Speech Masking System Based on Spatially Separated Multiple TTS Maskers With A Compact Circular Loudspeaker Array

Takuma Okamoto
National Institute of Information and Communications Technology, Japan
okamoto@nict.go.jp

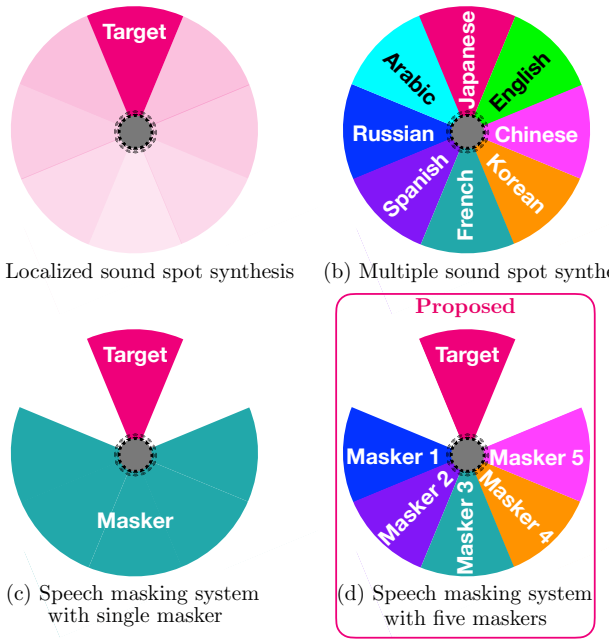
***Abstract*—Speech masking system with spatially separated maskers is proposed by utilizing multiple sound spot synthesis and multi-speaker neural text-to-speech (TTS) technologies. Although previous systems introduce time-reversed signals of target speech for maskers, these meaningless sounds are unpleasant to listeners. Conversely, the proposed method introduces maskers with the same voice quality as a target speech generated by multi-speaker neural TTS model with global style tokens. Additionally, these TTS-based maskers are synthesized by multiple sound spot synthesis in multiple directions. Then, the target speech can only be heard at the target direction while multiple TTS-based meaningful maskers can also be easily heard at the other directions without discomfort. We implement a speech masking demo system with a compact circular array of 16 loudspeakers carried out with a backpack. In the demonstration, the proposed speech masking system with spatially separated multiple TTS-based maskers is demonstrated.**

***Index Terms*—speech masking system, multiple sound spot synthesis, multi-speaker neural text-to-speech, global style tokens, loudspeaker array.**

I. Introduction

Compared to parametric arrays of ultrasonic loudspeakers [1], localized sound spot synthesis [2]–[16] (Fig. 1(a)), which can realize audible and inaudible areas by using multiple loudspeakers, is superior in terms of the synthesis sound quality and produced sound pressure level. Additionally, multiple sound spot synthesis (Fig. 1(b)), which can present different sounds in different zones simultaneously, is also an important technology for multilingual speech communication, museums, and other speech applications. We have proposed spatial Fourier transform [17]-based multiple sound spot synthesis methods [4], [6], [7], and implemented with a compact circular array of 16 loudspeakers [18]. Additionally, we have implemented a demo system integrating multiple sound spot synthesis and multilingual simultaneous speech-to-speech translation on-premise on a laptop without network connection. Finally, the complete demo system can be carried out with a backpack [19].

Similar to localized sound spot synthesis and multiple sound spot synthesis technologies, speech privacy [20] and sound masking [21] are also important for speech communication.[1]



(a) Localized sound spot synthesis (b) Multiple sound spot synthesis

(c) Speech masking system with single masker (d) Speech masking system with five maskers

Fig. 1. (a) Localized sound spot synthesis system. (b) Multiple sound spot synthesis system. (c) Speech masking system with single masker. (d) Proposed speech masking system based on spatially separated multiple maskers generated by multi-speaker text-to-speech with global style tokens of arbitrary target speech. All systems are implemented with a circular array of 16 loudspeakers.

In these techniques, maskers (masking signals) are introduced to make a target sound inaudible.[2] Compared with noise-based maskers [22], maskers generated from a target speech is more effective [23], and time-reversed signals of a target speech are introduced [24], [25]. However, the time-reversed signals are insufficient because these meaningless sounds are unpleasant to listeners. Additionally, the target speech can be sometimes heard because the temporal and frequency structures of the time-reversed signals are different from those of the target speech.

To introduce meaningful maskers while keeping the masking performance, we propose a speech masking system by uti-

This study was partly supported by JSPS KAKENHI under Grant Number JP23K11177.

[1]Several localized sound spot synthesis methods with loudspeaker arrays considering auditory masking have also been investigated [9]–[11].

[2]As shown in Fig. 1(a), localized sound spot synthesis without masker cannot be directly used for speech privacy because it is difficult to realize a completely silent direction, and a target sound can be heard even slightly in the non-target direction.

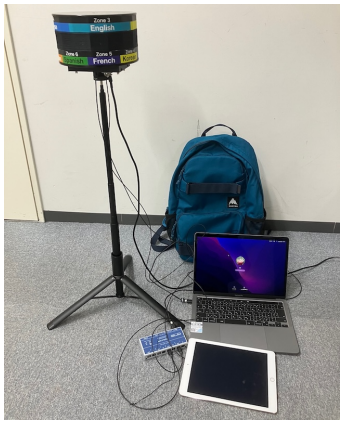


Fig. 2. Compact speech masking demo system with a circular array of 16 loudspeakers carried out with a backpack.

lizing multiple sound spot synthesis and multi-speaker neural text-to-speech (TTS) technologies. In the proposed method, multiple maskers with the same voice quality as arbitrary target speech are generated by multi-speaker neural TTS. Both speech masking system with single masker (Fig. 1(c)) or that with multiple maskers (Fig. 1(d)) can be realized by multiple sound spot synthesis. By the proposed speech masking system with multiple TTS-based maskers, the target speech can only be heard at the target direction while multiple TTS-based meaningful maskers can also be easily heard at the other directions without discomfort. Similar to the multiple sound spot synthesis demo system [19], the proposed speech masking demo system is also implemented with a compact circular array of 16 loudspeakers carried out with a backpack (Fig. 2).

II. Proposed method

A. Multi-speaker neural text-to-speech model

To generate maskers with the same voice quality as arbitrary target speech, multi-speaker neural TTS with global style tokens (GSTs) [26] is introduced. For stable and faster inference, a non-autoregressive neural TTS acoustic model based on [27] combined with GST (Fig. 3(a)) is introduced although the original TTS model with GST is autoregressive [26]. In the acoustic model, Conformer [28]-based encoder [29], ConvNeXt [30]-based decoder [31], and Gaussian upsampling [32], [33] are introduced. Additionally, phoneme embedding skip connection is also introduced for stable phoneme duration control [34], [35]. In the training, alignment between phoneme sequence and acoustic feature sequence is gradually trained with monotonic alignment search [33], [36]. The fundamental frequency estimated in the variance adapter is analyzed by Harvest [37] in the training. MS-FC-HiFi-GAN [38] is introduced for the neural vocoder. The acoustic model and neural vocoder are separately trained and jointly finetuned [27]. By using the acoustic model and neural vocoder trained with multi-speaker corpus (e.g. LibriTTS-R [39]), maskers with the same voice quality as arbitrary target speech can be generated (Fig. 3(a)).

B. Speech masking system with multiple sound spot synthesis

In the proposed speech masking system, multiple maskers with the same voice quality as a target speech are first generated by the multi-speaker neural TTS model with multiple text sequences and the target speech. Then, the target speech and multiple maskers are synthesized by spatial Fourier transform-based multiple sound spot synthesis using a circular loudspeaker array (Fig. 3(b)). Compared with the previous speech masking systems with meaningless time-reversed signals, the target speech can only be heard at the target direction while multiple TTS-based meaningful maskers can also be easily heard at the other directions without discomfort by the proposed speech masking system.

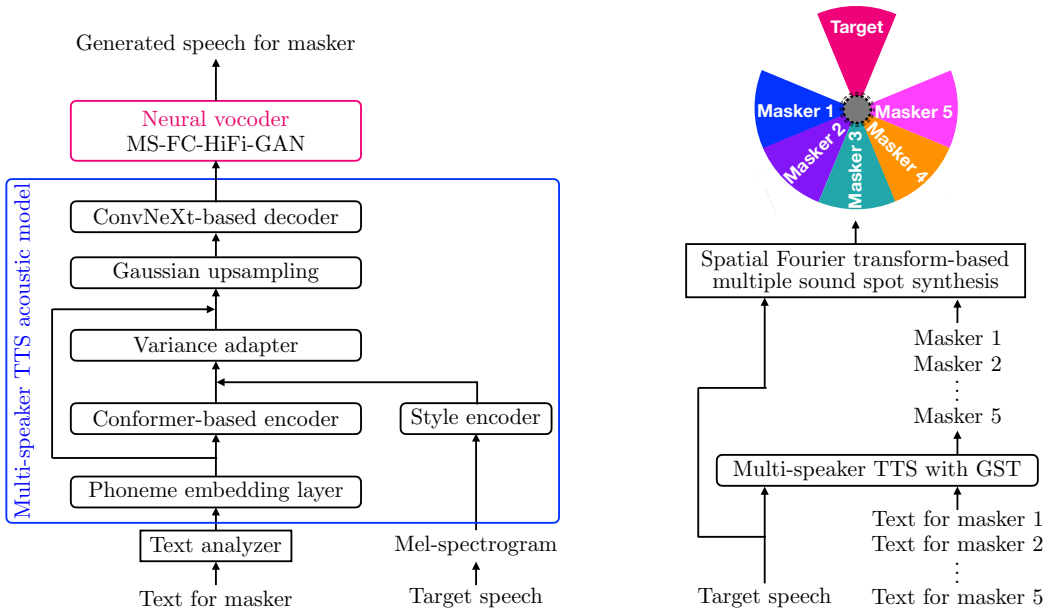
III. Implementation

The multi-speaker English neural TTS acoustic model and neural vocoder were trained using LibriTTS-R corpus [39]. These models were implemented by ESPnet2-TTS [40] on PyTorch [41], and were trained using an NVIDIA Tesla A100 GPU with 40 GB of memory. The sampling frequency was 24 kHz. The input acoustic features were 80-dimensional mel-spectrograms band-limited to 7,600 Hz. The STFT length and shift length were 1024 and 256 samples, respectively. The test set speech samples of the AudioMOS Challenge 2025 Track 3 are generated by this neural TTS model [42].

For the demo system, the multi-speaker neural TTS model was driven on a laptop (Apple MacBook Air M2 15 inch, Apple M2, 24 GB memory). In the demo system, not only speech corpora but also actual recorded speech can be used for the target speech. For the target speech from speech corpora, the test sets of English female and male speakers from Hi-Fi-CAPTAIN corpus [43], which is not included in the training, were used. For multiple sound spot synthesis, the compact system with a small 16-channel amplifier used in [19] was introduced. To easily control the target speech signals, maskers and their output powers, the demo system was implemented with PureData (Pd) [44] and controlled by a tablet (Apple iPad Air) via open sound control (Fig. 4). Then, multiple maskers with the same voice quality as the target speech can be generated on the laptop. Finally, the target speech and multiple maskers can be synthesized by multiple sound spot synthesis using a circular array of 16 loudspeakers. In the implementation for the proposed speech masking system with multiple TTS-based maskers, eight-direction sound spot synthesis was introduced, and single target speech ($\pi/4$ rad.) and five maskers ($\pi/4$ rad. $\times$ 5) can be synthesized as shown in Fig. 1(d). In the speech masking system with single masker, eight-direction sound spot synthesis was also introduced, and single target speech ($\pi/4$ rad.) and single masker ($5\pi/4$ rad.) can be synthesized as shown in Fig. 1(c).

In the implementation, the following speech masking systems can be demonstrated.

Single masker of noise or environmental sounds with Fig. 1(c): Although it can be easily realized, the output power of the masker should be high for high masking performance.



(a) Multi-speaker TTS with global style tokens

(b) Proposed speech masking system

Fig. 3. (a) Multi-speaker neural text-to-speech (TTS) with global style tokens (GSTs) for generating speech for masker. (b) Proposed speech masking system based on spatially separated multiple maskers generated by multi-speaker neural TTS with GST of arbitrary target speech.

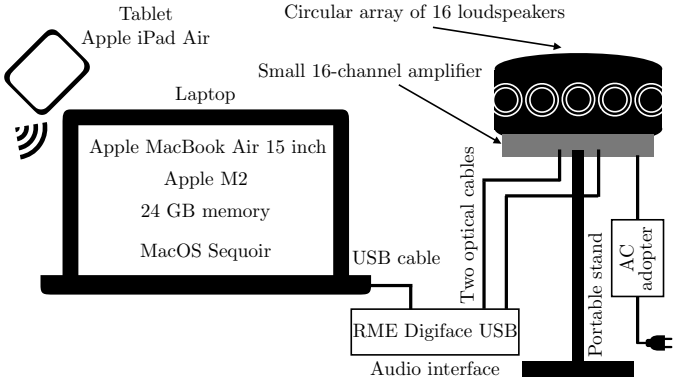


Fig. 4. Configuration of implemented compact demo system carried out with a backpack.

Therefore, the target speech is noisy, and listeners at non-target direction are suffer from the high sound pressure of the masker.

Single masker of time-reversed signals with Fig. 1(c): The output power of the masker should also be high for high masking performance. However, the time-reversed signals are insufficient because these meaningless sounds are unpleasant to listeners. Additionally, the target speech can be sometimes heard because the temporal and frequency structures of the time-reversed signals are different from those of the target speech.

Single masker generated by multi-speaker TTS with Fig. 1(c): The output power of the masker should also be high for high masking performance. Although the TTS-based masker is meaningful, the target speech can also be sometimes heard because the temporal and frequency structures of the masker generated by multi-speaker TTS are also different from those of the target speech.

Proposed multiple maskers generated by multi-speaker TTS with Fig. 1(d): High masking performance can be realized with low output powers of the maskers because multiple TTS-based maskers can efficiently mask the target speech by the superposition of various maskers with different temporal and frequency structures. Then, the target speech can only be heard at the target direction while multiple TTS-based meaningful maskers can also be easily heard at the other directions without discomfort.

IV. Demonstration

By using the implemented compact demo system using a circular array of 16 loudspeakers carried out with a backpack, the following four demonstrations are provided. In the demonstration, not only speech corpora but also actual recorded speech can be used for the target speech.

- Localized sound spot synthesis (Fig. 1(a))
- Multiple sound spot synthesis (Fig. 1(b))
- Speech masking system with single masker (white noise, pink noise, environmental sounds [45], time-reversed signals and speech generated by multi-speaker TTS) (Fig. 1(c))
- Proposed speech masking system with spatially separated multiple maskers generated by multi-speaker TTS (Fig. 1(d))

Participants can confirm the effectiveness of the proposed method through demonstrations by freely changing the target speech signals, maskers and their output powers.

REFERENCES

[1] P. J. Westervelt, "Parametric acoustic array," *J. Acoust. Soc. Am.*, vol. 35, no. 4, pp. 535–537, Apr. 1963.

[2] J.-W. Choi and Y.-H. Kim, "Generation of an acoustically bright zone with an illuminated region using multiple sources," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1695–1700, Apr. 2002.

[3] M. Shin, S. Q. Lee, F. M. Fazi, P. A. Nelson, D. Kim, S. Wang, K. H. Park, and J. Seo, "Maximization of acoustic energy difference between two spaces," *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 121–131, July 2010.

[4] T. Okamoto, "Generation of multiple sound zones by spatial filtering in wavenumber domain using a linear array of loudspeakers," in *Proc. ICASSP*, May 2014, pp. 4733–4737.

[5] T. Betlehem, W. Zhang, M. Poletti, and T. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015.

[6] T. Okamoto, "Analytical methods of generating multiple sound zones for open and baffled circular loudspeaker arrays," in *Proc. WASPAA*, Oct. 2015.

[7] T. Okamoto and A. Sakaguchi, "Experimental validation of spatial Fourier transform-based multiple sound zone generation with a linear loudspeaker array," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1769–1780, Mar. 2017.

[8] F. Olivieri, F. M. Fazi, S. Fontana, D. Menzies, and P. A. Nelson, "Generation of private sound with a circular loudspeaker array and the weighted pressure matching method," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1579–1591, Aug. 2017.

[9] J. Donley, C. H. Ritz, and W. B. Kleijn, "Multizone soundfield reproduction with privacy and quality based speech masking filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1041–1055, June 2018.

[10] T. Lee, J. K. Nielsen, and M. G. Christensen, "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2412–2426, 2020.

[11] D. Wallace and J. Cheer, "Design and evaluation of personal audio systems based on speech privacy constraints," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2271–2282, Apr. 2020.

[12] L. Shi, T. Lee, L. Zhang, J. K. Nielsen, and M. G. Christensen, "Generation of personal sound zones with physical meaningful constraints and conjugate gradient method," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 823–837, 2021.

[13] T. Okamoto, "2D multizone sound field synthesis with interior-exterior Ambisonics," in *Proc. WASPAA*, Oct. 2021, pp. 276–280.

[14] V. Moles-Cases, S. J. Elliott, J. Cheer, G. Pinero, and A. Gonzalez, "Weighted pressure matching with windowed targets for personal sound zones," *J. Acoust. Soc. Am.*, vol. 151, no. 1, pp. 334–345, Jan. 2022.

[15] M. Hu, H. Zou, J. Li, and M. G. Christensen, "Maximizing the acoustic contrast with constrained reconstruction error under a generalized pressure matching framework in sound zone control," *J. Acoust. Soc. Am.*, vol. 151, no. 4, pp. 2571–2759, Apr. 2022.

[16] T. Abe, S. Koyama, N. Ueno, and H. Saruatari, "Amplitude matching for multizone sound field control," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 656–669, 2023.

[17] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*. London: Academic Press, 1999.

[18] T. Okamoto, "Multilingual sound spot synthesis systems," in *Proc. Internoise*, Aug. 2023, p. 5861–5865.

[19] T. Okamoto and M. Kono, "Simultaneous speech translation integrated compact multiple sound spot synthesis system on a laptop carried out with a backpack," in *Proc. Interspeech*, Aug. 2025.

[20] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle, and B. G. Watters, "Speech privacy in buildings," *J. Acoust. Soc. Am.*, vol. 34, no. 4, p. 475–492, Apr. 1962.

[21] A. W. Bronkhost and R. Plomp, "Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Am.*, vol. 92, no. 6, p. 3132–3139, Dec. 1992.

[22] J. S. Bradley, "The acoustical design of conventional open plan offices," *Can. Acoust.*, vol. 31, no. 2, pp. 32–31, June 2003.

[23] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 110, no. 5, p. 2527–2538, Nov. 2001.

[24] T. Arai, "Masking speech with its time-reversed signal," *Acoust. Sci. Tech.*, vol. 31, no. 2, p. 188–190, Mar. 2010.

[25] Y. Hioka, J. James, and C. I. Watson, "Masker design for real-time informational masking with mitigated annoyance," *Appl. Acoust.*, vol. 159, p. 107073, Feb. 2020.

[26] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, July 2018, pp. 5167–5176.

[27] T. Okamoto, Y. Ohtani, and H. Kawai, "Mobile PresenTra: NICT fast neural text-to-speech system on smartphones with incremental inference of MS-FC-HiFi-GAN for low-latency synthesis," in *Proc. Interspeech*, Sept. 2024, pp. 997–998.

[28] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036–5040.

[29] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by Conformer," in *Proc. ICASSP*, June 2021, pp. 5874–5878.

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. CVPR*, June 2022, pp. 11 976–11 986.

[31] T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, "ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion," in *Proc. ICASSP*, Apr. 2024, pp. 12456–12460.

[32] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. ICLR*, May 2021.

[33] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, Sept. 2022, pp. 21–25.

[34] T. Okamoto, Y. Ohtani, S. Shimizu, T. Toda, and H. Kawai, "Challenge of singing voice synthesis using only text-to-speech corpus with FIRNet source-filter neural vocoder," in *Proc. Interspeech*, Sept. 2024, pp. 1870–1874.

[35] T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda, and H. Kawai, "Phoneme-level duration controllable neural text-to-speech with phoneme embedding skip connection and modified Gaussian duration modeling," *IEEE Access*, vol. 13, pp. 118 369–118 380, 2025.

[36] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.

[37] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.

[38] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, "Fast neural speech waveform generative models with fully-connected layer-based upsampling," *IEEE Access*, vol. 12, pp. 31409–31421, 2024.

[39] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech*, Aug. 2023, pp. 5496–5500.

[40] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv:2110.07840*, 2021.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

[42] W.-C. Huang, H. Wang, C. Liu, Y.-C. Wu, A. Tjandra, W.-N. Hsu, E. Cooper, Y. Qin, and T. Toda, "The AudioMOS Challenge 2025," in *Proc. ASRU*, Dec. 2025.

[43] T. Okamoto, Y. Shiga, and H. Kawai, "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.

[44] M. S. Puckette, "Pure data," in *Proc. ICMC*, Sept. 1997.

[45] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. MM*, Oct. 2015, pp. 1015–1018.