High Sample Rate Evaluation of Audio Playback Quality Using Deep Learning Audition Models

Baicheng Huang
Tsinghua University
Beijing, China
huangbc22@mails.tsinghua.edu.cn

Haiwei Chai
Conceptual Computing
Cambridge, Massachusetts
chaihw@conceputing.com

Feng Zhu
Conceptual Computing
Cambridge, Massachusetts
zhuf@conceputing.com

Dong Liu
Conceptual Computing
Cambridge, Massachusetts
liud@conceputing.com

Xiaoyong Pan
Conceptual Computing
Cambridge, Massachusetts
panxy@conceputing.com

***Abstract*—Computer audition models based on deep learning provide convenient alternative tools for evaluating audio playback quality beyond human listening test. Utilizing higher sampling rate compared to conventional digital audio, these models show promising results on various evaluation tasks.**

***Keywords*—deep learning, computer audition, audio playback quality, high sample rate**

I. INTRODUCTION

Deep learning computer audition models [1,2] provide a convenient method for audio playback quality evaluation. Before the emergence of these computer audition models, audio playback quality evaluation was administered as subjective human listening tests. Most evaluation tasks involve subtle discriminative tasks that can be hard to discern for human listeners. Thus these human listening experiments cannot provide stable or convincing statistical or descriptive metrics [3].

We constructed two deep learning-based computer audition systems utilizing two typical deep learning architectures. One architecture is based on computer vision systems [4,5]. We represent audio segments as images and use the computer vision architecture to classify auditory pictures. The second architecture is based on the large language models (LLMs) [6,7]. We embed audio representations as token sequences and process these token sequences as auditory language.

The proposed deep computer audition models are applied on several audio playback quality evaluation tasks. Specifically, we tested the effects of using high sample rate audio recordings (192.0 kHz/ 24 bit) and compared with the same processing architecture using conventional lower sample rate (44.1 kHz/ 24 bit). These comparative experiments demonstrated the advantages of using high sample rate audio signal as the input data for deep learning system.

II. EVALUATION SYSTEM DESIGN

The playback outputs of audio devices is connected to the audio interface for direct output signal evaluation. For acoustical evaluation tasks, the sound is collected by a microphone and connected to a microphone pre-amplifier before routed to the audio interface. Then the digital audio data is processed by deep learning models. The system configurations are illustrated in Figure 1(a) and (b) corresponding to the direct output signal scenario and the acoustical scenario.

III. AUDIO SIGNAL REPRESENTATIONS

The audio data captured from audio interface is represented as time-frequency representations. Time-frequency analysis tools combine the analyses in both time domain and frequency domain. A joint analysis across time and frequency enables the signal components to be separated in both time and frequency. This signal separation is especially essential for the analyses of non-stationary signals, whose signal structure is changing over time.

A. Short-Time Fourier Transform Magnitude

The short-time Fourier transform magnitude (STFTM) is usually the starting point for time-frequency analysis [8]. STFTM analysis first slices the signal $s(t)$ into short time frames (e.g., a slice of 1/20 the signal length). Then we compute the Fourier transform for each short time frame and calculate the magnitude. The STFTM is the concatenation of these Fourier transform magnitudes from consecutive time frames:

$$|S_w(u,w)| = \left|\int_{-\infty}^{\infty} s(t)\mathrm{win}(t-u)e^{-jwt}\,dt\right|, \qquad (1)$$

where $\mathrm{win}(t)$ is a window function that selects a small segment from signal $s(t)$. w is the radial frequency index $w = 2\pi f$ for f in Hz. $|\cdot|$ denotes the magnitude. u is the time location of the short-time window.

For a specific window location u_1, we calculate the frequency profile for $s(t)\mathrm{win}(t-u_1)$ (a slice of $s(t)$) as:

$$|S_w(u_1,\mathbf{w})| = \begin{bmatrix} |S_w(u_1,w_1)| \\ |S_w(u_1,w_2)| \\ \cdots \\ |S_w(u_1,w_N)| \end{bmatrix}, \qquad (2)$$

where $w_1, w_2, \cdots, w_N$ is a sequence of frequency points. Then we combines the frequency profiles from multiple time steps $u_1, u_2, \cdots, u_M$ to form the STFTM matrix:

$$\begin{aligned} &|S_w(\mathbf{u},\mathbf{w})| \\ &= [|S_w(u_1,\mathbf{w})|, |S_w(u_2,\mathbf{w})|, \cdots, |S_w(u_M,\mathbf{w})|], \end{aligned} \qquad (3)$$

where $|S_w(u_M,\mathbf{w})|$ denote a column vector by concatenating together the frequency profiles of multiple time slices, we can track the changes of frequency contents across different time frames.

B. Wigner-Ville Distribution

The Wigner-Ville Distribution (WVD) [9] of a signal $s(t)$ is denoted as:

$$W_s(t,w) = \int_{-\infty}^{\infty} s\left(t+\frac{\tau}{2}\right) s^*\left(t-\frac{\tau}{2}\right) \exp\{-iw\tau\}\, d\tau, \quad (4)$$

where w is the radial frequency as $w = 2\pi f$, f as physical frequency in Hz. $s^*(t)$ denotes the complex conjugate of $s(t)$. For real signal $s^*(t) = s(t)$. $\exp\{-iw\tau\}$ is the complex exponential.

The signal components inside the integral is composed of two parts:

$$S_a(t,\tau) = s\left(t+\frac{\tau}{2}\right), \quad (5)$$

$$S_b(t,\tau,w) = s^*\left(t-\frac{\tau}{2}\right) exp\{-iw\tau\}, \quad (6)$$

where $S_b(t,\tau,w)$ is a time shifted and frequency modulated ("pitch shifted") version of $S_a(t,\tau)$. The $W_s(t,w)$ is the correlation between $S_a(t,\tau)$ and $S_b(t,\tau,w)$. Recall that the Fourier transform calculates the correlation of $s(t)$ and $\exp\{iw\tau\}$ to depict the frequency component at w. $W_s(t,w)$ at a specific t location is the correlation of $s(t)$ abd $s^*(t-t_d)\exp\{iw\tau\}$ so it considers a $\exp\{iw\tau\}$ -type correlation information but also a correlation type of $s^*(t-t_d)$ where d is the delay. Thus $W_s(t,w)$ include more information about the self-similarity of $s(t)$ compared to Fourier analysis. This self-similarity item usually leads to high analytical resolution and feature information in many analytical applications.

C. Discrete Wavelet Transform

The wavelet transform of signal $s(t)$ is denoted as:

$$T_s(a,b) = w(a)\int_{-\infty}^{\infty} s(t)\psi^*\left(\frac{t-b}{a}\right) dt, \quad (7)$$

where $w(a)$ is a weighting function for scaling the energy according to scale a (can be interpreted as a frequency term). b is the time shift. $\psi(t)$ is the wavelet function. $*$ denotes complex conjugate.

The wavelet transform can be treated as a natural extension to Fourier Transform. The Fourier transform of a signal is the correlation between the signal and the complex exponential $\exp\{iw\tau\}$. The wavelet transform provides more choices of $\psi(t)$ beyond a complex exponential signal. Usually the choice of $\psi(t)$ is tailored to the waveform characteristics of the signal and the application scenarios. For example, when signal has many abrupt changes or the application is looking for rapid edges, a rectangular-shaped wavelet function would be ideal.

Discrete wavelet transform (DWT) selects an array of discrete (a,b) values for calculating the $T_s(a,b)$. As a (scale) stands for frequency and b (shift) stands for time location, $T_s(a,b)$ form a matrix or image of signal characteristics (e.g, energy) over time and frequency [10]. In practice, we choose an (a,b) array to form an orthonormal basis in the correlation items (Dyadic Grid Coefficients) so the sample points include the full information while the different (a,b) points are non-redundant.

IV. Deep Learning Models

A. Computer Vision Based Model

Our pattern recognition engine is based on a deep computer vision system, which utilizes Convolutional Neural Networks (CNNs) to model image patterns [11]. CNNs are the architecture of choice for modern computer vision tasks, as their pattern recognition performance has achieved levels comparable to, or higher than, human experts in many applications. The key features of the CNN architecture are its partially connected structure and its parameter-sharing mechanism. In a CNN, an output neuron connects only to a local rectangular region of the input layer. Moreover, for each output feature map, the same filter weights and bias term are applied across all spatial locations. This parameter-sharing mechanism significantly reduces the computational load during training and inference and enables effective learning from limited data.

B. Large Language Models

To convert continuous audio waveforms into discrete token sequences suitable for LLMs, we implemented a tokenization scheme [12] based on the Descript Audio Codec (DAC) [13], which served as the audio encoder. This DAC model leverages a hierarchical quantization architecture comprising 9 parallel codebooks, each containing 1024 unique tokens. Before encoding, the original audio files underwent a rigorous preprocessing sequence. Each audio channel was

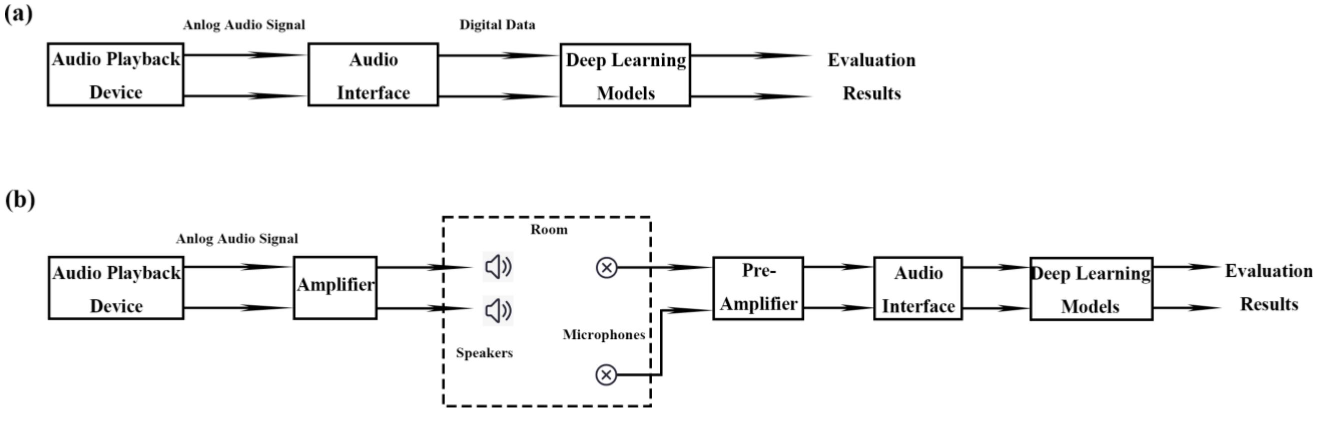

Fig. 1. Evaluation System Configuration. (a) The audio interface directly receives the analog audio signal transmitted from the audio playback device and inputs it into the deep learning model for evaluation. (b) Audio signals are played in the room by different audio amplifiers and other devices, collected by microphones, converted into digital signals, and then input into deep learning models for evaluation.

TABLE II. COMPARE DEVICE PLAYBACK SIGNALS (EVALUATION ACCURACY IN PERCENTAGE)

Playback files	Device 1/2					Device 2/3					Device 1/3				
	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test
	192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz	
Mp3 low quality	96.6	95.8	96.8	95.6	Middle	96.2	96.0	97.4	95.4	Difficult	96.0	96.4	97.4	95.6	Difficult
Mp3 high quality	98.0	97.0	98.8	98.0	Easy	98.0	96.4	99.0	97.8	Easy	98.0	97.0	99.0	97.4	Middle
44.1 kHz/ 24 bit uncompression	98.2	98.2	99.0	98.4	Easy	99.2	98.4	98.4	96.6	Easy	98.2	97.8	98.6	98.2	Easy
96.0 kHz/ 24 bit uncompression	98.0	96.0	98.0	96.4	Middle	98.0	96.6	97.4	96.0	Middle	97.6	95.8	97.4	96.6	Middle

TABLE I. COMPARE SOUND FROM DIFFERENT DEVICES (EVALUATION ACCURACY IN PERCENTAGE)

Playback files	Device 1/2					Device 2/3					Device 1/3				
	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test
	192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz	
Mp3 low quality	91.0	86.0	94.0	91.2	Difficult	90.4	85.2	94.6	88.2	Difficult	92.6	83.6	93.6	90.0	Difficult
Mp3 high quality	95.8	91.2	97.0	95.4	Easy	95.8	90.2	97.8	92.2	Easy	94.8	92.2	98.8	93.2	Easy
44.1 kHz/ 24 bit uncompression	98.4	91.8	99.8	94.8	Easy	97.4	92.8	98.6	96.4	Middle	95.4	93.6	97.8	94.0	Easy
96.0 kHz/ 24 bit uncompression	93.6	89.0	96.0	92.8	Middle	94.4	88.8	94.4	93.2	Middle	93.2	88.4	97.0	90.8	Easy

extracted independently. All waveforms were uniformly resampled to the model's required and normalized. A critical step involved precisely padding the signal to ensure its total length was an integer multiple of the model stride (512 samples), guaranteeing frame alignment for subsequent processing.

To accommodate this multi-codebook architecture, we constructed a custom "word-level" tokenizer. We mapped the vocabularies of the 9 codebooks (totaling 9,216 base tokens) into a flat, non-overlapping integer space. Specifically, we used the formula $q_{m,t} = 1024 \cdot m + p$ to assign a unique ID $q_{m,t}$ to the p-th token in the m-th codebook, where t is the time index of the sample sequence. The preprocessed waveforms were fed into the DAC model's encoder in chunks. For each audio frame, the encoder output 9 parallel discrete codes—one token from each of the 9 codebooks.Finally, to generate the one-dimensional sequence stream required by the LLMs, we performed an interleaving process on these parallel codes: the 9 tokens from a single timestep were flattened sequentially according to their codebook order (e.g., $[q_{1,t=1}, q_{2,t=1}, \cdots, q_{9,t=1}, q_{1,t=2}, \ldots]$). This resulting one-dimensional token stream constituted the final data format for input into LLMs for training.

V. APPLICATION EXAMPLES

To validate the effectiveness of the proposed deep learning audition models, particularly the advantages of using a high sample rate, we designed three distinct evaluation experiments. These experiments assess the models' ability to discriminate between different audio playback devices and acoustical environments.

A. Comparative Device Playback Evaluation

This experiment assesses the system's capacity to differentiate between three media players playing identical audio source files. The evaluation employed the direct signal assessment configuration illustrated in Figure 1(a). In this setup, the analog audio output from each playback device was connected directly to the audio interface, where it was digitized and subsequently fed into the deep learning models for classification.

The results presented in Table 1 demonstrate that models using 192.0 kHz sample rate consistently achieved higher classification accuracy than those using the 44.1 kHz sample rate, with both the Computer Vision and LLMs architectures.

TABLE III. COMPARE ACOUSTICAL ENVIRONMENTS (EVALUATION ACCURACY IN PERCENTAGE)

Playback files	Environment 1/2					Environment 2/3					Environment 1/3				
	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test	Computer Vision Model		Large Language Model		Human listening test
	192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz		192.0 kHz	44.1 kHz	192.0 kHz	44.1 kHz	
Mp3 low quality	93.0	84.0	94.6	90.6	Middle	92.8	98.4	94.0	98.8	Difficult	92.6	84.4	94.6	90.0	Difficult
Mp3 high quality	96.4	91.8	97.8	94.4	Easy	95.6	92.0	97.4	94.8	Easy	95.8	92.2	97.6	94.4	Middle
44.1 kHz/ 24 bit uncompression	97.0	93.8	98.4	96.4	Easy	97.0	93.2	98.8	96.4	Middle	96.6	93.6	98.0	96.6	Easy
96.0 kHz/ 24 bit uncompression	94.0	88.0	96.4	92.8	Middle	94.2	88.6	96.0	92.8	Middle	94.6	88.2	95.6	92.6	Middle

B. *Compare Acoustical Signals from Different Devices*

This experiment compares the acoustical characteristics of three different audio amplifiers when the same signal source and loudspeakers. The acoustical evaluation configuration shown in Figure 1(b) was utilized. The source signal was fed through each amplifier. The sound from the speakers is captured by a microphone array before being input to the models for evaluation.

In an acoustical evaluation that introduces variables such as room reflections and microphone characteristics, the richer data provided by a high sample rate enables the model to more precisely identify the sonic signatures of different amplifiers (Table 2).

C. *Compare Acoustical Environments*

This experiment tests the system's sensitivity to subtle changes in the acoustical environment, specifically by differentiating between the effects of three different sound absorption panels installed in the listening room. The audio source, amplifier, and speaker system were held constant for this test. The experimental setup again followed the acoustical evaluation protocol shown in Figure 1(b).

These results indicate that the system is highly sensitive to small variations in room acoustics, such as changes in reverberation time and frequency response caused by different absorptive materials. The harmonic overtones and transient details preserved at the 192.0 kHz sample rate are critical for enabling the model to classify these environmental differences (Table 3).

VI. CONCLUSIONS

Our proposed computer audition systems based on deep learning architectures show convincing performances for various audio device playback quality evaluation tasks. The evaluation results concur with human rating results. The computer audition systems using higher audio sample rate shows better classification metrics for perceptual tasks of discriminating different parallel test configurations comparing to computer audition systems based on standard 44.1 kHz/ 24 bit signal interfaces.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath, "Deep learning for audio signal processing," IEEE J. Sel. Top. Signal Process., vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[2] G. Mittag and A. Möller, "Deep learning-based non-intrusive speech quality assessment," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Brighton, UK, 2019, pp. 6905-6909, doi: 10.1109/ICASSP.2019.8683515.

[3] F. E. Toole, Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms, 3rd ed. New York, NY, USA: Focal Press, 2017.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in Proc. 25th Int. Workshop Neural Netw. Signal Process. (NNSP), Boston, MA, USA, 2015, pp. 1-6, doi: 10.1109/NNSP.2015.7300902.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780, doi: 10.1109/ICASSP.2017.7952261.

[6] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in Proc. Interspeech, Brno, Czechia, 2021, pp. 571-575, doi: 10.21437/Interspeech.2021-1006.

[7] Z. Borsos et al., "AudioLM: a language modeling approach to audio generation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 2505-2516, 2023, doi: 10.1109/TASLP.2023.3276013.

[8] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE, vol. 65, no. 11, pp. 1558–1564, Nov. 1977, doi: 10.1109/PROC.1977.10770.

[9] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "The Wigner distribution—a tool for time-frequency signal analysis, Part I: Continuous-time signals," Philips J. Res., vol. 35, no. 3, pp. 217–250, 1980.

[10] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pp. 674–693, Jul. 1989, doi: 10.1109/34.192463.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[12] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL), Berlin, Germany, 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.

[13] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," 2023, arXiv:2306.06546. [Online]. Available: https://arxiv.org/abs/2306.06546