDURES

Please keep in mind that this test is all about your **personal preferences and subjective quality perceptions**.

The procedure is a slightly modified version of MUSHRA. The rating scale is again 0…100 but there are two fixed anchors: the low anchor (at 10) and the high anchor (at 90). The **conditions should be rated against these two fixed anchors and it is also possible to rate them above the high or below the low anchor** (i.e. the whole range 0…100 may be used freely). In other words, it is possible that you think that a condition sounds better than the high anchor.

Each item includes **a duplicate of the high anchor which should be rated to 90** (just like the hidden reference in MUSHRA but it is to be rated to 90, not 100).

WHERE AND HOW CAN I TAKE THE TEST?

It is available in **Hendrix**.

To start the test:

1 Go to folder /wavswitch/paunonli/gvalleytest/
2 Start the ListeningTestGUI version **located in the above mentioned folder***
3 Click "open config -file"
4 Navigate to /wavswitch/paunonli/gvalleytest/gtestvalley.ltg

* This test procedure won't work with the basic version of ListeningTestGUI that is found on the desktop

Thanks for your time,

Lari (6781, paunonli@iis...)

Figure B.5: Written instructions for the formant enhancement test (Section 7.2).