IMPROVED PORTABLE MULTIPLE SOUND SPOT SYNTHESIS SYSTEM WITH A BAFFLED CIRCULAR ARRAY OF 16 LOUDSPEAKERS

Takuma Okamoto[1]*, *Katsushi Ueno*[2], *Tsukasa Okabe*[2], *Kentaro Tani*[2], *Yasuhiko Yoshikata*[2], *Miyuki Sudo*[1], *Manae Kuwahara*[1], *Keita Hikita*[1]

[1]National Institute of Information and Communications Technology, Japan, [2]Kitanihon Onkyo, Japan

ABSTRACT

To make multiple sound spot synthesis technology widely available, we have previously implemented a portable demonstration system based on mode-matching method using a compact circular array of 16 loudspeakers. However, the previous system has two critical problems for practical applications. 1) The maximum output sound pressure level is only 72 dB, which is too small for demonstration in a large conference venue. 2) The output sound quality is not high because low frequency components below 500 Hz cannot be produced. Additionally, 3) a large space is required because 4) participants must basically walk around the array to listen to the synthesized sound field. To improve the previous system, we developed an improved system using a baffled circular array of 16 loudspeakers combined with an electric turntable. The improved system can 1) produce a sound pressure level of 80 dB, which is sufficient for demonstration in a large conference venue, and 2) drastically improve the synthesis sound quality with low frequency components above 200 Hz. Additionally, it can be demonstrated 3) in a small space because 4) participants do not have to walk around the array by the electric turntable. It can realize eight-language sound spot synthesis for eight directions. The improved system including the turntable can be carried out with a single suitcase. In WASPAA 2023 Demonstrations, we demonstrate both four- and eight-language sound spot synthesis for four and eight directions using the improved system.

Index Terms— Circular loudspeaker array, localized sound spot synthesis, multiple sound spot synthesis, portable demo system, text-to-speech

1. INTRODUCTION

Compared to parametric arrays of ultrasonic loudspeakers [1], localized sound spot synthesis [2–13], which can realize audible and inaudible areas using multiple loudspeakers, is superior in terms of the synthesis sound quality and produced sound pressure level. Additionally, multiple sound spot synthesis [4–6, 8, 14], 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.

Typical sound field synthesis methods, such as wave field synthesis [15–17], spectral division method [18, 19], and higher-order Ambisonics (HOA) [20–22], introduce planar, linear and surrounding arrays of loudspeakers. However, these arrays are difficult to transport easily compared with applications for speech communication technologies, such as automatic speech recognition (ASR) [23] and text-to-speech (TTS) [24], that can be easily demonstrated elsewhere using a smartphone.

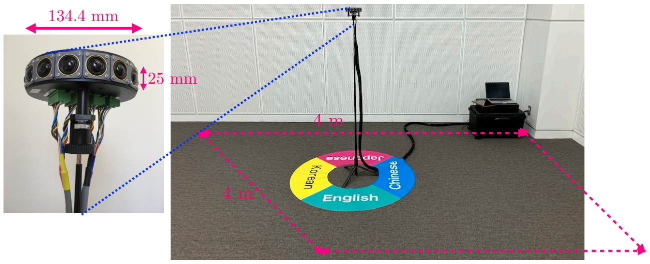


Figure 1: Previous demo system.

2. PREVIOUS SYSTEM

2.1. Overview of previous system

To make sound field control technology, especially multiple sound spot synthesis technology widely available, we have implemented a portable multiple sound spot synthesis system based on mode-matching method [5] using a compact circular array of 16 loudspeakers as small as possible (Fig. 1) [25, 26].[1][2] Each loudspeaker driver is 25 mm, and the diameter of the circular array is only 134.4 mm. The spatial Nyquist frequency is then about 6.5 kHz. The implemented system, constructed from the compact loudspeaker array, an amplifier for 16 loudspeakers including D/A, a loudspeaker stand, a laptop, a tablet and cables, can be carried out with a single suitcase. The demo system is implemented with Pure-Data (Pd) [27] and controlled by the tablet via open sound control. The synthesized sound field can be interactively rotated by the tablet in real-time with HOA-based panning implemented in Pd. The previous demo system realized four-language sound spot synthesis for four directions combined with multilingual neural TTS [28] (Fig. 3(a)), where neural TTS models for English and Japanese are trained using Hi-Fi-CAPTAIN corpus [29].

2.2. Problems of previous system

Although the previous array is compact, it has two critical problems for practical applications because the size of the array is too small. 1) The maximum output sound pressure level is 70 dB, which is too

*This study was partly supported by JSPS KAKENHI Grant Number JP23K11177, and JST A-STEP Tryout Grant Number JPMJTM22D8.

[1] https://ast-astrec.nict.go.jp/en/MultipleSoundSpotSynthesis/

[2] https://youtu.be/In8AfVcoTC4

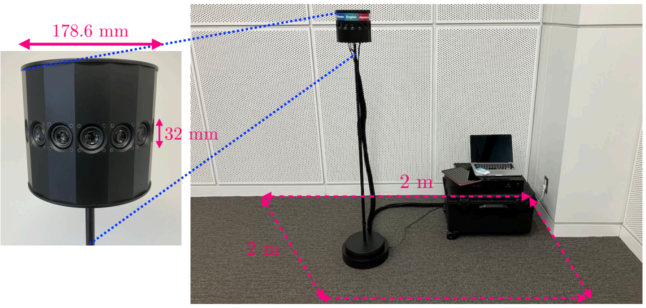
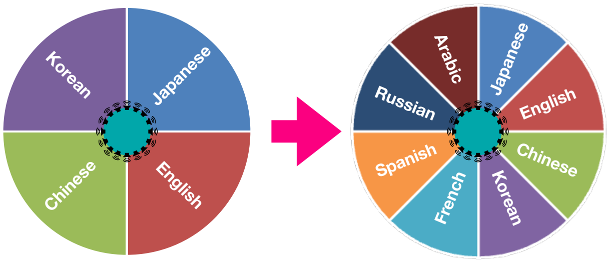


Figure 2: Improved demo system with an electric turntable. It can be carried out with a single suitecase.



(a) Previous
Four spot synthesis

(b) Improved
Eight spot synthesis

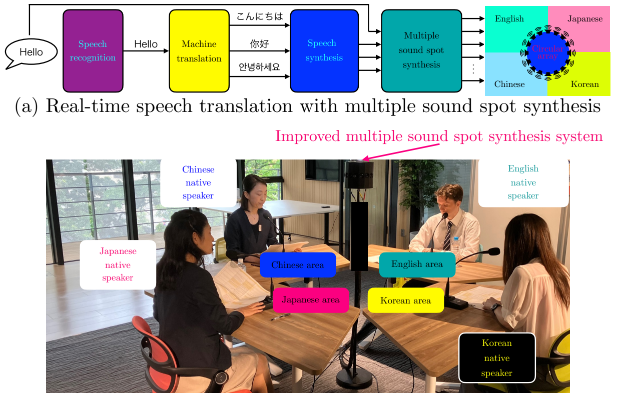
Figure 3: (a) Four-language sound spot synthesis for four directions by previous system. (b) Eight-language sound spot synthesis for eight directions by improved system.

small for demonstration in a large conference venue. 2) The output sound quality is not high because low frequency components below 500 Hz cannot be produced. Additionally, 3) a large space is required for demonstration because 4) participants must basically walk around the array to listen to different sounds although the synthesized sound field can be interactively rotated by the tablet (Fig. 1).

3. IMPROVED SYSTEM

3.1. Overview of improved system

To improve the previous system, we developed an improved system using a baffled circular array of 16 loudspeakers combined with an electric turntable (Fig. 2). To improve the total output power and low frequency response, we developed a slightly larger loudspeaker driver with a diameter of 32 mm, and the diameter of the improved circular array is 178.6 mm. The spatial Nyquist frequency is then about 4.9 kHz. For avoiding the diffraction of sound from other directions, a cylindrical rigid baffle was introduced. Although the spatial Nyquist frequency of the improved system is lower than that of the previous system, the improved system can 1) produce a sound pressure level of 80 dB, which is sufficient for demonstration in a large conference venue, and 2) drastically improve the synthesis sound quality with low frequency components above 200 Hz. Additionally, it can be demonstrated 3) in a small space, and 4) participants do not have to walk around the array by the electric turntable. Furthermore, it can realize eight-language sound spot synthesis for eight directions (Fig. 3(b)). The improved system including the turntable can also be carried out with a single suitcase. The detailed results of experiments will be submitted to a journal paper.



(a) Real-time speech translation with multiple sound spot synthesis

(b) Actual implemented real-time four-language speech translation system

Figure 4: (a) Real-time multilingual speech translation system combined with multiple sound spot synthesis. (b) Actual implemented real-time four-language speech translation system combined with improved multiple sound spot synthesis system.

3.2. WASPAA 2023 Demonstrations

In WASPAA 2023 Demonstrations, we demonstrate both four- and eight-language sound spot synthesis for four and eight directions using the improved system (Fig. 3). The input eight-language speech sounds are also generated by multilingual TTS based on high-fidelity and real-time methods [28, 30, 31]. Additionally, copyright-free music audio signals are also synthesized. The driving signals of 16 loudspeakers are calculated from the input audio signals and FIR filters for localized sound spot synthesis for each area in real-time by FFT-based convolution. The demo system is consructed from the baffled circular array, amplifier for 16 loudspeakers including D/A, loudspeaker stand, laptop, tablet, cables, and turntable (Fig. 2). All the equipment can be carried out with a single suitcase. A single electrical outlet (100 V in USA) is sufficient. The minimum space required for the demonstration is 2 m × 2 m (Fig. 2). If a larger space is available, it is better. Participants can experience the synthesized eight-spot sound field that they have never heard before, and they can interactively rotate the synthesized sound field with the tablet. Additionally, they can understand that demonstration systems for sound field control can also be carried out elsewhere.

3.3. Actual implemented application (not included in demo)

Finally, we briefly introduce an actual implemented application using the improved system, which has been presented in NICT openhouse 2023. By combining the improved multiple sound spot synthesis system based on acoustic signal processing with real-time multilingual speech translation system (ASR + machine translation + TTS) based on machine learning (Fig. 4(a)), we have actually implemented a real-time four-language speech translation system with multiple sound spot synthesis (Fig. 4(b)). In the system, each native speaker speaks only his or her own language and listens only to the synthesized speech sounds translated from other languages into his or her own language, where each translated speech sound is heard only by each native speaker through multiple sound spot synthesis.

4. 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] ——, “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.*, vol. 141, no. 3, pp. 1769–1780, Mar. 2017.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] S. Spors, R. Rabenstein, and J. Ahrens, “The theory of wave field synthesis revisited,” in *Proc. 124th Conv. Audio Eng. Soc.*, May 2008.

[16] A. J. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[17] G. Firtha, P. Fiala, F. Schultz, and S. Spors, “Improved referencing schemes for 2.5D wave field synthesis driving functions,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1117–1127, May 2017.

[18] J. Ahrens and S. Spors, “Sound field reproduction using planar and linear arrays of loudspeakers,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2038–2050, Nov. 2010.

[19] G. Firtha, P. Fiala, F. Schultz, and S. Spors, “On the general relation of wave field synthesis and spectral division method for linear arrays,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2393–2043, Dec. 2018.

[20] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, Nov. 2005.

[21] J. Ahrens and S. Spors, “An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions,” *Acta Acust. Acust.*, vol. 94, no. 6, pp. 988–999, Nov. 2008.

[22] F. Winter, J. Ahrens, and S. Spors, “On analytic methods for 2.5-D local sound field synthesis using circular distributions of secondary sources,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 914–926, May 2016.

[23] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[24] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[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.

[26] T. Okamoto, “Multilingual sound spot synthesis systems,” in *Proc. Internoise*, Aug. 2023.

[27] M. S. Puckette, “Pure data,” in *Proc. ICMC*, Sept. 1997.

[28] T. Okamoto, T. Toda, and H. Kawai, “Multi-stream HiFi-GAN with data-driven waveform decomposition,” in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[29] 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.

[30] 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.

[31] T. Okamoto, T. Toda, and H. Kawai, “E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion,” in *Proc. Interspeech*, Aug. 2023, pp. 2043–2047.