

HAAQI-Net: A Non-intrusive Neural Music Audio Quality Assessment Model for Hearing Aids

Dyah A. M. G. Wisnu, *Student Member, IEEE*, Stefano Rini, *Senior Member, IEEE*, Ryandhimas E. Zezario, *Member, IEEE*, Hsin-Min Wang, *Senior Member, IEEE*, and Yu Tsao, *Senior Member, IEEE*

Abstract—This paper introduces HAAQI-Net, a non-intrusive deep learning-based music audio quality assessment model tailored for hearing aid users. Unlike traditional methods such as the Hearing Aid Audio Quality Index (HAAQI), which relies on intrusive comparisons to a reference signal, HAAQI-Net provides a more accessible and computationally efficient alternative. By leveraging a bidirectional Long Short-Term Memory (BLSTM) architecture with attention mechanisms and incorporating features extracted using the pre-trained BEATs model, HAAQI-Net can predict HAAQI scores directly from assessed music audio clips and hearing loss patterns. Experimental results demonstrate the effectiveness of HAAQI-Net. Compared with the true HAAQI scores, the predicted scores have a Linear Correlation Coefficient (LCC) of 0.9368, a Spearman's Rank Correlation Coefficient (SRCC) of 0.9486, and a Mean Squared Error (MSE) of 0.0064, and the inference time is significantly reduced from 62.52 seconds to 2.54 seconds. However, although HAAQI-Net has high performance, feature extraction through the large BEATs model will cause computational overhead. To address this, a knowledge distillation strategy is applied. This involves constructing a student distillBEATs model that distills information from the teacher BEATs model during training of HAAQI-Net, thereby reducing the number of parameters required for feature extraction. The distilled HAAQI-Net model maintains strong performance with an LCC of 0.9071, an SRCC of 0.9307, and an MSE of 0.0091, while the number of parameters is reduced by 75.85%, and the inference time is reduced by 96.46%. The reduction in computational complexity and inference time improves the efficiency and scalability of HAAQI-Net, making it an efficient and scalable solution for real-world music audio quality assessment in hearing aid settings. It opens avenues for further research into optimization techniques for deep learning models tailored to specific application domains. Furthermore, this work contributes to the broader field of audio signal processing and quality assessment, providing insights into developing efficient and accurate models for practical applications in hearing aid technology.

Index Terms—Non-intrusive music audio quality assessment; HAAQI; Hearing aids; HAAQI-Net; BEATs, Knowledge Distillation

Dyah A. M. G. Wisnu is with the Taiwan International Graduate Program—Social Network and Human Centered Computing, Institute of Information Science, Academia Sinica, Taipei, Taiwan, and also with the College of Informatics, National Chengchi University, Taipei, Taiwan (e-mail: dyahayumgw@iis.sinica.edu.tw)

Stefano Rini is with Institute of Communications Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan (email: stefano@nctu.edu.tw)

Ryandhimas E. Zezario is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan (email: ryandhimas@iti.sinica.edu.tw)

Hsin-Min Wang is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan (email: whm@iis.sinica.edu.tw)

Yu Tsao are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan (e-mail: yu.tsao@iti.sinica.edu.tw)

I. INTRODUCTION

DESPITE advances in audio processing technology, music perception remains a complex challenge for individuals using hearing aids [1]. In audiology, hearing aids are tailored to the specific hearing loss characteristics of each user [2]. Given the subjective nature of music perception, objective measurement methods are urgently needed to ensure reliable assessment. Numerical methods such as mean squared error (MSE), signal-to-noise ratio (SNR), and signal-to-distortion ratio (SDR), while providing quantitative metrics, often fail to align closely with human auditory perception [3]. The Hearing Aid Audio Quality Index (HAAQI) [4] provides a music quality score that is more congruent with human perception. However, HAAQI has several limitations: (i) it requires the availability of ground truth signals, which makes it unusable in situations where such signals are unavailable; (ii) its computational complexity restricts its applicability, especially in real-time processing scenarios or low-resource environments; (iii) since it is non-differentiable, it cannot be directly incorporated into deep learning models for downstream applications as a loss function. These properties limit its applicability in practical scenarios, such as when an objective assessment must be made on the fly, or when it is used as a loss function in training deep learning models for music audio enhancement.

In this paper, we propose HAAQI-Net, a deep learning-based neural counterpart of HAAQI that overcomes the above limitations of HAAQI. HAAQI-Net is designed to be (i) non-intrusive, enabling assessment without the need for ground truth signals; (ii) efficient, allowing real-time processing and reducing computational overhead; and (iii) differentiable, facilitating its integration into deep learning frameworks for training and optimization. By addressing these challenges, HAAQI-Net provides a promising solution for objective music audio quality assessment for hearing aid users, offering improved accessibility, efficiency, and compatibility with modern computational frameworks [4]. This innovation has huge potential to enhance the overall listening experience for people with hearing loss, helping to improve quality of life and participation in social activities [2].

The proposed HAAQI-Net model involves a bidirectional Long Short-Term Memory (BLSTM) model followed by an attention mechanism. We use the pre-trained BEATs model [5] as the feature extractor. Although HAAQI-Net has high performance, feature extraction through the large BEATs model results in a certain computational overhead. Therefore, we reduce the number of parameters in the HAAQI-Net architecture

by incorporating knowledge distillation to transfer expertise from a large teacher model (HAAQI-Net with BEATs) to a compact student model (HAAQI-Net with distillBEATs). Furthermore, we also consider an adaptive distillation strategy to dynamically adjust the loss weight of each training sample based on its difficulty, enhancing the student model’s learning process. These extensions improve the efficiency and effectiveness of HAAQI-Net in real-world applications, particularly in scenarios requiring real-time processing or low-latency inference.

The effectiveness of HAAQI-Net is investigated through extensive numerical experimentation. When compared with the true HAAQI scores, the predicted scores have a Linear Correlation Coefficient (LCC) of 0.9368, a Spearman’s Rank Correlation Coefficient (SRCC) of 0.9486, and a Mean Squared Error (MSE) of 0.0064, and the inference time is significantly reduced from 62.52 seconds to 2.54 seconds. Furthermore, the distilled HAAQI-Net not only maintains high-quality predictions but also significantly reduces the runtime under different settings. This improvement makes HAAQI-Net more useful in practical applications of music audio quality assessment for hearing aid users, where efficient processing is crucial for providing timely and accurate feedback to users.

The remainder of this paper is organized as follows. Section II reviews related work. Section III illustrates the proposed methodology. Section IV outlines the experimental setup and reports the results. Finally, Section V provides conclusions and discusses future work.

II. LITERATURE REVIEW

This section presents a brief review of important literature on music audio quality assessment, particularly in the context of hearing aid applications.

Although there are many speech assessment metrics, such as Mean Opinion Score (MOS) [6], Perceptual Evaluation of Speech Quality (PESQ) [7], Perceptual objective listening quality analysis (POLQA) [8], speech transmission index (STI) [9], normalized-covariance measure (NCM) [10], short-time objective intelligibility (STOI) [11], extended STOI (eSTOI) [12], spectrogram orthogonal polynomial measure (SOPM) [13], neurogram orthogonal polynomial measure (NOPM) [14], and neurogram similarity index measure (NSIM) [15], there are relatively few dedicated measures for music audio quality assessment, especially for hearing aids. Music (and various other audio) quality assessment is broadly divided into intrusive and non-intrusive methods. Intrusive methods involve comparing a corrupted or processed signal to be evaluated with the original signal. Common intrusive methods include PEAQ [16], PEMO-Q [17], and HAAQI [4]. PEAQ and PEMO-Q do not take hearing loss into account, whereas HAAQI is designed to predict music quality for hearing aid users. While compensating for hearing loss, hearing aids pose distinct challenges, such as degradation of sound quality due to factors such as nonlinear processing and amplification. HAAQI utilizes an auditory model attuned to impaired hearing. It then assesses the quality by comparing the outputs of this auditory model for both the degraded signal and the reference

signal. By evaluating differences in signal characteristics, such as envelope modulation and temporal fine structure, HAAQI addresses challenges associated with background noise, non-linear processing, and varied listening environments of hearing aid users. Traditional non-intrusive assessment methods of audio quality include 3SQM and ITU-T Recommendation P.563 [18]. A learning-to-rank (LTR) method for music audio quality assessment is proposed in [19]. Recently, non-intrusive neural models for speech assessment have been proposed based on deep learning architectures, such as BLSTM [20], Convolutional Neural Network (CNN) [21], CNN-BLSTM [22], and Transformer [23]. Meanwhile, many studies have also focused on developing speech assessment models for hearing aid users [24]–[29]. However, to the best of our knowledge, there are no neural models specifically designed for non-intrusive music audio quality assessment for hearing aid users. This gap in the literature motivates the development of HAAQI-Net, a deep learning-based neural counterpart of HAAQI, which aims to address the need for efficient and accurate non-intrusive music audio quality assessment methods tailored for hearing aid users.

III. HAAQI-NET

This section first introduces the BEATs model for feature extraction, then the model architecture and training objective of the proposed HAAQI-Net model, and finally the knowledge distillation strategy to improve the efficiency of HAAQI-Net.

A. BEATs

BEATs [5] is a pre-trained framework that addresses key challenges in audio processing with its innovative acoustic tokenizer and audio Self-Supervised Learning (SSL) model. This framework iteratively enhances performance by using an acoustic tokenizer for generating discrete labels from unlabeled audio and optimizing the audio SSL model with masking and a discrete label prediction loss. Moreover, this audio SSL model achieves state-of-the-art performance on a variety of audio classification benchmarks, surpassing predecessors that utilize broader training data and a larger number of model parameters [5].

In this work, we utilize BEATs as the feature extractor. The process starts with pre-processing to extract Mel-Frequency Cepstral Coefficients (MFCCs). Then, it employs patch-based embedding through a convolutional layer, followed by normalization. Processed through a Transformer encoder, these embedded features are adept at capturing both local and global dependencies. The whole process is expressed as:

$$\mathbf{X}_{\text{BEATs}}^i = TE(LN(PE(Prep(\mathbf{X}))), \quad (1)$$

where TE denotes the Transformer encoder, LN is layer normalization; PE stands for patch embedding; $Prep$ represents the pre-processing operation; \mathbf{X} is the input waveform; and $\mathbf{X}_{\text{BEATs}}^i$ is the output features from the i -th layer of BEATs’ Transformer encoder.

To ensure we capture detailed information from all layers of BEATs’ Transformer encoder, we utilize the outputs of all layers instead of just the last one. This approach prevents

the loss of nuanced details present in the earlier layers. By extracting and incorporating the outputs of all layers, we can comprehensively grasp meaningful information across the entire sequence. We apply a weighted sum operation to these outputs, allowing us to blend them effectively and ensure that no valuable information is overlooked, which is expressed as:

$$\mathbf{X}_{w_sum} = \sum_{i=1}^L (LN(\mathbf{X}_{BEATs}^i) \times \sigma(w_i)), \quad (2)$$

where L is the number of layers in the Transformer encoder; σ denotes the softmax activation function; and w_i is the learnable weight associated with \mathbf{X}_{BEATs}^i . The element-wise multiplication of the normalized BEATs features and their respective weights ensures that more attention is given to certain parts of the input features during processing. Finally, the weighted features are summed together to obtain the final input feature vector, \mathbf{X}_{w_sum} .

B. Network Architecture

The overall architecture of HAAQI-Net is illustrated in Fig. 1. The input waveform is passed through the pre-trained BEATs model to obtain audio features. An adapter layer implemented as a dense layer is used to adapt these features specifically for music audio quality assessment. This adaptation enhances the model's ability to learn more compact and salient feature representations and capture intricate relationships in the data. The input to HAAQI-Net includes the adapted BEATs features and the hearing loss pattern. The core of HAAQI-Net consists of a BLSTM layer that captures the unique time-varying characteristics of music signals. The BLSTM layer is followed by a fully connected layer with 256 nodes activated by the Rectified Linear Unit (ReLU). A multi-head attention mechanism with 16 heads captures temporal dependencies in the data. A linear layer followed by a sigmoid activation produces frame-level scores. The frame-level scores are then aggregated by a global average pooling layer to form an overall clip quality assessment.

In the training phase, the model processes input tensors with dimensions $[B, T, F]$, where B denotes the batch size, T is the frame number per training music clip, and F is the dimension of the feature vector. The objective function for training HAAQI-Net is the sum of the clip-level loss and the averaged frame-level loss:

$$L_{Qual} = \frac{1}{B} \sum_{n=1}^B \left[(\hat{Q}_n - Q_n)^2 + \frac{1}{T_n} \sum_{t=1}^{T_n} (\hat{Q}_n - q_{n,t})^2 \right], \quad (3)$$

where \hat{Q}_n and Q_n represent the true and estimated clip-level quality scores for the n -th training music clip, respectively; T_n is the number of frames in the n -th training music clip, and $q_{n,t}$ denotes the estimated frame-level quality score of the t -th frame of the n -th training music clip. During training, the pre-trained BEATs model is fixed.

C. Knowledge Distillation

In our HAAQI-Net model, the pre-trained BEATs model plays a significant role in deploying informative acoustic

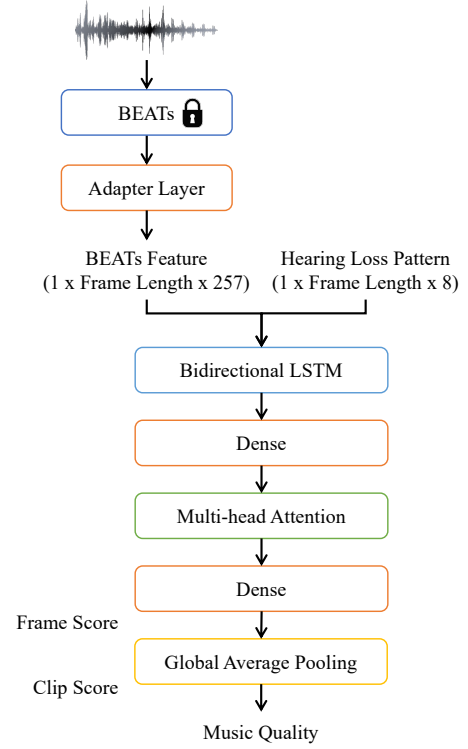


Fig. 1. The architecture of HAAQI-Net.

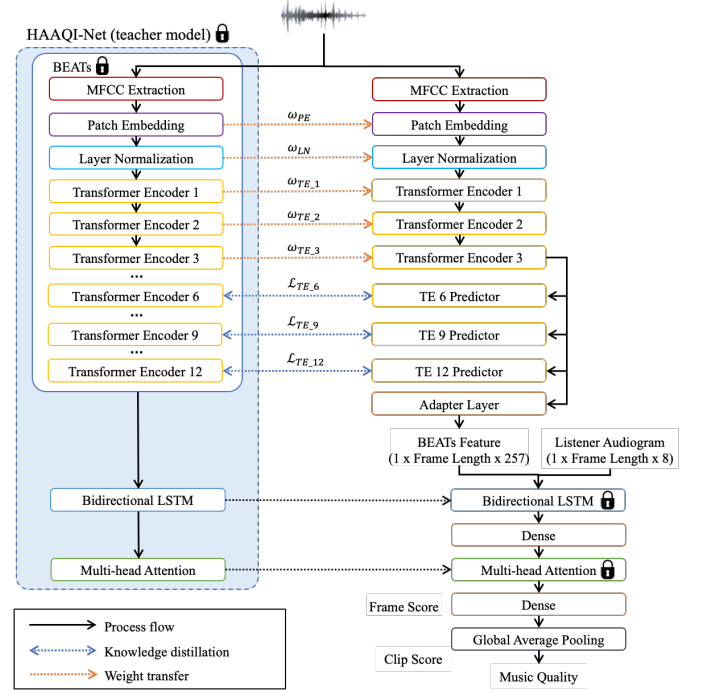


Fig. 2. The architecture of HAAQI-Net with knowledge distillation.

features. However, due to its model structure, BEATs require computationally intensive resources to perform feature extraction. This dependency introduces significant computational overhead, especially in scenarios that require real-time processing or low-latency inference. To this end, our focus shifted to developing a more compact and efficient version of HAAQI-

Net, but still ensuring satisfactory performance in music audio quality assessment for hearing aid users.

Knowledge distillation [30] emerges as a pivotal technique in our endeavor to transfer expertise from a large, cumbersome model (teacher) to a learner, more lightweight counterpart (student). Our goal is to distill the rich representations acquired by the BEATs model in the teacher network into the student model while ensuring that performance is not affected. To achieve this goal, as shown in Fig. 2, we perform a multi-stage distillation process beyond the final output layer to cover multiple intermediate layers of the teacher model. This strategic approach ensures that the student model fully grasps the teacher model’s ability to extract features and potentially maintains generalization.

Furthermore, we leverage the power of transfer learning [31] to initialize the weights of the student model using pre-trained parameters of the corresponding layers of the teacher model. This strategic integration allows the student model to exploit the knowledge encoded in the teacher model parameters, resulting in faster convergence of the learning process. In addition, as shown in Fig. 2, we retain fundamental modules such as pre-processing, patch embedding, layer normalization, and a subset of Transformer encoder layers while discarding redundant elements. In this way, we assume that effective knowledge distillation can be achieved, potentially optimizing performance in music audio quality assessment tasks for hearing aid users. In the following part, we will introduce our adaptive distillation strategy.

1) *Adaptive Distillation*: Adaptive distillation is beneficial in situations where the complexity or difficulty of samples in the training data varies significantly. In traditional distillation methods, the student model learns from a fixed teacher model, and all training samples are treated equally in terms of their contribution to the loss function. However, this may not be optimal when dealing with diverse datasets, where some samples are more difficult than others. Adaptive distillation solves this problem by dynamically adjusting the loss weight based on the estimated difficulty of each training sample.

In our model, the adaptive distillation loss function works by first calculating a difficulty measure for each training sample, typically based on a similarity measure between the student model’s prediction and the ground truth label, such as cosine similarity. Training samples with a higher similarity measure are considered easier, while those with a lower similarity measure are deemed more difficult. The loss for each training sample is then weighted according to its difficulty measure. More difficult training samples receive higher weights in the loss function, allowing the model to focus more on learning from challenging training samples. Finally, the weighted losses are aggregated to calculate the distillation loss for the batch:

$$L_{Distil} = \frac{1}{B} \sum_{n=1}^B \left(\frac{1}{3} \sum_{i=6,9,12} L_{TE_i,n} \right) \times d_n, \quad (4)$$

where d_n is the difficulty weight for training sample n , and $L_{TE_i,n}$ is the layer-wise distillation loss for the i -th layer of

training sample n calculated as

$$L_{TE_i,n} = L_{L1,n}^i + L_{cos,n}^i. \quad (5)$$

The layer-wise L1 loss $L_{L1,n}^i$ is calculated as

$$L_{L1,n}^i = \left| \mathbf{X}_{BEATs,n}^i - \mathbf{X}_{TE_i_Predictor,n} \right|, \quad (6)$$

where $\mathbf{X}_{BEATs,n}^i$ and $\mathbf{X}_{TE_i_Predictor,n}$ are the output features from the i -th layer of BEATs’ Transformer encoder and the TE i Predictor of the distilled model for training sample n , respectively. $L_{cos,n}^i$ is the layer-wise sigmoid cosine loss for the i -th layer of training sample n calculated as,

$$L_{cos,n}^i = -s_{n,i} \times \log(\hat{s}_{n,i}) - (1 - s_{n,i}) \times \log(1 - \hat{s}_{n,i}), \quad (7)$$

where $s_{n,i}$ is the cosine similarity between $\mathbf{X}_{BEATs,n}^i$ and $\mathbf{X}_{TE_i_Predictor,n}$, and $\hat{s}_{n,i}$ is its sigmoid transformation calculated as,

$$\hat{s}_{n,i} = \frac{1}{1 + e^{-s_{n,i}}}. \quad (8)$$

By combining L1 loss and cosine loss in Eq. (5), we aim to exploit their complementary properties to capture different aspects of the difference between predicted and true BEATs features. The overall loss for training HAAQI-Net with knowledge distillation is as follows:

$$L = L_{Qual} + L_{Distil}. \quad (9)$$

IV. EXPERIMENTS

This section details the experiments performed in this work, covering aspects such as data preparation, experimental setup, and obtained results.

A. Data preparation

Before delving into the evaluation of HAAQI-Net, we first explain how music samples are (i) selected, (ii) processed, and (iii) how hearing loss patterns are generated.

1) *Music Samples*: The music dataset is based on the small split of the FMA (FMA-small) dataset [32] and the MTG Jamendo dataset [33]. FMA-small is a balanced dataset for genre classification. From the eight genres available in FMA-small, we selected five genres: hip-hop, instrumental, international, pop, and rock. Considering that people with hearing loss are more likely to be older adults who regularly listen to classical music and orchestral music [34], we added additional music samples of these two genres from the MTG-Jamendo dataset. Through random selection, we collected 600 mono-audio samples in 7 genres, each lasting 30 seconds.

2) *Music Signal Processing*: The HAAQI-Net model is trained using data processed according to the method outlined in [35]. Their study aimed to explore how various signal processing techniques, typical in commercial hearing aids, affect judgments of music quality. Their experiments involved 100 distinct processing conditions in three groups:

- 32 conditions involving noise addition and nonlinear processing under different SNRs and parameter settings.

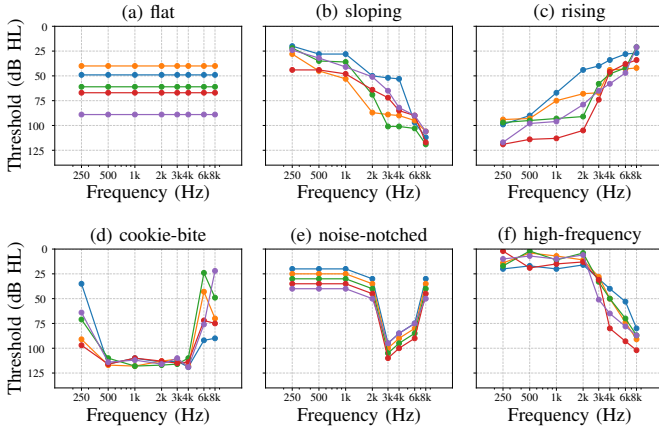


Fig. 3. Some examples of hearing loss audiograms: the y -axis represents the hearing threshold in dB, and the x -axis represents the frequency in Hz.

- 32 conditions employing linear filtering under different settings.
- 36 conditions combining noise addition, nonlinear processing, and linear filtering under different settings.

Noise conditions included music embedded in stationary speech-shaped noise and multi-talker babble. Nonlinear processing encompassed symmetric peak clipping, amplitude quantization, wide dynamic range compression (WDRC), spectral subtraction noise suppression, and various combinations. Linear processing involved low-pass and high-pass filters, spectral tilt adjustment, resonance peaks, and combinations of band-pass filters with resonance peaks. The combined processing conditions included all possible combinations of the above conditions.

3) *Hearing Loss Patterns*: When generating hearing loss patterns, we followed the approach in [36]. Hearing loss is described through an audiogram, which is a graphical display showing the degrees of hearing loss in different frequency regions, as shown in Fig. 3. A threshold above 20dB at any frequency is considered hearing loss. We represent a specific pattern of hearing loss using an 8-dimensional vector, where each dimension represents the threshold at that specific frequency. The eight frequencies used are 250, 500, 1K, 2K, 3K, 4K, 6K, and 8KHz. There are 6 types of audiograms: flat, sloping, rising, cookie-bite, noise-notched, and high-frequency, as illustrated in Fig. 3. In total, we established 300 hearing loss patterns, 50 patterns per hearing loss category. The patterns of each category were divided into two groups, 40 patterns for training and the remaining 10 patterns for testing.

Table I summarizes the data used to evaluate HAAQI-Net. Each music sample was paired with a randomly selected pattern of hearing loss. The test set was divided into seen and unseen subsets, where “seen” means the processing type was seen in training. Note that, as mentioned above, the hearing loss patterns in testing are always unseen in training. The distribution of HAAQI scores for all music samples across different data processing conditions and patterns of hearing loss is shown in Fig. 4.

TABLE I
STATISTICS OF THE DATA USED TO EVALUATE HAAQI-NET.

| Genre | Training Data | Testing Data | |
|---------------|---------------|--------------|--------------|
| | | Seen Noise | Unseen Noise |
| FMA-small | | | |
| Hip-hop | 3,048 | 552 | 116 |
| Instrumental | 3,890 | 538 | 104 |
| International | 3,368 | 484 | 104 |
| Pop | 3,632 | 496 | 110 |
| Rock | 4,644 | 396 | 74 |
| MTG-Jamendo | | | |
| Classical | 3,830 | 206 | 82 |
| Orchestral | 3,388 | 768 | 130 |
| Total | 25,800 | 3,440 | 720 |

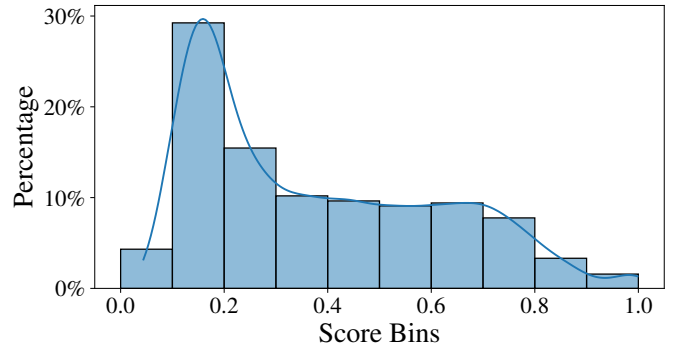


Fig. 4. Distribution of HAAQI scores for the music samples used to evaluate HAAQI-Net.

B. Experimental Setup

In this section, we introduce the experimental setup, including input configurations and various experimental scenarios.

1) *Inputs and Configurations*: In our experiments, we explored different settings to identify the most effective features for music quality prediction. These settings include spectrogram, speech SSL models, and BEATs features. For the spectrogram features, the power spectral features are extracted via a 512-point short-time Fourier transform (STFT) with a Hamming window size of 512 points and a hop size of 256 points, resulting in a 257-dimensional magnitude spectrum. The sampling rate for this input is 32 kHz. For speech SSL models, we selected three well-known models: Wav2Vec 2.0 Large [37], HuBERT Large [38], and WavLM Large [39]. Specifically, we used the last layer of these models as the acoustic features. As for the BEATs [5] features, we used four different configurations: “BEATs (Last)”, “BEATs (Last) + Win Avg”, “BEATs (Last) + Adapter”, and “BEATs (WS) + Adapter”. In detail, “BEATs (Last)” uses the last layer of the Transformer encoder as acoustic features with dimension 768. “BEATs (Last) + Win Avg” applies a moving average for every three elements so that the feature dimension is comparable to that of the spectrogram features. “BEATs (Last)

TABLE II
PERFORMANCE OF MUSIC QUALITY PREDICTION OF HAAQI-NET USING DIFFERENT INPUT FEATURES.

| Input Features | All | | | Seen | | | Unseen | | |
|--------------------------|----------------|-----------------|------------------|----------------|-----------------|------------------|----------------|-----------------|------------------|
| | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow |
| Spectrogram | 0.6701 | 0.7084 | 0.0301 | 0.6848 | 0.7106 | 0.0298 | 0.5999 | 0.6612 | 0.0314 |
| Wav2Vec 2.0 Large (Last) | 0.6483 | 0.6643 | 0.0302 | 0.6651 | 0.6809 | 0.0304 | 0.5030 | 0.5268 | 0.0295 |
| HuBERT Large (Last) | 0.6392 | 0.6645 | 0.0336 | 0.6396 | 0.6642 | 0.0343 | 0.5717 | 0.5478 | 0.0305 |
| WavLM Large (Last) | 0.7809 | 0.7834 | 0.0202 | 0.7806 | 0.7862 | 0.0214 | 0.7050 | 0.6757 | 0.0145 |
| BEATs (Last) | 0.9274 | 0.9342 | 0.0071 | 0.9234 | 0.9332 | 0.0081 | 0.9014 | 0.8762 | 0.0026 |
| BEATs (Last) + Win Avg | 0.8765 | 0.8908 | 0.0123 | 0.8724 | 0.8879 | 0.0135 | 0.8119 | 0.8275 | 0.0061 |
| BEATs (Last) + Adapter | 0.9368 | 0.9486 | 0.0064 | 0.9327 | 0.9455 | 0.0073 | 0.9282 | 0.9188 | 0.0024 |
| BEATs (WS) + Adapter | 0.9456 | 0.9603 | 0.0055 | 0.9410 | 0.9568 | 0.0063 | 0.9518 | 0.9417 | 0.0014 |

+ Adapter” employs an adapter (i.e., a dense layer) to reduce the dimensionality of the BEATs features (the last layer output) to 257. The adapter also plays a role in adapting the BEATs features for music audio quality assessment tasks. “BEATs (WS) + Adapter” employs an adapter to reduce the dimensionality of the BEATs features (the weighted sum of all layers) to 257.

Each type of acoustic input is concatenated with the hearing-loss pattern to form the final input to HAAQI-Net. The corresponding ground-truth quality score is calculated by the HAAQI method, ranging from 0 to 1, with 0 indicating poor quality and 1 representing perfect quality. The stimuli were amplified using the National Acoustics Laboratories revised (NAL-R) [40] linear fitting prescriptive formula based on individual hearing loss patterns. We trained HAAQI-Net using the Adam optimizer with a learning rate of 10^{-4} and an early stopping technique. To evaluate the performance, three criteria are used: LCC, SRCC, and MSE.

2) *Experimental Scenarios*: We explored two scenarios to evaluate the generalization ability of HAAQI-Net: the seen set and the unseen set. The seen set contains data with the same music processing conditions as the training set, while the unseen set contains data with different music processing conditions than the training set. We selected 82 conditions for the seen set, leaving 18 conditions exclusive for the unseen set. The unseen set comprises conditions such as “compression + babble”, “compression + spectral subtraction + babble”, “multiple resonance peaks + low pass filter”, “babble + compression + high pass filter”, “babble + compression + low pass filter”, “babble + compression + positive spectral tilt”, “babble + compression + negative spectral tilt”, “babble + compression + single resonance peak”, and “babble + compression + multi-resonance peak”. The distribution of training and testing data for each genre, as well as the seen and unseen sets, are summarized in Table I. 80% of the training data is used for training and 20% for validation.

C. Experimental Results

For a thorough grasp of the model’s performance, we analyze (i) overall performance, (ii) scenario-based performance,

and (iii) efficiency.

1) Overall Performance with Different Input Features:

Table II shows the music quality prediction performance of HAAQI-Net using different input features. We first focus on the overall performance (see “All” in the table) and make the following observations. First, the spectrogram provides a standard representation of audio signals but lacks context, resulting in moderate performance. Second, Wav2ec 2.0 and HuBERT, although effective in many speech processing tasks, show inferior performance, possibly due to their general-purpose nature. Third, WavLM, designed for automatic speech recognition, outperforms other speech SSL models by capturing higher-level semantic information. Fourth, BEATs trained with different audio data outperforms spectrogram and all three speech SSL models. Fifth, dimension reduction based on moving average will reduce the performance (“BEATs (Last) + Win Avg” vs “BEATs (Last)”), but the low-dimensional features are still more effective than the spectrogram (“BEATs (Last) + Win Avg” vs spectrogram). Sixth, a simple adapter can effectively adapt the BEATs features for music quality prediction (“BEATs (Last) + Adapter” vs “BEATs (Last)”). Seventh, “BEATs (WS) + Adapter” is the most effective among all input feature configurations compared here.

As shown in the scatter plots in Fig. 5, it is clear that among all the features compared here, the predictions of HAAQI-Net using the “BEATs (WS) + Adapter” features are most concentrated near the optimal diagonal. In addition, it can be seen that the performance of various BEATs features is better than that of spectrogram and three speech SSL model features.

2) *Scenario-Based Performance*: To fully evaluate the generalization ability of HAAQI-Net, we examine its performance on the seen and unseen test sets (see “Seen” and “Unseen” in Table II). From the table, we can see that under different input feature configurations, HAAQI-Net generally performs worse on the unseen test set than on the seen test set as expected. Various BEATs features outperform the spectrogram and three SSL model features on both seen and unseen test sets. The adapter for adapting the BEATs features for music quality prediction appears to be effective in reducing the performance gap between the seen and unseen test sets. Surprisingly, HAAQI-Net with “BEATs (WS) + Adapter” performs even

TABLE III
PERFORMANCE OF HAAQI-NET UNDER DIFFERENT TYPES OF HEARING LOSS.

| Hearing loss type | Spectrogram | | | BEATs (Last) | | | BEATs (Last) + Adapter | | | BEATs (WS) + Adapter | | |
|-------------------|----------------|-----------------|------------------|----------------|-----------------|------------------|------------------------|-----------------|------------------|----------------------|-----------------|------------------|
| | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow |
| Flat | 0.6129 | 0.6445 | 0.0407 | 0.9320 | 0.9387 | 0.0074 | 0.9503 | 0.9552 | 0.0054 | 0.9589 | 0.9630 | 0.0047 |
| Sloping | 0.6520 | 0.6919 | 0.0332 | 0.9263 | 0.9373 | 0.0072 | 0.9234 | 0.9424 | 0.0080 | 0.9411 | 0.9593 | 0.0060 |
| Rising | 0.7104 | 0.7432 | 0.0175 | 0.9174 | 0.9248 | 0.0056 | 0.9279 | 0.9501 | 0.0050 | 0.9497 | 0.9608 | 0.0035 |
| Cookie-bite | 0.7192 | 0.7401 | 0.0111 | 0.8101 | 0.8169 | 0.0058 | 0.8791 | 0.8547 | 0.0039 | 0.8757 | 0.9007 | 0.0040 |
| Noise-notched | 0.6439 | 0.6726 | 0.0278 | 0.9350 | 0.9479 | 0.0057 | 0.9326 | 0.9494 | 0.0064 | 0.9413 | 0.9568 | 0.0050 |
| High-frequency | 0.5864 | 0.5862 | 0.0400 | 0.9149 | 0.9263 | 0.0090 | 0.9401 | 0.9432 | 0.0068 | 0.9412 | 0.9584 | 0.0069 |

slightly better on the unseen test set than the seen test set in terms of MSE and LCC. Again, “BEATs (WS) + Adapter” is the most effective among all input feature configurations compared here.

3) Performance under Different Hearing Loss Patterns:

The performance of HAAQI-Net with different input features under different hearing loss patterns is shown in Table III. We can see that the “BEATs (Last)” features are more effective than the spectrogram across all hearing loss types. The adapter for adapting the “BEATs (Last)” features for music quality prediction appears to be effective at boosting the performance (“BEATs (Last) + Adapter” vs “BEATs (Last)”). The weighted-sum approach provides more informative features than using the last layer directly, as “BEATs (WS) + Adapter” outperforms “BEATs (Last) + Adapter” for most types of hearing loss. It is worth noting that HAAQI-NET performs slightly worse under the cookie-bite loss type than other loss types.

4) Performance under Different Signal Processing Conditions: Fig. 6 shows the performance of HAAQI-Net with the “BEATs (WS) + Adapter” features under different signal processing conditions.

The 8 processing conditions in Fig. 6(a) correspond to the noise addition and non-linear processing in [35]. For each condition, the result is an average of different settings, such as different SNRs and clipping thresholds. “LTASS” represents music with additive stationary speech-shaped noise, “Babble” denotes music with multi-talker babble, “Peak Clip” indicates music subjected to symmetric instantaneous peak clipping, “Quant” represents music quantized with reduced bit depth, “Comp” refers to music processed through multi-channel WDRC, “Comp+Babble” signifies music processed through WDRC after adding babble, “SSub+Babble” denotes music processed by spectral subtraction after adding babble, and “Comp+SS+Babble” refers to music processed by WDRC and spectral subtraction after adding babble. In Fig. 6(a), the “Comp” condition shows relatively high true HAAQI mean and standard deviation values, while its predicted HAAQI values are closely aligned. This suggests that the model performs consistently and accurately under the “Comp” condition. The results under the “Quant” condition are similar to those under the “Comp” condition. The results under the “LTASS” and “Peak Clip” conditions are close to each other, but the HAAQI scores are lower compared to

the “Comp” and “Quant” conditions. In contrast, conditions including “Babble”, “Comp+Babble”, “SSub+Babble” and “Comp+SS+Babble” show lower true and predicted HAAQI means, as well as lower standard deviations, indicating that the corresponding signal processing methods consistently have a large impact on music quality. Overall, the results highlight the varying HAAQI scores (music quality) under noise addition and non-linear processing, and the close alignment between predicted and true HAAQI values demonstrates the superior performance of HAAQI-Net.

The 8 processing conditions in Fig. 6(b) pertain to the linear filtering conditions in [35]. For each condition, the result is an average of different settings, such as different pass bands. “HP Filt” represents music processed by a high-pass filter, “LP Filt” denotes music processed by a low-pass filter, “BP Filt” represents music processed by a band-pass filter, “Pos Tilt” represents music processed through a filter with positive spectral tilt, “Neg Tilt” denotes music processed through a filter with negative spectral tilt, “Single Peak” indicates music processed through a filter with a single spectral peak, “Multi Peak” represents music processed through a filter with three spectral peaks, and “Multi Peak+LP Filt” represents music processed sequentially through a filter with three spectral peaks and a low-pass filter. From Fig. 6(b), we can also see some similar trends to Fig. 6(a). For example, different filtering conditions can result in significantly different HAAQI scores. Conditions with higher HAAQI means also show higher standard deviations. Except for the “LP Filt”, “HP Filt”, and “Single Peak” conditions, all other more complex filtering conditions result in lower HAAQI scores. Overall, except for the “HP Filt” condition, HAAQI-Net predictions are in good agreement with the true HAAQI scores.

The 36 processing conditions in Fig. 6(c-h) correspond to possible combinations of 6 “noise addition and nonlinear processing” methods and 6 “linear filtering” methods, as settings in [35]. Each plot shows the results of a subset that combines a specific “noise addition and nonlinear processing” method with the 6 “linear filtering” methods shown in the legend. Although some data points appear to be slightly off the diagonal, such as those corresponding to the combination of one of “noise addition and nonlinear processing” methods with “Pos Tilt” (e.g., the combination of “Quant” and “Pos Tilt” in Fig. 6(f)), the performance of HAAQI-Net is generally good.

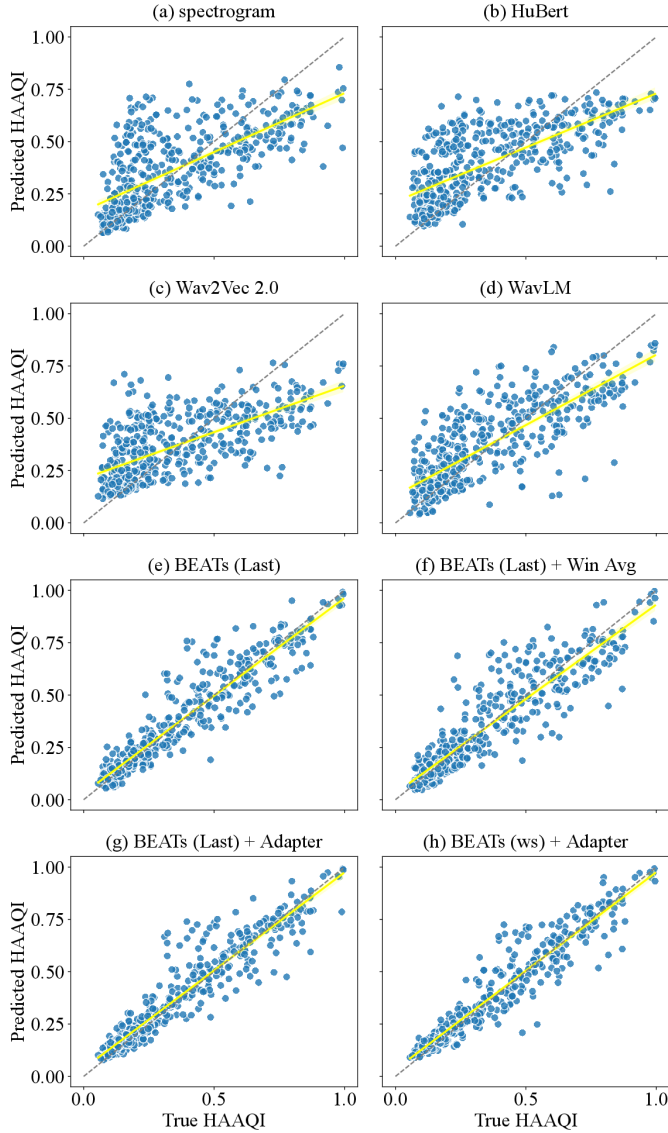


Fig. 5. Scatter plots of music quality prediction of HAAQI-Net using different input features. The dashed diagonal line represents the optimal prediction, while the yellow line represents the regression line for the model predictions. Data points below the dashed diagonal line indicate that the model’s predictions are lower than the true HAAQI scores, while data points above the line indicate that the model’s predictions are higher.

The noise and nonlinear conditions for the combined data mirror those of the noise and distortion data. In the “Comp+Babble” category, for instance, true and predicted HAAQI means are relatively close, with slight variations observed across different sub-conditions such as “Comp+HP Filtr”, “Comp+LP Filtr”, “Comp+Multi Peak”, “Comp+Neg Tilt”, “Comp+Pos Tilt”, and “Comp+Single Peak”. Notably, “Comp+Babble” sub-conditions generally exhibit similar patterns, with true and predicted means closely aligned. Similarly, in the “Babble” category, true and predicted HAAQI means are comparable across different sub-conditions such as “Babble+HP Filtr”, “Babble+LP Filtr”, “Babble+Multi Peak”, “Babble+Neg Tilt”, “Babble+Pos Tilt”, and “Babble+Single Peak”, suggesting consistent modeling across these scenarios. Moreover, in the “Peak Clip” and “Quant” categories, true

and predicted means exhibit close alignment, albeit with slight deviations observed in certain sub-conditions. However, in the “LTASS” category, while true and predicted means are relatively close, some sub-conditions such as “LTASS+Pos Tilt” show slightly higher deviations. Overall, the analysis suggests that the model effectively predicts HAAQI scores across various processing conditions, with a generally close alignment observed between true and predicted means. This consistency in performance underscores the model’s robustness and reliability across different scenarios.

5) *Performance across Different Genres*: Figure 7 shows the scatter plots of music quality prediction of our best-performing model (HAAQI-Net with the “BEATs (WS) + Adapter”) across various music genres. We can see that the model performs well in all music genres except pop music. In fact, for hip-hop, instrumental, international, pop, rock, classical, and orchestral music, the LCC values are 0.9678, 0.9407, 0.9321, 0.8960, 0.9603, 0.9593, and 0.9615, respectively. The model’s relatively poor performance in pop music may be due to its complex arrangements and electronic instrumentation. For international music, its diverse cultural styles may also pose slight challenges to the model. Overall, the experimental results show that the model has certain generalization capabilities across genres.

6) *Performance of HAAQI-Net with Knowledge Distillation*: We conduct an extensive analysis of the performance of HAAQI-Net under various distillation strategies. The results are shown in Table IV.

The original HAAQI-Net, whose predictions exhibit high correlation and low MSE value with true HAAQI scores in above experiments, serves as the benchmark. Two loss functions are studied: “L1 + Cosine” loss and “Adaptive” loss. The former combines “L1” loss and “Cosine” loss, aiming to exploit their complementary properties to capture different aspects of the difference between predicted and true BEATs features. It is calculated using Eq. (4) without the difficulty weights of the training samples. In contrast, the “Adaptive” loss is directly calculated by Eq. (4). It dynamically adapts to the difficulty of each training sample during training and has unique advantages in scenarios with varied sample complexity. In addition, we compare three architectures for distilling the transformer encoder. The first is single-layer distillation that uses only one prediction head (i.e., “TE 12 Predictor”) connected after “Transformer Encoder 3” to imitate the output of “Transformer Encoder 12” of the teacher model (see Fig. 2). The second is multi-layer distillation with independent prediction heads. As shown in Fig. 2, the three independent prediction heads (i.e., “TE 6 Predictor”, “TE 9 Predictor”, and “TE 12 Predictor”) connected after “Transformer Encoder 3” respectively simulate the outputs of “Transformer Encoder 6”, “Transformer Encoder 9”, and “Transformer Encoder 12” of the teacher model. The third is multi-layer distillation with sequential prediction heads, i.e., “TE 6 Predictor” imitates the output of “Transformer Encoder 6” based on “Transformer Encoder 3”, “TE 9 Predictor” imitates the output of “Transformer Encoder 9” based on “TE 6 Predictor”, and “TE 12 Predictor” imitates the output of “Transformer Encoder 12” based on “TE 9 Predictor”. We

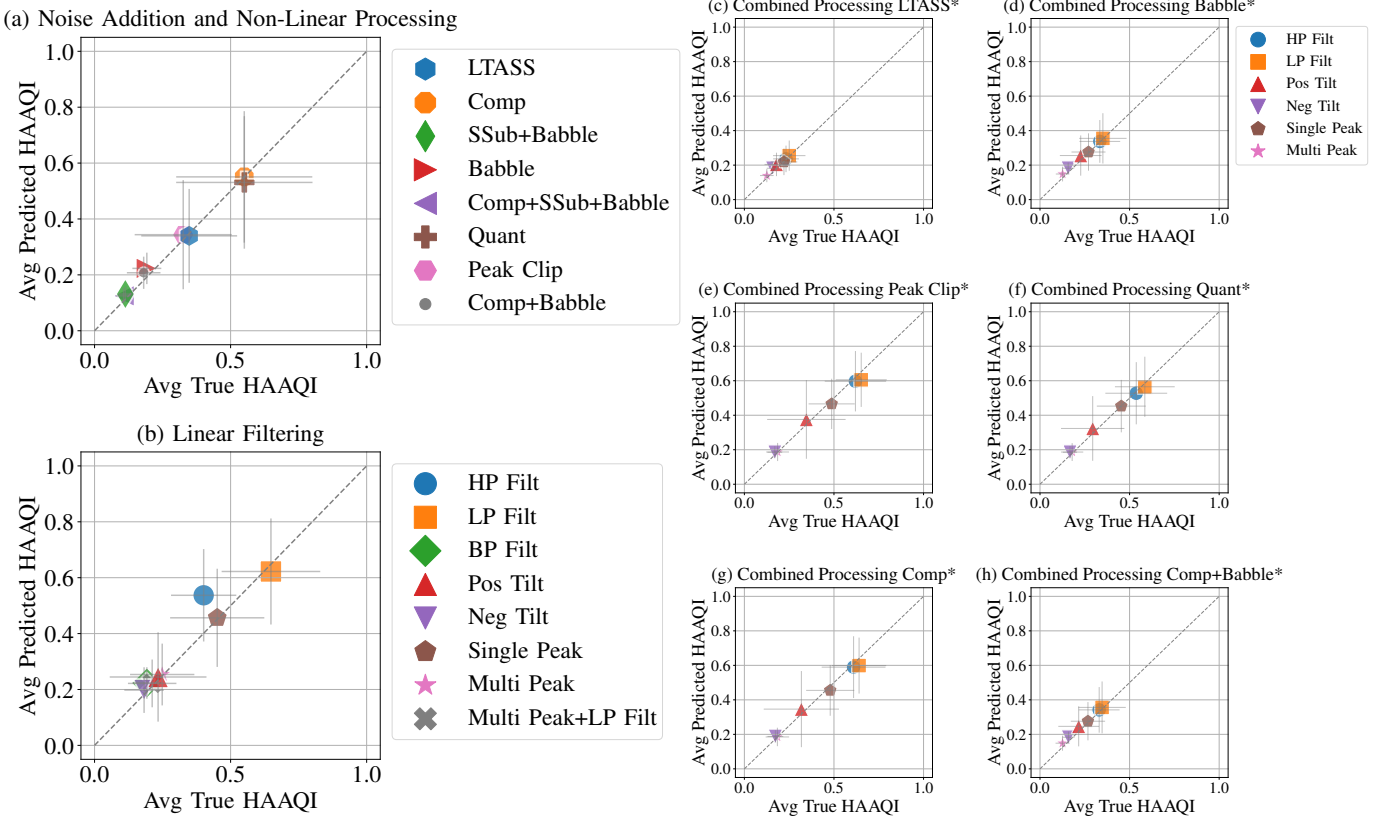


Fig. 6. Performance of HAAQI-Net with “BEATs (WS) + Adapter” under different signal processing conditions. Each data point in the plot represents a specific signal processing condition, averaged across all combinations of music samples and hearing loss patterns. The horizontal and vertical error bars represent the standard deviation of the true HAAQI scores and the predicted HAAQI scores, respectively. In (c), LTASS* represents the combination of LTASS with 6 of the linear filtering methods in (b), and similarly for (d)-(h).

also investigate whether the BLSTM and ATT layers should be fixed or fine-tuned during the distillation process, and whether we should use the weighted sum of the outputs of “TE 6 Predictor”, “TE 9 Predictor”, and “TE 12 Predictor” or the output of “TE 12 Predictor”.

Several observations can be drawn from Table IV. First, when both use the “L1 + Cosine” loss, multi-layer distillation outperforms single-layer distillation regardless of whether sequential prediction heads or independent prediction heads are used. Second, fine-tuning BLSTM and ATT layers during the distillation process does not lead to performance gains (see Multi-layer distillation (fine-tuning BLSTM & ATT) vs. Multi-layer distillation). Third, multi-layer distillation with independent prediction heads generally outperforms multi-layer distillation with sequential prediction heads. Fourth, the “Adaptive” loss is more effective than the “L1 + Cosine” loss (see Multi-layer distillation/Adaptive vs. Multi-layer distillation/L1+Cosine). Fifth, the weight-sum method is better than simply using the output of “TE 12 Predictor” (see Multi-layer distillation with ws/Adaptive vs. Multi-layer distillation/Adaptive). Overall, the performance of the distilled HAAQI-Net under the best configuration setting is very close to that of the original HAAQI-Net (see Multi-layer distillation with ws/Adaptive vs. Original HAAQI-Net), while the number of model parameters is reduced by 75.84%.

7) *HAAQI-Net Tested on the MUSDB18-HQ Dataset*: The MUSDB18-HQ dataset [41] is a widely used dataset in the field of music source separation and quality assessment. It is a high-quality version of the MUSDB18 dataset, which contains professionally produced music tracks with isolated musical sources (e.g., vocals, drums, bass, and other instruments). As shown in Table V, the MUSDB18-HQ dataset contains 50 music clips in 9 genres. It is worth noting that most of the music genres in the MUSDB18-HQ dataset are unseen in the training data of the HAAQI-Net models, which adds additional difficulty to the evaluation. In our experiment, the audio files are corrupted with randomly selected unseen noises. These noises include ambient sounds commonly encountered in urban environments, such as cafe chatter, street noise, bus rumble, and pedestrian footsteps. By incorporating these unseen noises, we aim to simulate real-world scenarios where music audio quality assessment models need to perform robustly in the presence of environmental disturbances.

Table VI shows the performance of three representative versions of HAAQI-Net tested on the MUSDB18-HQ dataset, including the original HAAQI-Net model (corresponding to Original HAAQI-Net in Table IV) and two reduced HAAQI-Net models, namely HAAQI-Net with distillBEATs[†] (corresponding to Multi-layer distillation/Adaptive in Table IV) and HAAQI-Net with distillBEATs (corresponding to Multi-layer

TABLE IV
PERFORMANCE OF HAAQI-NET WITH DIFFERENT DISTILLATION STRATEGIES.

| Distillation Methods | Distillation Loss | All | | | Seen | | | Unseen | | |
|--|-------------------|----------------|-----------------|------------------|----------------|-----------------|------------------|----------------|-----------------|------------------|
| | | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow | LCC \uparrow | SRCC \uparrow | MSE \downarrow |
| Original HAAQI-Net | - | 0.9368 | 0.9487 | 0.0064 | 0.9327 | 0.9455 | 0.0073 | 0.9283 | 0.9188 | 0.0024 |
| Single-layer distillation | L1 + Cosine | 0.8043 | 0.8393 | 0.1877 | 0.7907 | 0.8386 | 0.0217 | 0.7619 | 0.7697 | 0.0047 |
| Sequential Prediction Heads | | | | | | | | | | |
| Multi-layer distillation | L1 + Cosine | 0.8904 | 0.9087 | 0.0106 | 0.8865 | 0.9080 | 0.0117 | 0.8729 | 0.8514 | 0.0052 |
| Multi-layer distillation (fine-tuning BLSTM&ATT) | L1 + Cosine | 0.8750 | 0.9018 | 0.0135 | 0.8694 | 0.8938 | 0.0147 | 0.8724 | 0.8728 | 0.0081 |
| Multi-layer distillation | Adaptive | 0.8963 | 0.9155 | 0.0108 | 0.8923 | 0.9107 | 0.0118 | 0.8704 | 0.8802 | 0.0063 |
| Independent Prediction Heads | | | | | | | | | | |
| Multi-layer distillation | L1 + Cosine | 0.8988 | 0.9105 | 0.0098 | 0.8919 | 0.9089 | 0.0114 | 0.9054 | 0.8876 | 0.0024 |
| Multi-layer distillation (fine-tuning BLSTM&ATT) | L1 + Cosine | 0.8951 | 0.9217 | 0.0103 | 0.8898 | 0.9149 | 0.0115 | 0.8961 | 0.8927 | 0.0048 |
| Multi-layer distillation | Adaptive | 0.9071 | 0.9307 | 0.0091 | 0.8997 | 0.9250 | 0.0106 | 0.9116 | 0.8994 | 0.0019 |
| Multi-layer distillation with ws | Adaptive | 0.9151 | 0.9331 | 0.0083 | 0.9083 | 0.9281 | 0.0096 | 0.9127 | 0.8980 | 0.0022 |

TABLE V
STATISTICS OF THE MUSDB18-HQ DATASET.

| Genre | Count |
|---------------|-------|
| Electronic | 10 |
| Rock | 12 |
| Folk | 5 |
| Pop | 4 |
| Metal | 3 |
| Hip-Hop | 2 |
| Blues | 1 |
| International | 1 |
| Indie | 11 |
| Total | 50 |

distillation with ws/Adaptive in Table IV). Due to unseen genres and noises, we can see that the performance of all three HAAQI-Net models drops when tested on the MUSDB18-HQ dataset (e.g., the LCC of Original HAAQI-Net, HAAQI-Net with distillBEATs[†], and HAAQI-Net with distillBEATs drops from 0.9368, 0.9151, and 0.9071 to 0.7996, 0.6059, and 0.6432). The results indicate that to be more reliable, HAAQI-Net needs to be trained using training data covering more types of genres and noises. HAAQI-Net with distillBEATs outperforms HAAQI-Net with distillBEATs[†], this trend is consistent with the results in Table IV. Although the performance gap between HAAQI-Net with distillBEATs and Original HAAQI-Net becomes larger, considering the former’s reduced parameter size and important role as a distilled version of a large model, future research can focus on improving the distillation process.

8) *Efficiency Evaluation*: Fig. 8 provides a comprehensive comparison of the runtimes of different stages of HAAQI, HAAQI-Net, and HAAQI-Net with distillBEATs.

TABLE VI
PERFORMANCE OF HAAQI-NET TESTED ON THE MUSDB18-HQ DATASET.

| Models | LCC \uparrow | SRCC \uparrow | MSE \downarrow |
|--|----------------|-----------------|------------------|
| Original HAAQI-Net | 0.7996 | 0.7633 | 0.0343 |
| HAAQI-Net with distillBEATs [†] | 0.6059 | 0.5996 | 0.0482 |
| HAAQI-Net with distillBEATs | 0.6432 | 0.6561 | 0.0507 |

For the conventional HAAQI method, the pre-processing stage involves extensive computation, resulting in an average runtime of 27.29 seconds per audio. This time-consuming process mainly includes feature extraction and signal processing tasks necessary to derive relevant audio features. In comparison, HAAQI-Net significantly reduces pre-processing time, with an average feature extraction time of only 2.28 seconds per audio. This substantial improvement can be attributed to HAAQI-Net’s efficient neural network architecture, which uses BEATs for feature extraction, effectively reducing computational overhead. In addition, HAAQI-Net achieves remarkable speed enhancements in HAAQI score prediction, requiring only 0.26 seconds on average per audio, compared to 35.23 seconds required by the conventional HAAQI method. This efficiency results from HAAQI-Net’s ability to efficiently process extracted features and use optimized neural network layers and computational strategies to produce accurate predictions. Overall, HAAQI-Net exhibits outstanding efficiency, taking only 2.54 seconds per audio to complete the entire prediction process, while the conventional HAAQI method requires 62.52 seconds. This remarkable performance highlights the effectiveness of deep learning techniques in accelerating music quality prediction tasks while maintaining high accuracy.

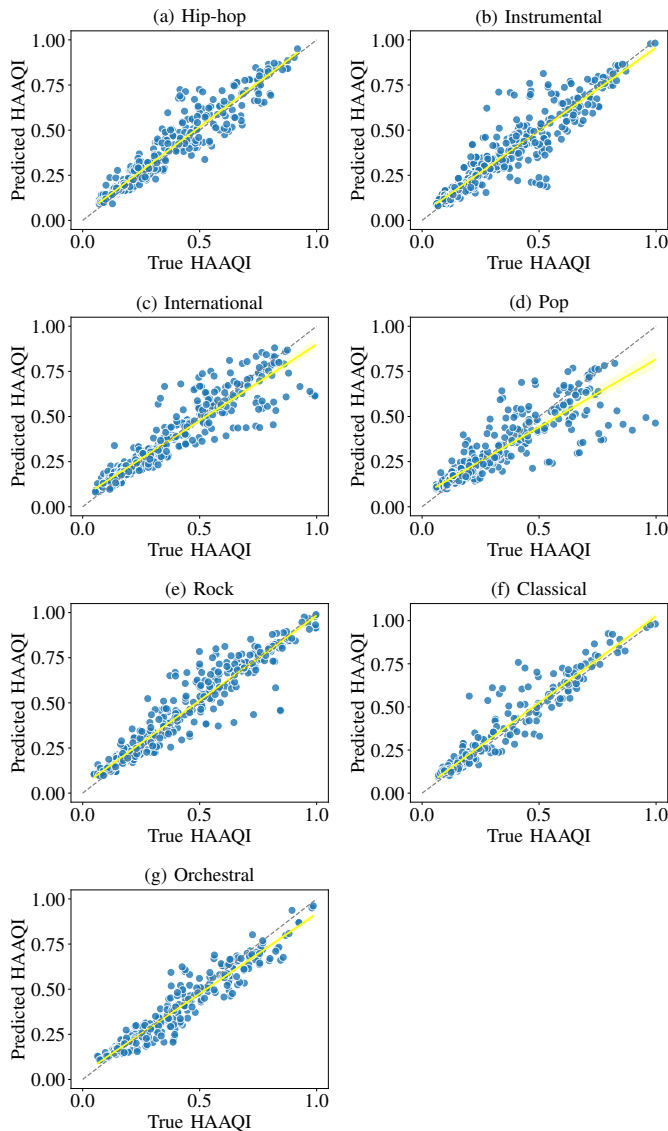


Fig. 7. Scatter plots of true HAAQI scores and HAAQI-Net’s predicted HAAQI scores for different music genres. The dashed diagonal line represents the optimal prediction, while the yellow line represents the regression line for the model predictions.

Furthermore, additional efficiency improvements can be achieved by integrating distillBEATs into HAAQI-Net. The runtime for pre-processing is further reduced to just 0.063 seconds per audio, while the calculation of HAAQI scores takes an astonishing average of 0.027 seconds. This significant improvement is attributed to the distillation process, which enhances model efficiency by transferring knowledge from a larger pre-trained model (BEATs) to a smaller target model (distillBEATs). As a result, HAAQI-Net with distillBEATs achieves unparalleled speed and scalability, making it well-suited for real-time music audio quality assessment applications where rapid processing is crucial.

These thorough evaluation results highlight the transformative impact of HAAQI-Net in revolutionizing music quality prediction, offering unprecedented efficiency and performance compared to traditional methods.

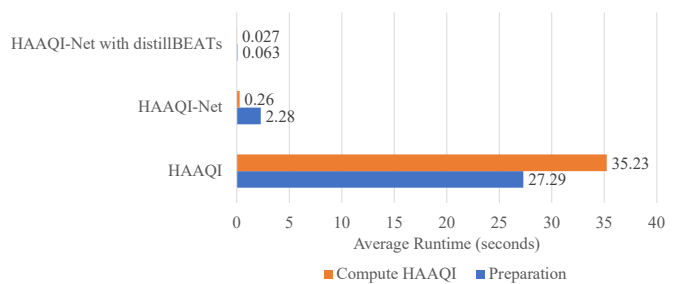


Fig. 8. Comparison of the runtimes of HAAQI, HAAQI-Net, and HAAQI-Net with distillBEATs&ws.

V. CONCLUSIONS

The development and evaluation of HAAQI-Net mark significant progress in music audio quality assessment for hearing aid users. Through systematic experimentation, our research demonstrates HAAQI-Net’s effectiveness, versatility, and efficiency across various scenarios and datasets. Our investigation into different input features reveals that pre-trained SSL models like WavLM and BEATs, particularly when paired with adapters, excel in capturing acoustic characteristics and semantic nuances, ensuring superior performance in music quality prediction.

Evaluations of HAAQI-Net’s generalization ability on seen and unseen datasets highlight its consistent performance, despite challenges such as overfitting with certain feature representations. The BEATs model consistently outperforms other models, effectively capturing and encoding general aspects of music quality. Additionally, HAAQI-Net’s adaptability across varied hearing-loss patterns and processing conditions underscores the importance of specialized feature representations in accurate music quality prediction, proving its effectiveness and reliability in different scenarios.

Moreover, HAAQI-Net achieves substantial efficiency improvements, significantly reducing processing time compared to traditional methods while maintaining high accuracy. The integration of distillation strategies further enhances this efficiency, positioning HAAQI-Net with distillBEATs as a promising solution for real-time music audio quality assessment. In summary, HAAQI-Net advances the assessment of music quality for hearing aid users by capturing musical nuances, generalizing across diverse scenarios, and delivering efficient performance, thereby enhancing the listening experience for individuals with hearing impairments. Future research will focus on optimizing HAAQI-Net’s performance and exploring its application in real-world settings.

REFERENCES

- [1] M. C. Leal, Y. J. Shin, M. Laborde, M. Calmels, S. Verges, S. Lugardon, S. Andrieu, O. Deguine, and B. Frayssé, “Music perception in adult cochlear implant recipients,” *Acta otolaryngologica*, vol. 123, no. 7, pp. 826–835, 2003.
- [2] B. Edwards, “The future of hearing aid technology,” *Trends in Amplification*, vol. 11, no. 1, pp. 31–45, 2007.
- [3] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

- [4] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 354–365, 2015.
- [5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATS: Audio Pre-Training with Acoustic Tokenizers," in *Proceedings of International Conference on Machine Learning (ICML)*, ser. ICML'23. JMLR.org, 2023.
- [6] P. ITU, "800: Methods for subjective determination of transmission quality," *Recommendation ITU-T*, 1996.
- [7] I.-T. Recommendation, "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.
- [8] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of The Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, June 2013.
- [9] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.
- [10] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *Journal of the Acoustical Society of America*, vol. 116, pp. 3679–3689, 2004.
- [11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [12] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [13] N. Mamun, M. S. A. Zilany, J. H. L. Hansen, and E. Davies-Venn, "An intrusive method for estimating speech intelligibility from noisy and distorted signals," *The Journal of the Acoustical Society of America*, vol. 150, no. 3, pp. 1762–1778, 2021.
- [14] N. Mamun, W. Jassim, and M. S. A. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 760–773, 2015.
- [15] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, p. 306–320, Feb 2012. [Online]. Available: <https://doi.org/10.1016/j.specom.2011.09.004>
- [16] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ—the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.
- [17] R. Huber and B. Kollmeier, "PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.
- [18] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 85–89.
- [19] Z. Li, J.-C. Wang, J. Cai, Z. Duan, H.-M. Wang, and Y. Wang, "Non-reference audio quality assessment for online live music recordings," in *Proceedings of ACM International Conference on Multimedia*, 2013, pp. 63–72.
- [20] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proceedings of Interspeech*, 2018, pp. 1873–1877.
- [21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proceedings of Interspeech*, 2021, pp. 2127–2131.
- [22] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [24] H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, and Y. Tsao, "HASA-Net: A non-intrusive hearing-aid speech assessment network," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 907–913.
- [25] Z. Tu, N. Ma, and J. Barker, "Exploiting hidden representations from a DNN-based speech recogniser for speech intelligibility prediction in hearing-impaired listeners," in *Proceedings of Interspeech*, 2022, pp. 3488–3492.
- [26] R. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids," in *Proceedings of Interspeech*, 2022, pp. 3944–3948.
- [27] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss," *Applied Acoustics*, vol. 214, p. 109663, 2023.
- [28] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.
- [29] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.
- [30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vision*, vol. 129, no. 6, p. 1789–1819, Jun 2021. [Online]. Available: <https://doi.org/10.1007/s11263-021-01453-z>
- [31] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.
- [32] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1768692>
- [33] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, Long Beach, CA, United States, 2019. [Online]. Available: <http://hdl.handle.net/10230/42015>
- [34] A. R. Bonneville, P. J. Rentfrow, M. K. Xu, and J. Potter, "Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood," *Journal of Personality and Social Psychology*, vol. 105, pp. 703–717, 2013.
- [35] K. H. Arehart, J. M. Kates, and M. C. Anderson, "Effects of noise, nonlinear processing, and linear filtering on perceived speech quality," *Ear and hearing*, vol. 31, no. 3, pp. 420–436, 2010.
- [36] W. B. Alshuaib, J. M. Al-Kandari, and S. M. Hasan, "Classification of hearing loss," *Update On Hearing Loss*, vol. 4, pp. 29–37, 2015.
- [37] A. Baevski, H. Zhou, A. Rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:219966759>
- [38] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Rahman Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:235421619>
- [39] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:239885872>
- [40] D. Byrne and H. Dillon, "The national acoustic laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257–265, 1986.
- [41] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," Dec. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3338373>