CHALMERS
UNIVERSITY OF TECHNOLOGY

Perceptual Assessment of Loudspeaker Performance A Comparison of In-Situ and Binaural Evaluation Methods

In cooperation with Axis Communication AB

Master's thesis in Msc, Sound and vibration

Minghui Li, Xin Su

DEPARTMENT OF Applied Acoustics, Architecture and Civil Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

MASTER'S THESIS 2026

Perceptual Assessment of Loudspeaker Performance - A Comparison of In-Situ and Binaural Evaluation Methods

Minghui Li & Xin Su

Department of Civil and Environmental Engineering
Division of Applied Acoustics
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2026

Perceptual Assessment of Loudspeaker Performance - A Comparison of In-Situ and Binaural Evaluation Methods

Master's Thesis 2026

Department of Civil and Environmental Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-41296 Göteborg
Sweden

Reproservice / Department of Civil and Environmental Engineering
Göteborg, Sweden 2026

Perceptual Assessment of Loudspeaker Performance - A Comparison of In-Situ and Binaural Evaluation Methods
Master's Thesis in the Master's programme in Sound and Vibration
Minghui Li & Xin Su
Department of Civil and Environmental Engineering
Division of Applied Acoustics
Chalmers University of Technology

Abstract

In practical industrial scenarios, conducting in-situ listening evaluations is often difficult due to limitations in environment and experimental conditions. As a result, recorded audio is frequently used as an alternative for subjective loudspeaker evaluation, although its perceptual validity remains unclear.

This study investigates whether simplified binaural recordings can preserve subjective loudspeaker evaluation trends observed under direct listening conditions. Subjective listening experiments were conducted using three loudspeakers with different spectral characteristics and four representative audio signals under both in-situ listening and binaural playback conditions. Pairwise comparison and ranking analyses were used to evaluate perceptual consistency between the two listening conditions.

The results show that binaural playback was generally able to preserve the subjective evaluation trends observed in direct listening experiments, particularly for bass and music signals. Additional experiments under background noise conditions further demonstrated that perceptual discriminability between loudspeakers decreased as noise level increased.

Based on the validated binaural recordings, this study further proposed several spectral-perceptual models linking spectral characteristics with subjective perception. Model-guided spectral modification experiments showed that the predicted perceptual tendencies were generally consistent with listeners' evaluations.

Overall, the results suggest that simplified binaural recordings can preserve perceptually relevant loudspeaker information to a considerable extent and may provide a practical alternative to large-scale in-situ listening tests during loudspeaker evaluation and design.

Keywords: Binaural recording; Loudspeaker evaluation; Subjective listening test; Spectral-perceptual analysis; Timbre perception; Speech clarity; Auditory urgency perception

Contents

Acknowledgements

We would like to sincerely thank our supervisors, Jens Ahrens and Visishta Kanthi, for their guidance and support throughout this project. Their feedback and insights greatly helped shape this work.

We are especially grateful to Roland Sottek for his generosity, patience, and willingness to spend his valuable time guiding us throughout this research. His guidance not only improved the quality of this thesis, but also deeply influenced the way we approached research and problem solving.

We would also like to express our heartfelt thanks to the entire Core Tech Audio team at Axis Communications. Throughout this project, we received tremendous support, encouragement, and kindness from the team. The discussions, feedback, and technical help we received created an environment where this work could truly grow. Special thanks are extended to all members at Axis Communications who participated in the listening tests and contributed their time and effort to this study. Without their participation, this research would not have been possible.

In addition, we would like to thank each other for the collaboration, trust, and shared perseverance throughout this journey. Working side by side through both setbacks and progress made this experience far more meaningful than research alone.

Finally, as the Swedish summer arrives, we hope our lives may also bloom with the same brightness, warmth, and color.

1. Introduction

This study focuses on the question of whether binaural recordings can replace in-situ listening tests under practical engineering conditions. Specifically, a simplified dummy head binaural recording system was constructed, and comparisons between in-situ listening experiments and binaural listening experiments were conducted to verify whether listeners' subjective evaluation trends for different loudspeakers remain consistent under the two listening conditions. Based on this, a prediction model for subjective evaluation trends was further developed using audio spectral features. In addition, modulation experiments with different types of audio signals and listening experiments under noisy conditions were carried out to analyze the applicability and robustness of binaural recordings in loudspeaker timbre discrimination.

1.1. Background

In the development of loudspeakers, audio product tuning, and subjective listening evaluation, in-situ listening tests are generally regarded as the most direct evaluation method. However, in practical industrial production and sales scenarios, in-situ tests are often limited by constraints related to space, equipment, time, and personnel. Therefore, using recordings to replace in-situ listening tests has important engineering significance.

Previous studies have shown that, under experimental conditions involving room scanning and the use of a Head and Torso Simulator (HATS), binaural recordings can effectively reproduce the in-situ listening experience. However, such methods require demanding experimental conditions and specialized equipment, making them difficult to apply widely in general engineering scenarios. Therefore, the focus of this study is not whether binaural recordings can perfectly reproduce the original sound field, but whether, under simplified recording conditions, binaural recordings can still preserve the key perceptual cues necessary for loudspeaker timbre discrimination and relative preference judgments.

From the perspective of auditory perception, human judgment of timbre does not rely entirely on the precise reconstruction of the physical spectrum, but rather on multidimensional perceptual integration. In subjective listening tasks, listeners also selectively attend to different acoustic cues depending on the evaluation criteria. Therefore, if binaural recordings can preserve the main temporal, spectral, and spatial information that influences loudspeaker evaluation, they may still produce evaluation trends consistent with those obtained from in-situ listening, even if the listening experience is not exactly identical to the original scene.

1.2. Aims

The main objectives of this study are as follows:

- To construct a simplified binaural recording system suitable for practical engineering conditions, capable of capturing the acoustic characteristics of different loudspeakers playing the same audio content in the same room.
- To verify, through in-situ listening tests and binaural listening tests, whether listeners' evaluation trends for different loudspeakers remain consistent under the proposed recording system.
- To define and quantify the consistency of evaluation trends between in-situ listening and binaural listening conditions.
- To develop a prediction model for subjective evaluation trends based on audio spectral features, and to validate the model predictions through modulation experiments using different types of audio signals.
- To investigate whether listeners can robustly discriminate loudspeaker timbre in noisy environments and determine the corresponding signal-to-noise ratio threshold.

1.3. Limitations

This study has several limitations. First, the validation results are based on specific audio content and experimental conditions, so the conclusions cannot be directly generalized to all types of audio signals or listening scenarios.

Second, the ability of binaural listening tests to replace in-situ listening tests in terms of evaluation trend consistency depends on the loudspeakers having perceptible differences under in-situ listening conditions. If the differences between loudspeakers are below the threshold of human auditory discrimination, consistent evaluation trends cannot be guaranteed.

In addition, the recording and playback chains must remain consistent throughout the experiments. Loudness calibration is required during the in-situ listening tests, the recording process, and the binaural playback stage to avoid subjective evaluation bias caused by loudness differences.

Finally, the results suggest that subjective evaluation metrics need clear and relatively stable semantic definitions. At the same time, the relationship between subjective attributes and spectral features is not universal across different types of audio content and must be analyzed in context. Moreover, subjective auditory perception is inherently context-dependent. Listeners' judgments are influenced not only by the audio signal itself, but also by factors such as loudspeaker reproduction characteristics, room acoustics, the recording chain, the preservation of binaural cues, headphone playback conditions, and the semantic definition of the evaluation task. These factors are not independent, but interact with one another throughout the perceptual process.

2. Theory

2.1. Timbre cue in binaural playback

Timbre is one of the most important subjective attributes in auditory perception. Even when two sounds have the same or similar pitch, loudness, and duration, human listeners are still able to distinguish them based on differences in timbre. Timbre can therefore be understood as the perceptual attribute that differentiates sounds beyond pitch, loudness, and duration.

In the playback chain considered in this study—including the original audio signal, loudspeaker reproduction, room propagation, binaural recording, and headphone playback—many of the acoustic processes involved in loudspeaker playback, room propagation, and the binaural recording/playback chain can generally be approximated as linear and time-invariant (LTI) systems under typical listening conditions and within their normal operating ranges.[Møl 92][Pau 09] Although loudspeaker frequency response, room acoustics, binaural filtering, and headphone playback may alter the signal spectrum to some extent, previous studies have shown that important spectrotemporal relational cues and higher-order statistical structures related to sound object recognition are often preserved throughout this process [McA 19, McD 11].

Current research on timbre constancy and auditory adaptation further suggests that human timbre perception does not rely entirely on the precise reconstruction of spectral content, but rather on the preservation of spectrotemporal organization, including temporal envelope structure, harmonic relationships, modulation patterns, and other relational acoustic cues [McA 19, McD 11]. Therefore, even when sounds undergo substantial spectral transformations during propagation and playback, the auditory system is still able to partially compensate for the overall coloration introduced by loudspeakers, room acoustics, and the playback chain, allowing relatively stable recognition of sound identity and timbral attributes [Pik 16, Too 88]. Pike further proposed that this compensation mechanism may be related to spectral normalization processes in speech perception, in which listeners adapt to spectral variations caused by differences in vocal tract structure [Pik 16]. Similar adaptation effects have also been observed for non-speech sounds, suggesting that the mechanism is not limited to speech processing but may generalize to broader auditory perception tasks [Pik 16].

However, timbre constancy does not imply that the auditory system completely eliminates the differences introduced by different playback chains. Instead, while maintaining the perceptual identity of a sound, human listeners remain sensitive to residual spectrotemporal differences caused by loudspeakers, room acoustics, and spatial radiation characteristics. In addition to preserving the basic spectral structure of the original audio signal, binaural recording systems also

capture important spatial and spectrotemporal information related to the loudspeaker and room, including spectral balance, transient response, directivity, early reflections, binaural coherence, and frequency-dependent spatial radiation patterns [Møl 92, Pau 09]. Previous studies have shown that the perception of loudspeaker differences depends not only on on-axis frequency response, but also to a large extent on off-axis response and its interaction with the room [Too 88, Too 08]. Because different loudspeakers exhibit different directivity characteristics, transient behavior, and room energy distributions, they produce different binaural spectrotemporal statistics. These differences can be partially preserved through dummy head binaural recordings [Møl 92, Pau 09, McD 11], and remain perceptually distinguishable to human listeners.

Therefore, even after the same audio content undergoes loudspeaker playback, room propagation, and binaural recording/playback, the auditory system is still able to perceive residual spatial and spectrotemporal differences between loudspeakers while partially compensating for the overall coloration introduced by the playback chain. For this reason, listeners may still form relatively stable timbre discrimination and subjective preference judgments based on the spectral balance, transient structure, and binaural spatial statistics preserved in binaural recordings [McD 11, Too 88, Too 08].

In listening tests, listeners do not typically perceive sound in a completely passive and holistic manner. Instead, they selectively attend to specific perceptual attributes depending on the listening task and the wording of the evaluation question. For example, when the question is "how deep is the bass," listeners tend to focus their attention on acoustic cues related to low-frequency perception, such as low-frequency extension, low-frequency energy distribution, room gain, temporal envelope characteristics, and other bass-related features, while paying less attention to other attributes such as high-frequency detail or spatial localization. This phenomenon is consistent with theories of Auditory Scene Analysis and selective auditory attention. Bregman proposed that the human auditory system does not simply receive sounds passively, but actively organizes, separates, and selects auditory objects that are relevant to the current task [Bre 90]. Further studies have shown that the auditory system can dynamically adjust the perceptual weighting of different acoustic cues according to the listening task, assigning different importance to different features under different listening conditions [Shi 08, Fri 07]. In audio evaluation research, this behavior is often referred to as attribute-based listening or analytical listening, in which listeners selectively extract acoustic information relevant to the perceptual attribute being evaluated [Too 08].

However, different perceptual attributes vary considerably in their level of abstraction. Questions such as "how deep is the bass" typically correspond to relatively low-level perceptual attributes, whose judgments can often be linked more directly to specific acoustic features such as low-frequency spectral energy, frequency extension, or temporal envelope characteristics. In contrast, higher-level and more abstract perceptual attributes—such as "naturalness" in music reproduction, "urgency" in alarm sounds, or "clarity" in speech perception—do not rely on a single acoustic cue. Instead, they are more closely related to semantic auditory perception or affective auditory perception. In these types of tasks, listeners usually integrate multiple percep-

tual cues simultaneously and dynamically adjust the perceptual weighting of different acoustic features in order to form higher-level perceptual interpretations.

For example, in the case of alarm urgency, previous studies have shown that humans tend to associate certain acoustic features with perceived danger, alertness, or behavioral urgency [Haa 08, Jus 03]. Therefore, when listeners are asked to evaluate "how urgent is the alarm," the auditory system not only increases the perceptual weighting of relevant acoustic cues, but also maps these low-level acoustic features onto semantically meaningful perceptual categories. In other words, the auditory system is performing not only selective auditory attention, but also higher-level perceptual mapping. This process is consistent with theories of ecological acoustics, auditory cognition, and affective sound perception, which suggest that human hearing is used not only to analyze sounds themselves, but also to infer behavioral meaning and potential events in the environment [Bre 90, Jus 03].

Therefore, from a theoretical perspective, although both "how deep is the bass" and "how urgent is the alarm" rely on selective auditory attention, they involve different levels of perceptual processing. The former depends more on the analysis of relatively explicit, local, and low-level acoustic cues, whereas the latter relies more heavily on higher-level perceptual integration and semantic interpretation across multiple temporal, spectral, spatial, and semantic cues.

2.2. Human Hearing

The human auditory system perceives the surrounding sound field through the coordinated operation of both ears. Compared with monaural hearing, binaural hearing not only enables perception of the spectral and loudness characteristics of sound itself, but also allows listeners to perceive the spatial location, direction, and acoustic environment of sound sources through the differences between the signals received at the two ears [Bla 97]. In real listening environments, the auditory system integrates temporal, level, and spectral information received at the left and right ears in order to separate and identify different sound sources within complex acoustic scenes.

Binaural hearing plays a critical role in human perception in complex acoustic environments. Even in the presence of background noise or multiple simultaneous sound sources, listeners are still able to use binaural spatial cues and spectral information to attend to target sounds and distinguish differences in timbre and other subjective attributes between sound sources [Bre 90]. Therefore, subjective auditory evaluation depends not only on the sound itself, but also on the spatial and spectral information perceived through the binaural auditory system.

Interaural Time Difference (ITD) refers to the difference in arrival time of the same sound source between the left and right ears. When a sound source is located directly in front of the listener, the sound reaches both ears almost simultaneously. However, when the sound source is positioned away from the center axis of the head, differences in propagation path length cause the sound to arrive earlier at the ear closer to the source, thereby producing an interaural time

difference.

2.2.1. Interaural Time Difference (ITD)

ITD is one of the most important binaural cues for spatial localization, particularly in the low-frequency range [Bla 97]. Because low-frequency signals have relatively long wavelengths, the auditory system is highly sensitive to interaural phase and time differences. As a result, ITD plays a major role in low-frequency sound localization, spatial perception, and source separation. In complex acoustic environments, listeners can also use ITD cues to attend to and distinguish target sound sources, thereby improving sound recognition performance [Bla 97].

For binaural recording systems, the preservation of ITD cues is essential for reconstructing spatial auditory information [Møl 92]. By separately recording the signals received at the positions of the left and right ears, binaural recording systems allow part of the interaural time difference information from the original sound field to be retained during headphone playback. When ITD cues are preserved accurately, listeners may still experience spatial perception and subjective auditory impressions similar to those obtained during in-situ listening [Møl 92].

2.2.2. Interaural Level Difference (ILD)

Interaural Level Difference (ILD) refers to the difference in sound pressure level between the signals arriving at the left and right ears from the same sound source. This difference is mainly caused by the head shadow effect. When a sound source is located to one side of the listener, the ear closer to the source typically receives a higher sound pressure level, while the opposite ear receives a relatively lower level.

ILD is one of the primary binaural cues for spatial localization, particularly in the high-frequency range. In binaural recording systems, the preservation of ILD cues helps maintain part of the spatial perception during headphone playback. In this study, however, ILD is mainly discussed as one of the theoretical foundations explaining how binaural recordings are able to preserve spatial auditory information.

2.2.3. Head-Related Transfer Function (HRTF)

The Head-Related Transfer Function (HRTF) describes the direction-dependent filtering effect that occurs when sound propagates from a location in space to the human ears, where it is shaped by the head, torso, and pinna structures [Bla 97]. HRTFs include interaural time differences, interaural level differences, and spectral modifications caused by the geometry of the head and outer ears, and therefore form an important foundation for binaural hearing and spatial perception [Bla 97].

In binaural recording systems, dummy heads or HATS devices simulate the structure of the human head and ears, allowing the recorded signals to preserve certain HRTF-related cues [Møl 92]. Although the HRTFs of dummy heads or HATS systems cannot perfectly match the

individual HRTF of every listener, the focus of this study is not precise spatial localization, but rather whether these binaural and spectral cues are sufficient to support listeners in discriminating timbral differences between loudspeakers [Møl 92].

subsectionHeadphone Transfer Function(HPTF)

The Headphone Transfer Function (HPTF) describes the acoustic transfer characteristics between the headphone driver and the listener's ear canal entrance or eardrum position. During binaural playback, headphones are not perfectly transparent reproduction devices. Their own frequency response, together with the acoustic coupling between the headphones and the ears, affects the final sound reaching the listener, thereby altering the spectral balance, timbre, and spatial perception of binaural recordings. As a result, HPTF is one of the key factors influencing the listening experience of binaural playback [Bla 97, Too 08].

Because HPTF is influenced by headphone model, wearing position, ear pad sealing, and individual ear geometry, different playback conditions may introduce additional spectral coloration. To minimize these effects, all binaural listening experiments in this study were conducted using the same headphones, the same audio interface output chain, and identical playback settings, with loudness calibration performed before the experiments. This study does not apply individualized HPTF compensation. Instead, playback conditions were kept consistent so that the influence of HPTF on different loudspeaker recordings remained as stable as possible, allowing the comparison to focus primarily on relative timbral differences between loudspeakers [Too 08, Koh 21].

The headphones used in this study were the Sennheiser HD650. From an HPTF perspective, open-back circumaural headphones such as the HD650 offer advantages including lower acoustic energy storage and relatively smooth mid-frequency reproduction. Because rear-cavity reflections and enclosed-cavity resonances are weaker, open-back headphones are generally less likely than closed-back headphones to introduce strong narrowband peaks, dips, or comb-filtering effects in the mid- and high-frequency range. In addition, the open-back design helps reduce additional coloration introduced by the headphones themselves, allowing listeners to perceive the timbral differences preserved in the binaural recordings more consistently [Too 08, Koh 21].

It should be noted, however, that the HPTF of open-back headphones is still affected by factors such as wearing position, ear pad compression, pinna shape, and ear canal geometry. In particular, low-frequency reproduction may be influenced by leakage paths and headphone fit. Therefore, this study does not assume that the HD650 completely eliminates the influence of headphone playback. Rather, its open-back design, relatively low coloration, and widespread use in professional listening evaluation make it suitable for minimizing headphone-related interference when comparing relative loudspeaker timbre under controlled playback conditions [Bla 97, Koh 21].

subsectionSpatial Cue Preservation in Binaural Recording The fundamental principle of binaural recording is to simulate the way human ears receive sound during recording by separately capturing the signals arriving at the left and right ear positions. During headphone playback, part of the binaural auditory information perceived by human listeners can then be reconstructed. Compared with traditional monaural or stereo recording, binaural recording is able to preserve

spatial cues and spectral characteristics of the original sound field to a certain extent, thereby providing a subjective listening experience that is closer to in-situ listening.

The effectiveness of binaural recording relies primarily on its ability to preserve spatially relevant perceptual cues. These cues include not only binaural spatial information such as Interaural Time Differences (ITD) and Interaural Level Differences (ILD), but also spectral and spatial characteristics shaped jointly by the head, pinna, and acoustic environment. When these perceptually important cues are sufficiently preserved throughout the recording and playback process, listeners may still experience spatial perception and subjective auditory characteristics similar to those of in-situ listening during headphone playback.

However, binaural recording does not imply a complete physical reconstruction of the original sound field. Factors such as headphone playback, individual HRTF differences, and the absence of head tracking generally prevent binaural recordings from fully reproducing all auditory information present in real acoustic environments. Therefore, the goal of this study is not to perfectly recreate the original sound field, but rather to investigate whether binaural recordings can preserve the key perceptual cues required for subjective evaluation.

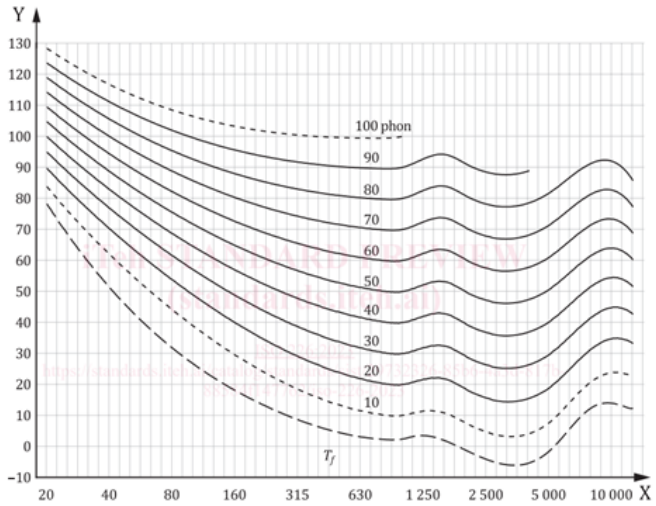
subsectionLoudness Zwicker defined loudness as "the auditory sensation corresponding to the perceived sound intensity" [Zwi 99]. Loudness therefore refers to the subjective perception of sound intensity by the human auditory system. It depends not only on the physical Sound Pressure Level (SPL) of a sound, but also on its frequency distribution. Even when sounds at different frequencies have the same physical sound pressure level, they may still be perceived as having different loudness levels by human listeners [Zwi 99].

Zwicker defined Equal Loudness Contours as "combinations of sound pressure levels and frequencies of pure continuous tones that are perceived as equally loud by human listeners" [Zwi 99]. Equal Loudness Contours describe the frequency-dependent sensitivity of human hearing to loudness perception. Studies have shown that the human auditory system is most sensitive in the frequency range of approximately 2–5 kHz, while sensitivity is relatively lower at low and very high frequencies. As a result, the same magnitude of spectral variation may produce different subjective perceptual effects across different frequency regions [Zwi 99][ISO 23].

2.2.4. Auditory Masking

Auditory masking is an important phenomenon in psychoacoustics, referring to the process by which the presence of one sound raises the hearing threshold of another sound, making the target sound more difficult to hear or distinguish. In masking phenomena, the sound producing the masking effect is typically referred to as the masker, while the masked target sound is referred to as the maskee. Masking reduces the auditory system's ability to perceive details and differences within the target sound.

In the listening experiments conducted in this study, the dominant form of masking was simultaneous masking, in which background noise and the target sound are present at the same time. Masking becomes more pronounced when the background noise and the target sound



Key

X frequency, expressed in Hz

Y sound pressure level, expressed in dB

T_f hearing threshold

NOTE 1 The hearing threshold under free-field listening condition, T_f, is indicated by a dashed line as defined in ISO 389-7[1].

NOTE 2 The contour at 10 phon is drawn by dotted lines because of the lack of experimental data between 20 phon and the hearing thresholds. Moreover, the 100-phon contour is also described by a dotted line because data from only one institute are available at this loudness level.

Figure 2.1.: Normal equal-loudness-level contours for pure tones (binaural, free-field listening, frontal incidence)

contain similar frequency components or fall within the same or adjacent critical bands. Because human auditory processing exhibits critical-band characteristics, the influence of noise on the target sound is not distributed uniformly across the spectrum, but instead depends strongly on the energy distribution of both the noise and the target signal within perceptual frequency bands.

Auditory masking also affects the perception of timbral differences. Timbre perception typically relies on acoustic cues such as spectral envelope characteristics, frequency-band energy distribution, and temporal detail. When these cues are partially masked by background noise, timbral differences between sounds become less distinct, reducing listeners' ability to discriminate between sound sources or differences in sound quality. In loudspeaker evaluation tasks, if spectral differences or timbral cues between loudspeakers are partially masked by background noise, the perceptual distinguishability between loudspeakers will decrease.

Therefore, as the Signal-to-Noise Ratio (SNR) decreases, the masking effect of background noise on the target sound becomes stronger, reducing the amount of timbral information available to listeners. As a result, the subjective differences between loudspeakers may become less stable or more difficult to distinguish.

2.3. Spectral Analysis

2.3.1. Short-Time Fourier Transform (STFT)

For time-domain audio signals, the conventional Fourier Transform can only provide overall spectral information and cannot describe how frequency components change over time. Because audio signals are typically non-stationary signals whose spectral characteristics vary with time, time-frequency analysis methods are required to perform localized spectral analysis of the signal.

The Short-Time Fourier Transform (STFT) is a commonly used time-frequency analysis method. Its basic principle is to divide a long-duration signal into multiple short-time windows and apply the Fourier Transform to the local signal within each window separately, thereby obtaining the distribution of the signal in both the time and frequency domains [All 77].

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

[All 77] where:

- $x[n]$ represents the input time-domain signal;
- $w[n]$ represents the sliding window function;
- m represents the time-window position;
- ω represents the angular frequency;
- $X(m,\omega)$ represents the joint time-frequency representation of the signal.

In practical computation, the audio signal is first divided into frames using a fixed window length, and a Fast Fourier Transform (FFT) is then applied to each frame separately. The analysis window continuously shifts along the time axis with a fixed step size, thereby forming a complete spectrogram. The STFT parameters used in this study are:

- FFT size: $n_{FFT} = 4096$
- Hop length: 1024

Here, the FFT size determines the frequency resolution, where a larger FFT size provides finer frequency analysis capability. The hop length determines the time resolution and controls the overlap between adjacent analysis windows.After taking the magnitude of the STFT result and squaring it, the power spectrum can be obtained:

$$P(m, \omega) = |X(m, \omega)|^2 \tag{2.2}$$

The power spectrum reflects the energy distribution of the signal across time and frequency, and is therefore widely used in audio feature extraction, spectral analysis, and psychoacoustic research.

2.3.2. Bark Scale

In human auditory perception, frequency resolution is not distributed uniformly along the linear frequency axis. Compared with physical frequency measured in Hz, human perception of frequency changes follows the psychoacoustic characteristics of critical bands more closely. Therefore, in audio perceptual analysis, it is often necessary to transform linear frequency into a perceptual frequency scale that better reflects human auditory characteristics.

The Bark Scale is a nonlinear perceptual frequency scale based on critical band theory, and provides a more accurate representation of the auditory system's frequency resolution across different frequency regions [Zwi 61]. Human hearing has relatively high frequency resolution in the low-frequency range, whereas frequency resolution gradually decreases at higher frequencies. The Bark Scale reflects this perceptual characteristic through nonlinear frequency mapping [Zwi 61].

$$z = 13 \arctan(0.00076 f) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{2.3}$$

[Tra 90] where:

- f represents the physical frequency (Hz);
- z represents the corresponding Bark-scale frequency value.

This transformation was proposed by Zwicker and is one of the commonly used approximation models for the Bark frequency scale in psychoacoustic research. The model maps the linear frequency axis onto a nonlinear perceptual frequency axis that better reflects the critical-band characteristics of human hearing, thereby improving the perceptual validity of audio analysis and auditory modeling.

2.3.3. Spectral Fluctuation Analysis

To characterize local spectral-envelope variations along the frequency axis, this study draws on concepts from spectral structure analysis [Gre 78] and log-spectral analysis [Gra 76], and adopts a statistical description method based on log-spectral differences to quantify the degree of local spectral variation. Previous studies have shown that spectral-envelope shape and local spectral variation reflect structural properties of sound and are closely related to auditory perception [Gre 78, McA 99].

The theoretical basis of this method originates from studies of spectral structure variation in spectral analysis and psychoacoustics [Gre 85]. Auditory perception research has shown that the human auditory system is highly sensitive to spectral envelopes and local frequency-energy variations [Gre 83]. Related studies often use spectral ripple analysis [Hen 03] to describe periodic fluctuations of spectral structure along the frequency axis in order to investigate the auditory system's ability to resolve spectral variation. Inspired by the concept of spectral ripple analysis, this study further adopts a statistical approach based on spectral differences to quantify local fluctuations in the spectral envelope. This method is not a direct implementation of traditional spectral ripple metrics, but rather a statistical description approach based on spectral variation characteristics.

The audio signal is first transformed into a time-frequency representation using the Short-Time Fourier Transform (STFT). After obtaining the spectral representation of the signal, the power spectrum is calculated.2.3.1

To obtain stable frequency-domain statistical features, averaging is performed along the time dimension:

$$\bar{P}(f) = \frac{1}{T} \sum_{t=1}^{T} P(f, t) \tag{2.4}$$

where T represents the number of time frames. Because human auditory perception is more sensitive to relative spectral variation on a logarithmic scale, a logarithmic transformation is further applied to the averaged power spectrum:

$$L(f) = \log(\bar{P}(f) + \epsilon) \tag{2.5}$$

where ϵ is a small constant introduced to prevent numerical instability caused by zero values. To quantify local spectral-envelope variation along the frequency axis, the log-spectral difference between adjacent frequency bins is calculated:

$$D(f_i) = L(f_{i+1}) - L(f_i) \tag{2.6}$$

Finally, the variance of the difference sequence is used to characterize the overall spectral fluctuation property:

$$R = \mathrm{Var}(D(f)) \tag{2.7}$$

A larger value of R indicates stronger local fluctuations and greater structural variation in the spectral envelope, whereas a smaller value of R corresponds to a smoother spectral distribution. This statistical feature is used to characterize spectral complexity and local structural variation across different frequency regions, providing a quantitative spectral basis for subsequent audio perceptual analysis and subjective evaluation studies.

2.3.4. Sharpness

In psychoacoustic research, sharpness is used to describe the perceived "sharp" or "piercing" quality of a sound and is considered one of the important perceptual parameters in sound quality evaluation [Zwi 99]. Compared with low-frequency components, high-frequency energy generally increases the perceived sharpness of a sound. Therefore, sharpness is closely related to the distribution of spectral energy in the high-frequency range.

The concept of sharpness was originally proposed by Zwicker and is theoretically based on the auditory critical-band concept and the Bark perceptual frequency scale [Zwi 61]. Because human auditory sensitivity is not uniform across different frequency regions, sharpness analysis is typically performed in the perceptual frequency domain. In the sharpness model, higher Bark regions correspond to higher-frequency components. Because high-frequency energy has a stronger influence on perceived sharpness, a frequency-weighting function $g(z)$ is applied to emphasize high-frequency regions:

$$g(z) = \begin{cases} 1, & z \leq 16 \\ 1 + 0.2(z - 16), & z > 16 \end{cases} \tag{2.8}$$

[Sot 16] When the Bark value exceeds 16, the weighting increases linearly with frequency, thereby enhancing the contribution of high-frequency regions to the overall sharpness. Based on the weighting scheme above, the sharpness of a single frame spectrum is defined as:

$$S_t = \frac{\sum P(f,t)\, z(f)\, g(z(f))}{\sum P(f,t)} \tag{2.9}$$

where:

- $P(f,t)$ represents the power spectrum at frequency f and time frame t;
- $z(f)$ represents the corresponding Bark frequency;

- $g(z)$ represents the high-frequency weighting function.

Finally, the sharpness values of all time frames are averaged to obtain the overall sharpness estimate of the audio signal:

$$S = \frac{1}{T} \sum_{t=1}^{T} S_t \tag{2.10}$$

where T represents the number of time frames. This method is able to characterize the distribution of high-frequency energy from a perceptual frequency perspective, and is therefore widely used in sound quality evaluation, environmental sound analysis, and psychoacoustic research.

2.3.5. spectral smoothness

The smoothness of the spectral envelope is one of the important descriptors of spectral structure characteristics in sound analysis [Van 85]. In auditory perception and timbre analysis, the continuity of spectral structure and local spectral variations can influence the perceived clarity, naturalness, and spectral complexity of sounds.

In general, the smoother the variation between adjacent frequency components in the spectrum, the more continuous the spectral envelope becomes, corresponding to higher spectral smoothness. Conversely, stronger local fluctuations along the frequency axis indicate greater irregularity and complexity in the spectral structure [Haj 07]. Therefore, the first-order variation of the spectral envelope along the frequency axis can be used to characterize spectral smoothness. Because human auditory perception of energy variation is more closely related to logarithmic relationships, spectral smoothness is typically analyzed based on the log-spectrum. For the log-spectrum $L(f)$ at frequency f, the variation between adjacent frequency bins can be expressed as:

$$D(f_i) = L(f_{i+1}) - L(f_i) \tag{2.11}$$

where $D(f_i)$ represents the local spectral variation between adjacent frequency bins.[Gra 76] To further characterize the overall smoothness of the spectral envelope, the statistical dispersion of the difference sequence can be used as an estimate of spectral structural complexity. A larger variance indicates stronger local fluctuations along the frequency axis, whereas a smaller variance corresponds to a more continuous and smoother spectral structure. Based on this concept, spectral smoothness can be represented as the negative correlation of spectral difference variation:

$$M = -\mathrm{Var}(D(f)) \tag{2.12}$$

[Pee 04]

2.3.6. Multivariate Regression and Regularization Theory

In audio perception research, subjective evaluation results are typically influenced by multiple acoustic features simultaneously. Therefore, the subjective perceptual process can be regarded as a mapping relationship between multidimensional acoustic features and perceptual ratings. For a given set of acoustic features, different features may contribute differently to subjective perception, making it necessary to apply statistical modeling methods to estimate the relationship between acoustic features and subjective evaluations [Has 09]. In linear regression theory, the target variable can be represented as a linear combination of multiple input features:

$$y = \mathbf{w}^T \mathbf{x} + b + \epsilon \tag{2.13}$$

where:

- $\mathbf{x}$ represents the input feature vector;
- $\mathbf{w}$ represents the feature weights;
- b represents the bias term;
- ϵ represents the random error.

This model assumes that subjective perceptual outcomes are jointly determined by the linear contributions of multiple acoustic features. Therefore, the regression coefficients can be used to describe the relative influence of different acoustic features on perceptual results. This model assumes that subjective perceptual outcomes are jointly determined by the linear contributions of multiple acoustic features. Therefore, the regression coefficients can be used to describe the relative influence of different acoustic features on perceptual results. However, in multidimensional acoustic feature analysis, different spectral features are often correlated with one another, resulting in multicollinearity. When strong correlations exist between features, traditional ordinary least squares regression may suffer from parameter instability and become overly sensitive to the training data, thereby reducing the generalization capability of the model [Hoe 00]. To improve model stability, statistical learning theory commonly applies regularization methods to constrain regression parameters. Among these methods, Ridge Regression introduces an L_2-norm penalty term into the objective function to regularize the regression weights:

$$\min_{\mathbf{w}} \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \|\mathbf{w}\|_2^2 \right) \tag{2.14}$$

[Hoe 70] where:

- the first term represents the model prediction error;
- the second term represents the weight regularization constraint;
- α is the regularization parameter.

2.3.7. Spectral Modulation

In the human auditory system, subjective sound perception is closely related to spectral structure. Psychoacoustic research has shown that human listeners are sensitive not only to the overall spectral energy distribution, but also to local spectral variations along the frequency axis. Periodic peak-and-valley structures in the spectrum, commonly referred to as spectral ripples, can significantly influence subjective perceptual attributes such as tension, sharpness, and perceived sound complexity [Pul 15].

Spectral ripple theory suggests that when the spectral envelope exhibits alternating enhancement and attenuation along the frequency axis, the auditory system perceives more pronounced spectral structural variation. The spacing between spectral peaks and valleys, modulation depth, and local spectral bandwidth all affect how the auditory system perceives spectral structure [Won 07]. Therefore, by controlling the local spectral energy distribution, it is possible to modify subjective perceptual characteristics of sound.

In frequency-domain signal processing, a parametric equalizer can modify spectral structure through localized frequency-band gain control. For the frequency region around the center frequency f_0, a peaking filter can be used to boost or attenuate spectral energy:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{2.15}$$

[Bri 21] where the filter parameters are jointly determined by the center frequency, gain, and quality factor (Q-factor). When alternating gain modulation is applied at multiple frequency locations, the spectral envelope forms a periodic fluctuation pattern, thereby producing spectral variation characteristics similar to spectral ripples. The spectral modulation depth determines the magnitude difference between spectral peaks and valleys, while the Q-factor influences the sharpness and bandwidth of the local spectral structure. Based on the theory described above, local spectral modulation can be used to actively modify spectral structure and further influence the corresponding subjective perceptual outcomes. Therefore, the spectral modulation process can essentially be regarded as a perception-driven spectral shaping method.

2.3.8. Spectral Similarity Constraint

To avoid excessive spectral distortion during spectral modulation, it is usually necessary to constrain the spectral difference between the modified audio and the original audio. In audio perception research, log-spectral distance is commonly used to describe the similarity of spectral structure between two audio signals [Gra 76].

$$D_{LS} = \left\{ \frac{1}{N} \sum_{n=1}^{N} \left| \log P(n) - \log \hat{P}(n) \right|^p \right\}^{1/p} \tag{2.16}$$

[Gra 76]. Inspired by this concept, this study further adopts the mean squared log-spectral difference in the discrete time-frequency domain as the spectral distance measure:

$$D = \frac{1}{N} \sum_{f,t} (L_1(f,t) - L_2(f,t))^2 \tag{2.17}$$

where:

- $L_1(f,t)$ and $L_2(f,t)$ represent the log-spectral representations of the two audio signals, respectively;
- f represents frequency;
- t represents time.

A larger spectral distance indicates greater differences between the modified audio and the original spectral structure, whereas a smaller distance corresponds to higher spectral similarity. Therefore, spectral distance can be used as a constraint in the spectral modulation process to limit excessive spectral deformation, thereby balancing perceptual enhancement with preservation of the original sound characteristics.

2.3.9. Modulation Structure and Clarity Perception Theory

In the human auditory system, sound perception depends not only on spectral energy distribution, but also strongly on temporal envelope structure. Psychoacoustic research has shown that the auditory system is highly sensitive to amplitude modulation within sound envelopes. As a result, temporal modulation structure can significantly influence sound intelligibility, clarity, and overall auditory perception characteristics [Drul 94].

For non-stationary audio signals, the temporal envelope typically contains modulation components across different frequency ranges. Modulation frequency describes the rate at which the sound envelope changes over time, and different modulation-frequency ranges correspond to different perceptual auditory characteristics. Temporal modulation analysis is typically based on the energy distribution of the sound envelope in the modulation-frequency domain. For the temporal envelope $e(t)$ of an audio signal, its modulation characteristics can be further represented as:

$$M(f_m) \tag{2.18}$$

[Dau 97] where:

- f_m represents the modulation frequency;
- $M(f_m)$ represents the modulation energy at the corresponding modulation frequency.

Previous studies have shown that variations in temporal envelope structure can influence how the auditory system perceives sound structure characteristics [Drul 94, Dau 97]. Therefore, the

relative differences between modulation structures can be used to describe differences between sounds at the temporal-structure level.

Because the human auditory system exhibits relatively stable perception of temporal envelope variation, modulation-frequency-domain analysis has been widely applied in research areas such as speech perception, sound quality analysis, and auditory modeling.

2.3.10. Hilbert Transform

For non-stationary audio signals, the temporal envelope reflects how sound energy varies over time. In auditory perception research, temporal envelope structure is closely related to speech clarity, rhythm perception, and modulation characteristics. Therefore, temporal envelope analysis is widely used in speech and audio signal processing [Opp 10]. The Hilbert Transform is a commonly used method for analytic signal construction. For a real-valued signal $x(t)$, its analytic signal can be expressed as:

$$z(t) = x(t) + j\hat{x}(t) \tag{2.19}$$

where:

- $\hat{x}(t)$ represents the Hilbert Transform of $x(t)$;
- j represents the imaginary unit.

Based on the analytic signal, the temporal envelope of the signal can be further obtained as:

$$A(t) = \sqrt{x^2(t) + \hat{x}^2(t)} \tag{2.20}$$

Because the temporal envelope reflects amplitude modulation structure, the Hilbert Transform is widely used in temporal modulation analysis and auditory modeling research [Opp 10].

2.4. Statistic Calculation

2.4.1. Confidence Interval Estimation

A Confidence Interval (CI) is used to quantify the reliability of a sample statistic in estimating a population parameter. A 95% confidence interval indicates that, under repeated sampling, approximately 95% of the calculated intervals would contain the true value of the population parameter.

This study adopts a 95% confidence interval because it is the most commonly used and widely accepted standard in statistical analysis. A 95% confidence level provides a reasonable balance between result reliability and estimation precision, thereby improving the credibility and interpretability of the research findings.

2.4.2. Pairwise comparison

Pairwise comparison is a research method used to analyze individual preferences and ranking consistency. In this method, different options are compared in pairs, and participants indicate their preferred choice for each comparison, from which an overall ranking can be constructed. In ranking consistency analysis, pairwise methods evaluate the stability and logical consistency of preferences by examining the relative preference relationships between options. Compared with directly asking participants to provide a complete ranking of all options, pairwise comparison reduces cognitive load and can more accurately reflect individual preferences.

In subjective listening evaluation, auditory perception inherently contains a certain degree of listener variability, meaning that different listeners do not necessarily produce identical preference rankings. Therefore, subjective evaluation results generally rely on group-level statistical analysis rather than on judgments from a single listener. ITU-R BS.1116-3, *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, states that subjective audio evaluation should be based on statistical analysis across listeners and reference-based comparison in order to improve the reliability and stability of evaluation results [ITU 15]. In addition, in psychophysics and observer agreement analysis, the degree of consistency within a group is commonly used to describe the stability of perceptual judgments [Fle 71, Lan 77].

Based on these concepts, this study defines the pairwise preference ratio to describe the degree of agreement within a listener group regarding the relative preference relationship between two loudspeakers. For any pair of loudspeakers, the pairwise preference ratio is defined as:

$$P(A > B) = \frac{N(A > B)}{N} \tag{2.21}$$

where:

- $N(A > B)$ represents the number of participants who judged loudspeaker A to be preferred over loudspeaker B;
- N represents the total number of participants.

A higher pairwise preference ratio indicates a stronger consensus within the group regarding the preference tendency for that pair, whereas a lower ratio suggests greater disagreement among listeners. Considering the inherent listener variability in subjective listening evaluation, this study adopts a 70% pairwise preference ratio as an operational majority-agreement criterion for identifying relatively stable preference tendencies. Specifically, when the preference ratio of a given pair reaches or exceeds 70%, the pair is considered to exhibit a relatively stable group preference tendency under the corresponding listening condition.

However, the pairwise preference ratio alone can only reflect consistency within a single listening condition and cannot directly evaluate whether binaural playback can replace in-situ listening. Because listener variability also exists under in-situ listening conditions, the key objective of binaural playback evaluation is not to achieve a particular absolute preference

ratio, but rather to determine whether the perceptual preference structure observed under in-situ conditions can be preserved. Based on this idea, this study further defines the pairwise preference deviation to describe perceptual consistency between binaural playback and in-situ listening:

$$\Delta P = |P_{\text{binaural}} - P_{\text{insitu}}| \tag{2.22}$$

where:

- P_{binaural} represents the pairwise preference ratio under binaural playback conditions;
- P_{insitu} represents the pairwise preference ratio under in-situ listening conditions.

A smaller ΔP indicates that binaural playback and in-situ listening share more similar perceptual preference tendencies, whereas a larger ΔP suggests greater perceptual deviation between the two listening conditions.

It should be noted that the preference-ratio and preference-deviation thresholds adopted in this study are intended primarily as operational criteria for describing perceptual consistency tendencies across different listening conditions, rather than as universally established perceptual boundaries. Because subjective listening evaluation inherently involves listener variability, and because no unified standard currently exists for evaluating perceptual equivalence in binaural playback, this study focuses more on the relative preservation of perceptual preference structure across listening conditions than on strict binary judgments based on absolute thresholds. For this reason, a 70% pairwise preference ratio is adopted as an empirical majority-agreement criterion for relatively stable preference tendencies, while $\Delta P \leq 0.15$ is used as a pragmatic tolerance criterion for acceptable perceptual consistency between binaural playback and in-situ listening, serving as an aid for interpreting group-level perceptual consistency trends in the experimental results.

2.4.3. Non-Parametric Statistics

The data obtained from subjective ranking experiments typically represent relative ordering relationships between objects, and are therefore fundamentally ordinal data. Because such data do not satisfy the assumptions of normal distribution and linear relationships required for statistical analysis of continuous variables, traditional linear statistical methods based on continuous-variable assumptions are not suitable. Instead, rank-based statistical methods are required to describe the consistency between rankings, which corresponds to non-parametric statistical analysis [Zar 14].

In ranking analysis, ranking results can first be mapped into corresponding rank sequences. For ranking tasks involving multiple evaluation objects, each object is assigned a rank according to its position in the ranking, thereby forming a rank vector that can be used for subsequent statistical analysis.

In rank-based non-parametric statistical analysis, Spearman's rank correlation coefficient is commonly used to measure the degree of consistency between two rank sequences [Spe 04]. Spearman's rank correlation coefficient describes the monotonic relationship between two ranked variables and is defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.23}$$

where d_i represents the rank difference of the same object between the two ranking sequences, and n represents the number of ranked objects. The value range of Spearman's correlation coefficient is $[-1, 1]$, and the coefficient reflects the degree of consistency between the two ranking sequences.[Spe 87] When multiple correlation coefficients need to be statistically aggregated, direct arithmetic averaging is not appropriate because correlation coefficients are not linearly additive. In particular, when the correlation coefficient approaches ± 1, its probability distribution becomes significantly skewed. Therefore, Fisher's z-transformation is commonly applied to normalize correlation coefficients and improve the stability of statistical analysis. The transformation is defined as:

$$z = \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \tag{2.24}$$

After Fisher's z-transformation, the correlation coefficients approximately follow a normal distribution, making them suitable for mean-value computation and overall statistical analysis. After statistical processing, the correlation coefficient can be recovered through the inverse transformation:

$$\rho = \tanh(z) \tag{2.25}$$

The methods described above provide a stable statistical basis for rank-correlation analysis in subjective ranking experiments, and are suitable for evaluating consistency across multiple experimental conditions.[Upt 14]

2.4.4. Generalized Estimating Equations Logistic Regression

Generalized Estimating Equations (GEE) is an extension of the Generalized Linear Model (GLM) that is primarily used for analyzing data with correlated or repeated-measurement structures. Unlike traditional Logistic Regression, which assumes that all observations are independent, GEE can account for correlated observations from the same subject under different experimental conditions by introducing a correlation structure [Lia 86].

In this study, GEE Logistic Regression was used to analyze the effects of noise level and listening method on loudspeaker timbre discrimination performance. The experiment has the following characteristics: first, the dependent variable is binary (`response = 0` for unsuccessful discrimination and `response = 1` for successful discrimination); second, the same participants were tested repeatedly under multiple noise conditions; third, observations within the same participant are correlated; and fourth, this study focuses more on population-level trends rather than

prediction for individual participants. These characteristics are consistent with the assumptions and application scenarios of GEE Logistic Regression, making it suitable for analyzing the experimental results in this study [Bal 04]. To analyze the overall effects of noise level and listening method on discrimination performance, a main-effects GEE Logistic Regression model was first established:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Method} + \beta_2 \cdot \text{Noise} \tag{2.26}$$

where:

- p represents the probability of correct discrimination;
- Method represents the listening method (in-situ or binaural);
- Noise represents the noise level;
- β_0 is the intercept term;
- β_1 represents the effect of listening method;
- β_2 represents the effect of noise level.

To further examine whether the effect of noise differs across listening methods, this study additionally established an interaction model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Method} + \beta_2 \cdot \text{Noise} + \beta_3 \cdot (\text{Noise} \times \text{Method}) \tag{2.27}$$

where:

- β_3 represents the interaction effect between noise level and listening method.

The interaction model allows in-situ and binaural conditions to have different noise slopes; in other words, the sensitivity of discrimination performance to changes in noise level may differ between the two listening methods.

If the interaction term is significant, it indicates that the two listening methods exhibit different trends in discrimination performance as noise level increases. In this study, the significance of the interaction term (`noise × method`) was used to determine whether the interaction model should be adopted.

When the interaction term was not significant ($p > 0.05$), it indicated that the two listening methods exhibited similar trends across noise levels; therefore, the Main Effects Model was adopted.

When the interaction term was significant ($p < 0.05$), it indicated that the two listening methods responded differently to changes in noise level; therefore, the Interaction Model was adopted.

3. Method

3.1. Experiment Overview

This chapter mainly focuses on whether people can make consistent subjective judgments of loudspeaker reproduction characteristics (or loudspeaker timbre) when using different listening methods, namely direct listening and binaural playback. More specifically, we investigate whether listeners produce similar ranking results for loudspeaker playback under these two listening conditions when answering different evaluation questions. To study this, listening experiments were designed to compare the ranking results obtained under both conditions.

Different from previous studies, this work does not aim to fully reproduce the real listening experience. For example, techniques such as head tracking or spatial scanning were not included. Instead, the study focuses on a more fundamental question: if the timbre information of a loudspeaker can be captured through binaural signals, can listeners still make evaluations similar to those made during direct listening? In addition, this chapter also consider more complex environments with background noise. In this case, we investigate whether listeners are still able to distinguish between different loudspeakers, and under what signal-to-noise ratio (SNR) conditions the differences between loudspeakers become difficult to perceive

Based on this, we further attempt to describe these perceptual differences using a more engineering-oriented model. Specifically, four different types of audio were selected, with each type corresponding to one perceptual parameter. We then investigated whether these perceptual parameters could be related to the spectral characteristics of the loudspeakers. In other words, after verifying that binaural recordings can to some extent replace direct listening evaluation, we analyze the spectral features of the binaural signals to better understand the basis of listeners' perceptual judgments. In this way, a relationship can be established between the spectral characteristics of loudspeaker playback and subjective perception. The purpose of this work is to reduce the dependence on large numbers of listening tests during the early stage of loudspeaker design, and to help loudspeaker designers quickly determine whether their tuning direction is appropriate.

To further verify whether this relationship is valid, controlled spectral modifications were applied to the recorded signals, and the signals were processed based on the proposed model. Afterwards, the rating trends predicted by the model were compared with the actual ranking trends obtained from listeners in the listening experiments for the modified audio samples. The experimental results show that, under the modulation conditions guided by the model, the changes in listeners' ranking results were generally consistent with the model predictions. This indicates

that the proposed model is able, to some extent, to reflect how changes in loudspeaker spectral characteristics influence subjective evaluations under specific listening questions.

Overall, the work in this chapter shows that, under the current experimental conditions and processing framework, binaural playback can preserve the evaluation trends of direct listening to a certain extent. At the same time, the spectral analysis method based on binaural signals is also able to describe the relationship between subjective perception and spectral changes, and can provide a certain level of prediction for changes in evaluation trends. Therefore, this method provides a possible way to reduce the need for large numbers of listening tests during the early stages of development. However, it should be noted that these conclusions are all based on specific experimental settings. The main purpose of this work is to verify the feasibility of the method itself, rather than to demonstrate its general applicability under broader conditions.

3.2. Experiment environment

3.2.1. Recording Room A

This room exhibited clear frequency-dependent reverberation characteristics. The reverberation time was unevenly distributed across frequencies. In the low-frequency range, energy decayed more slowly, resulting in stronger energy retention and accumulation effects. In the mid- and high-frequency ranges, the decay was relatively faster, although a moderate level of reverberation was still present overall. These acoustic characteristics can affect different types of audio signals in different ways. For speech signals, longer reverberation times may mask consonant transients and increase temporal overlap between syllables, which can reduce speech clarity. For music signals, reverberation can enhance sound field continuity, spatial envelopment, and overall spatial perception. Considering both its energy decay characteristics and spatial response structure, this environment can be classified as an indoor sound field dominated by reverberation.

3.2.2. Recording Room B

This laboratory is an indoor space with basic acoustic treatment. The walls are covered with a thin felt material, which provides some attenuation of mid- and high-frequency reflections (over 500 Hz). However, the low-frequency absorption performance is limited, meaning that the room does not meet free-field conditions overall. Environmental measurements showed that the low-frequency range was strongly affected by external environmental noise, such as nearby laboratories and structure-borne vibration, resulting in noticeable frequency fluctuations and instability. In the mid-frequency range, room modes and reflection interference effects were still present. Although the high-frequency range showed relatively better attenuation, reflections were not completely eliminated. Overall, this environment can be considered a semi-controlled sound field with weak sound absorption and limited sound isolation. It remains sensitive to external noise and has certain limitations for precise low-frequency measurements.

Although the two experimental rooms differed in terms of reverberation characteristics, background noise, and spatial response, both environments were closer to ordinary real-world indoor listening spaces rather than ideal free-field or anechoic conditions. Therefore, the results of this study mainly reflect subjective evaluation trends under realistic listening conditions, which may increase the practical relevance of the findings for real application scenarios.

3.2.3. Playback Room

In the binaural listening experiments, a quiet, well-lit, and acoustically isolated room was used to ensure that the participants felt comfortable during the tests and were not affected by external noise.

3.3. Audio Signal Selection

This study selected representative types of audio signals from three aspects: spectral characteristics, temporal structure, and signal semantics, in order to construct the audio stimulus set for the subjective evaluation experiments. Specifically, the selected signals included low-frequency-dominant signals, broadband signals, modulated signals, and speech signals.

From the spectral perspective, as shown in the Figure 3.1, these signals cover a wide frequency range from low to high frequencies and include different spectral distribution characteristics such as narrowband, broadband, and multi-harmonic structures. From the temporal perspective, both steady-state signals and non-stationary signals with clear modulation characteristics were included. From the semantic perspective, the signals represented several typical application types, including music signals, warning-type signals, and speech signals. Through this multi-dimensional combination, effective coverage of different acoustic characteristics could be achieved with a limited number of stimuli.

From the perspective of auditory perception, these four types of signals were used to represent or evoke different subjective perceptual dimensions, including low-frequency perception, timbre naturalness and spectral balance, urgency-related characteristics, and speech clarity. This provides the basis for establishing the relationship between spectral characteristics and subjective evaluation in the later analysis.

3.4. Equipment list

3.4.1. Selection of Loudspeakers

Loudspeaker A

Loudspeaker A has a frequency response range of 60 Hz to 20 kHz, with a nominal coverage angle of 140° horizontally and 90° vertically. It uses a two-way design consisting of a 4-inch low-frequency driver and a 1.25-inch high-frequency driver.

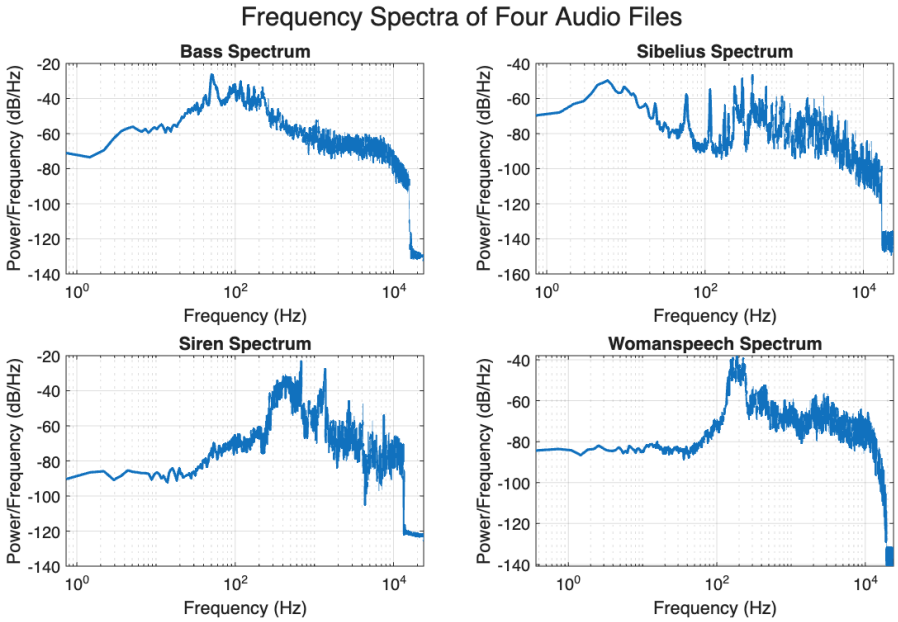


Figure 3.1.: Frequency Response of 3 Loudspeakers

In terms of directivity characteristics in Figure 3.2, the loudspeaker provides relatively wide coverage in the low- to mid-frequency range, resulting in relatively small differences between on-axis and off-axis responses. As the frequency increases, the directivity gradually becomes narrower. This effect is especially noticeable in the range of approximately 5 kHz to 10 kHz, where the high-frequency radiation becomes more directional and shows some irregular behavior.

Because of the two-way structure, driver interference can also occur around the crossover frequency, particularly in the vertical direction. This characteristic may reduce the spatial response uniformity in the vertical plane, making the high-frequency performance more sensitive to changes in listening position.

Overall, this loudspeaker shows the typical directivity characteristics of a two-way system: relatively wide coverage at low and mid frequencies, together with stronger directionality at high frequencies.

Loudspeaker B

Loudspeaker B is a horn-loaded loudspeaker with a frequency response range of 280 Hz to 12.5 kHz. Its nominal coverage angle at 2 kHz is 70° horizontally and 100° vertically. The main characteristic of this loudspeaker is the constant directivity behavior provided by its horn structure. Above approximately 2 kHz, the horizontal and vertical coverage patterns remain relatively stable across a wide frequency range, giving the loudspeaker strong control over sound coverage.

Table 3.1.: Selected Audio Signals and Corresponding Perceptual Tasks

Name	Type	Perceptual Description	Audio Source
bass	Low-frequency-dominant music signal	How deep the bass is?	Bass clip from the music "Factory of Faith"
music	Broadband music signal	How naturalness the music is?	Violin clip from Jean Sibelius
siren	Modulation signal, alarm signal	How urgency the alarm is?	Normal siren signal
speech	Speech signal	How clarity the speech is?	Female speech samples from the game "Elden Ring"

In terms of directivity distribution, as reference in Figure 3.3, The horizontal coverage is relatively stable, while the vertical direction shows more obvious lobing and null structures. This indicates that the response varies significantly across different vertical angles.

Overall, this loudspeaker exhibits the typical directivity characteristics of a horn-loaded system: concentrated coverage, strong directionality, and relatively high sensitivity to changes in vertical listening position.

Loudspeaker C

Loudspeaker C uses a single-driver design with a frequency response range of 200 Hz to 16 kHz. Its nominal coverage angle at 2 kHz is approximately 130°.

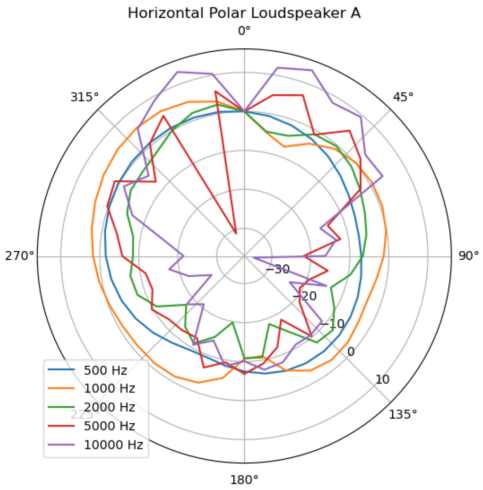
In terms of directivity characteristics shows in Figure 3.4, this loudspeaker shows the typical natural directivity behavior of a single-driver system. In the low-frequency range, the radiation pattern is close to omnidirectional. In the mid-frequency range, it provides relatively wide and smooth coverage, resulting in more continuous spatial response changes overall. In the high-frequency range, the directivity gradually becomes narrower as frequency increases, making the high-frequency response more sensitive to off-axis listening positions.

Since the loudspeaker does not use a crossover structure, there are no obvious multi-driver interference effects. As a result, the response consistency between the horizontal and vertical directions is relatively good.

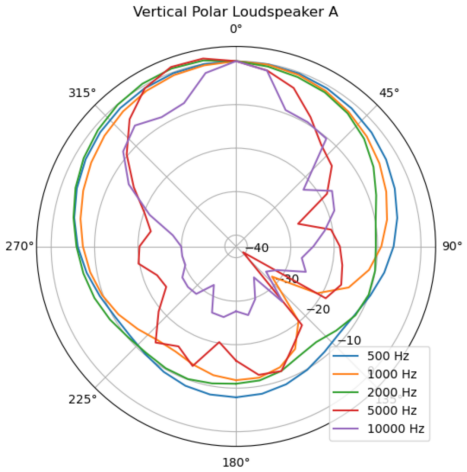
Overall, this loudspeaker exhibits relatively natural and continuous directivity characteristics.

Spectral Characteristics

From the overall spectral characteristics, as shown in the Figure 3.5, the three loudspeakers exhibit clear differences in their frequency response distributions. Loudspeaker A provides the

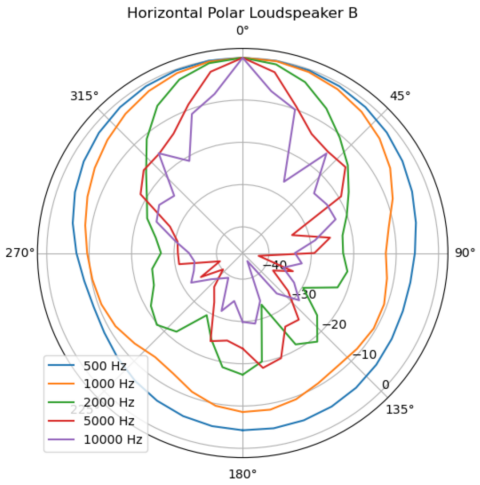


(a) Horizontal

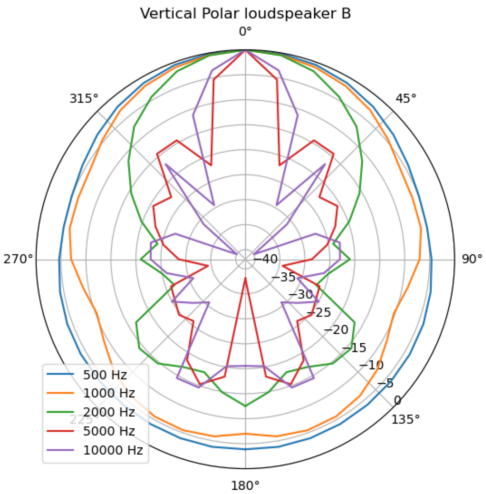


(b) Vertical

Figure 3.2.: Directivity of Loudspeaker A

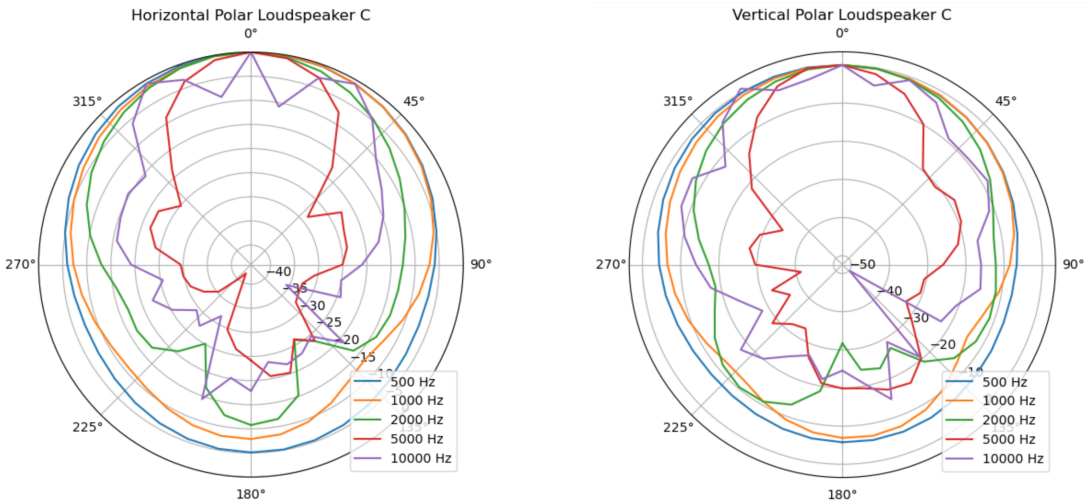


(a) Horizontal



(b) Vertical

Figure 3.3.: Directivity of Loudspeaker B



(a) Horizontal

(b) Vertical

Figure 3.4.: Directivity of Loudspeaker C

widest frequency coverage, maintaining relatively high energy in the low-frequency range while also showing a relatively smooth extension in the mid- and high-frequency regions. Loudspeaker B has the weakest low-frequency extension, with most of its energy concentrated in the mid-frequency range, while its high-frequency response shows more noticeable fluctuations. In comparison, Loudspeaker C exhibits relatively stronger energy in the mid- and high-frequency regions, although its low-frequency extension is relatively limited.

Overall, the three loudspeakers show clear differences in low-frequency extension, mid-frequency energy distribution, and high-frequency response characteristics. These differences provide the basis for the later subjective ranking experiments on spectral perception.

3.4.2. Recording Devices

B1-E Dummy Head with BE-P1 Microphone

This microphone(hereafter referred to as the “Dummyhead”) was used in all binaural recording experiments in this study. The recording device employed a fibreglass dummy head with dimensions similar to a human head, together with silicone artificial ears designed to resemble the shape and material properties of human ears. This setup was intended to simulate human head-related transfer function (HRTF) characteristics to a certain extent. Detailed specifications of the dummy head and the microphones are provided in the Appendix.

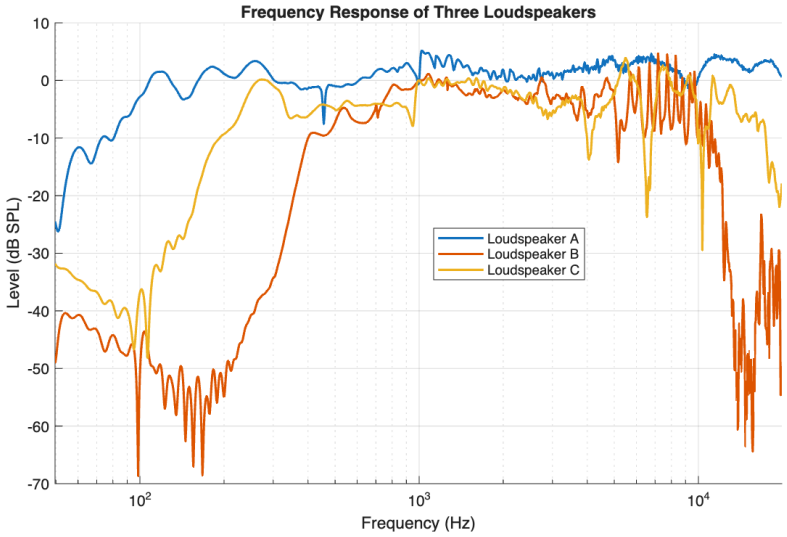


Figure 3.5.: Frequency Response of 3 Loudspeakers

Bruel And Kjaer 4128-C HATS

This microphone(hereafter referred to as the "HATS") was used for recording the noise-free audio samples in Room B

Audio Sound card

Audio recording and playback in this study were carried out using the MOTU M2 audio interface from MOTU.

3.4.3. Playback Devices

The Sennheiser HD 650 headphones (hereafter referred to as "HD650") are open-back, circumaural dynamic headphones commonly used for music production, monitoring, and subjective listening evaluation.

Compared with closed-back headphones, open-back headphones do not form a fully enclosed acoustic space behind the driver. The sound radiation from the rear side of the diaphragm can be released through the open structure, which helps reduce spectral coloration caused by pressure buildup, standing waves, and cavity reflections inside the earcups.

From the perspective of the Headphone Transfer Function (HPTF), the transfer characteristics of open-back headphones are mainly determined by the driver response, the coupling between the earpads and the front cavity, pinna scattering, and ear canal geometry, while being less influenced by pressure modes in a closed rear cavity. Therefore, under controlled listening conditions, open-back headphones such as the HD650 can usually provide a more natural and less colored acoustic

coupling between the headphones and the ears. This helps reduce the influence of the headphones themselves on timbre, spatial perception, and transient characteristics, allowing listeners to focus more on evaluating the experimental audio material itself.

The frequency response range of the HD650 is approximately 10 Hz–39.5 kHz. In this study, no additional EQ was applied to the headphones. The same headphones were used both during the recording process and in the binaural listening experiments.

3.4.4. Software

Audacity software

Audacity is an open-source audio processing software that can be used for audio recording, editing, and basic signal processing. In this study, Audacity was mainly used for binaural signal recording, audio segment trimming, and audio file organization.

webMUSHRA Software

WebMUSHRA is a web-based subjective listening test platform that supports several types of listening experiments, including MUSHRA, ranking, and rating tests. In this study, WebMUSHRA was mainly used to build the listening test workflow, randomize the playback order of audio samples, and collect listeners' subjective evaluation results.

Other software

Microsoft Excel was used for data organization and analysis.

MATLAB and Python were used for signal processing, visualization, and AI-assisted analysis.

ChatGPT was used for language polishing, figure generation, and LaTeX formatting support.

3.4.5. Background Noise Generation System

The background noise loudspeaker array consisted of four Genelec loudspeakers and was used for background noise playback during both the direct listening experiments and the audio recording process with background noise. The loudspeakers were positioned around the listening location in order to create a relatively uniform diffuse noise field. Before the experiments, the playback sound pressure level of the background noise was calibrated to ensure a consistent listening level across all experimental conditions.

3.5. Experimental Design Concept

Based on the objectives described above, the experiment adopted a multi-stage design to gradually verify the effectiveness of binaural playback and establish the relationship between acoustic characteristics and subjective evaluation.

First, in a real sound field environment, four selected audio signals were played through three loudspeakers with clearly different spectral characteristics. Participants were asked to rank the playback results of the same audio reproduced by different loudspeakers according to predefined perceptual questions. These original subjective ranking data were used to obtain the evaluation results under real listening conditions.

Next, binaural recordings were made using the same setup as the direct playback condition, and the first binaural playback listening experiment was carried out based on these recordings. By comparing the ranking results between direct listening and binaural playback conditions, we examined whether listeners could maintain consistent ranking trends under binaural playback, thereby verifying the feasibility of using binaural recordings as the basis for later analysis.

After confirming the reliability of binaural playback, spectral feature analysis was further applied to the binaural signals in order to extract key acoustic features related to changes in ranking results. This step was intended to explore the relationship between loudspeaker spectral response characteristics and subjective perception.

In addition, to investigate the influence of environmental noise on loudspeaker timbre perception, the direct listening experiments were repeated under background noise conditions with different signal-to-noise ratios (SNRs). While keeping the audio content and loudspeaker conditions unchanged, the experiments examined whether listeners could still distinguish differences between loudspeakers in noisy environments, and attempted to estimate the threshold at which listeners could no longer perceive differences in loudspeaker timbre. Binaural listening experiments were then also conducted using the recordings made under noisy conditions.

Finally, based on the previous analysis results, model-based spectral modulation was applied to the binaural recordings, followed by additional binaural playback listening experiments. By comparing the ranking results before and after modulation, as well as under different noise conditions, the study further evaluated whether the proposed relationship between acoustic features and subjective perception could effectively explain the observed changes in ranking trends.

Overall, through the multi-stage framework of "direct listening – binaural playback validation – feature extraction – noise disturbance – model verification" which is shown in Figure 3.6, this study systematically investigated the relationship between loudspeaker spectral characteristics and subjective perceptual ranking, while also evaluating the stability and interpretability of this relationship under different listening conditions.

3.5.1. Quantitative Evaluation and Decision Criteria

This study used a pairwise comparison method to analyze the subjective ranking results obtained under both in-situ listening and binaural playback conditions. When the pairwise preference ratio exceeded 70%, the loudspeaker pair was considered to show a stable preference tendency. In addition, this study used $\Delta P \leq 0.15$ as the acceptable perceptual consistency criterion to determine whether binaural playback could preserve the perceptual preference tendency observed

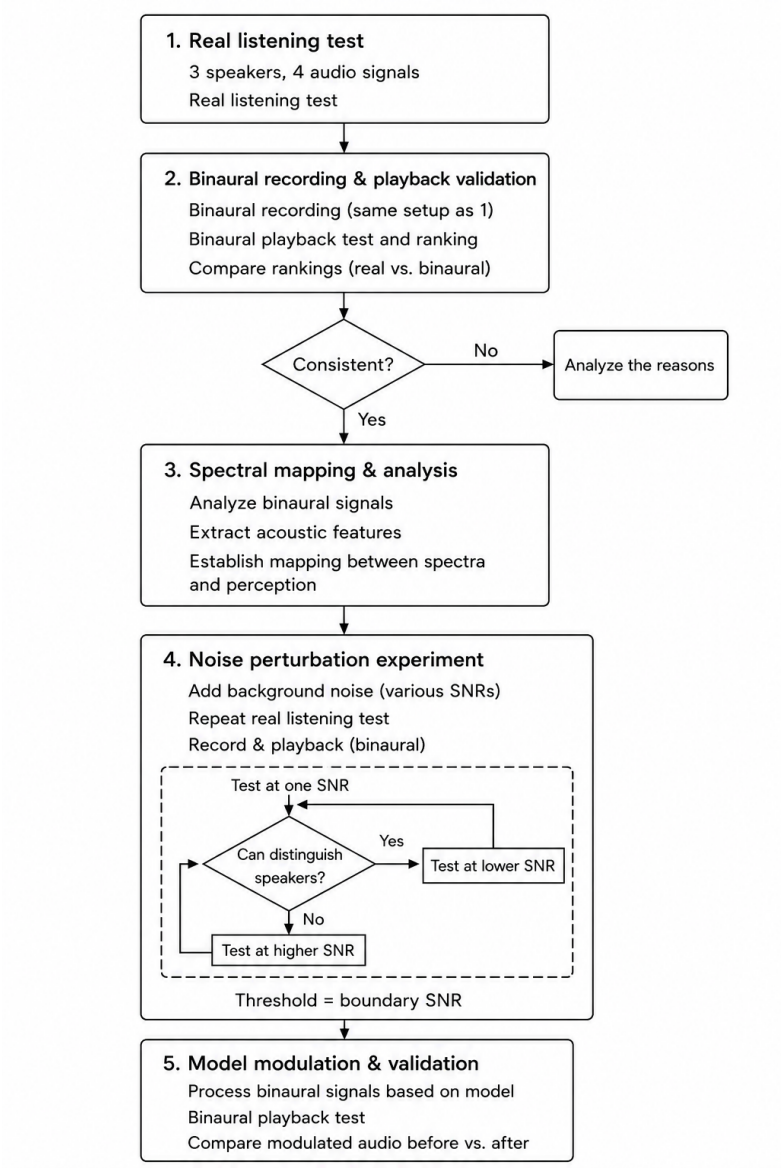


Figure 3.6.: Experiment Design Concept

under direct listening conditions. The related definitions and theoretical background have already been introduced in the Theory section.

3.5.2. Experimental Procedure for Subjective Listening Tests

The subjective listening experiments in this study were designed to analyze the consistency of loudspeaker perception under different playback conditions and to ensure the reproducibility and interpretability of the results.

All subjective evaluation tasks were divided into two categories: rating tasks and ranking tasks. In rating tasks, participants independently rated each loudspeaker according to the given perceptual criteria, using a predefined continuous scale (e.g., 1–9 or 1–100), a method consistent with standardized auditory evaluation procedures defined in ITU-R BS.1534 [ITU 01] and related psychoacoustic scaling methodologies. In ranking tasks, participants ordered multiple loudspeakers from best to worst based on overall perceived sound quality, which is commonly analyzed using inter-rater agreement frameworks such as Fleiss' kappa and related categorical agreement metrics [Fle 71][Lan 77].

In the direct listening experiments, the interface was presented using paper-based questionnaires (designed from Google Forms). In the binaural playback experiments, the interface was implemented using the WebMUSHRA platform, which is based on ITU-R BS.1534 recommendations for multi-stimulus continuous rating assessment. Before the experiments began, participants were given detailed instructions to ensure consistent understanding of the tasks across different experimental conditions, in line with established loudspeaker subjective evaluation methodologies [Too 88].

The sample of the listening test are shown in Figure 3.11. The detailed experimental procedures are provided in the appendix.

3.5.3. Definitions and Notation Standardization

To avoid conflicts with standard terminology in existing literature, this study defines the subjective evaluation metrics corresponding to the four types of audio signals as follows:

Bass Definition

In this study, "bass" is operationally defined as a deep perception dimension that reflects listeners' comparative judgments of low-frequency performance between loudspeakers, specifically in terms of perceived low-frequency extension and low-frequency energy strength.

This definition is grounded in established psychoacoustic and room acoustics research, which consistently demonstrates that perceived bass strength is primarily governed by the level and spectral distribution of low-frequency energy rather than frequency content alone. Bradley and Soulodre showed that subjective bass perception in concert halls is strongly correlated with

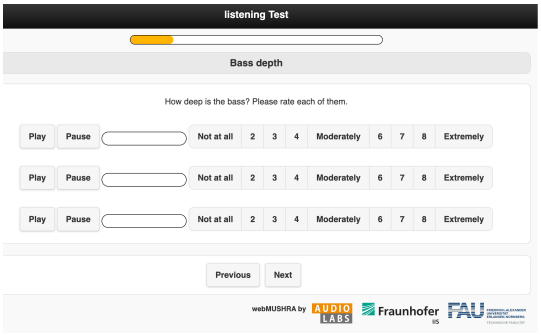


Figure 3.7.: Sample 1

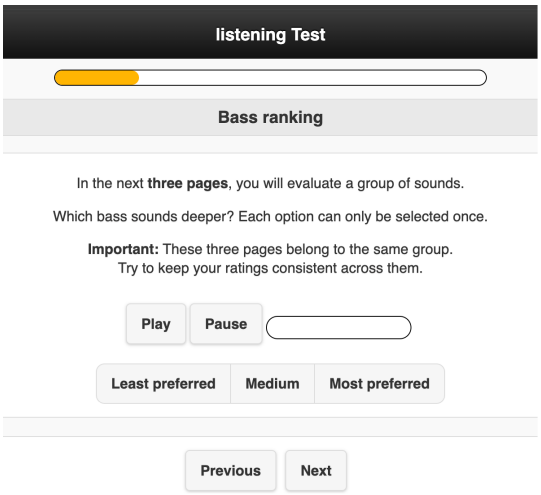


Figure 3.8.: Sample 2

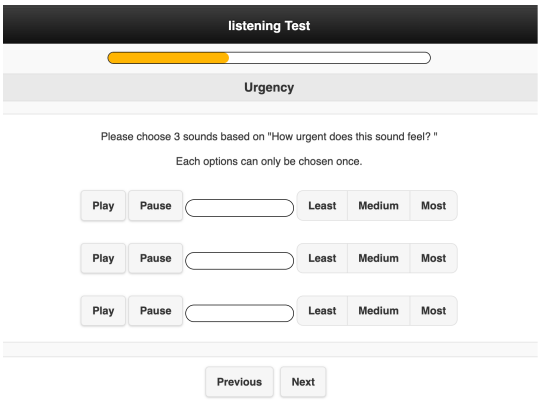


Figure 3.9.: Sample 3

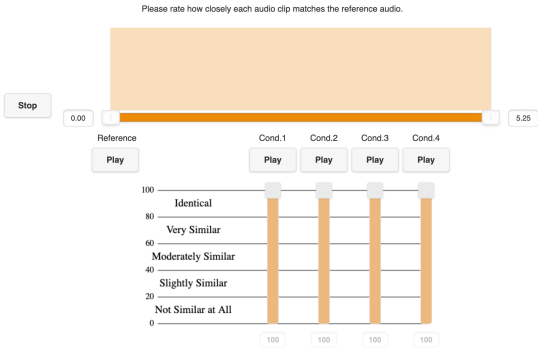


Figure 3.10.: Sample 4

Figure 3.11.: Listening Test Surface Samples

early-arriving low-frequency energy levels, while traditional reverberation-time-based metrics exhibit weak predictive power for perceived bass strength [Bra 97].

Furthermore, research in concert hall acoustics has demonstrated that variations in low-frequency energy distribution over time and space significantly affect perceived bass strength, particularly through early and late arriving sound components [Bar 03]. Paired-comparison listening experiments further confirm that perceived bass strength is a robust perceptual attribute that can be consistently elicited across listeners when comparing loudspeaker or room responses with controlled low-frequency spectral differences [Tah 14].

Therefore, the bass metric used in this experiment should be interpreted as a perceptual construct rooted in low-frequency spectral energy distribution and its perceptual salience, rather than as a simple frequency-band descriptor. This ensures that the definition is theoretically grounded, experimentally validated, and free from ambiguity in interpretation.

Naturalness Definition

In this study, "music naturalness" is defined as a perceptual attribute describing the overall spectral balance and perceived realism of loudspeaker reproduction when presenting broadband musical signals. It reflects listeners' global judgment of whether the reproduced sound exhibits a smooth and well-balanced spectral distribution without noticeable frequency-band distortions or local spectral emphasis.

This definition is grounded in established psychoacoustic and loudspeaker evaluation literature, which shows that perceived sound quality and listener preference are strongly associated with spectral smoothness and frequency-response linearity [Oli 04]. Toole further demonstrated that deviations in spectral balance introduce audible coloration, reducing perceived naturalness in reproduced sound systems [Too 08].

Additionally, auditory perception research indicates that timbral naturalness is fundamentally derived from spectral structure cues embedded in complex audio signals [Bla 97]. Therefore, music naturalness in this work is interpreted as a spectral balance–based perceptual judgment of overall audio fidelity rather than a single-frequency or isolated-band descriptor.

Urgency Dfinition

"Urgency" in this study refers to a perceptual attribute describing listeners' subjective judgment of the level of urgency and alertness elicited by siren-like alarm sounds, reflecting the perceived intensity of psychological emergency induced by the auditory stimulus.

This construct is grounded in psychoacoustic research on auditory warning design, which demonstrates that perceived urgency is systematically influenced by acoustic features such as intensity, temporal modulation, and spectral roughness [Haa 08].

Alarm sound design studies further indicate that urgency perception is strongly associated with aggressive and high-arousal acoustic characteristics, suggesting a robust mapping between acoustic parameters and perceived emergency level [Edw 01].

Additionally, auditory scene analysis and attention-based auditory processing theories suggest that salient acoustic events are preferentially detected and prioritized by the auditory system, contributing to heightened perceived urgency in complex auditory environments [Bre 90][Fri 07].

Clarity Definition

"Speech" signals in this study are uniformly defined as female speech signals. In this context, the term "woman speech" is used interchangeably with "female speech" throughout the study to avoid ambiguity and ensure consistent terminology. This definition serves as an experimental control to guarantee comparability across all experimental conditions and to eliminate variability arising from speaker-related acoustic differences.

The "clarity" metric refers to the perceived intelligibility and perceptual transparency of the speech signal, reflecting the listener's ability to perceive speech content without degradation or masking effects.

This construct is grounded in psychoacoustic research showing that speech clarity is strongly dependent on the preservation of temporal envelope structure and spectral modulation cues, where distortions in these domains lead to reduced intelligibility [Drul 94].

Furthermore, auditory perception studies indicate that speech understanding relies on modulation-based representations of acoustic signals, suggesting that clarity emerges from the fidelity of spectro-temporal structures in the auditory system [Sin 03][**?**].

3.6. In-situ and Binaural Playback Consistency Experiment

In this study, 3 loudspeakers were used to play the same audio in the same room, and all recordings were made using the same recording chain. The evaluation results of listeners under direct listening and recorded listening conditions were then compared to examine whether consistent evaluation trends could be observed across the two conditions. If such consistency can be verified, it would indicate that even under simplified recording conditions, binaural recordings are still able to preserve the key acoustic cues that influence subjective evaluation. This would further suggest that, under the proposed recording method, binaural listening tests can serve as a substitute for direct listening tests.

3.6.1. Audio Recording Part I

To verify the robustness of the proposed recording method across different rooms and recording devices, recordings were carried out in both Room A and Room B using both the HATS and the dummy head systems. The Dummy Head and HATS were calibrated before the experiments began.

The experimental setup is shown in Figure 3.12. The three loudspeakers were placed in front of the Dummy Head/HATS, with the acoustic center of each loudspeaker aligned with the acoustic

center of the binaural microphones. The distance between each loudspeaker and the Dummy Head/HATS was fixed at 1 m.

Before the formal recordings, pink noise was used for sound pressure level(SPL) calibration. A reference microphone and sound level meter were used to measure the SPL of each loudspeaker at the reference position. The output gain of each loudspeaker was then adjusted so that all loudspeakers produced the same SPL when playing the same audio signal, while also keeping the playback level within a comfortable listening range.

During recording, the MOTU M2 audio interface was used, and the input gain of the recording chain remained unchanged throughout the experiments to ensure consistency between recordings. Considering that the binaural recording system itself introduced some background noise, the loudspeaker playback level was increased as much as possible without causing recording clipping, while the recording gain of the audio interface was kept as low as possible. This was done to obtain recordings with a higher signal-to-noise ratio.

All recordings were made at a sampling rate of 48 kHz using 32-bit float format. Before the formal recordings, calibration between dB SPL and digital recording level (dBFS) was carried out. Using pink noise tests, the recording level was controlled within approximately -18 dBFS to -12 dBFS in order to balance dynamic range usage and clipping headroom.

After recording, the Sennheiser HD 650 headphones were connected to the output of the audio interface and mounted on the Dummy Head for binaural playback calibration. By adjusting the playback gain in Audacity, the headphone playback SPL measured by the sound level meter was matched to the SPL of the original loudspeaker playback condition. This ensured consistency in playback loudness between the binaural playback condition and the direct listening condition.

The loudness of all recordings was measured using LUFS-based loudness calculation. The results are shown in Table 3.2, in order to ensure consistent perceived loudness across the recordings obtained using this method.

Table 3.2.: LUFS Loudness Measurements of Different Audio Samples in Two Rooms

Room	**Audio**	**Loudspeaker A**	**Loudspeaker B**	**Loudspeaker C**
Room A	bass	-46.281	-46.276	-47.615
	music	-49.382	-50.370	-50.461
	siren	-42.294	-42.649	-42.893
	woman speech	-43.658	-44.016	-44.755
Room B	bass	-59.904	-56.988	-56.613
	music	-61.927	-58.384	-58.023
	siren	-56.185	-55.922	-52.218
	woman speech	-56.552	-53.230	-53.670

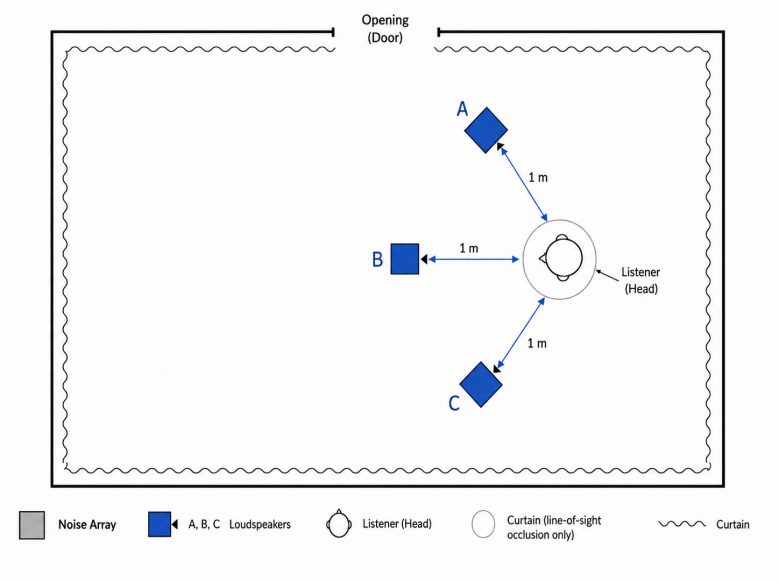


Figure 3.12.: Layout of the in-situ listening experiment. Three loudspeakers (A, B, and C) were positioned approximately 1 m from the listener. Curtains were placed around the room to reduce direct visual identification of the loudspeakers while maintaining the original room acoustic condition. The listener remained at a fixed position during the experiment.

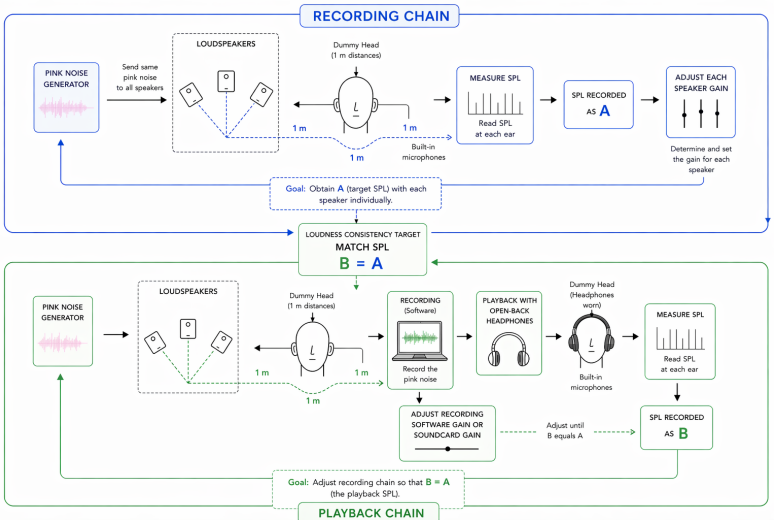


Figure 3.13.: Recording Chain Diagram

3.6.2. In-situ Listening Test Part I

The direct listening experiments were conducted in Room A. The experimental setup is shown in Figure X. The three loudspeakers were placed at the same height, with their acoustic centers facing the listener. The distance between the listener and the loudspeakers was fixed at 1 m. An acoustically transparent but visually opaque curtain was placed between the listeners and the loudspeakers in order to conduct the experiment under blind listening conditions. During the experiment, the same audio sample was played through different loudspeakers in varying orders by the experimenter. While the audio was being played, the listener was asked to face the loudspeaker currently producing sound in order to reduce the influence of loudspeaker directivity and the listener's HRTF on the evaluation results. For each audio sample, after listening to a single loudspeaker playback, the listener was required to provide a subjective rating. After all loudspeakers had been evaluated individually, the listener further performed a relative ranking among all loudspeakers. The detailed experimental setup is shown in Figures 3.14.

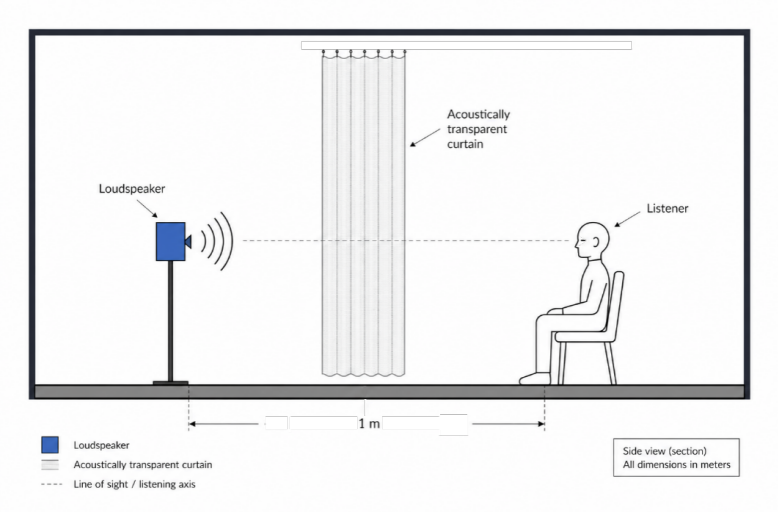


Figure 3.14.: Side-view of the in-situ listening setup. The loudspeaker acoustic center was aligned with the listener's ear position at a distance of approximately 1 m. An acoustically transparent curtain was used to block visual identification of the loudspeaker, enabling a double-blind listening condition.

3.6.3. Binaural Listening Test Part I

In the binaural listening tests, all binaural recordings were played back at a sampling rate of 48 kHz using 32-bit float format. Playback was carried out using the MOTU M2 audio interface and the Sennheiser HD 650 connected to a computer. The listening test questionnaire was designed using WebMUSHRA. The procedure was consistent with the direct listening experiment. For

each audio sample, after listening to playback from a single loudspeaker, the participant was asked to provide a subjective rating. After all loudspeakers had been evaluated individually, the participant further performed a relative ranking among all loudspeakers. For each question, the audio samples corresponding to the three loudspeakers were presented in random order to minimize the influence of playback order on the test results.

The binaural listening experiments included recordings made in different rooms and with different recording systems, including recordings made in Room A using the dummy head and recordings made in Room B using the HATS. This design was intended to examine whether the proposed recording method could still be applied under different room conditions and recording devices. After the experiments, short interviews were also conducted with the participants.

3.7. Noise Condition Experiment

To determine the applicability of the proposed recording method under noisy conditions, this experiment investigated whether listeners could still distinguish loudspeaker timbre during both binaural listening and in-situ listening, as well as the background noise threshold at which differences in loudspeaker timbre could no longer be perceived.

3.7.1. In-situ Listening Test Part II - With Background Noise

The direct listening experiments with background noise were conducted in Room A. During the experiments, the Background Noise Generation System was used to create a background noise field and play everyday environmental noise, which is shown in Figure 3.15 and Figure 3.16. All other experimental settings remained the same as those used in the noise-free experiments(3.6.2).

Under clean listening conditions, participants were generally able to perceive subjective differences between loudspeakers when reproducing the same audio signal. Therefore, subjective listening experiments were further carried out under background noise conditions in order to investigate the perceptual discriminability of different listening tasks in noisy environments.

Because participants were aware of which loudspeaker was playing during the direct listening experiments, it was difficult to reliably quantify perceptual ability using simple binary tasks such as "whether a difference could be perceived." For this reason, the present study adopted a ranking task as an indirect measurement method. By observing whether participants could maintain the same ranking trends as those obtained under clean audio conditions, it was possible to determine whether they could still stably perceive subjective differences between loudspeakers.

During the experiments, if a participant could still produce the same ranking trend as in the clean audio condition under the current noise level, it was considered that the participant could still reliably distinguish perceptual differences between loudspeakers, and the background noise level was gradually increased. Otherwise, the background noise level was reduced.

Finally, the background noise level immediately before the participant first failed to reliably distinguish perceptual differences between loudspeakers was defined as the perceptual discrim-

ination threshold for that perceptual task. This threshold was used to represent the perceptual limit for discriminating loudspeaker differences under background noise conditions.

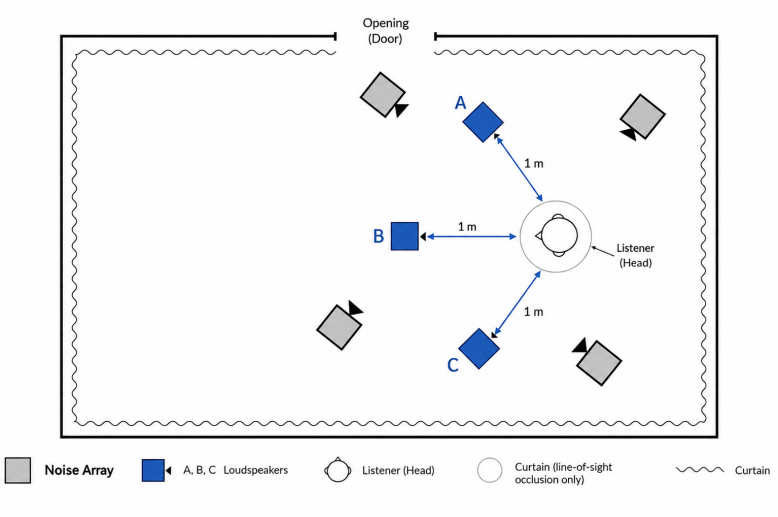


Figure 3.15.: Layout of the background noise listening experiment. Three loudspeakers (A, B, and C) were positioned approximately 1 m from the listener, while additional loudspeakers were arranged around the room as a noise array to generate the background noise sound field. Curtains were used to block direct visual identification of the playback loudspeakers during the experiment.

3.7.2. Audio Recording Part II

Based on the perceptual thresholds obtained from the direct listening experiments under background noise conditions, binaural recordings of noisy audio samples at different background noise levels were made using the same binaural recording chain as in the previous experiments(3.6.1), while keeping the experimental setup consistent with the direct listening experiments with background noise.

3.7.3. Binaural Listening Test Part II

The binaural listening experiments were conducted by the same participants who took part in the direct listening experiments. Considering that preference-based ranking results tend to become unstable when the discriminability between different loudspeakers decreases, an additional ABX test was introduced to further verify loudspeaker discriminability.

In the experiment, the clean-audio binaural recording of one loudspeaker was used as the reference. Participants were then asked to choose, from two noisy recordings with the same

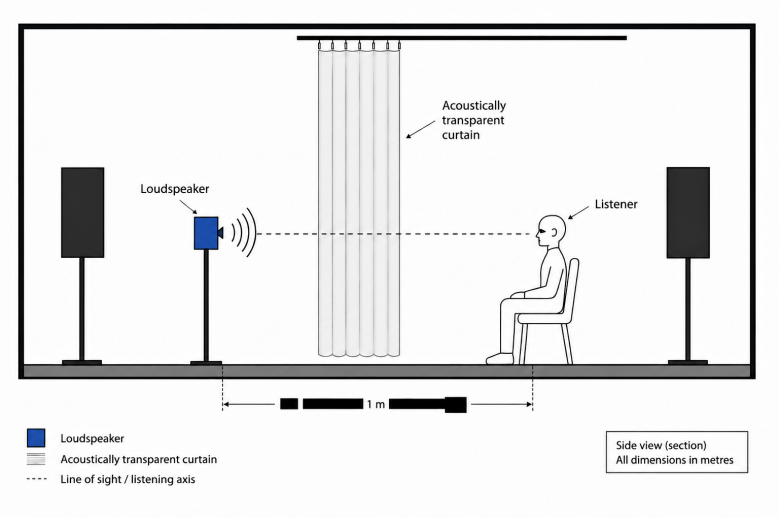


Figure 3.16.: Side-view of the in-situ listening setup with noise array

signal-to-noise ratio but reproduced by different loudspeakers, the audio sample that was more similar to the reference.

When the ABX test results were no longer significantly higher than the chance level(50% accuracy), it was considered that the participants could no longer reliably distinguish between different loudspeakers under that signal-to-noise ratio condition. The corresponding SNR was then regarded as a reference threshold at which the consistency of evaluation trends began to fail.

3.8. Spectral-Perceptual Modeling and Validation

After verifying the consistency of ranking results, this study further constructed a mapping model between acoustic features and subjective perception based on listeners' subjective rankings and the spectral information of the corresponding binaural signals. The main assumption of the model is that the spectral features contained in binaural recordings can, to some extent, represent listeners' subjective rankings of different loudspeaker playback results.

The modeling process in this study was based on binaural recordings obtained in real acoustic environments. These signals contain not only the original audio information, but also the combined effects of loudspeaker spectral characteristics, room acoustics, and the recording system. Therefore, they can be regarded as the acoustic output received before the sound enters the human ears.

Under this framework, the object analyzed in this study is not the ideal loudspeaker frequency response measured under controlled conditions, but the overall acoustic performance produced by the loudspeaker when operating in a specific acoustic space. Since loudspeakers in real applications are always coupled with room acoustics, analyzing them separately in anechoic

environments may reveal their intrinsic device characteristics, but cannot fully represent their subjective performance in real listening situations. Compared with traditional methods that focus on intrinsic loudspeaker characteristics measured in anechoic conditions, the proposed approach describes auditory input from the perspective of system-level acoustic output, which is more closely related to actual listening experiences.

It should also be noted that the spectral features obtained in this framework are the result of the combined effects of multiple factors, including the loudspeaker, the acoustic space, and the receiver position. Therefore, these features no longer correspond to the independent spectral response of a single loudspeaker. This means that the study does not attempt to separate the individual contributions of different physical factors, but instead focuses on how their combined effects influence human perception. In other words, this work investigates the relationship between system-level acoustic output and subjective listening perception in realistic usage scenarios.

Based on this framework, spectral energy analysis was applied to the binaural recordings in order to compare the actual acoustic differences produced by different loudspeakers under the same spatial conditions, and further examine how these differences influenced listeners' subjective ranking results. Combined with the subjective evaluation data, the extracted spectral energy features were then modeled to establish a mapping relationship between acoustic features and subjective perception, allowing the model to predict the corresponding perceptual ranking or preference for a given input signal.

Finally, the model outputs predicted subjective evaluation results for different loudspeaker playback conditions, including rankings or scores. It should be further emphasized that, because the model is built under specific acoustic conditions, its conclusions are mainly applicable to comparative analysis in the same or similar acoustic environments. However, the model still has more direct practical relevance for predicting real listening experiences.

3.8.1. Bass Perception Model

Model Construction

To quantify the perceived low-frequency strength of audio signals, this study defined a spectral-energy-based metric called the bass score. This metric was used to represent the overall energy level of the signal within the low-frequency range, allowing comparison and ranking between different loudspeakers.

From the perspective of human hearing, the audible frequency range is approximately 20 Hz to 20 kHz, while low frequencies are generally considered to cover the range from about 20 Hz to 200 Hz. However, under practical listening conditions, different low-frequency sub-bands do not contribute equally to subjective low-frequency perception. Specifically, frequency components below approximately 50 Hz are perceived with lower sensitivity by the human auditory system. In addition, in real room environments, this very low-frequency range is strongly affected by room modes and the playback limitations of audio equipment, resulting in relatively unstable

behavior. Therefore, the contribution of this frequency range to stable subjective ranking is relatively limited.

On the other hand, the frequency components most closely related to subjective low-frequency perception are mainly concentrated in the range of approximately 50–150 Hz. This range is usually associated with important perceptual dimensions such as bass extension and impact. In order to avoid truncating the transition region between low and low-mid frequencies, while still maintaining a more complete representation of the low-frequency structure, the upper frequency limit in this study was extended to 180 Hz.

Therefore, in the proposed model, the low-frequency analysis band is defined as:

$$F_{\text{bass}} = \{f \mid 50\,\text{Hz} \leq f \leq 180\,\text{Hz}\}$$

In the implementation process, the input audio signal was first resampled to 48 kHz and converted into a mono signal. The signal was then processed using the Short-Time Fourier Transform (STFT) to obtain its power spectrum representation:

$$P(f,t) = |\text{STFT}(y(t))|^2$$

where f represents frequency and t represents the time frame.

Within the defined low-frequency band, the energy across all time frames was averaged to obtain the bass score:

$$S_{\text{bass}} = \frac{1}{N_t N_f} \sum_{t=1}^{N_t} \sum_{f \in F_{\text{bass}}} P(f,t)$$

where N_t represents the number of time frames, and N_f represents the number of frequency bins within the selected frequency band.

This score represents the overall energy level of the signal within the low-frequency region. For the same audio content, the S_{bass} values obtained from different loudspeakers can be used for ranking comparison. A larger value indicates stronger low-frequency energy and corresponds to a more prominent bass perception. Table 3.3 shows the model-based scores of the recordings obtained when the bass audio was played by the three loudspeakers in Room A. It can be observed

Table 3.3.: Objective Evaluation Scores of Different Audio Clips

Audio Clip	**Score**
Audio_A	0.29613504
Audio_B	5.6923152×10^{-5}
Audio_C	0.0060145957

that, according to the model prediction results, the ranking order is $A > C > B$.

Signal Modification and Validation

After defining the low-frequency score metric S_{bass}, this study further applied targeted spectral modulation to the audio signals based on the proposed scoring model.

More specifically, given the input signal $y(t)$, the signal was first transformed using the Short-Time Fourier Transform (STFT) to obtain its complex spectrum representation:

$$S(f,t) = |S(f,t)|e^{j\phi(f,t)}$$

where $|S(f,t)|$ represents the magnitude spectrum and $\phi(f,t)$ represents the phase information.

According to the previous definition, the low-frequency score S_{bass} is determined by the energy within the frequency band

$$F_{\text{bass}} = [50, 180] \text{ Hz}$$

Therefore, during the modulation process, only the magnitude spectrum within this frequency band was scaled, while all other frequency components and the phase information were kept unchanged. The modulation operation was defined as:

$$|S'(f,t)| = \begin{cases} \alpha \cdot |S(f,t)|, & f \in F_{\text{bass}} \\ |S(f,t)|, & \text{otherwise} \end{cases}$$

where α is the low-frequency scaling coefficient ($0 < \alpha < 1$), which controls the attenuation level of the low-frequency energy.

The core idea of this modulation method is that it only modifies the spectral components involved in the calculation of S_{bass}. Therefore, the modulation directly affects the score itself. More specifically, when $\alpha < 1$, the overall energy within the low-frequency band decreases, resulting in a lower low-frequency score S_{bass}. In contrast, when enhancement is applied ($\alpha > 1$), the score increases accordingly.

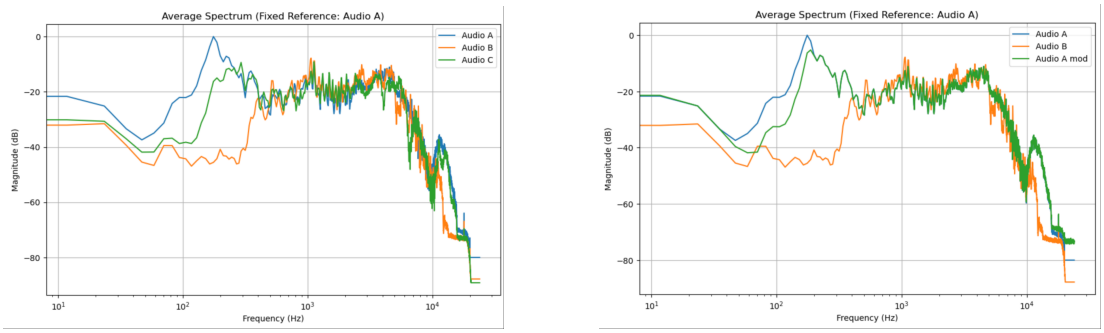
After the magnitude modulation was completed, the modified magnitude spectrum was recombined with the original phase information:

$$S'(f,t) = |S'(f,t)|e^{j\phi(f,t)}$$

Finally, the modified signal was transformed back into the time domain using the inverse Short-Time Fourier Transform (ISTFT), resulting in the modulated audio signal $y'(t)$.

The Figure 3.17 below shows the spectrogram of the bass recording before and after modulation.

Here, Audio A, B, and C refer to the recordings obtained when the audio was played through the three different loudspeakers. The suffix "mod" indicates the modulated version of the corresponding audio signal. For example, Audio A mod refers to the modulated version of Audio A. It should be noted that modulation was not applied to all audio samples in this study. Instead, the analysis mainly focused on the two audio samples that were perceptually more similar. In addition, after analyzing the spectra, it was found that Loudspeakers B and C contained almost



(a) Original Audio Spectrogram

(b) Modulated Audio Spectrogram

Figure 3.17.: Spectral Comparison Before and After Model-Guided Modification - Bass

no low-frequency energy below 200 Hz. Therefore, during the modulation process, it was not appropriate to artificially generate low-frequency components in frequency regions where no such spectral content originally existed.

As a result, the bass modulation was applied only to the recording from Loudspeaker A. If the modulated audio was later evaluated as perceptually worse in the listening experiments, the proposed model would be considered valid under this condition.

3.8.2. Music Perception Model

Model Construction

In this study, the perceptual metric defined for broadband signals represented by music signals is referred to as naturalness. From the perspective of human perception, naturalness means that the audio energy is relatively balanced across different frequency ranges, preserving sufficient detail while avoiding harsh high-frequency components.

From the perspective of spectral analysis, this metric is defined based on the spectral difference between the recorded signal and the original reference audio. The score is defined as the degree to which the loudspeaker changes the original timbral structure of the audio signal. In other words, the less the recording differs from the original audio, the more natural the perceived listening experience is considered to be.

Human perception of timbre mainly depends on the energy distribution across different frequency regions, rather than instantaneous phase information or fine spectral details. Therefore, in this study, the spectral envelope was used to represent the tonal balance and timbral characteristics of the sound. If the spectral envelope of the recorded signal remains close to that of the original audio across different frequency bands, it indicates that the loudspeaker preserves the original timbre relatively well, and therefore provides higher perceived naturalness.

To make the feature representation more consistent with human auditory perception, the Bark scale was used for spectral band division in this study. Since human frequency perception is

not linear, but is instead organized more closely according to critical bands, the Bark scale can describe human perception of spectral balance more accurately than linear frequency spacing.

To quantify the "tightness" and "pleasantness" of audio signals, the original audio signal was first transformed into quantifiable spectral features. The audio signal $y(t)$ was processed using the Short-Time Fourier Transform (STFT) to obtain its frequency–time magnitude spectrum:

$$S(f,t) = |\mathrm{STFT}(y(t))|$$

For multi-channel audio signals (such as stereo signals), the STFT was calculated separately for each channel, and the results were averaged during the feature extraction stage.

To obtain a spectral representation that better matches human auditory perception, this study constructed 24 Bark triangular filters. For the k-th Bark band, the energy was defined as:

$$E_k(t) = \sum_f H_k(f) P(f,t)$$

where $H_k(f)$ represents the k-th Bark filter, and $P(f,t)$ represents the power spectrum.

This step maps the original spectrum into Bark perceptual bands. Afterwards, averaging was performed along the time dimension to obtain the average energy distribution of the entire audio signal in each Bark band. Considering that human perception of loudness approximately follows a logarithmic relationship, logarithmic compression was further applied to the Bark-band energies. The final log-Bark spectral envelope vector was defined as:

$$\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,K}]$$

where $K = 24$.

Since there may be overall loudness differences between different recordings, while this study focuses more on spectral shape rather than overall loudness, mean normalization was applied to the spectral envelope vectors. First, the mean value of each audio spectral vector was calculated as:

$$\mu_i = \frac{1}{K} \sum_{k=1}^{K} x_{i,k}$$

Normalization was then performed as:

$$\hat{x}_{i,k} = x_{i,k} - \mu_i$$

Therefore, the original reference audio was used as the target spectral profile. For each loudspeaker under test, the spectral deviation between the recording and the reference audio was defined as:

$$d_{i,k} = \hat{x}_{i,k} - \hat{x}_{\mathrm{ref},k}$$

where $\hat{x}i, k$ represents the normalized energy of the i-th recording in the k-th Bark band, and $\hat{x}\mathrm{ref}, k$ represents the normalized energy of the reference audio in the corresponding Bark band.

To quantify the overall spectral difference between the recorded signal and the reference audio, the Root Mean Square (RMS) distance was used as the overall spectral deviation metric:

$$D_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d_{i,k}^2}$$

where K is the number of Bark bands, and $d_{i,k}$ represents the spectral deviation between the i-th recording and the reference audio in the k-th Bark band.

Finally, the score was defined as:

$$\text{Score}_i = D_i$$

As the evaluation metric for loudspeaker naturalness. A smaller Score indicates a smaller spectral envelope difference between the recorded signal and the original reference audio, meaning that the loudspeaker introduces less modification to the original timbre. Therefore, the listening experience is considered to be more natural. It can be observed that, according to the model

Table 3.4.: Objective Evaluation Scores of Different Audio Clips

Audio Clip	**Score**
Audio_A	1.704332
Audio_B	2.591925
Audio_C	1.786418

prediction, the ranking order is $A > C > B$.

Signal Modification and Validation

After defining the naturalness evaluation metric, this study further applied targeted spectral modulation to the audio signals based on the proposed scoring model. The modulation method was designed according to the Bark-band analysis results.

The modulation analysis focused on Loudspeaker A using the previously introduced violin stimulus. The analysis showed that, within Bark bands 5–17, the envelope energy vector difference between Loudspeaker A and the reference audio was clearly smaller than that of Loudspeakers B and C. In contrast, within Bark bands 1–4, the difference between Loudspeaker A and the reference audio was larger than that of Loudspeaker B. In the high-frequency range of Bark bands 18–24, all three loudspeakers showed relatively large differences from the reference audio, while the differences between the loudspeakers themselves were comparatively smaller.

Although the loudspeakers exhibited different spectral characteristics, listeners were still able to distinguish them based on timbral differences. Therefore, this study further increased the envelope energy vector differences between the audio signals and the reference audio in selected

Bark bands, in order to artificially enlarge the perceptual differences between loudspeakers and evaluate whether the proposed model could correctly predict changes in subjective evaluation trends.

Considering that human hearing is particularly sensitive to the mid-frequency range of approximately 2–5 kHz, while sensitivity to low frequencies (20 Hz–2 kHz) and high frequencies (5 kHz–16 kHz) is relatively lower, two different modulation strategies were designed based on auditory perception characteristics. These strategies were used to investigate how energy changes in different frequency regions influence subjective evaluation trends.

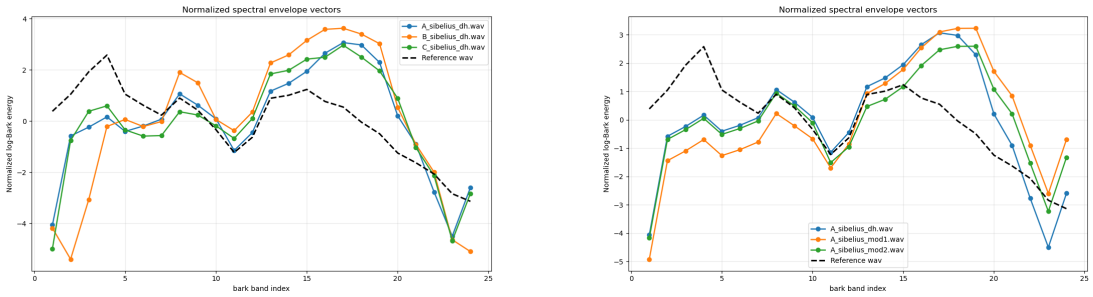
The first modulation strategy mainly adjusted the energy in the mid-frequency region (1500–5000 Hz) in order to emphasize timbral characteristics within the frequency range where human hearing is most sensitive. The gain was set as:

$$G_{\text{mid}} = +6 \text{ dB}$$

The second modulation strategy modified the overall tonal balance by attenuating low frequencies (below 500 Hz) while enhancing high frequencies (above 1500 Hz). The gains were defined as:

$$G_{\text{low}} = -6 \text{ dB}, \quad G_{\text{high}} = +6 \text{ dB}$$

The Figures 3.18 and 3.19 show the normalized spectral envelope vectors before and after model-based modulation, as well as their deviation distributions relative to the reference spectrum. These results are presented to illustrate the spectral feature changes introduced during the modulation process.

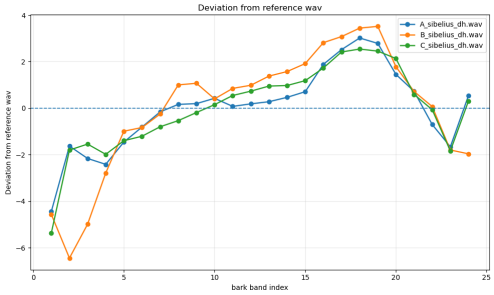


(a) Original Normalized Spectral Envelope Vectors

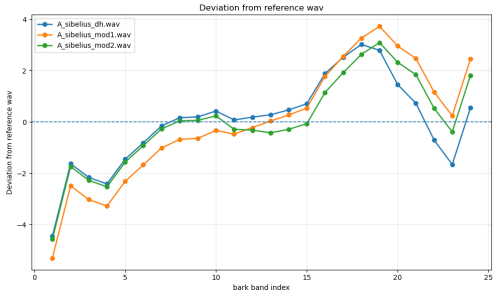
(b) Modified Normalized Spectral Envelope Vectors

Figure 3.18.: Normalized Spectral Envelope Comparison Before and After Model-Guided Modification - Music

Here, Audio A, B, and C refer to the recordings obtained when the audio was played through the three different loudspeakers. The suffix “mod” indicates the modulated version of the corresponding audio signal. For example, Audio A mod refers to the modulated version of Audio A.



(a) Original Deviation from the Reference Spectrum



(b) Modified Deviation from the Reference Spectrum

Figure 3.19.: Deviation from the Reference Spectrum Before and After Model-Guided Modification - Music

3.8.3. Urgency Perception Model

Model Construction

For the modeling of urgency perception in alarm signals, this study combined subjective perceptual evaluation with spectral acoustic features to analyze and model the perceived urgency of alarm sounds. The results showed that urgency perception is not determined by a single frequency component, but is closely related to spectral variations across different frequency regions.

First, in the low- to mid-frequency range (approximately 200–1000 Hz), stronger local spectral variations tend to increase the perceived intensity and tension of the sound. In addition, spectral variations in the high-frequency region (above approximately 4000 Hz) further enhance the perceived urgency of alarm signals. The concentration of energy in the high-frequency range, often related to sharpness, also influences listeners' perception of sound sharpness and alertness. On the other hand, when multiple audio samples were all perceived as highly urgent, listeners tended to prefer sounds that were subjectively more comfortable and more pleasant to listen to. Therefore, besides features related to tension and urgency, spectral smoothness in the mid-frequency region (mid smoothness) was further introduced into the model to represent the comfort level of the sound.

Finally, by analyzing listeners' subjective ratings obtained from the binaural listening experiments together with the spectral feature parameters extracted from the corresponding binaural recordings, a relationship between subjective perception and spectral acoustic features was established.

To quantify the perceived "urgency" and "pleasantness" of alarm sounds, the audio signals were first transformed into the frequency domain for analysis. For the input signal $y(t)$, the Short-Time Fourier Transform (STFT) was applied to obtain the frequency–time representation:

$$S(f,t) = |\mathrm{STFT}(y(t))|$$

For stereo audio signals, the features of the left and right channels were calculated separately and then averaged to obtain the final feature representation.

To better match human auditory perception of frequency, the linear frequency scale was converted into the Bark perceptual scale:

$$z = 13 \arctan(0.00076f) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right)$$

The Bark scale was later used for the calculation of the sharpness-related features, in order to improve the consistency between the extracted features and subjective auditory perception.

The spectral ripple feature was used to measure the degree of local spectral variation within a specific frequency range, in order to reflect the complexity and tension of the alarm sound spectrum. It was defined as:

$$\text{ripple} = \text{Var}\left(\Delta \log(E(f))\right)$$

where $E(f)$ represents the frequency energy obtained after averaging the power spectrum along the time dimension, and Δ represents the first-order difference between adjacent frequency bins.

According to different frequency ranges, the ripple feature was divided into low-mid ripple (200–1000 Hz) and high ripple (4000–10000 Hz). A larger ripple value indicates stronger local spectral variations and a less smooth spectral structure, corresponding to stronger stimulation and higher perceived urgency.

The sharpness feature was used to describe the concentration of sound energy in the high-frequency region, in order to reflect the perceived sharpness and stimulation of the sound. To reduce the coupling between the sharpness feature and the low-/mid-frequency ripple features, only the spectral characteristics within the high-frequency range (4–9 kHz) were considered, and Bark-scale weighting was applied:

$$S = \frac{\sum_{f \in HF} P(f)\, z(f)\, g(f)}{\sum_{f \in HF} P(f)}$$

where $P(f)$ represents the power spectrum, $z(f)$ represents the Bark frequency scale, and HF represents the high-frequency range (4–9 kHz). The high-frequency weighting function $g(f)$ was defined as:

$$g(f) = \begin{cases} 1, & z(f) \le 16 \\ 1 + 0.2\big(z(f) - 16\big), & z(f) > 16 \end{cases}$$

A larger sharpness value indicates that the high-frequency energy is more concentrated, resulting in a sharper and more tense perceived sound.

The mid smoothness feature was used to measure the smoothness of the spectrum within the 2–5 kHz frequency range, in order to reflect the perceived comfort and pleasantness of the sound:

$$\text{mid\_smoothness} = -\text{Var}\left(\Delta \log(E(f))\right)$$

Because a negative variance form was adopted, smaller spectral variations correspond to a smoother spectrum and therefore a larger mid smoothness value. Higher mid-frequency smoothness usually indicates that the sound is more stable and softer, resulting in a more comfortable subjective listening experience.

According to the subjective ratings given by listeners in the binaural listening experiments(3.6.3), a mapping relationship between acoustic features and human urgency perception was established. The experiments were conducted using both the dummy head and HATS recording conditions, and the scores from the two recording conditions were averaged to obtain the human score corresponding to each audio sample.

For each audio sample, four spectral features were first extracted: low-mid ripple, high ripple, sharpness, and mid smoothness. These features were then arranged into a feature matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix}$$

where each row corresponds to one audio sample, and each column corresponds to one spectral feature.

Since the different features have different numerical scales, standardization was first applied:

$$X_{\text{scaled}} = \frac{X - \mu_X}{\sigma_X}$$

where μ_X represents the feature mean and σ_X represents the feature standard deviation. After standardization, each feature has a mean value of 0 and a variance of 1, which helps avoid imbalance caused by differences in feature scales during regression.

Ridge Regression was then used to establish the relationship between the spectral features and the human scores:

$$\hat{y} = X_{\text{scaled}} w + \epsilon$$

where

$$w = [w_1, w_2, w_3, w_4]^T$$

represents the regression coefficients, ϵ represents the error term, and $\hat{y}$ represents the predicted urgency score generated by the model.

Since Ridge Regression introduces L2 regularization, the optimization objective can be written as:

$$\min_w \|X_{\text{scaled}} w - y\|_2^2 + \alpha \|w\|_2^2$$

where y represents the human scores obtained from the subjective experiments, and α is the regularization coefficient. By minimizing the above loss function, the contribution of different spectral features to urgency prediction can be estimated.

To further improve the interpretability of the model, the feature contributions were divided into two levels:

$$\text{Tension} = f(\text{low-mid ripple},\ \text{high ripple},\ \text{sharpness})$$

and

$$\text{Pleasantness} = f(\text{mid smoothness})$$

where Tension represents the contribution of stimulation and tension-related characteristics, while Pleasantness represents the contribution of comfort and pleasantness.

Finally, the urgency score was defined as:

$$\text{Urgency Score} = \alpha \cdot \text{Tension} + \beta \cdot \text{Pleasantness}$$

where α and β are the contribution weights obtained through regression analysis, and the Urgency Score represents the final predicted urgency level of the alarm sound.

This model shows that the perceived urgency of alarm sounds is related not only to spectral stimulation characteristics, but also to perceived comfort. Even when multiple audio samples have relatively high urgency, listeners may still prefer sounds that are subjectively more comfortable.

Table **??** and **??** shows detail the model-based scores of the recordings obtained when the siren audio was played by the three loudspeakers in Room A.

It can be observed that, according to the model prediction, the ranking order is $C > A > B$.

Signal Modification and Validation

After establishing the urgency model, this study further performed spectral modulation on alarm sounds based on the key spectral features identified in the model, in order to verify the model's ability to predict urgency perception.

Instead of modifying the overall spectrum arbitrarily, the modulation was specifically designed to target spectral features that showed strong relationships with perceived urgency. According to the previous analysis, low-mid ripple and sharpness contributed significantly to urgency perception. Therefore, the modulation mainly focused on the low- and mid-frequency range (approximately 200–1150 Hz).

To implement the spectral modulation, peaking EQ filters were applied to multiple center frequencies for enhancement or attenuation:

$$y_{\text{out}} = \text{lfilter}(b, a, y)$$

where f_0 represents the center frequency of the filter, gain represents the amplification applied to the corresponding frequency band, and Q represents the filter bandwidth.

Afterwards, alternating enhancement and attenuation structures were constructed within the low- and mid-frequency range:

$$f : \{200, +1\}, \{275, -1\}, \ldots, \{1150, -1\}$$

Table 3.5.: Feature Weights for Urgency Score Prediction

Feature Weights	
Feature	**Weight**
low_mid_ripple	0.2918
high_ripple	-0.0663
sharpness	0.2177
mid_smoothness	0.1079
Spearman's Correlation	**1.0**
Feature Contribution	
low_mid_ripple	42.7%
high_ripple	9.7%
sharpness	31.8%
mid_smoothness	15.8%
Tension	84.2%
Pleasantness	15.8%
Internal Contribution to Tension	
low_mid_ripple	50.7%
high_ripple	11.5%
sharpness	37.8%

Table 3.6.: Objective Evaluation Scores of Different Audio Clips

Audio A	
low_mid_ripple	0.9333
high_ripple	1.1826
sharpness	0.9613
pleasant	1.0174
tension	0.9726
pleasantness	1.0174
Audio B	
low_mid_ripple	0.9144
high_ripple	0.8543
sharpness	0.9613
pleasant	0.9709
tension	0.9252
pleasantness	0.9709
Audio C	
low_mid_ripple	1.1523
high_ipple	0.9631
sharpness	1.0774
pleasant	1.0117
tension	1.1022
pleasantness	1.0117
Final Order	
1. Audio C Score	1.0879
2. Audio A Score	0.9797
3. Audio B Score	0.9324

Through periodic spectral enhancement and attenuation, the local spectral variations became more pronounced, thereby increasing the low-mid ripple feature. At the same time, larger Q values further increased local spectral sharpness, enhancing the sharpness feature.

To avoid severe distortion caused by excessive spectral modulation, a spectral similarity constraint (spectral distance) was further introduced:

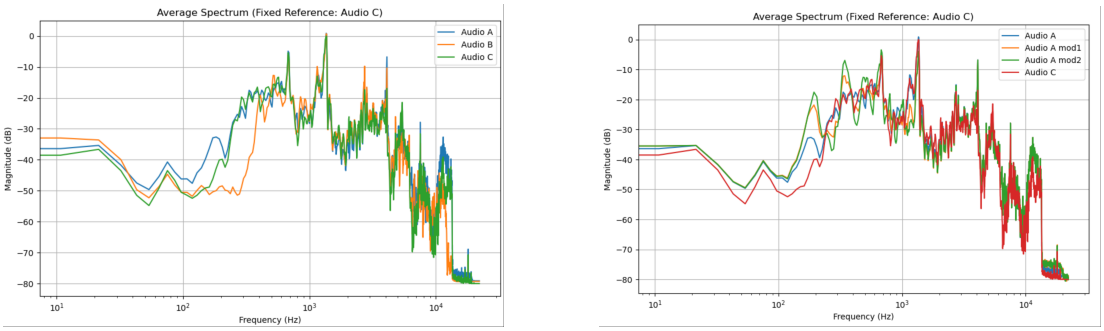
$$\text{dist} = \text{mean}\left(\left|\log|\text{STFT}(y_{\text{orig}})| - \log|\text{STFT}(y_{\text{mod}})|\right|^2\right)$$

When the spectral distance after modulation exceeded a predefined threshold, the modulation scheme was discarded to ensure that the modulated alarm sound still remained sufficiently similar to the original signal.

The modulated audio signals were then re-input into the urgency model, and the corresponding features and urgency scores were recalculated. By testing different modulation strength parameters, the modulation result with the highest predicted urgency score was selected.

The experimental results showed that the modulated audio exhibited clear changes in key features such as low-mid ripple and sharpness, while the urgency scores predicted by the model also increased accordingly. Compared with the original audio, the modulated alarm sounds were predicted by the model to produce stronger urgency perception. This result further supports that the proposed urgency model can effectively reflect the relationship between spectral structure changes and subjective urgency perception.

The Figure 3.20 shows the spectrogram of the siren recoding before and after modulation.



(a) Original Audio Spectrogram

(b) Modulated Audio Spectrogram

Figure 3.20.: Spectral Comparison Before and After Model-Guided Modification - Bass

In this study, the siren signals were not modified through simple spectral adjustment. Instead, the modulation process was guided by the proposed model, where key spectral features were selectively modified. Under the constraint of maintaining spectral similarity to the original audio, an optimization-based search was further applied to find the optimal modulation configuration. This approach was used to verify the model's ability to predict perceived urgency.

In the In-situ and Binaural Playback Consistency expriment(3.6), the urgency evaluations for the audio reproduced by Loudspeaker A and Loudspeaker C were found to be relatively similar.

Therefore, during the siren modulation experiment, Audio C was used as the reference, while Audio A was selected for modulation.

It should also be noted that Audio A, B, and C refer to the recordings obtained when the audio was played through the three different loudspeakers. The suffix "mod" indicates the modulated version of the corresponding audio signal. For example, Audio A mod refers to the modulated version of Audio A.

3.8.4. Speech Perception Model

Unlike the previous audio cases, the field of speech clarity already has relatively mature and widely validated perceptual models. Therefore, this study did not attempt to construct a new perceptual feature framework. Instead, existing modulation-based concepts from speech transmission analysis were adopted to evaluate and analyze the speech signals reproduced by different loudspeakers.

Previous studies have shown that human perception of speech clarity is closely related to the modulation structure of the speech envelope. Especially in the mid- and high-frequency regions, the preservation of speech modulation information significantly affects consonant recognition, speech intelligibility, and overall clarity perception. Based on this idea, this study adopted a modulation-based clarity model to compare the original speech signal with the binaural recordings reproduced by different loudspeakers. Speech clarity was then evaluated by analyzing how well the modulation information in different frequency bands was preserved.

Because the loudspeaker playback and recording process introduce propagation delay, temporal alignment between the original speech signal and the recorded signal is first required. Let the original speech signal be:

$$y_{\text{ref}}(t)$$

and the recorded signal be:

$$y_{\text{deg}}(t)$$

The time delay between the two signals is calculated using cross-correlation:

$$\tau = \arg\max_{k} \sum_{t} y_{\text{deg}}(t)\, y_{\text{ref}}(t-k)$$

The signals are then cropped and synchronized according to the estimated delay, in order to ensure temporal alignment for the subsequent modulation analysis.

Following conventional speech clarity models, the speech signal was divided into four key frequency bands:

$$\text{200–500 Hz}, \quad \text{500–1000 Hz}, \quad \text{1000–2000 Hz}, \quad \text{2000–4000 Hz}$$

These frequency bands cover the main formant regions and consonant information regions in speech signals, and therefore play an important role in speech intelligibility.

For each frequency band, a fourth-order Butterworth bandpass filter was first applied to extract the corresponding band-limited signal:

$$y_b(t) = H_b(y(t))$$

The envelope was then extracted using the Hilbert transform:

$$e_b(t) = |\mathcal{H}(y_b(t))|$$

where $\mathcal{H}(\cdot)$ represents the Hilbert transform.

To represent the modulation strength of speech signals, the modulation depth was defined as:

$$M_b = \frac{\sigma(e_b)}{\mu(e_b)}$$

where $\sigma(e_b)$ represents the standard deviation of the envelope, and $\mu(e_b)$ represents the mean value of the envelope. This metric reflects the temporal variation degree of the speech envelope.

For each frequency band, the modulation ratio between the recorded signal and the reference speech signal was then calculated as:

$$R_b = \frac{M_b^{\text{deg}}}{M_b^{\text{ref}}}$$

where M_b^{ref} represents the modulation depth of the original speech signal, and M_b^{deg} represents the modulation depth of the recorded signal reproduced by the loudspeaker. This ratio reflects how well the loudspeaker preserves the modulation structure of the original speech signal.

Finally, the speech clarity score was defined as the average modulation ratio across all frequency bands:

$$C = \frac{1}{N} \sum_{b=1}^{N} R_b$$

where $N = 4$ is the number of frequency bands.

A higher score indicates that the loudspeaker preserves the modulation structure of the original speech signal more effectively, corresponding to higher perceived speech clarity.

Table 3.7 shows the model-based scores of the recordings obtained when the womanspeech audio was played by the three loudspeakers in Room A.

Table 3.7.: Objective Evaluation Scores of Different Audio Clips

Audio Clip	**Score**
Audio_A	0.938
Audio_B	0.874
Audio_C	0.895

It can be observed that, according to the model prediction results, the ranking order is $A > C > B$

3.8.5. Model Validation

The model validation experiments were conducted using binaural listening. Since the validation process in this study was based on spectral modulation of binaural recordings, only binaural playback was used in the validation experiments in order to keep the experimental conditions consistent with the previous binaural listening experiments. Participants listened to the modulated clean-audio binaural recordings through the Sennheiser HD 650 headphones and completed the subjective evaluations using WebMUSHRA.

Only the ranking evaluation method was used in this experiment. Participants were asked to subjectively rank the playback results of the same audio signal under different modulation conditions according to the corresponding perceptual metric. The overall testing procedure remained consistent with the previous binaural listening experiments.

During the experiments, all test audio samples were presented in random order in order to minimize the influence of playback sequence on subjective judgment, thereby improving the reliability and consistency of the experimental results.

4. Results

4.1. Participants

The experiments in this study were divided into four parts. The first three parts aimed to verify the consistency of subjective ranking results under different playback conditions, and further investigate the influence of background noise on the ranking results. The fourth part was designed to validate the prediction ability of the proposed model.

In the first three experiments, participants were required to complete the listening tasks under specific room conditions and fixed experimental equipment settings. Due to experimental limitations, a total of 16 participants were recruited. Since this part of the study focused on within-subject ranking consistency across different playback conditions, all participants completed all experimental tasks in order to allow direct comparison of the results.

Considering that the main objective of this stage was to analyze within-subject consistency rather than to estimate population-level distribution characteristics, this number of participants was considered reasonable under controlled experimental conditions.

In contrast, the fourth experiment was mainly designed to validate the prediction performance of the model-based audio modulation. This stage no longer involved consistency comparisons between different playback media and did not rely on the specific experimental equipment used in the previous experiments. Therefore, in addition to the original 16 participants, the sample size was further expanded in order to improve the stability of the model prediction results and their applicability across different listeners. Finally, 14 additional participants were recruited at this stage, resulting in a total of 30 valid datasets for analysis.

The table 4.1 shows the distribution of all participant-related metrics.

Table 4.1.: Participant Information for the Listening Experiments

	First Three Experiments	**Fourth Experiment**
Gender	13 males, 3 females	18 males, 12 females
Age Range	26–56	25–56
Normal Hearing	All participants	All participants

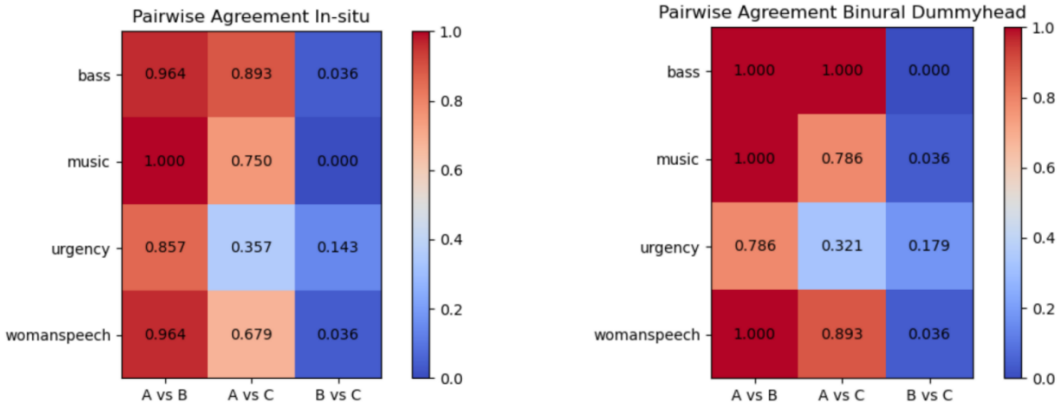
4.2. Consistency Evaluation Results

To compare the consistency of subjective evaluation results between direct listening and binaural playback conditions, the 16 participants completed both the noise-free direct listening experiment (3.6.2) and the first binaural playback experiment (3.6.3) described in the previous Methods section.

Both rating and ranking evaluation methods were used in the experiments, allowing subjective consistency under different playback conditions to be analyzed from both the scoring perspective and the ranking perspective. In addition, the two evaluation methods could also serve as mutual validation, thereby reducing the influence of random choices on the experimental results.

4.2.1. In-situ and Binaural Playback Consistency Results

Based on the pairwise comparison method introduced in the Theory section, a comparative analysis was conducted between the direct listening results obtained in Room A and the corresponding binaural playback results. The results are shown in Figure 4.1. As shown in the figure, participants exhibited relatively stable relative ranking relationships for all four types of audio signals under both the direct listening (in-situ) condition and the binaural recording playback (binaural dummy head) condition.



(a) In-situ

(b) Binaural

Figure 4.1.: Pairwise Results for In-situ and Dummy head Binaural Condition

According to the consistency criterion defined previously, when the pairwise preference ratio reached or exceeded 70%, the relative perceptual relationship between loudspeakers was considered to show stable consistency within the participant group.

The experimental results show that both the bass and music audio samples exhibited stable ranking trends under the direct listening (in-situ) and binaural playback conditions. The main

subjective ranking result was $A > C > B$, and the corresponding pairwise preference ratios all exceeded 70%.

For the urgency (alarm sound) audio, both the direct listening and binaural playback conditions generally showed the ranking trend $C > A > B$. Among these relationships, the pairwise preferences of $C > B$ and $A > B$ showed relatively high consistency. Although the preference ratio for $C > A$ did not reach 70%, the same overall trend was still observed under both listening conditions.

For the woman speech audio, the subjective ranking results under both the direct listening and binaural playback conditions still mainly followed the trend $A > C > B$. However, compared with the other audio samples, the $A > C$ preference trend was relatively weaker. In particular, the corresponding preference ratio in the direct listening experiment did not reach 70%, indicating that the consistency of listeners' perceptual differences between A and C was relatively limited.

Overall, the subjective ranking trends obtained under different experimental conditions showed relatively high consistency, indicating that binaural recordings were able to preserve the subjective perceptual relationships observed in the direct listening experiments to a large extent.

Further examination of the results between different loudspeaker pairs shows that the discriminability varied clearly across different audio conditions. In particular, the loudspeaker pairs A vs. B and C vs. B showed relatively high consistency for most audio samples. In both the direct listening and binaural playback experiments, the corresponding pairwise preference ratios were close to or reached 1, indicating that participants were able to reliably distinguish the subjective perceptual differences between these loudspeaker pairs.

In contrast, the overall consistency between A and C was relatively lower, indicating that the perceptual differences between these two loudspeakers were more difficult to distinguish reliably. However, under the bass and music conditions, the pairwise preference ratio between A and C still remained relatively high, suggesting that listeners were still able to stably perceive the differences between them under specific audio conditions.

Furthermore, the pairwise preference deviation was calculated, and the results are shown in the following Table 4.2 For the pairwise preference deviation, according to the criterion defined

Table 4.2.: Pairwise Preference Deviation

ΔP	**A vs B**	**A vs C**	**B vs C**
bass	0.036	0.107	0.036
music	0.000	0.036	0.036
urgency	0.071	0.036	0.036
woman speech	0.036	0.214	0.000

in the Methods section, when $\Delta P \leq 15\%(0.015)$, the binaural playback and in-situ listening conditions were considered to have acceptable perceptual consistency under that condition. The

experimental results showed that only the A vs. C pair under the woman speech condition produced a pairwise preference deviation exceeding this threshold, while the ΔP values for all other audio conditions and loudspeaker pairs remained within 15%. These results indicate that, under most experimental conditions, binaural playback was able to preserve the perceptual preference tendencies observed in the direct listening experiments relatively well.

4.2.2. Confidence Interval Analysis of Pairwise Preference Results

Furthermore, the rating results were analyzed, and the corresponding 95% confidence intervals were calculated. As shown by the 95% confidence intervals in the Figure 4.6, the bass condition produced the most stable results. In both the direct listening and binaural playback experiments, the rating of Loudspeaker A was consistently higher than that of Loudspeaker B, while Loudspeaker C stably remained between them. At the same time, the overall error ranges for each loudspeaker were relatively small, indicating relatively high agreement among participants in their bass evaluations. Therefore, the corresponding ranking results can be considered highly reliable.

A similar trend was also observed under the music condition. Although there was some variation in ratings among different participants, the overall trend of $A > C > B$ could still be clearly observed. Moreover, this trend remained consistent under both the direct listening and binaural playback conditions, indicating that the subjective evaluation of broadband music signals also showed relatively good stability.

In contrast, the urgency condition showed noticeably larger error ranges, while the rating differences between different loudspeakers were relatively smaller. This suggests that listeners exhibited greater individual variability when judging the "urgency" of alarm sounds, resulting in lower subjective evaluation stability compared with the bass and music conditions. Although the overall ranking trend was still present, the results showed a relatively higher level of dispersion.

For the woman speech condition, the overall trend also remained relatively stable. The ratings of Loudspeaker A were generally higher than those of Loudspeaker B, while Loudspeaker C remained between them. However, compared with the bass condition, the error ranges were slightly larger, indicating that the subjective evaluation of speech signals was more easily influenced by individual factors. As a result, the rating consistency among participants was relatively lower.

4.2.3. Intra-subject Consistency Across Listening Conditions

Subsequently, in order to further compare the consistency of subjective ranking results under different experimental conditions, the Spearman rank correlation coefficient was used to analyze the ranking results obtained from participants under the in-situ and binaural conditions. Since this metric reflects the monotonic relationship between different ranking results, it is suitable for evaluating the consistency of subjective ranking trends across different playback conditions.

From the distribution of the experimental results, it can be observed that the ranking results in the direct listening experiments (in-situ) showed relatively larger variability across participants.

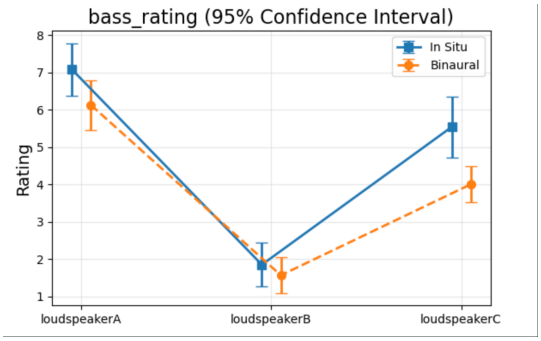


Figure 4.2.: Bass 95% CI

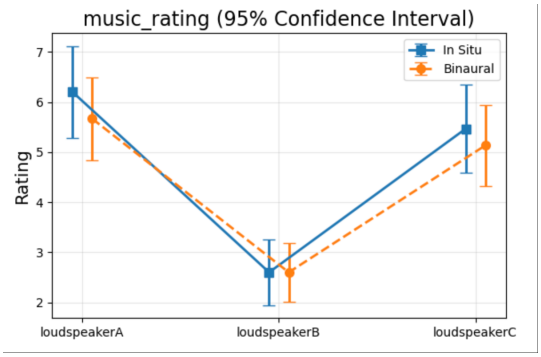


Figure 4.3.: Music 95% CI

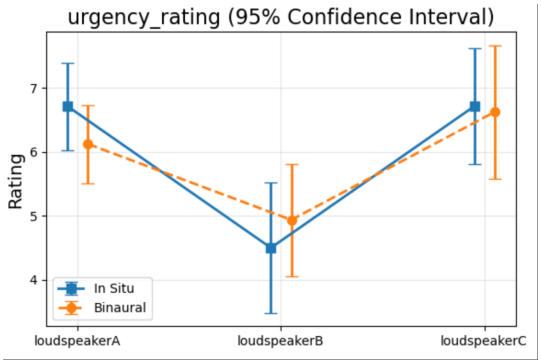


Figure 4.4.: Urgency 95% CI

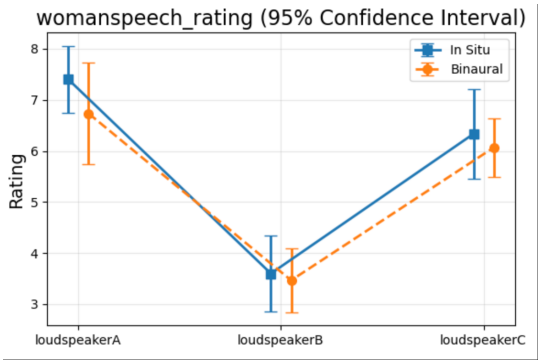


Figure 4.5.: Woman speech 95% CI

Figure 4.6.: 95% Confidence Interval of Four Audio Clip

In contrast, the ranking results obtained in the binaural playback experiments tended to be more concentrated, indicating higher inter-subject consistency. In other words, under the binaural condition, participants were more likely to form relatively stable subjective ranking trends.

The Spearman correlation analysis further quantified the observations described above. After obtaining the Spearman correlation coefficient for each audio condition, the overall correlation analysis was calculated using Fisher z-transform averaging, as introduced in the Theory section. The results are shown in the following Table 4.3:

Table 4.3.: Spearman Rank Correlation Coefficient

Audio Type	**In-situ**	**Binaural**
bass	0.857	1.000
music	0.857	0.393
urgency	0.429	0.107
woman speech	0.571	0.893
Total	0.725	0.894

The results show that both the in-situ and binaural conditions exhibited relatively strong positive correlations, indicating that participants were generally able to maintain stable subjective ranking trends under both direct listening and binaural playback conditions. However, compared with the direct listening condition, the participant group showed greater stability under the binaural playback condition.

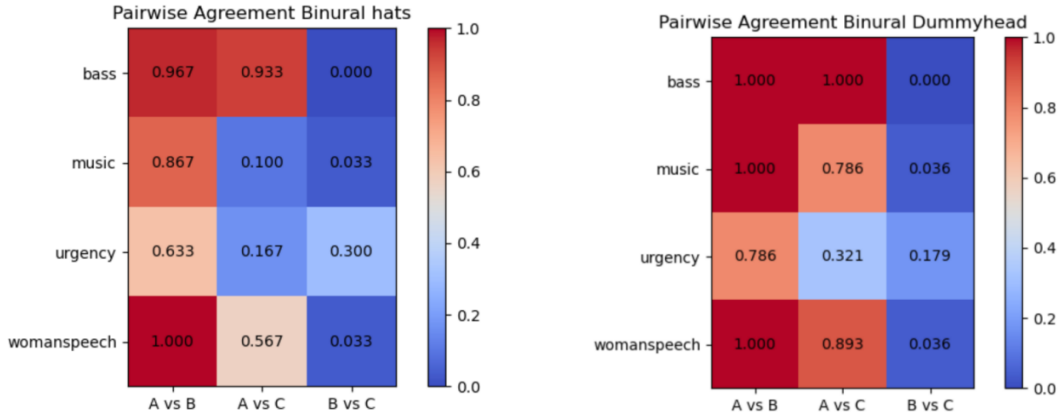
When comparing different audio types, some differences in consistency could still be observed across the test materials. Certain audio samples showed relatively high consistency under both binaural and in-situ conditions, while others exhibited more noticeable variability, suggesting that the stability of subjective ranking depends to some extent on the type of audio signal.

Overall, the binaural listening experiments showed higher ranking correlations, indicating that their subjective evaluation results were generally more consistent and stable than those obtained from the direct listening experiments. This phenomenon may be related to the fact that the playback chain, listening position, and reproduction system were more controllable under binaural conditions, thereby reducing the influence of random variations present in real listening environments on subjective judgments.

4.2.4. Comparison Between Different Binaural Playback Conditions

Furthermore, the binaural playback results obtained from recordings made in the two different rooms were also analyzed. The results are shown in Figure 4.7.

From the HATS results, the bass condition showed stable ranking trends under both binaural playback conditions. The main subjective ranking result was $A > C > B$, and the corresponding



(a) HATS

(b) Binaural

Figure 4.7.: Pairwise Results for HATS and Dummy head Binaural Condition

pairwise preference ratios all exceeded 70%.

For the music audio, the ranking relationship between A and C under the HATS condition changed noticeably compared with the dummy head condition. The subjective ranking trend became $C > A > B$, and the corresponding pairwise preference ratio exceeded 70%, indicating that participants showed a more stable preference for Loudspeaker C under the HATS condition.

For the urgency (alarm sound) audio, both binaural playback conditions generally showed the ranking trend $C > A > B$. Among these relationships, the pairwise preferences of $C > B$ and $A > B$ remained relatively consistent. Compared with the dummy head condition, the $C > A$ preference trend became stronger under the HATS condition, with the corresponding pairwise preference ratio exceeding 70%.

For the woman speech audio, the overall ranking trend still mainly followed $A > C > B$. However, compared with the dummy head condition, the $A > C$ preference trend became relatively weaker, and the corresponding pairwise preference ratio did not reach 70%.

According to the LUFS loudness statistics presented in the Methods section(Table 3.2), clear differences in loudness consistency can be observed between the dummy head and HATS conditions. For the same audio signal, the loudness variation range under the dummy head condition was approximately 0.599–1.339 dB. Such a difference is generally difficult for human listeners to perceive reliably and can therefore be considered approximately loudness-consistent. In contrast, under the HATS condition, the loudness variation increased to approximately 3.291–3.967 dB.

Further inspection showed that this additional loudness deviation was caused by abnormal gain changes in the recording chain during the recording process, possibly due to accidental adjustment of the audio interface gain knob. As a result, relatively large loudness differences occurred between recordings from different loudspeakers, and these differences had already reached a perceptually noticeable level.

Combining these results with the pairwise comparison analysis, it can be observed that different types of audio signals show clear differences in sensitivity to loudness variation.

For the bass signals, even with additional loudness differences, participants still tended to prefer the loudspeakers with stronger low-frequency performance. This suggests that the perceptual target of bass perception could, to some extent, suppress the influence caused by loudness differences. For the music signals, the ranking results showed much stronger sensitivity to loudness variation, and the original ranking trend even exhibited reversal behavior. For the siren signals, the overall ranking trend still remained relatively stable as $C > A > B$, although the corresponding preference tendencies became further strengthened. For the speech signals, although the loudness differences introduced some disturbance to the results, participants still tended to prefer the loudspeakers with higher speech clarity. However, the overall stability of the ranking results decreased to some extent.

To further verify whether the above trends were mainly caused by loudness differences, the recordings were later recalibrated for loudness, and additional listening tests were conducted. Since it was difficult to recall all original participants, five participants from the original group were selected for the repeated experiments.

After the loudness differences were reduced, the ranking trends under the HATS condition again showed relatively high consistency with those obtained under the dummy head condition. This result indicates that the abnormal loudness differences mentioned above indeed had a significant influence on the subjective ranking results for certain audio conditions.

4.3. Background Noise Experiment Results

In this experiment, the A-weighted equivalent continuous sound pressure level (LeqA) was calculated over a continuous 30-second playback period. During both the direct listening experiments with background noise and the recording process, the playback levels of the three loudspeakers were kept identical and remained the same as those used in the noise-free experiments (3.6.2). Therefore, the signal-to-noise ratio (SNR) could be adjusted by changing the level of the background noise.

Since the background noise used in the experiments was not uniformly distributed noise such as white noise or pink noise, but instead consisted of everyday environmental noise, its sound pressure level could not be represented accurately using a simple theoretical value. Therefore, this study used the measured LeqA value obtained over a 30-second playback period of the background noise to represent the background noise level. The signal-to-noise ratio was then calculated based on the LeqA values of both the signal and the noise.

In the table, "✓" indicates that, under the corresponding A-weighted equivalent continuous sound pressure level condition, the participants were able to correctly match the target loudspeaker with the loudspeaker represented by the reference audio, meaning that they could still distinguish the timbral differences between loudspeakers. In contrast, "–" indicates that, under

that LeqA condition, the participants either produced incorrect matching results or could no longer distinguish the timbral differences between loudspeakers.

4.3.1. Perceptual Discrimination Threshold Under Background Noise

Bass

The results of the direct listening and binaural listening experiments for the bass audio signal under background noise conditions with LeqA values of 41.7 dB, 53.7 dB, and 56.6 dB are shown in Table 4.4. The identification accuracies of the reference loudspeaker under the three noise levels are shown in Table 4.5. The identification accuracies (3.7.3) under all three noise levels were greater than 50%. According to the experimental design(3.7.3), under background noise conditions, the perceptual discrimination threshold at which listeners could still distinguish the playback differences among Loudspeakers A, B, and C for the bass signal was determined to be LeqA = 56.6 dB. The corresponding signal-to-noise ratio at this threshold was 14.09 dB.

Table 4.4.: Comparison Between In-situ and Binaural Evaluation Results Under Different Noise Conditions - Bass

	41.7		53.7		56.6	
Subject	**In-situ**	**Binaural**	**In-situ**	**Binaural**	**In-situ**	**Binaural**
1	✓	✓	✓	✓	✓	✓
2	✓	✓	–	✓	–	✓
3	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✓
8	✓	✓	–	✓	–	✓
9	✓	✓	✓	✓	–	✓
10	✓	✓	✓	✓	✓	✓
11	✓	✓	✓	✓	✓	✓
12	–	–	–	–	–	–
13	✓	✓	✓	–	✓	–
14	–	✓	–	–	–	–
15	✓	✓	✓	✓	✓	✓
16	✓	✓	✓	✓	✓	✓

Table 4.5.: Recognition Rates Under Different Noise Levels for Bass With Noise

Condition	Noise Level (dB)	Responses	Recognition Rate (%)
In-situ	41.7	14/16	87.5
	53.7	12/16	75.0
	56.6	11/16	68.8
Binaural	41.7	15/16	93.8
	53.7	13/16	81.2
	56.6	13/16	81.2

Music

The results of the direct listening and binaural listening experiments for the music audio signal under background noise conditions with LeqA values of 41.7 dB, 45.7 dB, and 49.5 dB are shown in Table 4.6. The identification accuracies of the reference loudspeaker under the three noise levels are shown in Table 4.7. In both the direct listening and binaural listening experiments, when the noise level was 41.7 dB, the identification accuracies were 56.2% and 81.2%, respectively, both exceeding 50%. However, when the noise level increased to 45.7 dB and 49.5 dB, the identification accuracies in both listening conditions dropped below 50%. According to the criterion used in this study, under background noise conditions, the perceptual discrimination threshold at which listeners could still distinguish the playback differences among Loudspeakers A, B, and C for the music signal was determined to be LeqA = 41.7 dB. The corresponding signal-to-noise ratio at this threshold was 11.805 dB.

Siren

The results of the direct listening and binaural listening experiments for the siren audio signal under background noise conditions with LeqA values of 41.7 dB, 46.6 dB, and 53.5 dB are shown in Table 4.8. The identification accuracies of the reference loudspeaker under the three noise levels are shown in Table 4.9. In the binaural listening experiment, only the condition with a noise level of 41.7 dB produced an identification accuracy greater than 50% for the reference loudspeaker timbre. In contrast, in the direct listening experiment, the identification accuracies at noise levels of 41.7 dB, 46.6 dB, and 53.5 dB all remained below 50%. At the noise condition of LeqA = 41.7 dB, listeners in the direct listening experiment were unable to reliably distinguish the loudspeaker timbre, while the binaural listening experiment still produced higher identification accuracy. Therefore, the binaural results were not consistent with the direct listening results under this noise level, indicating that binaural playback could not replace the direct listening experiment under this condition.

Table 4.6.: Comparison Between In-situ and Binaural Evaluation Results Under Different Noise Conditions - Music

	41.7		45.7		49.5	
Subject	**In-situ**	**Binaural**	**In-situ**	**Binaural**	**In-situ**	**Binaural**
1	✓	✓	–	✓	–	–
2	–	✓	–	–	–	–
3	✓	✓	✓	–	–	–
4	–	✓	–	✓	–	–
5	✓	–	✓	–	✓	–
6	✓	✓	–	–	–	–
7	–	✓	–	–	–	–
8	–	✓	–	✓	–	–
9	✓	✓	✓	✓	–	–
10	✓	–	✓	–	✓	–
11	✓	✓	✓	✓	–	–
12	–	–	–	–	–	–
13	–	✓	–	–	–	–
14	–	✓	–	–	–	–
15	✓	✓	✓	–	✓	–
16	✓	✓	–	✓	–	✓

Table 4.7.: Recognition Rates Under Different Noise Levels for Music With Noise

Condition	**Noise Level (dB)**	**Responses**	**Recognition Rate (%)**
In-situ	41.7	9/16	56.2
	45.7	6/16	37.5
	49.5	3/16	18.8
Binaural	41.7	13/16	81.2
	45.7	6/16	37.5
	49.5	1/16	6.2

Table 4.8.: Comparison Between In-situ and Binaural Evaluation Results Under Different Noise Conditions - Siren

	41.7		46.6		53.5	
Subject	**In-situ**	**Binaural**	**In-situ**	**Binaural**	**In-situ**	**Binaural**
1	✓	✓	✓	–	✓	–
2	–	–	–	–	–	–
3	–	✓	–	–	–	–
4	✓	✓	✓	✓	–	✓
5	–	✓	–	✓	–	✓
6	–	✓	–	✓	–	–
7	–	✓	–	–	–	–
8	–	✓	–	✓	–	✓
9	–	–	–	–	–	–
10	✓	–	✓	–	–	–
11	✓	✓	✓	✓	–	–
12	–	–	–	–	–	–
13	–	✓	–	✓	–	✓
14	–	–	–	–	–	–
15	✓	✓	✓	–	–	–
16	✓	–	✓	–	✓	–

Table 4.9.: Recognition Rates Under Different Noise Levels for Siren With Noise

Condition	**Noise Level (dB)**	**Responses**	**Recognition Rate (%)**
In-situ	41.7	6/16	37.5
	46.6	6/16	37.5
	53.5	2/16	12.5
Binaural	41.7	10/16	62.5
	46.6	6/16	37.5
	53.5	4/16	25.0

Woman Speech

The results of the direct listening and binaural listening experiments for the woman speech audio signal under background noise conditions with LeqA values of 41.7 dB, 49.7 dB, and 53.4 dB are shown in Table 4.10. The identification accuracies of the reference loudspeaker under the three noise levels are shown in Table 4.11. In the direct listening experiment, when the noise level was 41.7 dB, the identification accuracy of the reference loudspeaker timbre reached 50%. When the noise level increased to 53.4 dB, the identification accuracy dropped below 50%. In the binaural listening experiment, the identification accuracies of the reference loudspeaker timbre remained above 50% at noise levels of 41.7 dB, 49.7 dB, and 53.4 dB. According to the criterion used in this study, under background noise conditions, the perceptual discrimination threshold at which listeners could still distinguish the playback differences among Loudspeakers A, B, and C for the speech signal was determined to be LeqA = 49.7 dB. The corresponding signal-to-noise ratio at this threshold was 16.17 dB.

Table 4.10.: Comparison Between In-situ and Binaural Evaluation Results Under Different Noise Conditions - Woman Speech

	41.7		49.7		53.4	
Subject	**In-situ**	**Binaural**	**In-situ**	**Binaural**	**In-situ**	**Binaural**
1	–	✓	–	✓	–	✓
2	✓	✓	✓	–	✓	–
3	✓	✓	✓	✓	✓	✓
4	–	✓	–	✓	–	✓
5	–	✓	–	✓	–	✓
6	✓	✓	✓	✓	–	✓
7	✓	✓	✓	✓	–	✓
8	–	✓	–	–	–	–
9	–	✓	–	✓	–	✓
10	✓	✓	✓	–	–	–
11	✓	✓	✓	✓	✓	✓
12	–	–	–	–	–	–
13	–	✓	–	–	–	–
14	–	✓	–	–	–	–
15	✓	✓	✓	✓	✓	✓
16	✓	✓	✓	✓	–	✓

Table 4.11.: Recognition Rates Under Different Noise Levels for Speech With Noise

Condition	Noise Level (dB)	Responses	Recognition Rate (%)
In-situ	41.7	8/16	50.0
	49.7	8/16	50.0
	53.4	4/16	25.0
Binaural	41.7	15/16	93.8
	49.7	10/16	62.5
	53.4	10/16	62.5

4.3.2. Noise and Playback Effects on Timbre Recognition

To analyze the effects of background noise and listening condition on loudspeaker timbre identification accuracy, the main-effects GEE Logistic Regression model (see Theory section) was defined as:

$$\log\left(\frac{p}{1-p}\right) = 5.8879 - 0.5701 \times \text{Method} - 0.0798 \times \text{Noise}$$

where: Method = 0 represents the binaural listening condition; Method = 1 represents the in-situ listening condition.

For the bass signal, the interaction effect between noise level and listening condition was not significant ($p = 0.888$). Therefore, the main-effects GEE model was adopted, indicating that the in-situ and binaural listening conditions showed similar trends as the noise level increased. The noise coefficient was negative, indicating that the probability of correctly identifying loudspeaker timbre differences gradually decreased as the noise level increased. According to the model estimation, for every 1 dB increase in noise level, the probability of successful identification decreased by approximately 7.7%.

For the music signal, a significant interaction effect was found between noise level and listening condition ($p = 0.039$). This indicates that the identification accuracy changed differently with increasing noise under the two listening conditions. Therefore, the results for this signal should be interpreted using the interaction model. The interaction model for the music signal was expressed as:

$$\log\left(\frac{p}{1-p}\right) = 23.4001 - 14.0537 \times \text{Method} - 0.5246 \times \text{Noise} + 0.3074 \times (\text{Noise} \times \text{Method})$$

When Method = 0 (binaural condition), the model becomes:

$$\log\left(\frac{p}{1-p}\right) = 23.4001 - 0.5246 \times \text{Noise}$$

When Method = 1 (in-situ condition), the model becomes:

$$\log\left(\frac{p}{1-p}\right) = 9.3465 - 0.2172 \times \text{Noise}$$

The slope comparison shows that:

$$\beta_{\text{binaural}} = -0.5246$$

$$\beta_{\text{in\_situ}} = -0.2172$$

For this signal, the identification accuracy under the binaural condition was more sensitive to changes in background noise. As the background noise level increased, the identification probability under the binaural condition decreased more rapidly than under the in-situ condition.

For the siren signal, the interaction effect between noise level and listening condition was not significant ($p = 0.759$). The two listening conditions showed the same noise slope, indicating that the binaural and in-situ conditions responded similarly to increasing background noise. Therefore, the main-effects GEE model was adopted. The main-effects GEE model for the siren signal was expressed as:

$$\log\left(\frac{p}{1-p}\right) = 5.5940 - 0.5950 \times \text{Method} - 0.1262 \times \text{Noise}$$

The negative noise coefficient indicates that the probability of correctly identifying loudspeaker timbre differences decreased as the background noise level increased. Since the interaction effect was not significant, the binaural and in-situ conditions showed similar trends in response to increasing noise.

For the speech signal, the interaction effect between noise level and listening condition was not significant ($p = 0.126$). Therefore, the main-effects GEE model was adopted. The main-effects GEE model for the speech signal was expressed as:

$$\log\left(\frac{p}{1-p}\right) = 6.3534 - 1.4097 \times \text{Method} - 0.1110 \times \text{Noise}$$

where: Method = 0 represents the binaural listening condition; Method = 1 represents the in-situ listening condition.

Therefore, under the binaural condition:

$$\log\left(\frac{p}{1-p}\right) = 6.3534 - 0.1110 \times \text{Noise}$$

Under the in-situ condition:

$$\log\left(\frac{p}{1-p}\right) = 4.9437 - 0.1110 \times \text{Noise}$$

These results indicate that, as the noise level increased, the identification probability for the speech signal decreased. The estimated coefficient was $\beta_{\text{noise}} = -0.1110$. After conversion, $e^{-0.1110} \approx 0.895$ which means that for every 1 dB increase in noise level, the probability of successful identification decreased by approximately 10.5%. At the same time, $\beta_{\text{method}} = -1.4097$ indicates that, under the same noise level, the overall identification accuracy in the in-situ condition was lower than that in the binaural condition.

Since this was a main-effects model, the two listening conditions shared the same noise slope, indicating that they exhibited similar decreasing trends as the noise level increased, although their overall identification levels were different. For the speech signal, this result suggests that background noise reduced identification performance under both experimental conditions, while the binaural condition still produced overall higher and more stable identification results.

4.4. Model Validation Results

Audio A, B, and C represent the binaural audio recordings obtained from playback through the three different loudspeakers. Audio signals with the suffix "mod" indicate spectrally modulated versions of the original recordings. For example, Audio A mod represents the modulated version of Audio A.

The Table 4.12 presents the prediction scores generated by the bass, music, and urgency models for the audio samples used in the model validation experiments. To facilitate the subsequent pairwise comparison analysis, the audio samples were ranked according to the model prediction results. The best-performing audio was labeled as I, the second-best as J, and the worst as K.

It should be noted that the meanings of the scores differ across different tasks. For the bass and urgency models, a higher model score indicates that the audio better matches the target optimization direction. In contrast, for the music model, since the optimization objective focuses more on tonal balance and naturalness, a lower model score indicates better audio quality.

For the above audio samples, the model predictions consistently followed the trend $I > J > K$, which was also reflected in the Figure 4.8. For both the bass and music audio conditions, the model predictions were generally consistent with the listeners' subjective evaluations, indicating that the proposed models were able to reflect the relationship between spectral variations and subjective perception relatively well. However, compared with the bass condition, where the $I > J > K$ trend was highly dominant and stable, the trend observed in the music condition was noticeably weaker. This suggests that the subjective perception of music signals was more variable and less strongly separated between different modulation conditions.

However, for the siren alarm signals, the prediction performance of the model was relatively weaker compared with the bass and music conditions. Further analysis suggested that different participants had different interpretations of the semantic meaning of "urgency." Some participants tended to associate sharper sounds, corresponding to higher sharpness values, with stronger urgency perception. In contrast, other participants were more likely to consider the degree of

Table 4.12.: Objective Evaluation Scores for Different Audio Categories

Audio Clip	Score	Mark
Bass		
Audio_A	0.29613504	I
Audio_A_mod	0.05260677	J
Audio_B	5.6923152×10^{-5}	K
Music		
Audio_A	1.704332	I
Audio_A_mod1	1.746207	J
Audio_A_mod2	2.248731	K
Urgency		
audio_A_mod2	1.1926	I
audio_C	1.0289	J
audio_A_mod1	0.9636	K

spectral variation as the primary factor contributing to perceived urgency.

It should be noted that the modulated audio signals in this study mainly focused on modifying spectral structure variations, while the high-frequency components themselves were not specifically enhanced or processed. Based on the observations described above, the researchers further clarified the semantic definition of "urgency" for the additional 14 participants recruited in the later experiments before the listening tests began. Specifically, participants were instructed to interpret stronger spectral variations as indicating higher perceived urgency.

The two participant groups were then analyzed separately, and the resulting siren pairwise comparison results are shown in the Figure 4.9.

The results show that, for the participants who were not given a constrained definition of "urgency," no unified evaluation trend was formed in the comparison between I and J. Overall, two opposite tendencies could be observed. One group of participants considered sounds with stronger spectral variations to be more urgent, while another group tended to perceive sharper sounds as being more urgent. As a result, the pairwise comparison results showed relatively clear disagreement.

In contrast, for the participant group that was explicitly instructed to interpret "greater spectral variation" as "higher urgency," the evaluation results became much more consistent across listeners. The ranking relationships obtained from the pairwise comparison analysis were also more closely aligned with the model prediction results.

These findings further suggest that the perceived urgency of alarm sounds is influenced not only

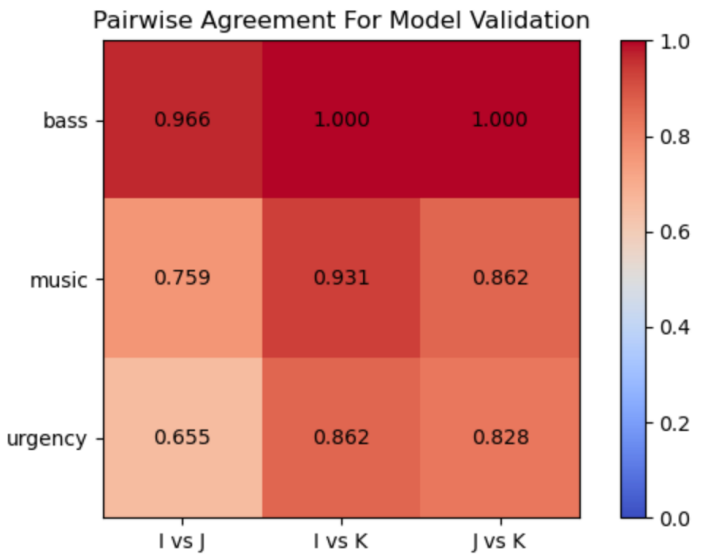


Figure 4.8.: Pairwise Results for Modulated Sound

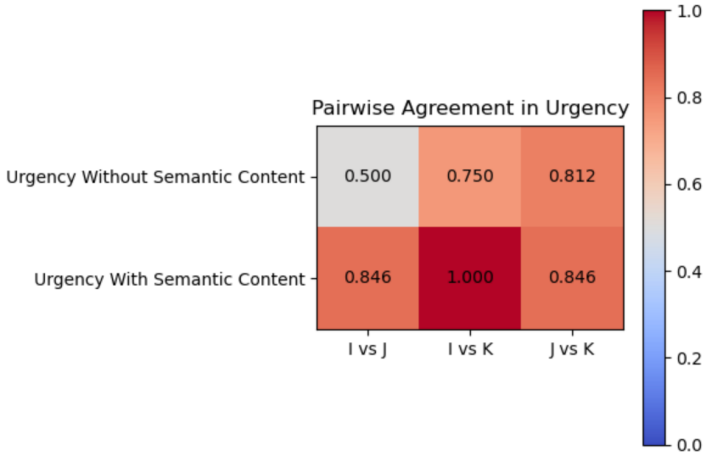


Figure 4.9.: Pairwise Results for Siren with/without Semantic Content

by the acoustic features themselves, but also by how listeners interpret the semantic meaning of "urgency." After the meaning of urgency was clearly defined, the prediction ability of the proposed model was further validated.

5. Discussion

5.1. Consistency of Subjective Evaluation

In this study, different types of audio content showed different levels of consistency between in-situ listening and binaural playback conditions. For the bass and music signals, the ranking trends were largely consistent across both listening conditions, suggesting that the key perceptual cues used to distinguish between loudspeakers could be reliably preserved through the binaural recording and headphone playback chain. However, some trend variations were observed for the siren and female speech signals. For the siren signal, listeners showed a clear C > A > B trend in response to the question of "which sound is more urgent" during the in-situ listening tests, while this trend became weaker in the binaural playback condition. In contrast, the female speech signal showed a more pronounced A > C > B trend during binaural playback, whereas the trend was less distinct under in-situ listening conditions. These results suggest that, although binaural recordings can preserve a large amount of loudspeaker-related information, listeners may rely on perceptual cues differently and assign different perceptual weights under different listening conditions.

For the siren signal, the observed changes are more likely related to differences in perceptual strategy and cue weighting. Siren-like sounds typically contain prominent frequency sweeps, strong high-frequency energy, and pronounced sharpness, making them highly salient perceptual cues. During in-situ listening tests, listeners usually have only limited opportunities for rapid comparison, and therefore tend to rely more on easily perceived cues such as brightness or sharpness, leading to the clear C > A > B trend. In the binaural playback condition, however, listeners were able to replay the audio repeatedly and make longer-term comparisons. As a result, their listening behavior may have shifted from an overall instantaneous judgment to a more analytical listening strategy, with greater attention paid to local timbral details and spectral variations. Consequently, the perceptual cues that dominated judgments during in-situ listening may have been reweighted, leading to the weaker ranking trend observed for the siren condition.

For the female speech signal, room reverberation may have been one of the main factors influencing the results. The listening room used in this study had a relatively long reverberation time, with an RT60 of approximately 600 ms in the mid–high frequency range and nearly 900–1000 ms at low frequencies. In such an environment, room reflections and late reverberation continuously modify the time-frequency structure of speech, resulting in temporal smearing and reverberant masking. Because speech clarity depends strongly on transient consonant information and formant structure in the mid-frequency range, long reverberation times can

reduce the perceptibility of subtle differences between loudspeakers. Under in-situ listening conditions, listeners were therefore more likely to form a general impression that "the speech from this loudspeaker is less clear," rather than further distinguishing finer timbral differences between loudspeakers. This may explain the weaker ranking trend observed for the female speech condition.

In contrast, under the binaural playback condition, although the room reverberation was still captured in the binaural signal, the spatial interaction process was effectively fixed within the recording chain. Listeners were no longer physically immersed in a dynamically changing reverberant sound field, and the room reflections did not vary with head movement or listening position. As a result, although the reverberation cues remained present, their masking effect on speech clarity judgment may have been weaker than in the in-situ listening condition. Under these circumstances, listeners were able to focus more consistently on the spectral structure and clarity of the speech itself, making it easier to distinguish differences between loudspeakers and resulting in the clearer A > C > B trend.

The experimental results obtained with background noise further demonstrate the influence of noise masking on the perception of loudspeaker differences. Unlike reverberation, which mainly affects temporal structure and speech clarity, background noise more directly reduces the effective signal-to-noise ratio (SNR), masking the weak perceptual cues that listeners rely on to distinguish between loudspeakers. As the noise level increases, listeners become less able to reliably perceive spectral details, transient changes, and local timbral structure, causing the differences between loudspeakers to become progressively less distinct. This masking effect is particularly pronounced for audio signals that depend on mid- and high-frequency details, since the corresponding perceptual cues are more easily masked by background noise.

At the same time, the differences between the in-situ and binaural experiments under noisy conditions also suggest that the listening task itself influences how listeners use the remaining perceptual cues. In noisy in-situ environments, listeners are usually limited to a small number of rapid judgments, while the added noise further increases perceptual uncertainty. As a result, listeners tend to rely more on overall impressions rather than stable local timbral cues. In contrast, binaural playback allows listeners to repeatedly replay the audio and gradually learn and reinforce the remaining spectral differences and timbral cues between loudspeakers. Therefore, even in the presence of background noise, listeners in the binaural playback condition are more likely to develop stable perceptual references and maintain more consistent identification trends.

Overall, these results suggest that the consistency between in-situ listening and binaural playback depends not only on whether the acoustic information is physically preserved, but also on changes in perceptual cue salience, auditory attention, room interaction, and listening strategy across different listening conditions. Reverberation mainly affects speech clarity and time-frequency structure through temporal smearing and reverberant masking, whereas background noise primarily masks subtle timbral differences between loudspeakers by reducing the effective SNR. When judgments rely more on stable spectral structure and local timbral detail, binaural playback is generally able to preserve loudspeaker differences more reliably. In contrast, when

judgments depend more heavily on perceptual cues such as speech clarity, brightness, or spatial salience, room acoustics and background noise are more likely to alter the weighting of these cues in subjective evaluation. These findings further suggest that, although the binaural recording system used in this study was not intended to perfectly reconstruct the complete real sound field, it was still able to preserve many of the key perceptual cues required for timbre-related loudspeaker evaluation.

5.2. Repeated Testing Is Required to Achieve Stable Results in In-Situ Listening

The experimental results show that, although both the in-situ and binaural conditions exhibited strong overall positive correlations, the subjective ranking results from the in-situ listening tests showed greater variability, whereas the binaural playback condition demonstrated higher consistency and stability. This suggests that different listening conditions not only affect how acoustic information is presented, but may also influence the judgment strategies listeners adopt during subjective evaluation.

In in-situ listening tests, subjective judgments in a real sound field are influenced by a larger number of dynamic factors. Since listeners simultaneously receive direct sound, room reflections, and spatial diffusion cues, small head movements, shifts in attention, and changes in reliance on different perceptual cues may all lead to different subjective judgments across repeated trials. In addition, human short-term auditory memory for complex timbral structure is inherently limited. When comparing multiple stimuli continuously, listeners are therefore more likely to rely on instantaneous overall impressions rather than consistently applying the same internal reference throughout the evaluation process. Together, these factors increase the variability of in-situ listening results.

In contrast, binaural playback experiments provide a more stable and repeatable listening environment. Because the binaural time-frequency structure is fixed within the recorded signal, listeners can repeatedly compare stimuli using a relatively consistent perceptual reference. At the same time, headphone playback reduces the dynamic spatial interactions present in real listening rooms, thereby minimizing the influence of environmental randomness on subjective judgments. As a result, listeners in the binaural condition are more likely to develop stable ranking trends and show higher inter-subject consistency.

These findings suggest that binaural recording methods are capable not only of preserving key perceptual cues from in-situ listening to a considerable extent, but also of providing greater experimental controllability and repeatability. For subjective listening tests focused on relative ranking and comparative evaluation, binaural recording has the potential to serve as an alternative evaluation method with high perceptual validity.

5.3. Loudness Consistency

The results from the two binaural recording experiments show that, despite differences in recording equipment, room conditions, and binaural capture methods between the dummy head and HATS setups, the overall ranking trends remained largely consistent for most audio signals. This suggests that, in loudspeaker subjective evaluation, certain spatial errors or geometric deviations in the binaural recording system do not necessarily disrupt listeners' overall perception of loudspeaker differences. For example, during the recording process, the acoustic center of each loudspeaker was manually aligned with the recording device, and the recording distance was controlled at approximately 1 m. However, small placement errors, angular deviations, or slight changes in recorder position could still introduce minor lateral shifts or changes in spatial centering during headphone playback. Nevertheless, the experimental results indicate that these spatial deviations did not significantly alter the overall ranking trends across listeners.

In contrast, loudness differences showed a much stronger effect. According to the LUFS measurements, the loudness differences between recordings in the dummy head condition remained relatively small, allowing the subjective ranking results to reflect the timbral differences between loudspeakers more consistently. In the HATS condition, however, abnormal gain changes in the recording chain introduced additional loudness deviations of nearly 3–4 dB between recordings, which is clearly within the perceptible range. Combined with the supplementary experiments conducted after loudness recalibration, the results further suggest that some of the observed changes in ranking trends were mainly caused by loudness differences rather than by the recording device, room conditions, or binaural recording method itself.

These findings indicate that, in loudspeaker subjective evaluation tasks, loudness consistency may be one of the most critical control variables if listeners are expected to judge primarily based on spectral characteristics, timbral structure, or spatial attributes of the loudspeakers themselves. Compared with small recording position errors, shifts in spatial centering, or differences between binaural recording systems, human listeners generally show much higher perceptual sensitivity to loudness variation. Once loudness differences exceed the perceptual threshold, subjective preference may shift noticeably toward the louder stimulus, thereby altering ranking results that would otherwise be based on timbral characteristics. The different sensitivity of various audio signals to loudness variation also reflects the competition between different perceptual cues during subjective evaluation. For bass signals, listeners mainly relied on low-frequency energy when making judgments. As a result, loudness changes could not introduce spectral structures that were not originally present, and therefore did not significantly alter the overall ranking trend. For music signals, however, loudness changes were more strongly coupled with overall preference judgments, as listeners often perceived louder audio as containing more detail, leading to reversals in the original ranking trends. In contrast, although the speech and siren conditions were also affected by loudness to some extent, listeners still retained their primary judgments related to speech clarity or urgency-related cues. Consequently, loudness differences mainly strengthened or weakened existing trends rather than completely reversing the ranking

order.

Overall, these results further suggest that, in binaural-recording-based loudspeaker evaluation experiments, strict loudness matching between recordings may be more important than achieving extremely precise spatial geometric consistency. As long as the binaural recording system can reliably preserve the main time-frequency structures and spatial cues related to loudspeaker reproduction, small spatial deviations generally do not significantly reduce subjective consistency at the group level. In contrast, once loudness differences exceed the perceptual threshold, they may become the dominant factor influencing subjective ranking results.

5.4. Differences Between Expert Evaluators and Ordinary Listeners

Further analysis of individual listener results revealed clear differences in the stability of subjective evaluations across participants with different backgrounds. The participants in this study included both subjects with audio engineering experience and ordinary listeners without professional training in the field. The results showed that audio engineers were generally able to clearly describe the perceptual cues they used to distinguish between different loudspeakers or audio samples, and their ranking results remained more consistent across repeated experiments. In contrast, non-expert listeners relied more on overall intuition or instantaneous impressions, resulting in larger variations across repeated tests.

This effect was particularly evident in the siren condition. Compared with bass or speech tasks, which involve relatively clear perceptual targets, the concept of "urgency" in the siren condition is a more abstract perceptual attribute. Unlike loudness, clarity, or bass strength, it does not have a single well-defined or universally shared perceptual definition. As a result, different listeners may rely on completely different perceptual cues when judging urgency. For example, in the follow-up modulation experiments based on different urgency definitions, some audio engineers regarded spectral variation and frequency dynamics as the primary contributors to perceived urgency. These listeners therefore tended to prefer the C > A > B trend in the original experiments, and after modulation that enhanced the relevant spectral features, they further preferred the modified audio over the original best-performing sample C. Meanwhile, other audio engineers considered sharpness itself to be the dominant cue for urgency perception, and therefore consistently selected C across all three binaural playback experiments.

In contrast, non-expert listeners lacked a stable and shared perceptual reference for judging urgency. Without clearly defined evaluation criteria or perceptual training, their judgments were more easily influenced by momentary attention, overall impressions, or whichever cue happened to be most salient at the time. As a result, their ranking results showed greater randomness across repeated experiments. This suggests that, for subjective evaluation of abstract perceptual attributes, the consistency and repeatability of experimental results may depend strongly on whether listeners possess stable cue interpretation strategies and listening habits.

These findings further suggest that, for loudspeaker evaluation tasks involving highly spe-

cialized or abstract perceptual attributes, appropriate listening training and the establishment of an expert evaluation panel with shared evaluation criteria may help improve the stability and repeatability of subjective test results. Under such conditions, highly consistent results may still be achieved even with a relatively small number of participants. In contrast, if the goal of the experiment is to understand the general subjective impressions of ordinary consumers, extensive listener training may not be necessary, but a larger participant pool may be required to reduce the influence of individual variability on the overall statistical results.

5.5. Broadband Spectral Structure and Perceived Naturalness

The experimental results suggest that listeners judge perceived naturalness mainly based on the global auditory impression formed by the overall spectral structure, rather than on changes within a single local frequency band. Compared with localized spectral modifications, broad spectral shifts across a wide frequency range were much more likely to reduce perceived naturalness. This indicates that, when evaluating naturalness, listeners pay greater attention to whether the sound preserves the overall spectral coherence and broadband balance of the original source.

The observed trends also suggest that, although the mid-frequency region plays an important role in timbre perception and sound identification, local spectral changes alone do not necessarily disrupt the overall sense of naturalness. In other words, the perceptual processes involved in recognizing "what a sound is" may operate differently from those involved in judging "whether a sound is natural." The former relies more heavily on local timbral cues, whereas the latter appears to depend more on whether the overall statistical structure of the sound matches real-world auditory experience.

These findings imply that perceived naturalness depends strongly on the relative balance across the entire audible frequency range. When the spectral structure is modified over a broad frequency range, the original relationships within the spectral envelope become disrupted, causing the sound to deviate from the long-term spectral characteristics typically associated with natural sound sources. Under such conditions, even if some local timbral cues are preserved, listeners may still perceive a reduction in overall realism or naturalness.

From the perspective of auditory cognition, prolonged exposure to natural acoustic environments allows humans to gradually develop internal statistical representations of real-world sound sources. As a result, when the overall spectral distribution deviates from these learned patterns, listeners may intuitively perceive the sound as "unnatural" or "artificial," even if they cannot explicitly identify the exact cause of the change. This further suggests that naturalness judgment involves not only low-level spectral perception, but also long-term auditory learning and cognitive expectations related to the statistical properties of natural sounds.

5.6. Semantic and Context Dependency in Acoustic–Perceptual Mapping

The main goal of the acoustic–perceptual mapping developed in this study was to establish a relationship between human subjective perception and the spectral features contained in recorded audio signals. However, the experimental results suggest that this relationship is not a fixed or universal physical mapping. Instead, it depends strongly on the semantic definition of the perceptual descriptor itself, the audio content, and the specific listening context.

First, if a stable and predictive acoustic–perceptual mapping is to be established, the subjective evaluation descriptors themselves must have clear and relatively consistent semantic definitions. For example, during the early stages of the siren experiments, listeners did not share a consistent understanding of "urgency." Some listeners focused more on sharpness and brightness, while others paid greater attention to frequency variation itself. As a result, different listeners relied on different perceptual cues during judgment, leading to large variations in the pairwise comparison results. After the definition of urgency was clarified further in the experiment, the ranking results became more consistent across listeners, and the model predictions aligned more closely with the subjective evaluations. This suggests that when a perceptual descriptor corresponds to multiple potential acoustic cues, listeners may base their judgments on different perceptual interpretations, reducing the stability of the acoustic–perceptual mapping. Therefore, the perceptual descriptors used for predictive modeling should correspond as closely as possible to acoustic dimensions that are clearly defined, perceptually stable, and consistently reproducible.

At the same time, the spectral features analyzed in this study do not represent the ideal spectrum of the original audio signal itself. Instead, they represent the binaural recording spectrum obtained after the audio had passed through loudspeaker playback, room propagation, and binaural recording. As a result, the spectrum inherently contains the combined effects of the audio content, loudspeaker characteristics, room acoustics, and the recording chain. In other words, the mapping established in this study corresponds to the relationship between subjective perception and the overall acoustic result reaching the listener's ears, rather than to a simple one-to-one relationship involving only a specific loudspeaker or frequency band.

This also highlights that loudspeaker subjective evaluation cannot be completely separated from the listening environment itself. In practical use, loudspeakers always operate within specific rooms and listening contexts, and the same audio content may be perceived differently depending on the loudspeaker and acoustic environment. For example, the same spectral modification may produce different perceptual effects for speech, music, and siren signals. Likewise, reverberation, spatial diffusion, and masking effects in different rooms can further alter the perceptual salience of these spectral features. Therefore, the acoustic–perceptual relationships associated with different audio content, subjective evaluation tasks, and playback conditions should be analyzed separately, rather than assuming that a single universal spectral mapping can apply across all listening scenarios.

In addition, the results further demonstrate the strong context dependency of subjective audi-

tory perception. Listeners' judgments are influenced not only by spectral variation itself, but also by the listening task, the semantic definition of the evaluation criteria, and the listening condition. When the perceptual task relies on clear and stable acoustic cues, the model predictions tend to show high agreement with subjective ranking results. In contrast, when the perceptual descriptor is more abstract or involves multiple perceptual dimensions simultaneously, listeners are more likely to develop different cue interpretations, increasing the uncertainty of subjective evaluation.

Therefore, the acoustic–perceptual mapping developed in this study is more suitable for predicting subjective evaluation results under specific listening conditions, playback chains, and perceptual tasks, rather than being considered a universal auditory model independent of experimental conditions and listening context.

6. Conclusion

This study proposed and validated a loudspeaker subjective evaluation method based on binaural recording and headphone playback. The experimental results show that, even without fully reconstructing the real sound field, the proposed method can still preserve the key perceptual cues required for timbre discrimination with relatively high stability, while maintaining strong consistency with in-situ listening in relative ranking and comparative evaluation tasks. Therefore, binaural-recording-based subjective evaluation has the potential to serve as a practical supplement to traditional in-situ listening tests, particularly for rapid subjective screening during early-stage loudspeaker design, remote listening tests, and highly repeatable comparative evaluations.

The results also demonstrate that the stability of loudspeaker subjective evaluation is influenced not only by playback conditions, but also by the type of perceptual cue, room acoustics, background noise, loudness consistency, and listener background. Among these factors, loudness consistency appears to be one of the most critical variables affecting subjective ranking results. In contrast, evaluation tasks involving more abstract perceptual definitions, such as speech clarity, sharpness, or urgency, are more easily influenced by room interaction, noise masking, and listening strategy. Therefore, subjective listening tests should carefully control room conditions, loudness matching, and listener training according to the specific perceptual task in order to improve the reliability and repeatability of the results.

This study further suggests that the relationship between human subjective perception and spectral features is not a fixed or universal physical mapping, but rather depends jointly on the audio content, listening task, and listening context. The acoustic–perceptual mapping established in this work essentially describes the relationship between subjective perception and the overall acoustic result reaching the listener after loudspeaker playback, room interaction, and binaural recording. In this sense, the proposed approach should be viewed more as a practical acoustic–perceptual analysis framework for studying the relationship between spectral structure and subjective perception under specific listening conditions and perceptual tasks, rather than as a universal predictive model that can be directly generalized to all sound scenarios and perceptual descriptors. Different types of audio content, playback environments, and subjective evaluation goals still require relatively independent analysis and modeling within their own listening contexts.

At the same time, the conclusions of this study are subject to certain experimental conditions and limitations. First, the validation results are based primarily on specific audio content and experimental procedures, and therefore do not imply that the proposed binaural recording method can be directly generalized to all types of audio signals or listening tasks. Second, the effec-

tiveness of binaural recording as a substitute for in-situ evaluation depends on the loudspeakers themselves having perceptible timbral differences that can be reliably detected by human listeners. If the differences between loudspeakers approach the threshold of auditory discrimination, stable and consistent subjective evaluation results may become difficult to achieve even with binaural recording methods. In addition, the experimental consistency observed in this study relied on strict control of both the recording and playback chains, including the use of standardized recording equipment, consistent playback conditions, and rigorous loudness calibration. Significant changes in these conditions may alter the corresponding acoustic–perceptual relationships.

Overall, this study supports the concept of perceptually relevant fidelity: the key to loudspeaker subjective evaluation is not the complete reconstruction of the real sound field itself, but rather the effective preservation of the perceptual cues and overall time-frequency structure required for human auditory judgment. Under appropriately controlled experimental conditions, binaural recording combined with headphone playback can provide a highly repeatable and practically useful alternative approach for loudspeaker subjective evaluation.

References

[Møl 92] Møller, H.: *Fundamentals of Binaural Technology*, Applied Acoustics, Vol. 36, No. 3–4, pp. 171–218, 1992

[Pau 09] Paul, S.: *Binaural Recording Technology: A Historical Review and Possible Future Developments*, Acta Acustica united with Acustica, Vol. 95, No. 5, pp. 767–788, 2009

[McA 19] McAdams, S.: *The Perceptual Representation of Timbre*, in *Timbre: Acoustics, Perception, and Cognition*, pp. 23–57, Springer, 2019

[McD 11] McDermott, J. H. and Simoncelli, E. P.: *Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis*, Neuron, Vol. 71, No. 5, pp. 926–940, 2011

[Pik 16] Pike, C. D.: *Timbral Constancy and Compensation for Spectral Distortion Caused by Loudspeaker and Room Acoustics*, Doctoral Dissertation, University of Surrey, 2016

[Too 88] Toole, F. E. and Olive, S. E.: *The Modification of Timbre by Resonances: Perception and Measurement*, Journal of the Audio Engineering Society, Vol. 36, No. 3, pp. 122–142, 1988

[Too 08] Toole, F. E.: *Sound Reproduction: Loudspeakers and Rooms*, Focal Press, 2008

[Bre 90] Bregman, A. S.: *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[Shi 08] Shinn-Cunningham, B. G.: *Object-based Auditory and Visual Attention*, Trends in Cognitive Sciences, Vol. 12, No. 5, pp. 182–186, 2008.

[Fri 07] Fritz, J., Elhilali, M. and Shamma, S.: *Adaptive Changes in Cortical Receptive Fields Induced by Attention to Complex Sounds*, Journal of Neurophysiology, Vol. 98, No. 4, pp. 2337–2346, 2007.

[Too 08] Toole, F. E.: *Sound Reproduction: Loudspeakers and Rooms*, Focal Press, 2008.

[Haa 08] Haas, E. C., Edworthy, J. and Hellier, E.: *Music, Acoustic, and Psychoacoustic Influences on Perceived Urgency in Auditory Warnings*, Human Factors, Vol. 50, No. 6, pp. 917–925, 2008.

[Jus 03] Juslin, P. N. and Laukka, P.: *Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?*, Psychological Bulletin, Vol. 129, No. 5, pp. 770–814, 2003.

[Bla 97] Blauert, J.: *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[Bre 90] Bregman, A. S.: *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[Koh 21] Kohnen, M. et al.: *Variabilities in Relation to Loudness Balancing*, Acta Acustica, Vol. 5, p. 40, 2021.

[Zwi 99] Zwicker, E. and Fastl, H.: *Psychoacoustics: Facts and Models*, 2nd ed., Springer, 1999.

[ISO 23] ISO: *ISO 226:2023 Acoustics — Normal Equal-Loudness-Level Contours*, International Organization for Standardization, 2023

[All 77] Allen, J. B. and Rabiner, L. R.: *A Unified Approach to Short-Time Fourier Analysis and Synthesis*, Proceedings of the IEEE, Vol. 65, No. 11, pp. 1558–1564, 1977.

[Zwi 61] Zwicker, E.: *Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)*, The Journal of the Acoustical Society of America, Vol. 33, No. 2, p. 248, 1961.

[Tra 90] Traunmüller, H.: *Analytical Expressions for the Tonotopic Sensory Scale*, The Journal of the Acoustical Society of America, Vol. 88, No. 1, pp. 97–100, 1990.

[Gre 78] Grey, J. M. and Gordon, J. W.: *Perceptual Effects of Spectral Modifications on Musical Timbres*, Journal of the Acoustical Society of America, Vol. 63, No. 5, pp. 1493–1500, 1978.

[Gra 76] Gray, A. H., Jr. and Markel, J. D.: *Distance Measures for Speech Processing*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 24, No. 5, pp. 380–391, 1976.

[McA 99] McAdams, S.: *Perspectives on the Contribution of Timbre to Musical Structure*, Computer Music Journal, Vol. 23, No. 3, pp. 85–102, 1999.

[Gre 85] Green, D. M. and Mason, C. R.: *Auditory Profile Analysis: Frequency Discrimination for Composite Spectra*, Journal of the Acoustical Society of America, Vol. 77, No. 2, pp. 453–459, 1985.

[Gre 83] Green, D. M. and Kidd, G.: *Further Studies of Auditory Profile Analysis*, Journal of the Acoustical Society of America, Vol. 73, No. 4, pp. 1260–1265, 1983.

[Hen 03] Henry, B. A. and Turner, C. W.: *The Resolution of Complex Spectral Patterns by Cochlear Implant and Normal-Hearing Listeners*, Journal of the Acoustical Society of America, Vol. 113, No. 5, pp. 2861–2873, 2003.

[Sot 16] Sottek, R.: *Sharpness Perception and Modeling of Stationary and Time-Varying Sounds*, Journal of the Acoustical Society of America, Vol. 140, pp. 2953–2954, 2016.

[Van 85] van Veen, T. M. and Houtgast, T.: *Spectral Sharpness and Vowel Dissimilarity*, Journal of the Acoustical Society of America, Vol. 77, No. 2, pp. 628–634, 1985.

[Haj 07] Hajda, J. M.: *The Effect of Dynamic Acoustical Features on Musical Timbre*, in *Analysis, Synthesis, and Perception of Musical Sounds*, Modern Acoustics and Signal Processing, Springer, New York, NY, 2007.

[Pee 04] Peeters, G.: *A Large Set of Audio Features for Sound Description (Similarity and Classification) from Temporal and Spectral Representations*, 2004.

[Has 09] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.

[Hoe 00] Hoerl, A. E. and Kennard, R. W.: *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, Vol. 42, No. 1, pp. 80–86, 2000.

[Hoe 70] Hoerl, A. E. and Kennard, R. W.: *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, Vol. 12, No. 1, pp. 55–67, 1970.

[Pul 15] Pulkki, V.: *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*, John Wiley & Sons, 2015.

[Won 07] Won, J. H., Drennan, W. R. and Rubinstein, J. T.: *Spectral-Ripple Resolution Correlates with Speech Reception in Noise in Cochlear Implant Users*, JARO, Vol. 8, No. 3, pp. 384–392, 2007.

[Bri 21] Bristow-Johnson, R.: *Audio EQ Cookbook*, W3C Working Group Note, 2021.

[Dru 94] Drullman, R., Festen, J. M. and Plomp, R.: *Effect of Temporal Envelope Smearing on Speech Reception*, The Journal of the Acoustical Society of America, Vol. 95, No. 2, pp. 1053–1064, 1994.

[Sin 03] Singh, N. C. and Theunissen, F. E.: *Modulation Spectra of Natural Sounds and Ethological Theories of Auditory Processing*, The Journal of the Acoustical Society of America, Vol. 114, No. 6, pp. 3394–3411, 2003.

[Dau 97] Dau, T., Kollmeier, B. and Kohlrausch, A.: *Modeling Auditory Processing of Amplitude Modulation*, Journal of the Acoustical Society of America, Vol. 102, No. 5, pp. 2892–2905, 1997.

[Opp 10] Oppenheim, A. V. and Schafer, R. W.: *Discrete-Time Signal Processing*, 3rd ed., Pearson, 2010.

[ITU 15] ITU-R BS.1116-3: *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, International Telecommunication Union, 2015.

[Fle 71] Fleiss, J. L.: *Measuring Nominal Scale Agreement Among Many Raters*, Psychological Bulletin, Vol. 76, No. 5, pp. 378–382, 1971.

[Lan 77] Landis, J. R. and Koch, G. G.: *The Measurement of Observer Agreement for Categorical Data*, Biometrics, Vol. 33, No. 1, pp. 159–174, 1977.

[Spe 04] Spearman, C.: *The Proof and Measurement of Association between Two Things*, The American Journal of Psychology, Vol. 15, No. 1, pp. 72–101, 1904.

[Spe 87] Spearman, C.: *The Proof and Measurement of Association between Two Things*, The American Journal of Psychology, Vol. 100, No. 3–4, pp. 441–471, 1987.

[Zar 14] Zar, J. H.: *Spearman Rank Correlation: Overview*, in *Wiley StatsRef: Statistics Reference Online*, 2014.

[Upt 14] Upton, G. and Cook, I.: *Fisher's z-Transformation*, in *A Dictionary of Statistics*, 3rd ed., Oxford University Press, 2014.

[Lia 86] Liang, K. Y. and Zeger, S. L.: *Longitudinal Data Analysis Using Generalized Linear Models*, Biometrika, Vol. 73, No. 1, pp. 13–22, 1986.

[Bal 04] Ballinger, G. A.: *Using Generalized Estimating Equations for Longitudinal Data Analysis*, Organizational Research Methods, Vol. 7, No. 2, pp. 127–150, 2004.

[Bra 97] Bradley, J. S., Soulodre, G. A.: *Factors Influencing the Perception of Bass in Concert Halls*, Proceedings of the Acoustical Society of America, Vol. 101, No. 5, pp. 3080–3090, 1997.

[Bar 03] Barron, M.: *Auditorium Acoustics and Architectural Design*, Spon Press, 2003.

[Tah 14] Tahvanainen, H., Pãtynen, J., Lokki, T.: *Perception of Bass with Musical Instruments in Concert Halls*, International Symposium on Musical Acoustics (ISMA), 2014.

[Oli 04] Olive, S. E., Welti, T., McMullin, E.: *Listener Preferences in Loudspeaker Sound Quality*, Journal of the Audio Engineering Society, Vol. 52, No. 5, pp. 434–454, 2004.

[Edw 01] Edworthy, J., Hellier, E.: *Alarm Sounds Should Be Loud and Angry: A Guide to Designing Auditory Warnings*, Human Factors, 2001.

[Drul 94] Drullman, R., Festen, J. M., Plomp, R.: *Effect of Temporal Envelope Smearing on Speech Reception*, The Journal of the Acoustical Society of America, Vol. 95, No. 2, pp. 1053–1064, 1994.

[ITU 01] ITU-R BS.1534-3: *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems (MUSHRA)*, International Telecommunication Union, 2001.

[Fle 71] Fleiss, J. L.: *Measuring Nominal Scale Agreement Among Many Raters*, Psychological Bulletin, Vol. 76, No. 5, pp. 378–382, 1971.

[Lan 77] Landis, J. R., Koch, G. G.: *The Measurement of Observer Agreement for Categorical Data*, Biometrics, Vol. 33, No. 1, pp. 159–174, 1977.

A. Appendix - Equipment

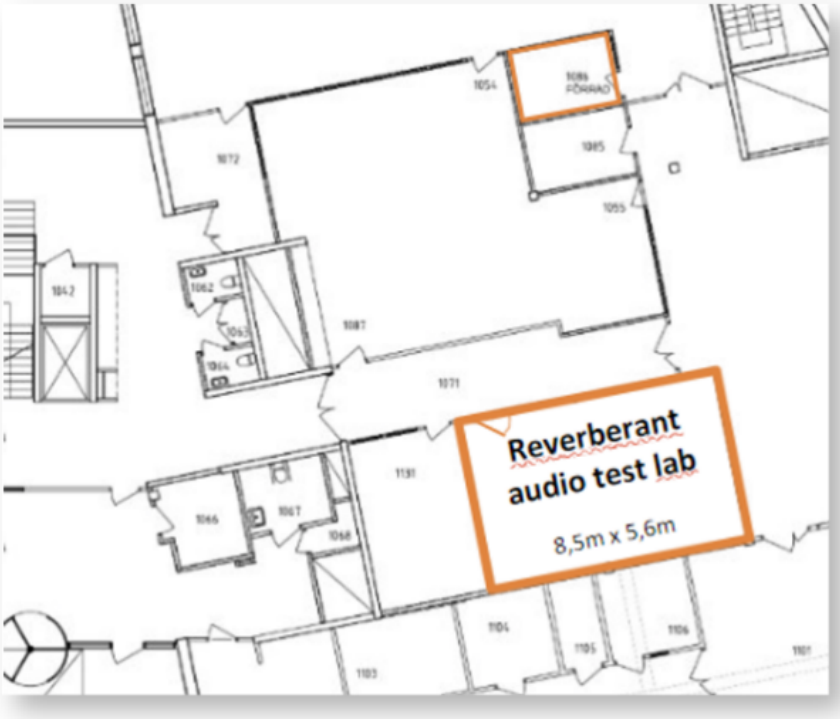


Figure A.1.: Room A Size

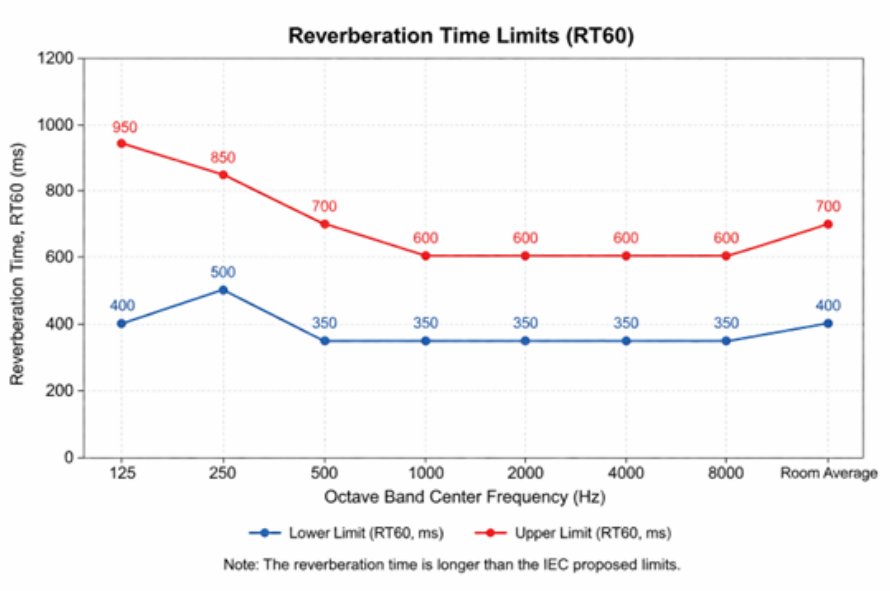


Figure A.2.: Reverberant Time of Reverberant Room

Acoustic environment	Noise, dBA	Noise, dB unweighted	Reverberation time
J025 (Golden ear)	8.7	45	

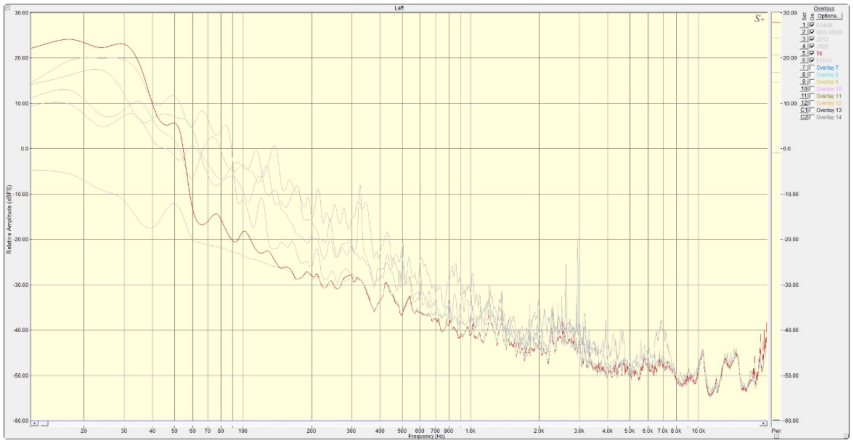


Figure A.3.: Room B dataset

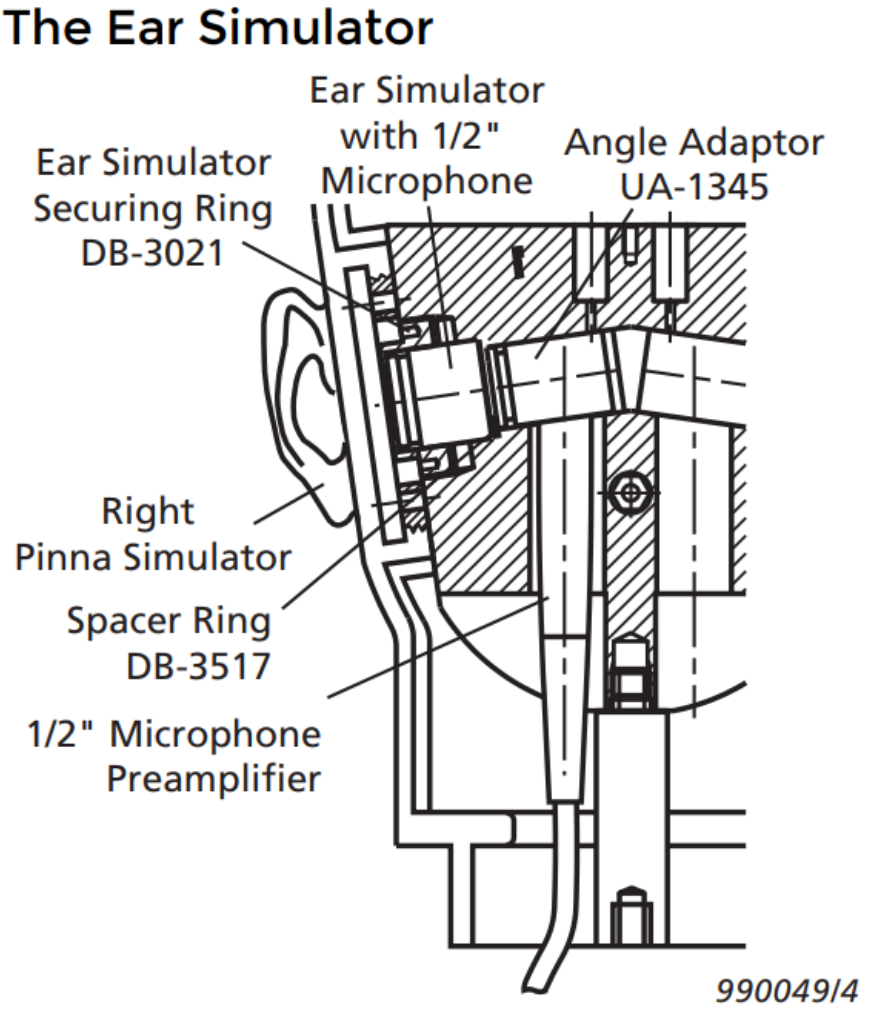


Figure A.4.: HATS simulator

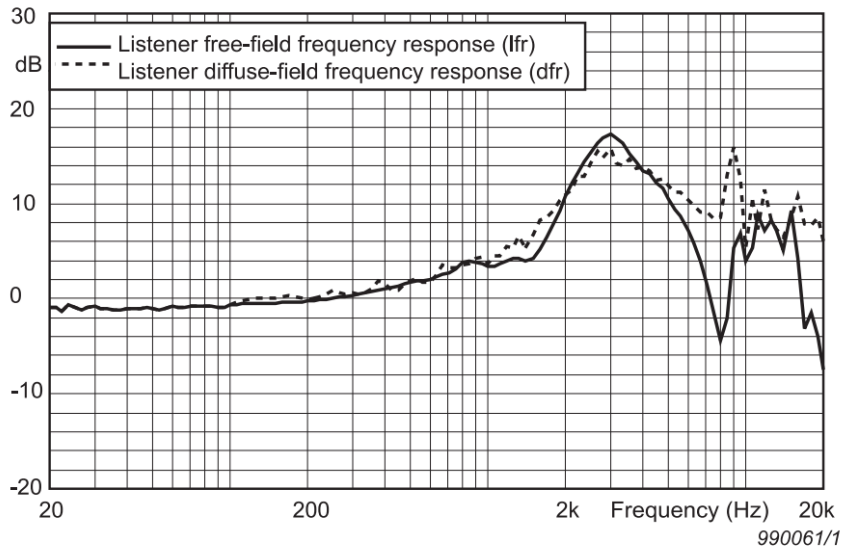


Figure A.5.: Microphone In HATS

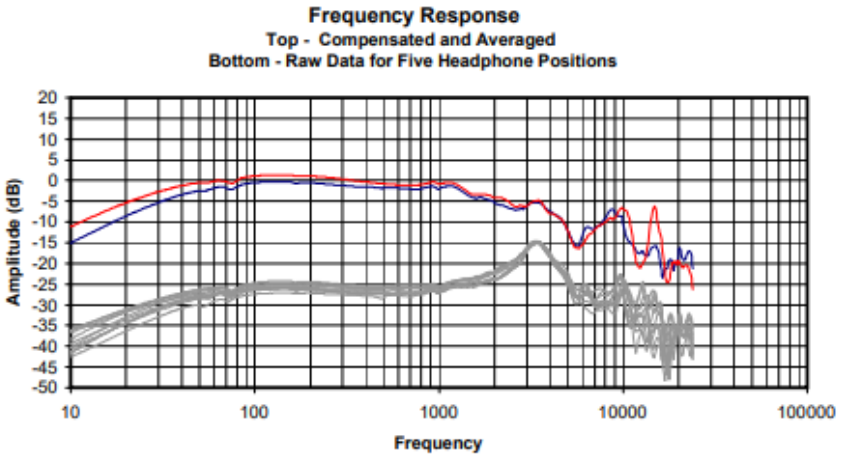


Figure A.6.: Frequency Response of HD650 Headphone

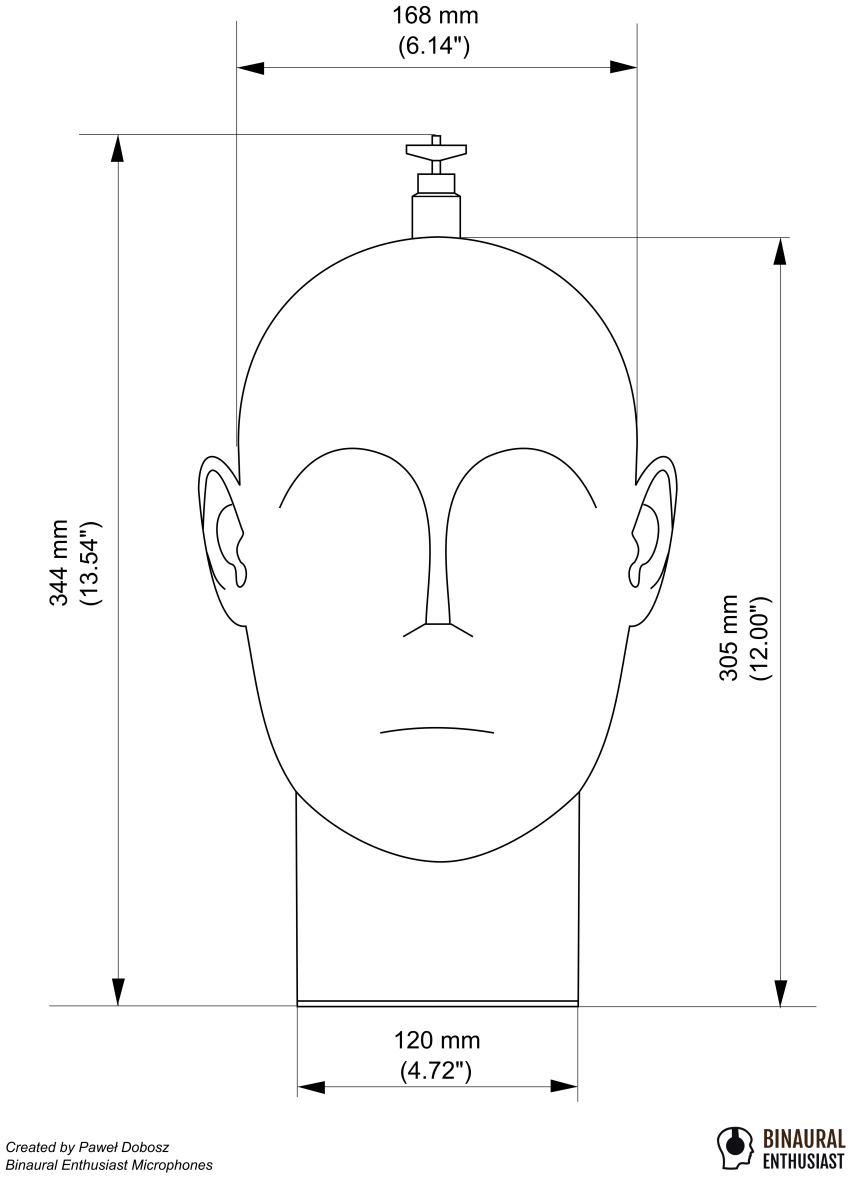


Figure A.7.: Front View of Dummy Head Microphone

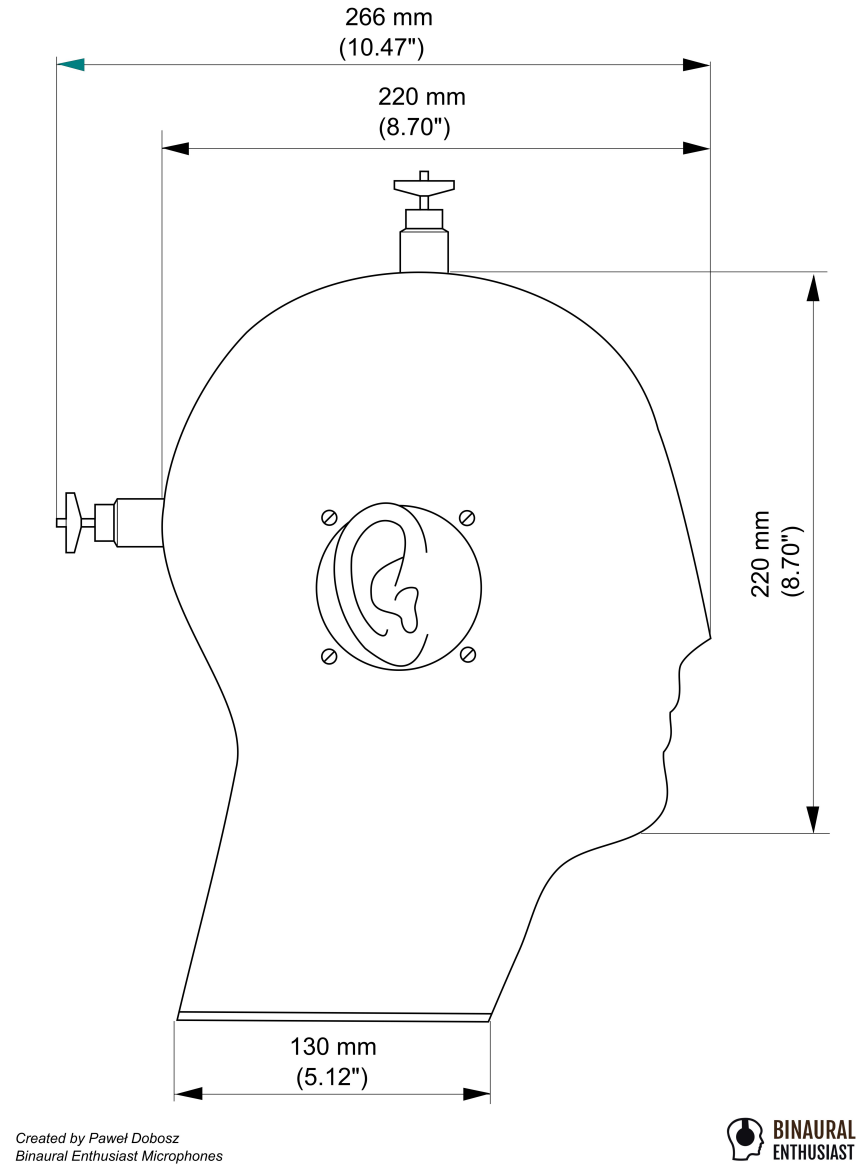


Figure A.8.: Side View of Dummy Head Microphone

B. Appendix - Listening Test

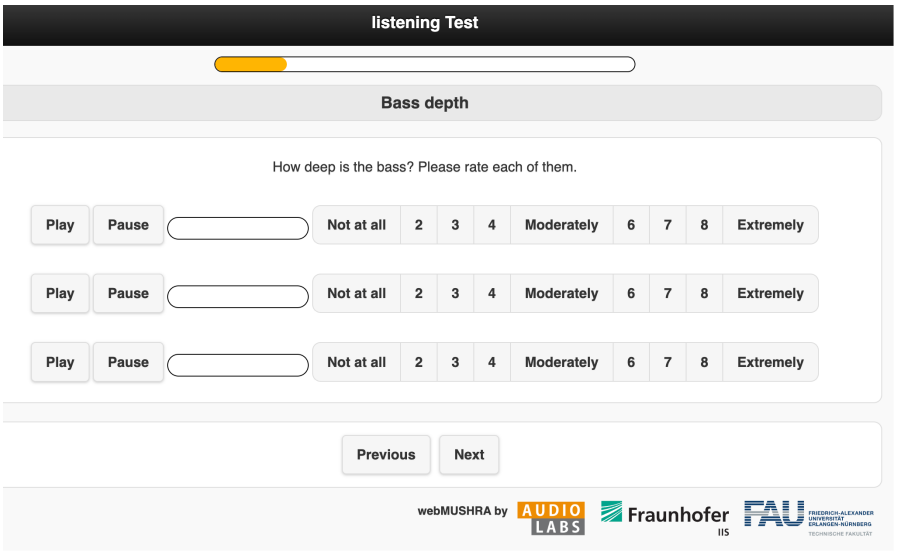


Figure B.1.: Binaural Listening Test 1 Rating Sample

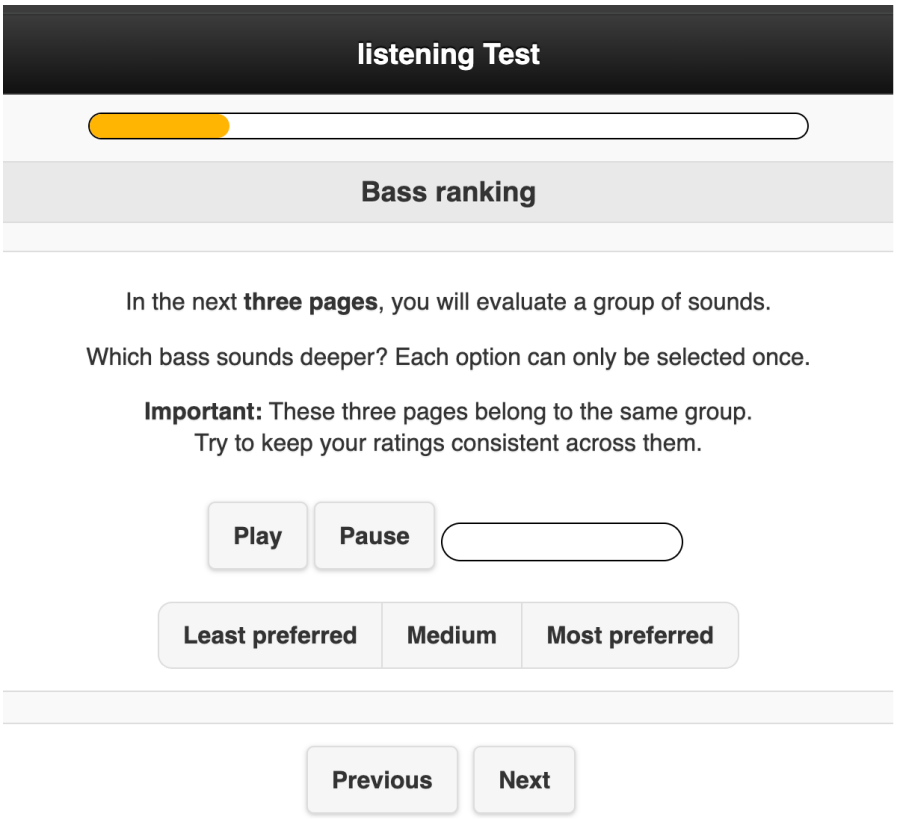


Figure B.2.: Binaural Listening Test 1 Ranking Sample 1

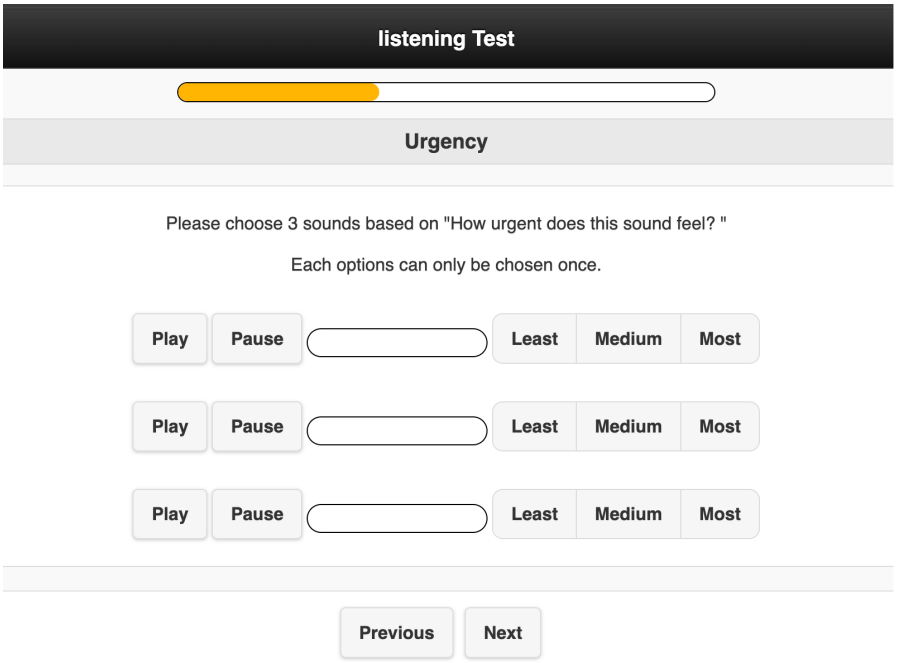


Figure B.3.: Binaural Listening Test Ranking Sample 2

Figure B.4.: Binaural Listening Test with Background Noise Sample

Test with background noise

In this section, you will hear some sound clips with background noise. As in the previous round of the experiment, please answer the questions based on your own perception

14. **Bass ranking**
Which bass sounds deeper? Please rank them from deepest to least deep.(eg.A>B>C)

15. **Music ranking**
Which music sounds most natural? Please rank them from most natural to least natural.(eg.A>B>C)

Figure B.5.: Noise Questionnaire Sample