

Supplementary Materials for TAS: Personalized Text-guided Audio Spatialization

Anonymous ACMMM Submission

1 INTRODUCTION

In this supplementary material, we provide details of the evaluation metrics in the quantitative results and the sound pressure levels in the personalized experiments. Firstly, we introduce the equations of evaluation metrics in Section 2. Then, the detailed expression process of sound pressure level in personalized experiments is provided in Section 3. **Finally, the binaural audio generated by our text-guided audio spatialization method can be viewed in the video in the supplementary materials.**

2 DETAILS OF EVALUATION METRICS

In this section, we introduce the evaluation metrics of this paper in detail, which provide a comprehensive evaluation of binaural audio. In the main text, we utilize six evaluation metrics to comprehensively evaluate binaural audio generated by different methods, including STFT Distance [1], Envelope (ENV) Distance [4], Wave L2 (WAV $\times 10^{-3}$) [6], Amplitude L2 (AMP), Phase L2 (PHA), and Signal-to-Noise Ratio (SNR) [5]. They are described as follows:

STFT Distance: It measures binaural audio on the spectrogram domain, which is the Euclidean distance between the predicted left and right channel spectrograms and their ground truth:

$$\mathcal{D}_S = \|S_l - S'_l\|_2 + \|S_r - S'_r\|_2. \quad (1)$$

ENV Distance: It measures binaural audio on the raw waveform domain, which is the Euclidean distance between the envelope of the left and right channels of the waveform and the ground truth:

$$\mathcal{D}_E = \|E[A_l] - E[A'_l]\|_2 + \|E[A_r] - E[A'_r]\|_2, \quad (2)$$

where $E[\cdot]$ denote the envelope of signal. It can capture the perceptual similarity of the waveform well.

Wave L2: It is the mean squared error between the predicted binaural audio and the ground truth binaural recording:

$$\mathcal{L}_2^{wav} = 10^3 \times ((A_l - A'_l)^2 + (A_r - A'_r)^2). \quad (3)$$

Amplitude L2 and Phase L2: Amplitude L2 and Phase L2 are the mean squared errors between the predicted binaural audio and the real binaural recording on the amplitude and phase after STFT [2] on the waveform:

$$\mathcal{L}_2^{amp} = (|S_l| - |S'_l|)^2 + (|S_r| - |S'_r|)^2, \quad (4)$$

and

$$\mathcal{L}_2^{pha} = (\angle(S_l) - \angle(S'_l))^2 + (\angle(S_r) - \angle(S'_r))^2, \quad (5)$$

where $|\cdot|$ and $\angle(\cdot)$ denote the modulu and phase angle of the complex number, respectively.

Signal-to-Noise Ratio: It is the power ratio of audio signal to noise. Audio signal refers to the ground truth binaural recording, while noise refers to the differential between the ground-truth and the predicted signal. The average signal-to-noise ratio of the two

channels of binaural audio can be described as:

$$\text{SNR} = \frac{10 \times (\log 10(\frac{A_l}{A_l - A'_l}) + \log 10(\frac{A_r}{A_r - A'_r}))}{2}. \quad (6)$$

3 DETAILS OF SOUND PRESSURE LEVEL

In the personalization experiment, we employ SPL to visualize the spatial variation of the binaural audio, which represents the difference in sound pressure level between the left and right channels of the binaural audio [3]. The sound pressure levels of the left and right channels of binaural audio are defined as follows:

$$\text{SPL}_l(t) = 20 \times \log_{10}(\frac{\|A_l(t)\|_2}{p_{ref}}), \quad (7)$$

and

$$\text{SPL}_r(t) = 20 \times \log_{10}(\frac{\|A_r(t)\|_2}{p_{ref}}), \quad (8)$$

where $A_l(t)$ and $A_r(t)$ represent the left channel signal and the right channel signal, respectively, $p_{ref} = 2 \times 10^{-5}$. Then, the spatial perception of binaural audio is defined as the difference in sound pressure level between the left and right channels:

$$\text{SPL}(t) = \text{SPL}_l(t) - \text{SPL}_r(t). \quad (9)$$

Finally, the direction of spatial perception of binaural audio can be expressed as:

$$\text{Direction}(t) = \begin{cases} \text{Left,} & \text{SPL}(t) > 0 \\ \text{Center,} & \text{SPL}(t) = 0 \\ \text{Right,} & \text{SPL}(t) < 0 \end{cases} \quad (10)$$

REFERENCES

- [1] Ruohan Gao and Kristen Grauman. 2019. 2.5D Visual Sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 324–333.
- [2] Daniel Griffin and Jae Lim. 1984. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243.
- [3] Zhaojian Li, Bin Zhao, and Yuan Yuan. 2023. Cross-modal Generative Model for Visual-Guided Binaural Stereo Generation. *arXiv preprint arXiv:2311.07630* (2023).
- [4] Yan-Bo Lin and Yu-Chiang Frank Wang. 2021. Exploiting Audio-Visual Consistency with Partial Supervision for Spatial Audio Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2056–2063.
- [5] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. 2022. Beyond Mono to Binaural: Generating Binaural Audio from Mono Audio with Depth and Cross Modal Attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 3347–3356.
- [6] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando De la Torre, and Yaser Sheikh. 2021. Neural Synthesis of Binaural Speech From Mono Audio. In *Proceedings of the International Conference on Learning Representations*.