Multilingual sound sopt synthesis systems

Takuma Okamoto[1]
National Institute of Information and Communications Technology
3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan

ABSTRACT

Localized sound zone synthesis, which can generate acoustically bright and dark zones using loudspeakers, is gaining attention as one of important acoustic communication techniques. By superposing multiple localized sound zones, multiple sound spot synthesis can simultaneously deriver different sound signals at different zones using a loudspeaker array. Additionally, it can be implemented as a user interface for real-time speech translation by combining multilingual speech synthesis technology. This paper introduces multilingual sound spot synthesis systems based on spatial Fourier transform implemented by using compact circular array of 16 loudspeakers and linear array of 64 loudspeakers combined with multilingual neural speech synthesis.

1. MULTIPLE SOUND SPOT SYNTHESIS TECHNOLOGY

Compared to parametric arrays of ultrasonic loudspeakers, localized sound spot synthesis [1,2] (Fig. 1(a)), which can realize audible and inaudible areas using multiple loudspeakers, is superior in terms of the synthesis sound quality and sound pressure level, and is an important technology as a sound presentation system [3]. Additionally, multiple sound spot synthesis [4–7] (Fig. 1(b)), which can simultaneously present different sounds in different areas by superposing multiple localized sound spots, is also an important sound presentation technology for multilingual communication, museums, and other speech and audio applications.

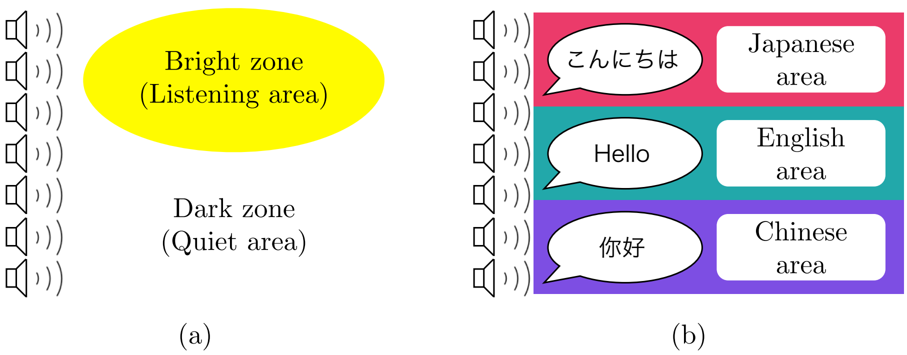


Figure 1: (a) Generating acoustically bright and dark zones with a loudspeaker array, (b) Generating multiple sound zones with a loudspeaker array.

Previously, we have proposed localized sound spot synthesis and multiple spot spot synthesis schemes based on spatial Fourier transforms [8] that can control continuous sound pressure without

[1]okamoto _ _at _ _ nict.go.jp

discrete control points [4–6,9–14], instead of multiple control points used in conventional methods [1, 2, 15, 16].

In spatial Fourier transform-based localized sound spot synthesis methods, the sound pressure distribution in reference line or circle is modelled as a rectangular window or a Hann window where the sound pressure in the audible region is 1 and that in the inaudible region is 0, as shown in Fig. 2(a) and (b). The Fourier transforms of the rectangular and Hann windows can then be analytically calculated. The spatial Fourier transform can explicitly remove the evanescent wave component, which is unstable in the acoustic inverse problem [8], and thus can calculate a more stable driving signal than the conventional methods using multiple control points. In the case of using linear and circular arrays, the proposed spatial Fourier transform-based methods [4–6] achieved more accurate control performance than the conventional method using multiple control points [1, 2, 15, 16].

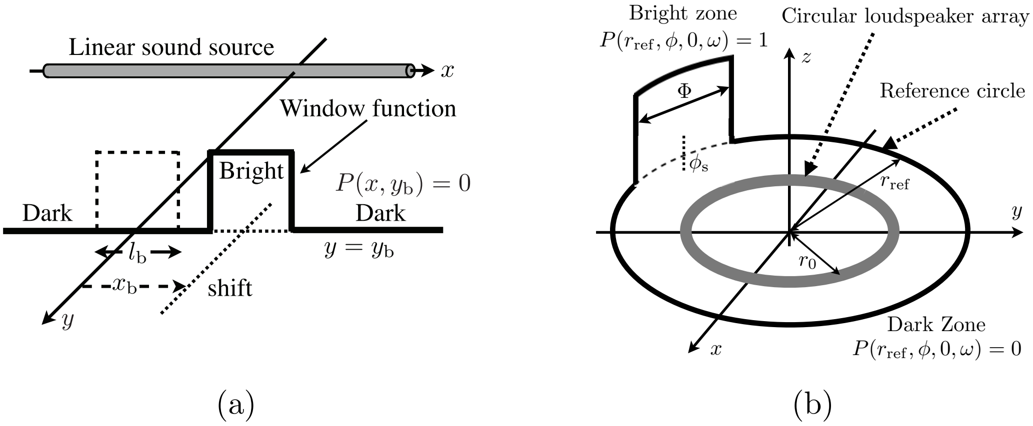


Figure 2: Generating acoustically bright and dark zones based on continuous rectangular sound pressure modeling (a) for a linear array and (b) for a circular array.

2. MULTILINGUAL SPEECH SYNTHESIS TECHNOLOGY

With the development of neural network technology, text-to-speech (TTS) synthesis technology has become capable of producing high-quality speech that is comparable to natural speech [17–19]. Recently, end-to-end models that can generate high-quality speech waveforms directly from text input using a single neural network have been proposed [20–22].

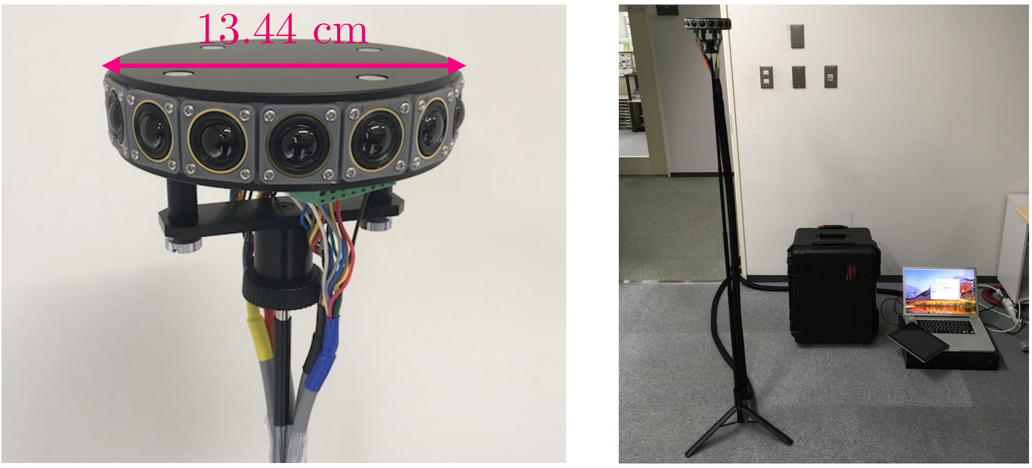
In NICT, we have been conducting research on neural network-based speech waveform generative models and TTS models [23, 24], and currently high-quality neural speech synthesis in 10 languages is available through our speech translation application, VoiceTra.[2]

3. IMPLEMENTATION OF MULTILINGUAL SOUND SPOT SYNTHESIS SYSTEMS

To promote the use of multiple sound spot synthesis technology, we have implemented a portable multiple sound spot synthesis system with a compact circular array of 16 loudspeakers [25] (Fig. 3 and Fig. 4(a)). Additionally, we have also implemented a multilingual sound spot synthesis system as a novel user interface for speech translation by integrating the spatial Fourier transform-based multiple sound spot synthesis technology and multilingual neural speech synthesis ((Fig. 4(b))).

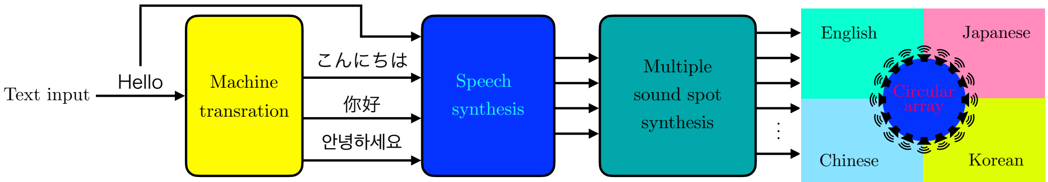
A demonstration experiment using the implemented multilingual sound spot synthesis system was conducted at Miraikan in December 2022 ((Fig. 5(a))). In the demonstration experiment, an alternative multilingual sound spot synthesis system with a linear array of 64 loudspeakers was additionally implemented ((Fig. 5(b))).

[2] https://voicetra.nict.go.jp

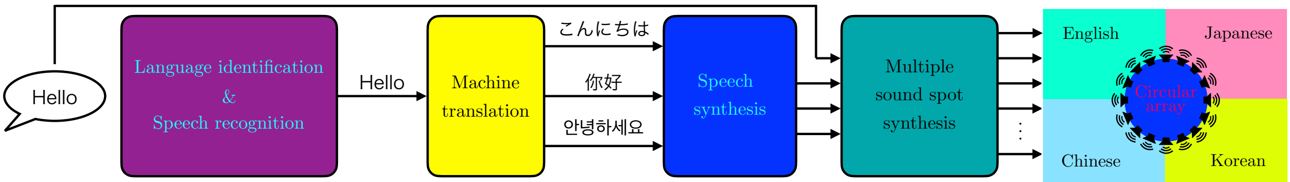


(a) (b)

Figure 3: Implemented small circular array of 16 loudspeakers.



(a) with text input



(b) as user interface for real-time speech translation

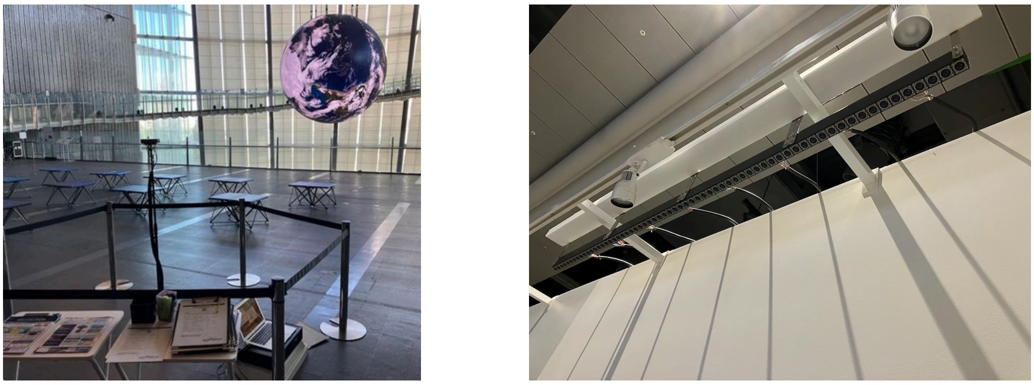
Figure 4: Implemented multilingual multiple sound spot synthesis systems.

ACKNOWLEDGEMENTS

This research was partly supported by JSPS KAKENHI Grant Numbers JP18K11387 and JP23K11177, and JST A-STEP Tryout Grant Number JPMJTM22D8. The development of the prototypes of the compact circular array of 16 loudspeakers and linear array of 64 loudspeakers was supported by JST Program Manager (PM) Development and Promotion Program (PI: Keita Hikita), and were developed in collaboration with Kitanihon Onkyo Co. Ltd. The demonstration test of multiple sound spot synthesis systems using the compact circular array and linear array was conducted at Miraikan in Tokyo, Japan. Finally, we would like to thank Keita Hikita, Miyuki Sudo, and Manae Kuwabara for their support of NICT Innovation Design Initiative (IDI), Project of Co-Creation Design (PoCC) for the social deployment of multiple sound spot synthesis technology.

REFERENCES

1. J.-W. Choi and Y.-H. Kim. Generation of an acoustically bright zone with an illuminated region using multiple sources. *J. Acoust. Soc. Am.*, 111(4):1695–1700, Apr. 2002.
2. 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.*, 128(1):121–131, July 2010.



(a) (b)

Figure 5: Demonstration experiments with (a) implemented compact circular array of 16 loudspeakers and (b) implemented linear array of 64 loudspeakers in Miraikan.

3. T. Okamoto. Spatial Fourier transform-based localized sound zone generation methods with loudspeaker arrays. *J. Acoust. Soc. Am.*, 146(4):2761–2762, Oct. 2019.
4. T. Okamoto. Generation of multiple sound zones by spatial filtering in wavenumber domain using a linear array of loudspeakers. In *Proc. ICASSP*, pages 4733–4737, May 2014.
5. T. Okamoto. Analytical methods of generating multiple sound zones for open and baffled circular loudspeaker arrays. In *Proc. WASPAA*, Oct. 2015.
6. 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.*, 141(3):1769–1780, Mar. 2017.
7. 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.*, 28:2412–2426, 2020.
8. E. G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*. Academic Press, London, 1999.
9. T. Okamoto. Near-field sound propagation based on a circular and linear array combination. In *Proc. ICASSP*, pages 624–628, Apr. 2015.
10. T. Okamoto. Horizontal local sound field propagation based on sound source dimension mismatch. *J. Inf. Hiding Multimed. Signal Process.*, 8(5):1069–1081, Sept. 2017.
11. T. Okamoto. Localized sound zone generation based on external radiation canceller. *J. Inf. Hiding Multimed. Signal Process.*, 8(6):1335–1351, Nov. 2017.
12. T. Okamoto. 2.5D localized sound zone generation with a circular array of fixed-directivity loudspeakers. In *Proc. IWAENC*, pages 321–325, Sept. 2018.
13. T. Okamoto. 3D localized sound zone generation with a planar omni-directional loudspeaker array. In *Proc. WASPAA*, pages 110–114, Oct. 2019.
14. T. Okamoto. 2D multizone sound field synthesis with interior-exterior Ambisonics. In *Proc. WASPAA*, pages 276–280, Oct. 2021.
15. O. Kirkeby and P. Nelson. Reproduction of plane wave sound fields. *J. Acoust. Soc. Am.*, 94(5):2992–3000, Nov. 1993.
16. T. Betlehem and T. D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *J. Acoust. Soc. Am.*, 117(4):2100–2111, Apr. 2005.

17. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. In *Proc. SSW9*, page 125, Sept. 2016.
18. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*, pages 4779–4783, Apr. 2018.
19. J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, pages 17022–17033, Dec. 2020.
20. J. Kim, J. Kong, and J. Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. ICML*, pages 5530–5540, July 2021.
21. D. Lim, S. Jung, and E. Kim. JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech. In *Proc. Interspeech*, pages 21–25, Sept. 2022.
22. T. Okamoto, T. Toda, and H. Kawai. E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion. In *Proc. Interspeech*, pages 2043–2047, Aug. 2023.
23. T. Okamoto, T. Toda, Y. Shiga, and H. Kawai. Noise level limited sub-modeling for diffusion probabilistic vocoders. In *Proc. ICASSP*, pages 6014–6018, June 2021.
24. T. Okamoto, T. Toda, and H. Kawai. Multi-stream HiFi-GAN with data-driven waveform decomposition. In *Proc. ASRU*, pages 610–617, Dec. 2021.
25. T. Okamoto, K. Ueno, T. Okabe, K. Tani, Y. Yoshikata, M. Sudo, M. Kuwahara, and K. Hikita. Portable multilingual sound spot synthesis system with a compact circular array of 16 loudspeakers. In *ICASSP 2023 Show & Tell Demo*, June 2023.