# From Cheap to Chic: Enhancing Music Playback Quality of Budget Earphones via Hardware-Aware Learning

Changshuo Hu*
cs.hu.2023@phdcs.smu.edu.sg
Singapore Management University
Singapore, Singapore

Hung Manh Pham*
hm.pham.2023@phdcs.smu.edu.sg
Singapore Management University
Singapore, Singapore

Ting Dang
ting.dang@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Jiannan Li
jiannanli@smu.edu.sg
Singapore Management University
Singapore, Singapore

Rajesh Balan
rajesh@smu.edu.sg
Singapore Management University
Singapore, Singapore

Dong Ma†
dm878@cam.ac.uk
University of Cambridge
Cambridge, United Kingdom

## Abstract

Low-end earphones are widely spread due to their affordability, but their limited speaker hardware often leads to poor music playback quality. This raises a key question: Can we compensate for hardware limitations to enhance the listening experience without modifying the device? Existing EQ-based approaches attempt this, but they rely on frequency response curves (FRCs) measured under ideal conditions, which fail to capture real-world distortions such as harmonic and intermodulation effects.

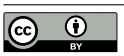
To address this, we propose Cheap2Chic, a two-stage, speaker-aware framework that improves music playback quality on low-end earphones using deep learning. In stage one, Cheap2Chic trains a *speaker characterization model* to capture the playback properties of the target earphone. In stage two, it introduces a *digital audio enhancement model*, prepended before the frozen *speaker characterization model*, that learns to adapt the input audio to compensate for hardware-induced distortions. To reduce the data collection burden required for accurate speaker characterization, Cheap2Chic also incorporates a data synthesis module. We built a prototype and collected a dataset to evaluate Cheap2Chic under diverse real-world conditions. Results show that Cheap2Chic significantly improves perceived music quality both objectively and subjectively, while remaining efficient on mobile devices.

## CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

*These authors contributed equally to this work.
†Corresponding author.

Authors' Contact Information: Changshuo Hu, cs.hu.2023@phdcs.smu.edu.sg, Singapore Management University, Singapore, Singapore; Hung Manh Pham, hm.pham.2023@phdcs.smu.edu.sg, Singapore Management University, Singapore, Singapore; Ting Dang, ting.dang@unimelb.edu.au, The University of Melbourne, Melbourne, Australia; Jiannan Li, jiannanli@smu.edu.sg, Singapore Management University, Singapore, Singapore; Rajesh Balan, rajesh@smu.edu.sg, Singapore Management University, Singapore, Singapore; Dong Ma, dm878@cam.ac.uk, University of Cambridge, Cambridge, United Kingdom.




ACM 2474-9567/2026/0-ART000
https://doi.org/10.1145/3774906.3800476

## Keywords

Earables, Audio Quality, Audio Processing

**ACM Reference Format:**
Changshuo Hu, Hung Manh Pham, Ting Dang, Jiannan Li, Rajesh Balan, and Dong Ma. 2026. From Cheap to Chic: Enhancing Music Playback Quality of Budget Earphones via Hardware-Aware Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 000 ( 2026), 13 pages. https://doi.org/10.1145/3774906.3800476

## 1 Introduction

Earphones have become the primary audio interface for modern digital experiences, supporting both immersive media playback (e.g., music delivery) and real-time communication (e.g., phone calls and online learning). In addition to the external form factor, such as over-head [16], in-ear [1], ear-clip [11], different earphones also distinguish themselves in their internal audio hardware, particularly in the quality of the speaker. This disparity is particularly noticeable during music listening, where premium earphones offer accurate frequency reproduction and rich timbre, while low-end models often suffer from weak bass, recessed midrange, or harsh treble. These limitations stem primarily from cost-driven design trade-offs, such as inferior speaker drivers, simplified amplification circuits, and poorly-designed acoustic enclosures.

Despite their degraded audio performance, low-end earphones remain one of the most widely distributed classes of audio devices. For instance, they are often provided as disposable in-flight earphones by airlines and serve as a practical, affordable option for lower-income populations. According to a Futuresource report [9], global earphone shipments reached 566 million units annually by 2024, with sub-USD 50 models accounting for 46% of the market [7]. This proportion is expected to be even higher in low-income regions. However, while affordable, these devices often fail to deliver satisfactory audio quality, particularly for music playback. Additionally, prolonged exposure to distorted or unbalanced audio may induce increased listening effort and contribute to long-term auditory strain or damage [26]. This raises a natural question: *Can we find an effective and practical way to compensate for the hardware limits of low-end earphones and enhance the listening experience*?

Addressing this requires two critical steps: (1) understanding the audio characteristics of the low-end speaker, and (2) based on that, designing a compensation scheme to modify the original digital audio, such that the resulting playback through the low-end speaker better preserves the intended audio quality. Existing

approaches for compensating speaker deficiencies primarily rely on equalization (EQ) [54]. It leverages the frequency response curve (FRC) as the speaker's characteristic, and adjusts the energy of different frequency bands in the digital audio signal to improve the perceived audio quality during playback. EQ methods generally fall into two categories. Graphic EQ [27, 43, 49] uses fixed frequency bands with sliders to boost or cut each band, offering simplicity at the cost of precision. In contrast, parametric EQ [19, 44, 45] provides fine-grained control over each band's frequency range and gain, enabling more precise audio tuning. More recently, automated systems such as AutoEQ [3] attempt to match a device's FRC to a target response (e.g., the Harman curve [41]) by optimizing a sequence of parametric filters.

However, EQ inherently has limitations in both key steps. First, FRC offers only a coarse, steady-state characterization of a device's behavior and fails to fully capture the true nature of the speaker. Specifically, FRC is typically measured using swept sine waves (i.e., one frequency at a time) under controlled conditions, which cannot capture behaviors such as harmonic distortion and intermodulation that arise when multiple frequency components are present, as in music playback. We elaborate on these limitations with experiments in Section 2.3. Second, EQ-based methods typically operate on a limited number of frequency bands (e.g., 10 or 16), offering only coarse granularity in adjusting the digital audio to compensate for speaker deficiencies. Furthermore, by relying solely on linear gain adjustments across bands, EQ fails to account for nonlinear distortion, transient smearing, and other time-varying artifacts present in continuous music streams.

To address these limitations, we propose Cheap2Chic, a speaker-aware framework designed to enhance music playback quality on low-end earphones. First (Stage 1), to accurately and comprehensively capture speaker characteristics, we propose a deep learning-based method that models the playback properties of low-end speakers by fitting a neural network (Section 3.2) in a data-driven manner. Trained on real-world data collected from commercial earphones, the model (denoted as *speaker characterization model*) effectively learns both steady-state and dynamic properties of the hardware. To reduce the burden of collecting large training data, we also develop a data synthesis module that combines coarse-grained hardware information extracted from the FRC using a vision-language model (VLM), with a small amount of real device data to generate high-quality synthetic training samples (Section 3.4).

Second (Stage 2), to effectively compensate for low-end hardware, we propose refining the digital audio using a deep neural network so that the modified signal delivers improved audio quality when replayed through the target low-end speaker (Section 3.3). Specifically, we freeze the *speaker characterization model* trained in Stage 1 simulating the low-end hardware properties, and append it to another model (denoted as *digital audio enhancement model*). By training this cascaded architecture end-to-end from raw digital audio to high-quality recordings captured from high-end earphones, the *digital audio enhancement model* learns to adapt the input signal to better suit the low-end hardware for producing high-quality audio. Notably, the *speaker characterization model* is used only during training to guide learning and only the *digital audio enhancement model* is required during inference.

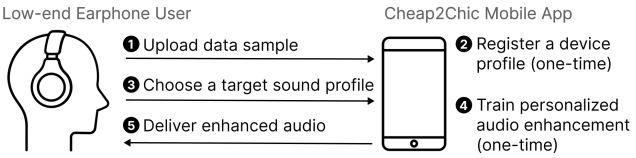



Fig. 1: **A typical five-step workflow for low-end earphone users to use Cheap2Chic to enhance audio quality.**

To evaluate Cheap2Chic, we first developed a prototype and collected an audio dataset spanning a variety of music genres using eight commercial earphones across different price points. We then conducted comprehensive experiments to assess the impact of various factors in practical settings, using both objective and subjective evaluation metrics. Results show that Cheap2Chic significantly enhances the perceived audio quality of low-end earphones and operates efficiently on mobile devices. It requires no extra hardware or sensors on the earphones, and any existing earphones can be used without modification, as all computation runs on the companion smartphone. Overall, this paper makes the following contributions:

- We proposed Cheap2Chic, a novel framework that enhances music delivery quality on low-end earphones. Cheap2Chic innovates with (1) a deep learning-based approach for speaker characterization and (2) a cascaded architecture that learns effective compensation strategies to address speaker deficiencies.
- We designed a data synthesis module that accurately simulates earphone recordings by leveraging coarse-grained FRC data. This approach innovatively treats the FRC as an image and extracts features using a vision-language model (VLM).
- We built a prototype and collected a dataset of playback recordings from eight commercial earphones across ten music genres (totaling 40 hours). *We plan to release this dataset publicly to facilitate and encourage broader research on earphone audio quality.*
- We conducted extensive experiments to evaluate the effectiveness and system overhead of Cheap2Chic under various real-world factors.

## 2 Background and Motivation

### 2.1 Application Scenario

Figure 1 illustrates the envisioned user experience of Cheap2Chic through a typical five-step usage flow. Alice, who regularly listens to music using a pair of low-end earphones, aims to enhance her listening experience with the Cheap2Chic mobile app. She begins by (1) connecting her low-end earphones to the mobile phone and performing a one-time device registration by uploading a small amount of sample data to the Cheap2Chic mobile app. (2) The app uses the sample data to build a device-specific profile (i.e., the *speaker characterization model*). (3) Alice can then select a preferred sound profile as the target of the enhancement in the app. For example, she may choose the profile of a high-end earphone model that features strong bass. (4) The app prepares personalized audio enhancement to match the sound profile of the selected device through a brief one-time training phase. (5) Once the training is completed (i.e., obtaining the *digital audio enhancement model*), Alice can play music directly through the Cheap2Chic mobile app and enjoy a significantly improved audio experience (e.g. stronger bass and other enhancements) tailored to her device. The entire

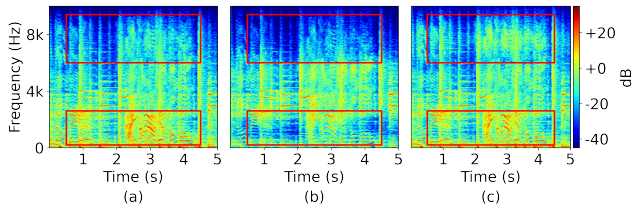



Fig. 2: **Spectrograms of (a) the original digital audio, the recording from (b) a low-end earphone, and (c) a high-end earphone.**

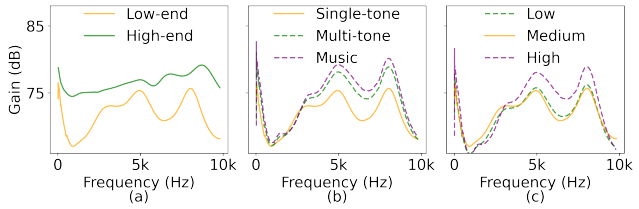



Fig. 3: **(a) FRC comparison of a high-end (Sennheiser IE 300) and a low-end (Fransun E20) earphone; FRCs of Fransun E20 measured under (b) different input audio types and (c) different playback volumes.**

process is fast, intuitive, and requires no technical expertise, offering perceptual upgrades without any hardware changes.

## 2.2 Low-end vs. High-end Earphones

A key distinction between low-end and high-end earphones lies in the capability of the playback component, namely the *speaker* [51]. To illustrate this gap, we play the same music clip with a Fransun E20 earphone (low-end, 5 USD) and a Sennheiser IE 300 earphone (high-end, 220 USD), using the prototype described in Section 4.1. Figure 2 presents the spectrograms of (a) the original digital signal, (b) the recording from the Fransun E20 earphone, and (c) the recording from the Sennheiser IE 300 earphone. It is evident that the Sennheiser IE 300 earphone produces a recording that closely matches the spectral profile of the original digital audio, while the output from the Fransun E20 low-end earphone shows significant spectral deviation. In particular, the lower red box highlights a marked reduction in energy below 2 kHz, indicating weaker bass reproduction. The upper red box shows that the high-frequency harmonics are attenuated and smeared, revealing a loss of harmonic detail and potential phase distortion.

To further explain these differences, we examine the frequency response curves (FRCs) of the two earphones in Figure 3(a). The low-end earphone exhibits a pronounced dip in the sub-2 kHz range, leading to substantial energy loss in the low-to-mid frequency band. This range is where music typically carries the most energy, so its suppression is likely to degrade the perceived audio quality. In contrast, the high-end earphone maintains the overall spectral structure, preserving harmonic clarity and producing playback that more faithfully reflects the original signal.

## 2.3 Limitations of FRC

The FRC is a standard acoustic characterization that describes how a playback device amplifies or attenuates different frequency components. It is typically measured using single-tone sine sweep inputs played at fixed sound pressure levels [6]. The recorded signal is then analyzed to compute the gain across frequencies, resulting in a device-specific response curve under controlled conditions. However, based on experimental results from the low-end Fransun E20 earphone in Figure 3(b) and 3(c), we observe that *FRCs measured under such controlled conditions fail to accurately capture playback behavior in more realistic scenarios*, such as when the input audio contains complex mixtures of frequency components or is played at varying volume levels.

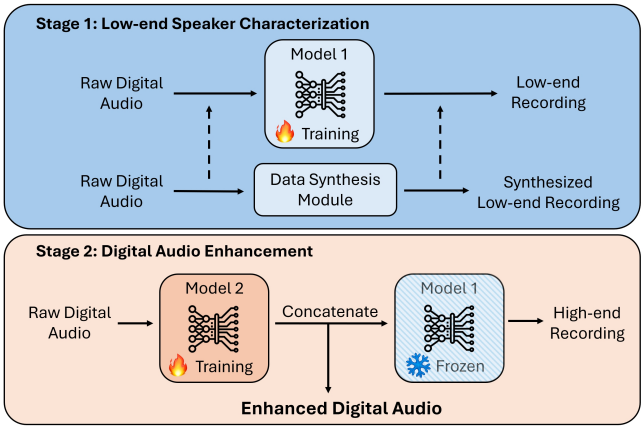



Fig. 4: **System overview of Cheap2Chic. The *Low-end Speaker Characterization* stage maps raw audio to low-end recordings via Model 1, aided by synthesized data. The *Digital Audio Enhancement* stage then transforms raw audio into enhanced audio, supervised by high-end recordings through the frozen Model 1.**

Specifically, when the digital audio includes multiple tones (e.g., in multi-tone signals or music), these components interact with each other during playback, introducing harmonic distortion[1] and intermodulation distortion[2][33, 48]. Additionally, when playing tones at different volumes, the speaker exhibits distinct forms of non-linear behavior [23] . At high volumes, excessive diaphragm excursion leads to distortion due to overdrive[3], while at low volumes, insufficient driving force results in weak low-frequency output and nonlinear attenuation of low frequencies[4]. As a result, the effective frequency response in these conditions deviates from the *standard FRC* commonly reported in device datasheets, as illustrated in Figure 3 (b) and (c). This discrepancy suggests that traditional EQ-based approaches, which rely on standard FRCs, may yield suboptimal audio enhancement performance in real-world use.

# 3 System Design

## 3.1 Overview

Cheap2Chic is a two-stage framework that enhances the perceived playback quality on low-end earphones through deep learning. As shown in Figure 4, Cheap2Chic comprises three key components.

In Stage 1, Cheap2Chic characterizes the playback behavior of the low-end earphone by training a deep neural network on its recordings (Section 3.2). The resulting *speaker characterization model* (denoted as Model 1) effectively captures various playback

[1]Harmonic distortion introduces energy at integer multiples of the input frequency.
[2]Intermodulation distortion generates additional frequency components at the sums and differences of input frequencies.
[3]At higher volumes, stronger magnetic forces may push the diaphragm beyond its linear range, resulting in increased harmonic and intermodulation distortion.
[4]At lower volumes, the driving force may be insufficient to overcome the resistance of the diaphragm, resulting in attenuation, particularly in low-frequency components.

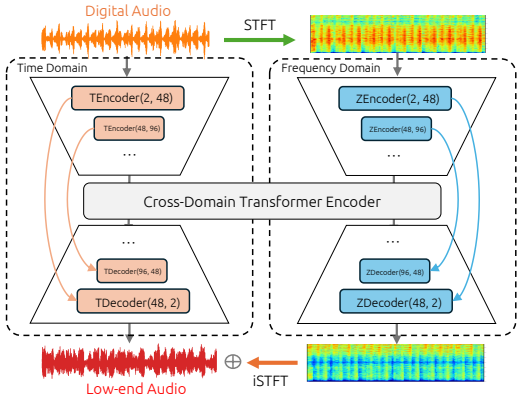



Fig. 5: **The Base Model used in Cheap2Chic, which is a dual-branch hybrid architecture adapted from Demucs v4, consists of a time-domain U-Net and a frequency-domain U-Net connected via a cross-domain transformer.**

characteristics of the target device, such as frequency response, harmonic distortion, intermodulation distortion, and etc. However, training a reliable and accurate *speaker characterization model* requires extensive recordings from the target device. To mitigate the burden of data collection, Cheap2Chic further incorporates a data synthesis module that generates synthetic audio recordings closely resembling real recordings from the low-end device (Section 3.4).

In Stage 2, Cheap2Chic introduces a *digital audio enhancement model* (denoted as Model 2), which is prepended before the frozen Model 1. This cascaded architecture is trained end-to-end using high-quality recordings as supervision. Through this process, the enhancement model learns to adapt the input digital audio such that, when played through the low-end earphone, the output closely matches the high-quality reference (Section 3.3). The knowledge embedded in Model 2 serves as a transformation strategy that compensates for the hardware limitations of the low-end speaker. Notably, while Model 1 is essential during training in Stage 2 to guide learning, only Model 2 is needed during actual deployment to modify the digital audio before streaming to the low-end speaker.

## 3.2 Low-end Speaker Characterization

The low-end speaker characterization phase (Stage 1) aims to capture the distortion characteristics of low-end earphones by learning how these devices alter clean digital audio signals. Specifically, the model aims to internalize the playback-related properties of the device that affect the signal. To achieve this, we propose training a neural network (i.e., *speaker characterization model*) that maps clean digital audio inputs to the corresponding low-end recordings captured after playback through the target devices.

Specifically, the *speaker characterization model* adopts a dual-branch hybrid architecture, combining both time domain and frequency domain modeling, adapted from Demucs v4 [46], as illustrated in Figure 5. This design aims to capture the complex distortion characteristics and temporal-spectral dependencies of low-end earphones. Each branch uses a U-Net as the base model to learn detailed signal representations in its respective domain. The time-domain branch operates directly on raw waveforms using a five-layer encoder-decoder structure. The frequency-domain branch first converts the waveform into a spectrogram via Short-Time Fourier Transform (STFT) with a 4096-point FFT, window

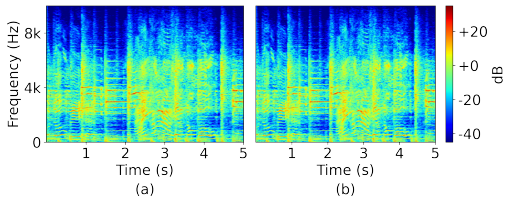



Fig. 6: **Spectrograms of (a) the true low-end recording, and the outputs of (b) our *speaker characterization model*.**

size of 4096, and hop size of 1024. This spectrogram is then processed through the U-Net architecture and reconstructed back to a waveform using inverse STFT (iSTFT), before merging with the time-domain output. Additionally, a Cross-Domain Transformer Encoder is applied to both branches in the latent space to facilitate effective feature fusion and capture cross-domain dependencies, improving the model's ability to represent complex distortions. Eventually, the reconstructed waveforms from both branches are merged via waveform-level summation to produce the final outputs, which have the same shape as the input waveforms.

Notably, since the original Demucs v4 is designed for music source separation with multiple output streams, we adapt it to our task by modifying the last convolutional layer in both branches. Specifically, we reduce the out-channel dimension from multiple sources (i.e., 4) to 1, resulting in a single audio waveform. This modification simplifies the model for our single-target mapping scenario and removes unnecessary outputs. Additionally, by adapting the model to output a single target branch, we also streamline the alignment between time and frequency representations, enabling more direct and effective cross-domain attention compared to the original multi-source setting.

To validate whether the *speaker characterization model* effectively captures the complex characteristics of low-end speakers, we examine the spectrograms of the true low-end recording and the outputs of our *speaker characterization model*. As shown in Figure 6, the spectrum produced by our model (b) closely matches the true low-end recording (a) across both low- and high-frequency regions. The reconstructed spectrogram preserves bass energy, harmonic structures, and fine spectral details, suggesting the effectiveness of the proposed learning-based hardware characterization strategy.

## 3.3 Digital Audio Enhancement

The objective of this stage is to train a *digital audio enhancement model* that modifies a clean digital input such that, after playback on a target low-end device, the perceived audio approximates the high-fidelity output of a reference high-end device. To achieve this, we adopt a cascaded architecture as shown in Figure 4: we first learn an enhancement mapping from the original digital audio to its adapted version, and then simulate the playback quality of low-end earphones by passing the adapted audio through the *speaker characterization model* trained in Stage 1. Specifically, only the enhancement model is trained in this phase, while the Stage 1 *speaker characterization model* is kept frozen to simulate the fixed distortion characteristics of the low-end device.

We reuse the same dual-branch hybrid architecture for the enhancement model to ensure structural compatibility with the speaker simulator and maintain stable, consistent learning across stages. The enhancement model takes a digital audio input and generates

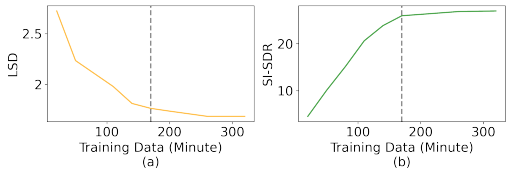



Fig. 7: **Effect of training data size on low-end speaker characterization. (a) LSD and (b) SI-SDR.**

an intermediate enhanced signal, which is then passed through the frozen simulator to produce a simulated playback. The model is trained by minimizing the perceptual discrepancy between this simulated output and the reference high-end recording, which is obtained by playing the same input through a high-end earphone and recording it using the same setup as in Stage 1. This supervision encourages the enhancement model to learn a pre-distortion function that offsets the degradation introduced by the hardware.

While the entire cascaded model is trained end-to-end to optimize the simulated output, i.e., the enhanced audio passed through the frozen speaker characterization model, to match the quality of high-end playback, only the enhancement model will be used during deployment. The speaker simulator serves solely as a training-time proxy to model device-specific distortions, allowing the enhancement model to learn how to pre-compensate for playback artifacts without requiring the real device.

Once training is complete, the system requires no further calibration or adaptation. At inference time, any clean digital waveform can be passed through the *digital audio enhancement model* alone to produce an audio signal optimized for playback on low-end devices.

## 3.4 Data Synthesis Module

*3.4.1 Motivation.* Different from traditional EQ-based methods, Cheap2Chic adopts a learning-based approach to model the playback characteristics of low-end earphones (Section 3.2). Due to its data-driven nature, this approach typically requires a substantial amount of training data to achieve comprehensive and accurate modeling. To assess the data requirements, we train the *speaker characterization model* using varying amounts of audio data and evaluate the resulting audio quality improvements using two metrics: Log-Spectral Distance (LSD) and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), as defined in Section 5.3. As shown in Figure 7, performance improves with increasing training data, and approximately 170 minutes of audio is required to train a reliable model. However, collecting nearly 3 hours of recordings for every new low-end device places a significant burden on users, thereby limiting the practicality of Cheap2Chic for real-world deployment.

*3.4.2 Solution overview.* To address this challenge, and with the potential to reduce real data dependency, Cheap2Chic introduces a data synthesis module that can transform input digital audio into synthetic recordings mimicking actual recordings from a specific low-end speaker. As the synthesis must be device-specific, *the core insight of Cheap2Chic is to leverage the coarse cues provided by the device's FRC to guide the generation process.* Figure 8 illustrates the overall pipeline of the synthesis module. First, Cheap2Chic innovatively employs a Vision-Language Model (VLM) to extract coarse embeddings from the FRC. These embeddings are then processed by a Feature-wise Linear Modulation (FiLM) block to derive two conditioning factors. Secondly, the generation model takes the digital

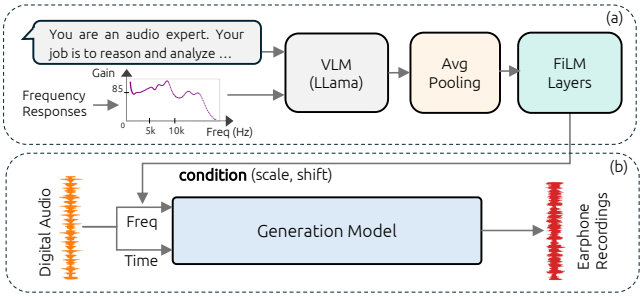



Fig. 8: **Overview of the data synthesis module: (a) extracting coarse information from the FRC graph using VLM, and (b) synthetic data generation with FiLM-based conditioning.**

audio as input and uses the FiLM outputs as conditions to guide the synthesis of device-specific recordings. This model is trained with supervision from real recordings of the earphones corresponding to the given FRCs. Once the model is trained or further fine-tuned when necessary, we can generate synthetic data corresponding to the conditioning FRC. Next, we describe the detailed process.

*3.4.3 Visualizing FRC and extracting coarse information using VLM.* While FRC fails to capture complex and fine-grained properties such as harmonics and intermodulation of the playback device (as discussed in Section 2.3), it still provides coarse-grained information about how an earphone shapes audio across frequencies. Therefore, we propose to incorporate FRC as a high-level, device-specific feature to guide the synthesis process. To do so, we first need to extract meaningful and representative information from the FRC. This task aligns naturally with the capabilities of Large Language Models (LLMs), which can reason over structured patterns and infer relationships that conventional numerical methods often miss. However, since the FRC is basically a high-dimensional numerical array representing frequency-dependent scaling factors, LLMs struggle to interpret them effectively such as extracting meaningful patterns like trends, resonance peaks, or correlations between different frequency ranges.

Inspired by how human audio experts interpret FRCs, typically by visualizing them as line graphs, we convert each FRC into a line-graph image that preserves its overall shape and makes key patterns more interpretable. The X- and Y-axis scales further provide context for understanding frequency and gain variations. To capture these perceptually relevant cues beyond raw numerical features, we employ a vision-language model (VLM; Llama-3.2-11B-Vision-Instruct [13]), which can reason over structured visual inputs and extract compact embeddings that better describe device-specific characteristics.

Specifically, after plotting the FRC graph with labeled X and Y axes and appropriate scales, we introduce a prompt to further guide the model's reasoning: *"You are an audio expert. Your job is to reason and analyze the given frequency response of an earphone device."*. As the VLM's output is spatial and high-dimensional (multiple token-level embeddings), we apply an average pooling layer to aggregate tokens and form a compact 4096-dimensional embedding that captures coarse device-level characteristics. Following that, the embedding is injected into the generation model via the FiLM [42] module, which consists of two dense layers to produce learnable shift and scale parameters. These parameters serve as conditioning inputs for the subsequent data generation module

training stage. Compared to using the full 4096-dimensional embedding directly, the scale and shift parameters offer a lightweight and more interpretable way to modulate the spectrogram (e.g., amplifies/suppresses and supplements features), enabling simple yet efficient adaptation to different device characteristics.

*3.4.4 Synthetic data generation with FiLM-based conditioning.* Using the FRC-derived shift and scale parameters as conditioning signals, we build a data generator that takes digital audio input, conditioned on these device-specific factors, to generate low-end device recordings. We adopt the same base model structure (Figure 5) as in Stage 1 and Stage 2 because they all share the similar audio-to-audio transformation task. Specifically, the shift and scale parameters extracted from the FiLM layers are used to modulate the frequency branch, affecting the spectrogram after the STFT and normalization layers. This approach is motivated by the fact that the VLM-derived FRC embeddings primarily capture coarse spectral characteristics, making them better suited to guide frequency-domain representations rather than time-domain features. The cross-modulation module can still integrate information from both frequency and time branches to produce device-specific recordings. This model is trained using real pairs of digital audio and their corresponding recordings from various devices, conditioned by their paired FRCs.

*3.4.5 Model fine-tuning for unseen devices.* The data generator described above can already synthesize high-quality data for low-end devices seen during training. However, its generalizability to unseen devices depends heavily on the diversity of the device training set. With a sufficiently diverse pretraining pool, the generator could better generalize from frequency response curves, laying the groundwork for eliminating real data altogether. In practice, collecting recordings from a large and varied set of earphones (e.g., 100 or 1000 models) is highly challenging. To address this limitation, we adopt a two-stage data generation approach: a pretraining phase, where a general data generation pipeline is learned using all available recordings from a limited range of devices, followed by a fine-tuning phase, in which the model is adapted using recordings from the target low-end device to achieve high-quality synthesis.

In our experiment, the pretraining phase is conducted on a dataset encompassing all devices listed in Table 1, excluding the target device, i.e., 7 devices. The fine-tuning phase for a target device is performed using a small set of real recordings (e.g., 30 minutes) to develop a device-specific generation model that better captures its unique playback characteristics. After fine-tuning, this model synthesizes additional training data from external digital audio corpora. The synthetic data is then combined with the limited real recordings to train the low-end *speaker characterization model*. This could significantly reduce the user burden by 82% (from 170 minutes to 30 minutes), making Cheap2Chic more practical and scalable for real-world applications, as motivated in 3.4.1.

# 4 Prototyping and Data Collection

## 4.1 Prototype

As Cheap2Chic aims to enhance the *perceived* music quality on low-end earphones, it is essential to train the models using recordings that *reflect how audio is perceived by human ears*. To this end, we developed a customized prototype, as illustrated in Figure 9. We employ an artificial ear model [10], which conforms to the IEC-711 standard and is typically used for earphone testing, to simulate human ears. It consists of an anatomically shaped pinna, a molded ear canal that replicates acoustic reflections, and a calibrated microphone positioned at the location of the tympanic membrane. For in-ear earphones, we insert them into the ear canal until a stable seal is achieved, while over-ear earphones are placed to fully cover the pinna, mimicking typical usage. Both playback and recording are managed by a Bela Mini development board [4] to ensure synchronization. Specifically, the external earphone receives digital audio via a 3.5 mm jack, and the microphone output captured at the tympanic membrane is streamed back to the board through a separate 3.5 mm jack. Recordings are sampled at 44.1 kHz with 16-bit resolution and saved in lossless WAV format for downstream analysis. Our prototype setup offers a more accurate reflection of human auditory perception than free-field microphone recordings, while also providing greater stability and eliminating inter-individual variability compared to collecting with real human subjects.

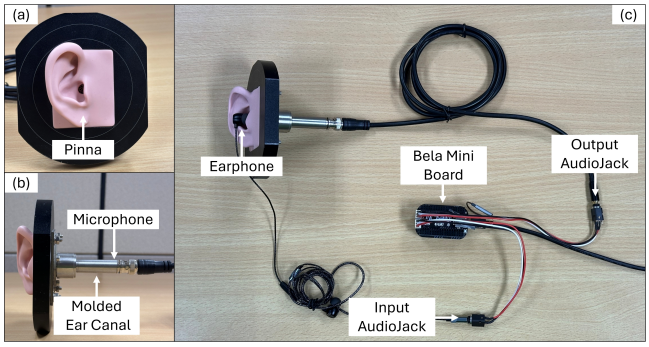



Fig. 9: **Illustration of the prototype and recording setup. (a) Front view of the artificial ear with an anatomically shaped pinna. (b) Side view showing the molded ear canal and microphone at the eardrum position. (c) Full setup including the tested earphone, artificial ear, and Bela Mini board for synchronized playback and recording.**

Table 1: **Summary of selected commercial earphones for dataset collection and performance evaluation.**

| Label | Model | Type | Price |
|---|---|---|---|
| IE-L1 | Fransun E20 [8] | In-ear | 5 USD |
| IE-L2 | JBL C100SI [12] | In-ear | 15 USD |
| IE-H1 | Aurvana Air [2] | In-ear | 200 USD |
| IE-H2 | Sennheiser IE 300 [15] | In-ear | 220 USD |
| OE-L1 | Mezone P2961 [14] | Over-ear | 7 USD |
| OE-L2 | Zihnic Foldable [17] | Over-ear | 20 USD |
| OE-H1 | Sony WH-1000XM5 [16] | Over-ear | 300 USD |
| OE-H2 | Beyerdynamic DT700 PRO X [5] | Over-ear | 320 USD |

## 4.2 Dataset Collection

Using the prototype, we collected a dataset comprising diverse music genres recorded through various earphone models to evaluate the performance of Cheap2Chic.

*Earphone Models:* To cover a broad range of usage scenarios, we selected eight commercial earphone models, including both in-ear (IE) and over-ear (OE) types. For each type, we chose two low-end models (L1 and L2), priced below 20 USD, and two high-end models (H1 and H2), priced above 200 USD. The low-end models serve as the target devices whose perceived audio quality Cheap2Chic aims to enhance, while the high-end models act as references for the high-quality audio experience users desire. For example, if the owner of a Franrun E20 seeks audio quality comparable to that of

the Sennheiser IE 300, then Franrun E20 and Sennheiser IE 300 are treated as the low-end and high-end models, respectively, during the development of the models described in Section 3. Table 1 summarizes the selected earphones.

*Source Music Corpus:* To record music with the selected earphones, we used the GTZAN dataset [53], a widely adopted benchmark corpus in music information retrieval. Its broad genre coverage enables us to evaluate how earphones respond to a wide range of frequency content and musical styles. We utilized all 10 musical genres[5] included in the GTZAN dataset. For each genre, we selected 60 audio clips, each 30 seconds long. All clips were converted to mono, uncompressed WAV format at 44.1 kHz for playback.

*Dataset Recording:* During playback, we positioned the right earphone over the artificial ear and routed audio clips exclusively through the right channel (muting the left channel to eliminate crosstalk). All recordings were conducted in a quiet indoor environment to minimize external noise. In total, we recorded *8 (earphones) × 10 (genres) × 60 (clips) × 0.5 minutes = 2,400 minutes (or 40 hours)* of music data that reflect realistic human auditory perception. We plan to release this dataset to the community to facilitate broader research on earphone audio quality.

# 5 Experiment setup

## 5.1 Device Pair Selection

As different users may own different types of low-end devices and seek to enhance the perceived quality to match various high-end earphones, we evaluate Cheap2Chic using diverse earphone combinations listed in Table 1. Specifically, we used eight representative device pairs, each mapping a low-end source device to a high-end reference device. These include in-ear earphone pairs: (IE-L1→IE-H1), (IE-L2→IE-H1), (IE-L1→IE-H2), (IE-L2→IE-H2); and over-ear headphone pairs: (OE-L1→OE-H1), (OE-L2→OE-H1), (OE-L1→OE-H2), (OE-L2→OE-H2). Unless otherwise specified, evaluation results are averaged across all eight pairs.

## 5.2 Data Preprocessing and Model Training

*Data preprocessing:* We segment the audio clips using a sliding window approach, with each sample being 5 seconds long with a stride of 0.5 seconds. Eventually, in our two-stage pipeline (default), the final dataset comprises 21,276 training, 5,319 validation, and 2,995 testing digital-playback audio pairs. It is worth noting that in our default setting, the low-end speaker characterization stage utilizes both synthesis and real data. Specifically, 18,912 (training) and 4,137 (validation) pairs among the total are generated by the data generation module. Meanwhile, all real data in the digital audio enhancement stage is used because it is supposed that high-quality earphones are predefined in the test-time deployment and a large corpus of recordings are available.

*Model Training:* We train the models using the AdamW optimizer with a learning rate of $5 \times 10^{-5}$, weight decay of 0.01, and parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1 \times 10^{-6}$. The low-end speaker characterization stage is trained for 20 epochs with a batch size of 16, while the digital audio enhancement stage is trained for 10 epochs with a batch size of 8. Both stages optimize mean squared error (MSE) loss. Early stopping is applied based on the validation loss to prevent overfitting. All experiments are conducted using two NVIDIA A100 GPUs, each with 40 GB of memory.

[5]The 10 genres are: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock.

## 5.3 Evaluation Metrics

To quantitatively evaluate the performance of Cheap2Chic, we adopt four complementary metrics that capture waveform fidelity, perceptual quality, and spectral similarity. All metrics use the recordings from high-end playback in the respective pair as the reference.

**SI-SDR**: The Scale-Invariant Signal-to-Distortion Ratio measures the waveform reconstruction quality while remaining invariant to amplitude scaling. Higher values indicate better performance.

**PEAQ [50]**: Perceptual Evaluation of Audio Quality is a standardized metric that estimates subjective audio quality, with scores ranging from 1 (very annoying) to 5 (imperceptible). Higher scores indicate better perceived fidelity.

**LSD [24]**: Log-Spectral Distance measures the average Euclidean distance between the log-magnitude spectrograms of the reference and test signals. Lower values indicate better spectral alignment.

**FAD [31]**: Fréchet Audio Distance quantifies perceptual similarity between two sets of audio in an embedding space derived from a deep learning model (e.g., VGGish). Lower FAD values indicate greater similarity in timbre and texture.

## 5.4 Baselines

We benchmark the performance of Cheap2Chic against both traditional EQ-based methods and recent deep learning approaches. Notably, existing deep learning methods do not account for hardware characteristics; therefore, for a fair comparison, we integrate their network architectures into Cheap2Chic 's two-stage pipeline and train them under the same data conditions as Cheap2Chic, i.e., using 30 minutes of real data and 150 minutes of synthetic data.

**AutoEQ [3]**: AutoEQ is a representative earphone equalization method that designs a 16-band Finite Impulse Response (FIR) filter using biquad peaking equalizers to match the frequency response of a source (low-end) device to that of a reference (high-end) device.

**DeFTAN-II [34]**: This is a single-branch, multi-channel speech enhancement model that integrates CNNs and Transformers. It partitions feature maps into locally emphasized and original subgroups and processes them using a dual-path feedforward network.

**HS-TasNet [55]**: This model is designed for music source separation using a hybrid framework comprising a waveform branch and a spectrogram branch. It also incorporates stacked LSTM blocks in the latent space to capture the long-range temporal dependencies.

# 6 Performance Evaluation

By default, for each low-end device, Cheap2Chic requires 30 minutes of real recordings to fine-tune the data synthesis model. This model then generates 150 minutes of synthetic data, which, together with the original 30 minutes of real data, is used to train and validate the *speaker characterization model*. Afterwards, the *digital audio enhancement model* is developed using all the 180 minutes of data from a high-end device as high-quality audio can be provided with the Cheap2Chic service provider without user involvement.

To evaluate the *actual perceived* performance of Cheap2Chic, we follow a three-step procedure: (1) select additional digital audio files spanning various genres from the GTZAN corpus (i.e., previously unseen during model development), (2) process them using the trained *digital audio enhancement model* to obtain the modified digital signals (i.e., inference), (3) *replay the modified signals* through

Table 2: **Performance across different baselines and variants.**

| System | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|
| Low-end Playback | 0.05 | 2.43 | 7.41 | 1.40 |
| AutoEQ [3] | 9.75 | 3.03 | 6.60 | 0.78 |
| DeFTAN-II [34] | 16.18 | 3.22 | 4.71 | 0.64 |
| HS-TasNet [55] | 17.20 | 3.68 | 4.49 | 0.43 |
| Limited-RealData | 16.64 | 3.35 | 4.52 | 0.41 |
| Full-RealData | 20.74 | 4.04 | 2.66 | 0.11 |
| **Cheap2Chic** | **20.56** | **3.91** | **2.93** | **0.14** |

the low-end devices, and record the output using the same setup described in Section 4. All evaluation results are derived from this newly collected test set, which consists of 5 minutes of recording per genre for each device pair, totaling 10 hours of evaluation data.

## 6.1 Overall Performance

In addition to the baselines presented in Section 5.4, we also compare Cheap2Chic against several of its variants. Specifically, *Low-end Playback* refers to the original recordings from the low-end earphones without any enhancement, i.e., without applying Cheap2Chic. *Limited-RealData* denotes the case where only 30 minutes of real data are used to train the *speaker characterization model*, whereas *Full-RealData* represents the scenario where the full 180 minutes of real data from the low-end device are used for training.

Table 2 presents the average results across eight device pairs. The reported metrics reflect relative differences between low-end and high-end playback, lower scores indicate perceptual gaps rather than the low-end baseline being unlistenable. We can observe that: (1) *Low-end Playback* yields the poorest performance across all metrics, highlighting the low quality of the original audio and the necessity of Cheap2Chic. (2) *Limited-RealData* significantly improves performance over *Low-end Playback*, despite requiring only 30 minutes of device-specific data, demonstrating the effectiveness of the two-stage pipeline in Cheap2Chic. (3) *Full-RealData* further enhances performance, as the increased amount of real data allows for more accurate and comprehensive hardware characterization of the low-end earphone. (4) Cheap2Chic achieves performance that is only marginally lower than *Full-RealData*, indicating that the data synthesis module is capable of generating high-quality recordings. Together, the *Limited-RealData* and *Full-RealData* variants illustrate how synthesized data influence performance, reflecting the role of the data synthesis module and serving as an implicit ablation study.

When compared with the three baselines, Cheap2Chic consistently achieves superior performance. Specifically, AutoEQ, which applies a fixed set of filters to align the frequency response of the low-end device with that of a high-end target, fails to capture and compensate for the nuanced characteristics of low-end hardware. As a result, it performs worse than learning-based approaches. Although both HS-TasNet and DeFTAN-II adopt the same two-stage pipeline as Cheap2Chic, they still underperform. This is because Cheap2Chic jointly extracts time-domain and frequency-domain features and leverages a cross-attention mechanism to enable more effective representation learning.

To provide a visual illustration of Cheap2Chic's effectiveness, Figure 10 presents spectrograms from (a) the Low-end Playback, (b) the target high-quality playback, and (c) the playback enhanced

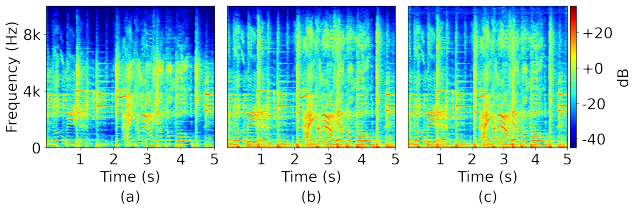



Fig. 10: **Spectrograms of (a) the low-end playback, (b) the target high-quality playback, and (c) the playback enhanced by Cheap2Chic.**

Table 3: **Performance comparison between Low-end Playback and Cheap2Chic across device pairs.**

| Device Pair | System | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|---|
| **Intra-type Pairs** | | | | | |
| (IE-L1→IE-H1) | Low-end Playback | 0.44 | 3.02 | 5.82 | 1.79 |
| | **Cheap2Chic** | **22.30** | **3.90** | **2.38** | **0.21** |
| (IE-L1→IE-H2) | Low-end Playback | -1.58 | 2.86 | 7.59 | 1.34 |
| | **Cheap2Chic** | **22.32** | **4.05** | **3.00** | **0.18** |
| (IE-L2→IE-H1) | Low-end Playback | -2.47 | 2.80 | 10.07 | 1.12 |
| | **Cheap2Chic** | **18.71** | **3.57** | **3.97** | **0.25** |
| (IE-L2→IE-H2) | Low-end Playback | -2.08 | 2.05 | 8.52 | 0.79 |
| | **Cheap2Chic** | **19.39** | **3.77** | **4.16** | **0.14** |
| (OE-L1→OE-H1) | Low-end Playback | 2.52 | 1.91 | 6.21 | 1.58 |
| | **Cheap2Chic** | **20.93** | **4.08** | **1.78** | **0.08** |
| (OE-L1→OE-H2) | Low-end Playback | -1.62 | 1.83 | 5.52 | 1.80 |
| | **Cheap2Chic** | **20.98** | **3.99** | **3.35** | **0.07** |
| (OE-L2→OE-H1) | Low-end Playback | 5.77 | 2.23 | 7.87 | 1.42 |
| | **Cheap2Chic** | **19.20** | **3.89** | **1.78** | **0.11** |
| (OE-L2→OE-H2) | Low-end Playback | -0.59 | 2.73 | 7.68 | 1.39 |
| | **Cheap2Chic** | **20.88** | **3.89** | **3.01** | **0.08** |
| **Cross-type Pairs** | | | | | |
| (IE-L1→OE-H1) | Low-end Playback | -6.46 | 3.06 | 11.32 | 2.06 |
| | **Cheap2Chic** | **16.24** | **3.71** | **3.56** | **0.15** |
| (IE-L2→OE-H1) | Low-end Playback | -14.33 | 2.09 | 15.30 | 1.38 |
| | **Cheap2Chic** | **17.54** | **3.37** | **6.36** | **0.26** |
| (OE-L1→IE-H2) | Low-end Playback | -2.09 | 3.02 | 6.83 | 1.99 |
| | **Cheap2Chic** | **19.86** | **3.83** | **2.93** | **0.11** |
| (OE-L2→IE-H2) | Low-end Playback | 2.52 | 2.13 | 9.58 | 1.70 |
| | **Cheap2Chic** | **20.20** | **3.89** | **3.16** | **0.09** |

by Cheap2Chic, using the (IE-L1→IE-H2) device pair. It is evident that, by modifying the digital audio, the low-end earphone is able to reproduce high-quality playback closely resembling that of the high-end reference, thereby improving the perceived audio quality.

## 6.2 Impact of Different Device Pairs

To assess whether Cheap2Chic consistently enhances playback quality across different device combinations, we evaluate its performance on a diverse set of low-end to high-end pairs. In addition to the intra-type pairs described in Section 5.1, we include cross-type cases where the low-end and high-end devices differ in form factor, such as in-ear to over-ear, or vice versa. These cross-type settings enable us to evaluate the generalizability of Cheap2Chic across distinct device architectures and to investigate whether physical differences between device types affect the fidelity of enhancement.

Table 3 summarizes performance across 12 device pairs. For each pair, we compare Cheap2Chic with *Low-end Playback*. Across all device pairs, Cheap2Chic consistently improves playback quality over the low-end baseline. This improvement holds even for

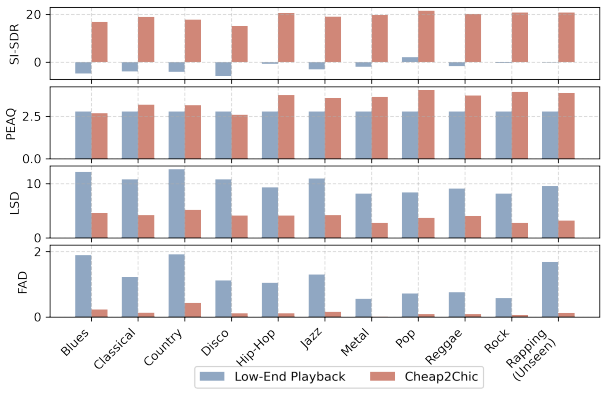



Fig. 11: **Performance comparison across ten GTZAN genres and one unseen genre.**

extremely low-cost models, for example, the $5 in-ear earphone (IE-L1) achieves SI-SDR gains exceeding 20 dB when paired with high-end targets. Both in-ear and over-ear categories benefit similarly, suggesting that Cheap2Chic effectively compensates for a wide range of hardware degradation profiles. However, performance on cross-type pairs is less uniform: enhancing over-ear playback to match in-ear targets yields strong results, while mapping from in-ear to over-ear is more challenging. This asymmetry likely arises from fundamental physical differences: compact in-ear devices are inherently limited in reproducing the expansive soundstage and broadband dynamics of over-ear headphones, due to constraints in speaker size and acoustic enclosure design [18]. Nevertheless, the gains in SI-SDR still exceed 16 dB.

## 6.3 Impact of Different Genres

In real-world scenarios, users listen to a wide variety of music genres that are characterized by distinct spectral and temporal properties. To evaluate how Cheap2Chic performs across such diverse content, we analyze its enhancement results over the ten music genres included in our collected corpus, as well as an unseen genre. As shown in Figure 11, all genres exhibit clear improvements over the *Low-end Playback* baseline, indicating that the system consistently enhances playback quality regardless of content.

Among them, genres such as Pop, Metal, and Hip-Hop show the most significant gains. These styles typically feature strong rhythmic structures, concentrated spectral energy, and regular patterns, which make them easier for the system to model and enhance. In contrast, Classical, Jazz, and Country display slightly smaller improvements, likely due to their soft dynamics, broader dynamic ranges, and more intricate spectral structures, which are more prone to distortion on low-end devices and harder to recover.

## 6.4 Impact of Music Playback Volume

Earphone users could listen to music at varying playback volumes depending on their environment, which may influence perceived sound quality. To evaluate the impact of playback volume, we conducted an experiment using the (IE-L1→IE-H2) device pair. Specifically, we applied a model trained at medium volume to enhance a set of digital audio signals, which were then played on IE-L1 at three common loudness levels [32]: low (65–75 dB), medium (75–85 dB), and high (85–95 dB). Each playback was compared against the reference recording from IE-H2 at the corresponding volume level.

Table 4: **Performance of Cheap2Chic across playback volume levels.**

| Volume Level | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|
| Low (65–75 dB) | 22.29 | 3.90 | 3.01 | 0.23 |
| **Medium (75–85 dB)** | **22.32** | **4.05** | **3.00** | **0.18** |
| High (85–95 dB) | 18.71 | 3.57 | 3.19 | 0.33 |

As shown in Table 4, Cheap2Chic still performs well across all tested volume levels. However, we observe that it achieves the best results when the playback volume matches the training condition, yielding the lowest spectral distortion and highest waveform fidelity. At lower volumes, performance degrades slightly, likely due to limited diaphragm excursion, which reduces signal energy and masks subtle spectral details. At higher volumes, the degradation is more pronounced across all metrics, which might be caused by excessive voltage driving the speaker beyond its linear operating range, thereby increasing harmonic distortion. Currently, Cheap2Chic does not incorporate volume variation; pretraining with clips across multiple loudness ranges or conditioning the model on volume-dependent features could further improve robustness.

## 6.5 Potential Improvement with Data Synthesis

As demonstrated in Section 6.1, the proposed data synthesis module is capable of generating high-quality synthetic recordings that accurately reflect the hardware characteristics of the target device, thereby reducing the user's data collection burden. This naturally raises an important question: *as a data-driven approach, can Cheap2Chic's performance be further improved by generating more synthetic data for model training?*

To explore this, we select the (IE-L1→IE-H2) device pair and augment the 30 minutes[6] of real data with varying amounts of synthetic data (ranging from 0 to 330 minutes) for training the *speaker characterization model*. As shown in Table 5, Cheap2Chic's performance improves steadily across all metrics as more synthetic data is added, eventually surpassing its default setting (30 min + 150 min). Notably, when using 220 minutes (or more) of synthetic data alongside 30 minutes of real data, Cheap2Chic even outperforms the *Full-RealData* setting, which uses 180 minutes of real recordings. When no real data were used for fine-tuning, performance degraded sharply, highlighting the pretrained model's limitation and the need for real data. However, adding excessive synthetic data increases training cost and reduces the relative contribution of real data, leading to diminishing reward. To verify the generality of this trend, we also evaluated another device pair (OE-L1→OE-H2) and observed a consistent pattern of gradual improvement followed by saturation. These results demonstrate that the data synthesis module produces high-quality samples and offers flexible scalability for model training, effectively serving as an ablation study that reveals how the balance between real and synthesized data influences overall performance.

## 6.6 Impact of Bluetooth Transmission

Some of the earphone models selected for data collection (Table 1) support both wired (via 3.5mm connection) and wireless (via Bluetooth) audio streaming. In the main evaluation, we used wired connections for all devices to ensure fair and consistent comparison.

[6] 30 and 180 minutes mark the lower and upper bounds of effective real-data usage, since performance drops sharply below 30 minutes and saturates beyond 180 minutes.

Table 5: **Impact of synthetic data volume.**

| Configuration (Real + Synthetic) | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|
| 30 min + 0 min | 18.22 | 3.49 | 5.37 | 0.64 |
| 30 min + 30 min | 19.66 | 3.78 | 4.48 | 0.39 |
| 30 min + 90 min | 21.98 | 3.92 | 3.57 | 0.23 |
| **30 min + 150 min (Cheap2Chic)** | **22.32** | **4.05** | **3.00** | **0.18** |
| 30 min + 210 min | 22.66 | 4.07 | 2.74 | 0.14 |
| 30 min + 330 min | 22.94 | 4.11 | 2.62 | 0.12 |
| 180 min + 0 min (Full-RealData) | 22.40 | 4.04 | 2.66 | 0.15 |

Table 6: **Impact of wired and wireless audio streaming.**

| Configuration | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|
| Low-end Playback (Wired) | −0.58 | 2.74 | 7.68 | 1.39 |
| Cheap2Chic (Bluetooth) | 19.41 | 3.77 | 1.95 | 0.17 |
| **Cheap2Chic (Wired)** | **20.93** | **4.08** | **1.78** | **0.08** |

However, since Bluetooth transmission is known to introduce issues such as packet loss [52], we conducted an additional experiment to assess whether Cheap2Chic's performance is affected under wireless conditions. Specifically, we evaluated the (OE-L1→OE-H1) pair, where OE-L1 (Mezone P2961) supports both playback modes. After developing the corresponding enhancement models, we played the same modified digital audio through OE-L1 using both the 3.5mm wired connection and Bluetooth, and recorded the outputs. These recordings were then compared to the high-end reference playback from OE-H1. To account for latency differences between wired and wireless channels, we aligned all recordings using cross-correlation.

As shown in Table 6, Cheap2Chic consistently improves perceived audio quality over the Low-end Playback baseline in both wired and Bluetooth playback modes. While Bluetooth playback exhibited a slight performance drop compared to the wired version, the degradation was minimal. This may be attributed to compression artifacts introduced during Bluetooth transmission, particularly affecting high-frequency content. However, since our enhancement primarily focuses on the mid- and low-frequency ranges, the overall impact on effectiveness remains marginal.

## 6.7 Subjective Evaluation

As the primary goal of Cheap2Chic is to enhance low-end earphone users' listening experiences, we conducted a subjective listening test to directly evaluate perceived audio quality from the listener's perspective. To this end, we designed a two-part evaluation procedure and recruited 10 participants (6 male and 4 female, with an average age of 28) who regularly listen to music, 4 of whom also had formal training in playing musical instruments. All evaluations were carried out in a quiet indoor environment using four representative device pairs: (IE-L1→IE-H1), (IE-L1→IE-H2), (OE-L1→OE-H1), and (OE-L1→OE-H2). The evaluation was approved by the Institutional Review Board (IRB) of the authors' institution and consists of two parts:

*Part 1 (MOS Rating).* For each device pair, participants listened to three 10-second music clips under three playback conditions: the original low-end playback, the version enhanced by Cheap2Chic, and the high-end reference. Each clip was long enough to allow reliable quality judgments while avoiding listener fatigue. The playback order was randomized, and participants could replay clips as needed. After listening, they rated the overall sound quality of each version immediately using a 5-point Mean Opinion Score (MOS) scale [29], where 5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad.

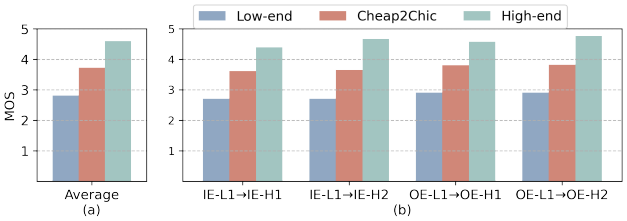



Fig. 12: **(a) MOS averaged across all device pairs and participants. (b) MOS for the individual device pair.**

*Part 2 (Post-experiment Interview).* After completing the listening tasks, participants were interviewed to share their subjective impressions of the enhanced audio and their perceived effectiveness of Cheap2Chic. This qualitative feedback aims to provide additional insights beyond the numerical ratings, helping us better understand how users perceived the improvements in listening scenarios.

Figure 12 presents the results of our subjective study, with (a) showing overall scores and (b) detailing results for individual device pairs. These results indicate that (1) Cheap2Chic consistently improves perceived audio quality over the original low-end playback, raising average MOS from below 3 (Fair) to nearly 4 (Good), demonstrating the effectiveness of Cheap2Chic; (2) although high-end devices still receive higher ratings compared to Cheap2Chic, this is expected for two main reasons: (i) it is impossible to fully compensate for the inherent hardware limitations of low-end devices using a software solution, and (ii) factors such as the shape, layout, and materials of the earphone design also influence perceived audio quality, areas where high-end devices naturally have an advantage; and (3) the improvement is consistent across different high-end targets, regardless of whether the devices are in-ear or over-ear. Moreover, the subjective scores may also reflect perceptual bias: participants were aware that they were wearing budget earphones during experiment, which could subconsciously lead them to assign slightly lower ratings even when the perceived quality was close to that of high-end playback. To let readers experience the perceptual improvements, we released an online demo page (https://contributor21.github.io/Cheap2Chic/) showcasing two device pairs across five music genres. All samples were recorded after playback through the respective earphones, where each column represents the low-end playback, the Cheap2Chic-enhanced version, and the high-end reference.

Seven participants explicitly described the improvement as *significant* compared to the original low-end version, indicating a strong preference for the enhanced playback. Across the interviews, several recurring keywords emerged, including *stronger bass*, *less harsh treble*, and *clearer instrumentals*, reflecting that participants perceived enhancements across diverse audio dimensions. Also, the majority of the participants found Cheap2Chic to be an interesting and novel solution for enhancing audio perception. In addition to general impressions, some participants offered more nuanced, genre-specific observations. For example, Participant 4,

who had received formal musical training, found the enhanced bass particularly satisfying in pop music but considered it slightly overpowering in classical tracks where subtler dynamics are preferred. Meanwhile, Participant 8, an audio equipment enthusiast, remarked that although high-end earphones typically offer better fidelity, the perceptual gap was less pronounced in country music, suggesting a reduced gap between Cheap2Chic and high-end devices.

### 6.8 System Performance

Since music playback on earphones is typically streamed from companion smartphones, we envision Cheap2Chic being deployed as a mobile application to deliver the enhancement service. The digital audio enhancement (i.e., model inference) can be performed either in the cloud or directly on the mobile device. For the latter, it is essential to assess whether Cheap2Chic can operate in real time while maintaining acceptable energy efficiency. To this end, we deployed our enhancement model on a Xiaomi 13 smartphone[7], featuring a Snapdragon 8 processor and a 4500 mAh battery.

To assess runtime efficiency, we continuously fed 5-second audio segments into the enhancement model and measured inference latency over 1,000 runs. The system achieved an average latency of 1.83 seconds per segment, demonstrating that Cheap2Chic supports real-time enhancements[8]. We further evaluated sustained resource usage by running the system continuously for one hour, during which a new segment was enhanced every five seconds to simulate realistic playback behavior. The measurement shows that battery consumption increased from 2% per hour (with audio playback only) to 8% per hour with enhancement enabled, corresponding to an average current draw of approximately 280 mA. The model maintained an average CPU utilization of 24% and a mean memory usage of 410 MB, reflecting a stable and moderate computational demand. These results suggest that Cheap2Chic can be feasibly deployed on smartphones and can operate continuously without imposing excessive battery drain or processor load.

## 7 Related Work

In the music domain, existing efforts to enhance the perceived listening experience can be broadly categorized based on whether the hardware characteristics of the playback device are taken into account. The first category, primarily EQ-based enhancement methods, leverages FRC to model device characteristics and applies filters to align the output with a target response, thereby improving playback quality. The second category, including approaches such as music coloring and stylization, focuses on achieving specific perceptual effects or styles (e.g., karaoke-style) by altering the digital audio directly, without considering the playback device's properties. In the following sections, we review each category in detail.

### 7.1 EQ-Based Enhancement Methods

Traditional EQ techniques can be categorized into three types. The first type is graphic equalizers [27, 43, 49], which divide the frequency spectrum into fixed bands and allow manual, user-controlled gain adjustments. This method offers an intuitive but relatively coarse level of control. The second group is parametric equalizers [19, 44, 45], which provide finer control over parameters such as center frequency, bandwidth, and gain for each band, enabling more precise tonal shaping. The third category consists of automated systems like AutoEQ [3], which automatically optimize a sequence of filters to match a playback device's FRC to a perceptual target, such as the Harman curve [39, 41]. Despite their differences, all these approaches rely on static FRCs measured under controlled conditions, which often fail to reflect the complex and variable playback behavior encountered in real-world playback scenarios. Moreover, the use of limited, overlapping filters constrains their ability to approximate detailed or content-dependent responses.

While still relying on FRCs and filter-based architectures, recent learning-based EQ frameworks aim to enhance response fitting and enable dynamic control. Martínez and Reiss [38] proposed a convolutional neural network (CNN) that selects among multiple filter types (e.g., shelving, peaking) to better match a target response. More recently, CONEqNet [28] introduced a content-aware system that adjusts filter parameters over time based on the evolving characteristics of the input audio. Additionally, adaptive EQ systems [21, 36] incorporate microphone feedback during playback to update filter gains and settings in real time, allowing the system to adapt to both the listener and the listening environment. While these approaches offer increased flexibility and automation, they remain within the traditional filter-based paradigm, focusing on aspects such as parameter selection [38], temporal adaptation [28], and improved target curve fitting [21, 36].

In contrast, rather than adjusting filters to fit a target response curve, Cheap2Chic uses neural networks to learn both device behavior and the corresponding waveform compensation, enabling end-to-end transformation grounded in real playback recordings rather than idealized targets.

### 7.2 Music Transformation

The music transformation literature can be further divided into two groups. The first group operates on symbolic-level representations, such as MIDI or note sequences. For example, MIDI-VAE [20] employs a variational autoencoder to perform conditional transformations over instrumentation and dynamics in multitrack sequences. Similarly, Nakamura et al. [40] propose an unsupervised melody conversion method by clustering pitch and rhythm patterns. These models primarily focus on the compositional structure of music and do not address audio signal fidelity or the realism of playback.

The second group performs timbre or style conversion directly at the audio level. Lu et al.[37] leverage learned timbre embeddings within a GAN-based framework to enable flexible one-to-many genre or instrument conversions. Groove2Groove[22] facilitates rhythmic style transfer for drum patterns using symbolic one-shot inputs. More recently, Li et al. [35] proposed a diffusion-based model that enables open-set music stylization from arbitrary reference recordings through mel-spectrogram inversion and style injection. While varying in scope and methodology, these works share a common goal: reshaping musical perception through neural models trained with unpaired or loosely supervised data.

While these works also employ deep models to modify digital audio, our approach differs in two key aspects. First, we explicitly model device-specific characteristics rather than assuming device-independent playback. Second, we optimize against real recorded

[7]Xiaomi 13 was originally released as a flagship model, and now retails for approximately 200 USD, making it a reasonable proxy for current mid-range devices. The cloud-based inference could be an alternative for resource-constrained smartphones.
[8]Here, *real-time* refers to buffered or locally stored music, where inference completes faster than the playback window. For latency-critical use cases such as live streaming, shorter windows and a redesigned processing pipeline would be required.

Table 7: **Comparison of enhancement performance under linear and logarithmic FRC scales for VLM reasoning.**

| FRC Scale | SI-SDR ↑ | PEAQ ↑ | LSD ↓ | FAD ↓ |
|---|---|---|---|---|
| **Linear** | **22.32** | **4.05** | **3.00** | **0.18** |
| Logarithm | 21.58 | 3.97 | 2.89 | 0.22 |

playback, ensuring that enhancements reflect actual listening outcomes rather than digital-domain alignment with abstract targets.

## 8 Discussion

**1. Extending to Speech-oriented Scenarios** Our current work focuses on music playback, as music is particularly sensitive to playback quality due to its rich spectral structure, wide dynamic range, and timbral complexity. However, earphones are also widely used for real-time communication scenarios such as phone calls and online learning, where playback quality can also affect the user experience. Extending our approach to speech-based applications is a promising direction for future work. However, unlike music, speech signals exhibit distinct properties: they are more temporally sparse, occupy narrower frequency bands, and are highly sensitive to latency. Consequently, optimization objectives in speech scenarios may shift toward intelligibility, low delay, and conversational clarity, rather than timbral richness or genre fidelity.

**2. Toward Zero-Shot Adaptation** While the proposed data synthesis module significantly reduces the user-side data collection burden by 82% through a pretrained generator, Cheap2Chic still requires about 30 minutes of real recordings to fine-tune the data generation model for each new device (Section 3.4.5), which may pose a practical burden. In practice, however, the same adapted model can be shared among all users of the same device model, avoiding repeated data collection. Moreover, with a broader and more diverse pretraining pool (e.g., around 100 devices), the generator may generalize better and even remove the need for this 30-minute data requirement, enabling zero-shot adaptation. In such cases, the model can synthesize plausible training data for unseen devices directly from their frequency response curves, without requiring any additional real recordings. This becomes especially feasible when the target device exhibits similar response characteristics to those observed during training, paving the way for future systems that can adapt to new devices with zero user burden.

**3. Impact of FRC Visualization Scale** When visualizing FRC line graphs, the frequency axis (X-axis) can be displayed using either a linear scale (e.g., [0, 1K, 2K, 3K, ...]) or a logarithmic scale (e.g., [0, 10, 100, 1K, ...]). These scaling choices introduce variations in frequency resolution across the spectrum: a logarithmic scale provides higher resolution at lower frequencies but reduced resolution at higher frequencies. Cheap2Chic uses a linear scale by default; however, we also conduct experiments using a logarithmic scale to examine its impact. As shown in Table 7, both linear and logarithmic scales achieve strong performance across all metrics, suggesting that the frequency scaling has minimal effect on the model's understanding. This is likely because both the curve and the corresponding axis labels are included in the image, enabling the VLM to correctly associate portions of the curve with specific frequency bands, thereby mitigating the effect of the scale used. Nevertheless, the logarithmic scale shows slightly lower performance, possibly due to its reduced resolution in the higher frequency range.

**4. Stereo Music and Spatial Effect** Most commercial music is produced in stereo and inherently encodes spatial information, including instrument placement, stereo width, and directional cues [30]. However, Cheap2Chic currently operates only on mono audio. A naive extension to stereo by enhancing the left and right channels independently may unintentionally amplify subtle inter-channel differences, disrupting the original stereo image and causing perceptual imbalance. As a result, users may lose important spatial cues during playback and experience reduced immersion or realism. To address this, future work could explore joint channel modeling to preserve inter-channel phase and amplitude relationships, incorporate the spatial coherence loss during training, or utilize Head-Related Transfer Function-aware [30] embeddings to maintain perceptual alignment. During evaluation, spatial quality should also be considered with binaural recordings or user studies.

**5. User-specific Audio Perception** Our prototype leverages a standardized ear simulator for playback recording, which ensures a consistent audio capture setup for fair performance comparison. However, individual differences in ear anatomy and auditory sensitivity can still slightly influence perceived audio quality. For example, variations in ear canal and pinna geometry may affect the frequency response and acoustic coupling [25]. Additionally, users may place different perceptual emphasis on certain frequency bands due to biological and experiential factors [47]. As a result, the generic enhancement model in Cheap2Chic may not equally satisfy all listeners, and it would be possible to further optimize individual performance by incorporating user-specific profiles, such as audiogram-based calibration, to achieve personalized enhancement.

**6. Genre-aware Optimizations** As shown in Figure 11, our enhancement performance exhibits slight variation across musical genres. This reflects genre-specific acoustic challenges, such as dense high-frequency content or rapid transients, that complicate accurate simulation and compensation. Additionally, these genre effects also interact with device-specific characteristics. For instance, we observe that while genres like Metal and Hip-Hop tend to yield lower scores overall, certain device pairs still perform well on them. This suggests that a uniform enhancement strategy may not generalize effectively across all musical types. Future systems could address this limitation by incorporating genre awareness, for instance, by deriving a genre embedding from the input audio and using it to condition the enhancement model at inference time, thereby enabling dynamic adaptation to genre-specific features.

## 9 Conclusion

In this work, we present Cheap2Chic, a two-stage audio enhancement framework that improves the perceived playback quality of low-end earphones. By modeling device-specific degradations with neural networks and learning a transformation to pre-compensate the digital audio, Cheap2Chic brings low-end playback experience closer to that of high-end models without requiring hardware modifications. Cheap2Chic also introduces a novel data generator that synthesizes training recordings, significantly reducing the data collection burden for users. Evaluations across diverse devices, genres, volumes, and transmission conditions confirm its effectiveness, even with limited real-world data for fine-tuning. Overall, Cheap2Chic provides a practical methodology toward democratizing high-quality audio experiences on affordable earphones.

## References


[1] Online. AirPods Pro 2 - Apple (SG). https://www.apple.com/sg/airpods-pro/.
[2] Online. Aurvana Air. https://www.amazon.com/Aurvana-Sweat-Resistant-Wired-Earphones-Sports/dp/B00CE2ELOS.
[3] Online. AutoEQ. https://autoeq.app/.
[4] Online. Bela Mini board. https://learn.bela.io/products/bela-boards/bela-mini/.
[5] Online. Beyerdynamic DT700 PRO X. https://www.lazada.sg/products/pdp-i2090958841-s11619391519.html.
[6] Online. CMS-28528N-L152 Datasheet - Miniature (10 mm 40 mm) | Speakers | Same Sky. https://www.sameskydevices.com/product/resource/cms-28528n-l152.pdf.
[7] Online. Earphones and Headphones Market Size, Report & Forecast Analysis 2030. https://www.mordorintelligence.com/industry-reports/earphones-and-headphones-market.
[8] Online. Fransun E20. https://shopee.sg/FRANSUN-FRANSUN-Original-Earphone-E20-Pink-Sauce-In-Ear-HiFi-Pink-Sleep-Earbuds-Game-with-Wheat-i.1561623237.42406301786.
[9] Online. Futuresource Consumer Headphones Market | Futuresource. https://www.futuresource-consulting.com/market-reports/futuresource-consumer-headphones-market.
[10] Online. GRAS 43AG Ear & Cheek Simulator. https://www.grasacoustics.com/products/ear-simulator-kit/constant-current-power-ccp-1/product/737-43ag.
[11] Online. HUAWEI FreeClip - HUAWEI Singapore. https://consumer.huawei.com/sg/headphones/freeclip/.
[12] Online. JBL C100SI. https://shopee.sg/JBL-In-Ear-Headphones-C100SI.-JBL-Signature-Sound-Superior-JBL-sound-with-bass-you-can-feel.-i.16732622.5956925644.
[13] Online. meta-llama/Llama-3.2-11B-Vision-Instruct · Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.
[14] Online. Mezone P2961. https://www.lazada.sg/products/mezone-head-mounted-wireless-headphone-p2961-bluetooth-50-foldable-headset-with-built-in-microphone-i3063925867.html.
[15] Online. Sennheiser IE 300. https://shopee.sg/Sennheiser-IE-300-High-Fidelity-In-Ear-Buds-(IEM)-i.56431667.10917339468.
[16] Online. Sony WH-1000XM5. https://shopee.sg/Sony-WH-1000XM5-WH1000XM5-1000XM5-Wireless-Noise-Cancelling-Headphones-i.439087721.21001553225.
[17] Online. Zihnic Foldable. https://www.lazada.sg/products/pdp-i3061645781-s20907853990.html.
[18] V Ralph Algazi and Richard O Duda. 2010. Headphone-based spatial sound. *IEEE Signal Processing Magazine* 28, 1 (2010), 33–42.
[19] Herwig Behrends, Adrian Von Dem Knesebeck, Werner Bradinal, Peter Neumann, and Udo Zölzer. 2011. Automatic equalization using parametric IIR filters. *Journal of the Audio Engineering Society* 59, 3 (2011), 102.
[20] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. 2018. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600* (2018).
[21] Adrian Celestinos, Elisabeth McMullin, Ritesh Banka, and Pascal Brunet. 2019. Personalized and self-adapting headphone equalization using near field response. In *Audio Engineering Society Convention 147*. Audio Engineering Society.
[22] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. 2020. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2638–2650.
[23] Andrzej Dobrucki. 2011. Nonlinear distortions in electroacoustic devices. *Archives of Acoustics* 36 (2011), 437–460.
[24] Augustine Gray and John Markel. 1976. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 5 (1976), 380–391.
[25] Dorte Hammersho/i and Henrik Mo/ller. 1996. Sound transmission to and within the human ear canal. *The Journal of the Acoustical Society of America* 100, 1 (1996), 408–427.
[26] Sarah Herrera, Adriana Bender Moreira de Lacerda, Debora Lurdes, Patricia Arruda Alcaras, Lucia Helena Ribeiro, et al. 2016. Amplified music with headphones and its implications on hearing health in teens. *The International tinnitus journal* 20, 1 (2016), 42–47.
[27] Martin Holters and Udo Zölzer. 2006. Graphic equalizer design using higher-order recursive filters. In *Proc. Int. Conf. Digital Audio Effects (DAFx-06)*. Citeseer, 37–40.
[28] Jesus Iriz, Miguel A Patricio, Antonio Berlanga, and José M Molina. 2023. CONEqNet: convolutional music equalizer network. *Multimedia Tools and Applications* 82, 3 (2023), 3911–3930.
[29] ITU-T. Online. Methods for subjective determination of transmission quality. https://www.itu.int/rec/T-REC-P.800.
[30] Gary S Kendall. 1995. A 3-D sound primer: directional hearing and stereo reproduction. *Computer music journal* 19, 4 (1995), 23–46.
[31] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466* (2018).
[32] Gibbeum Kim and Woojae Han. 2018. Sound pressure levels generated at risk volume steps of portable listening devices: types of smartphone and genres of music. *BMC Public Health* 18 (2018), 1–7.
[33] Paul W Klipsch. 1969. Modulation distortion in loudspeakers. *Journal of the Audio Engineering Society* 17, 2 (1969), 194–206.
[34] Dongheon Lee and Jung-Woo Choi. 2024. DeFTAN-II: Efficient multichannel speech enhancement with subgroup processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).
[35] Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2024. Music Style Transfer with Time-Varying Inversion of Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 547–555.
[36] Juho Liski, Vesa Välimäki, Sampo Vesa, and Riitta Väänänen. 2017. Real-time adaptive equalization for headphone listening. In *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 608–612.
[37] Chien-Yu Lu, Min-Xin Xue, Chia-Che Chang, Che-Rung Lee, and Li Su. 2019. Play as you like: Timbre-enhanced multi-modal music style transfer. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 33. 1061–1068.
[38] M Martinez Ramirez, Joshua Reiss, et al. 2018. End-to-end equalization with convolutional neural networks. (2018).
[39] Elisabeth McMullin. 2017. A study of listener bass and loudness preferences over loudspeakers and headphones. In *Audio Engineering Society Convention 143*. Audio Engineering Society.
[40] Eita Nakamura, Kentaro Shibata, Ryo Nishikimi, and Kazuyoshi Yoshii. 2019. Unsupervised melody style conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 196–200.
[41] Sean Olive, Todd Welti, and Omid Khonsaripour. 2017. A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 1–Listening Test Results and Acoustic Measurements. In *143rd Audio Eng. Convention, New York, USA (October 2017)*.
[42] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
[43] Jussi Rämö, Vesa Välimäki, and Balázs Bank. 2014. High-precision parallel graphic equalizer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 12 (2014), 1894–1904.
[44] Phillip Regalia and Sanjit Mitra. 1987. Tunable digital frequency response equalization filters. *IEEE transactions on acoustics, speech, and signal processing* 35, 1 (1987), 118–120.
[45] Joshua D Reiss. 2010. Design of audio parametric equalizer filters directly in the digital domain. *IEEE transactions on audio, speech, and language processing* 19, 6 (2010), 1843–1848.
[46] Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[47] A Scharine, K Cave, and T Letowski. 2009. Auditory perception and cognitive performance. *Helmet-mounted displays: Sensation, perception, and cognition issues* (2009), 391–489.
[48] Doron Shmilovitz. 2005. On the definition of total harmonic distortion and its effect on measurement interpretation. *IEEE Transactions on Power delivery* 20, 1 (2005), 526–528.
[49] Stéphan Tassart. 2013. Graphical equalization using interpolated filter banks. *Journal of the Audio Engineering Society* 61, 5 (2013), 263–279.
[50] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. 2000. PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society* 48, 1/2 (2000), 3–29.
[51] Floyd E Toole. 1985. Subjective measurements of loudspeaker sound quality and listener performance. *Journal of the Audio Engineering Society* 33, 1/2 (1985), 2–32.
[52] Jacopo Tosi, Fabrizio Taffoni, Marco Santacatterina, Roberto Sannino, and Domenico Formica. 2017. Performance evaluation of bluetooth low energy: A systematic review. *Sensors* 17, 12 (2017), 2898.
[53] George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing* 10, 5 (2002), 293–302.
[54] Vesa Välimäki and Joshua D Reiss. 2016. All about audio equalization: Solutions and frontiers. *Applied Sciences* 6, 5 (2016), 129.
[55] Satvik Venkatesh, Arthur Benilov, Philip Coleman, and Frederic Roskam. 2024. Real-time low-latency music source separation using Hybrid spectrogram-TasNet. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 611–615.