Simultaneous Speech Translation Integrated Compact Multiple Sound Spot Synthesis System On A Laptop Carried Out With A Backpack

Takuma Okamoto[1], *Michiyo Kono*[1]

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

okamoto@nict.go.jp

Abstract

Multiple sound spot synthesis, which can present different sounds in different zones simultaneously using a loudspeaker array, is an important spatial sound presentation technology for speech and audio applications. We have previously implemented a portable multiple sound spot synthesis system with a compact circular array of 16 loudspeakers, which can be carried out with a suitcase. To further improve the mobility, we implement a very small 16-channel amplifier directly mounted under the compact circular array. Additionally, we implement a 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. In the Show & Tell, we demonstrate four-language sound spot synthesis combined with multilingual simultaneous speech-to-speech translation using the compact demo system.

Index Terms: loudspeaker array, multilingual simultaneous speech-to-speech translation, multiple sound spot synthesis, portable demo system, sound field control

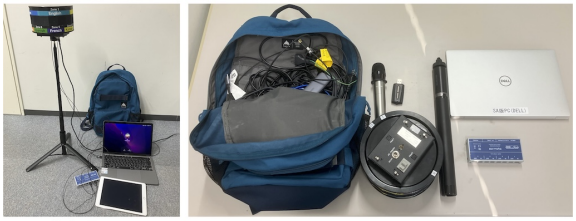
1. Introduction

Multiple sound spot synthesis, which can present different sounds in different zones simultaneously using a loudspeaker array, is an important spatial sound presentation technology for multilingual speech communication, museums, and other speech and audio applications. We have proposed spatial Fourier transform-based multiple sound spot synthesis methods with linear and circular loudspeaker arrays [1, 2].

2. Previous System

To make multiple sound spot synthesis technology widely available, we have implemented a portable multiple sound spot synthesis system using a compact circular array of 16 loudspeakers [3,4].[1][2] Each loudspeaker driver is 32 mm, and the diameter of the circular array is only 178.6 mm. The spatial Nyquist frequency is then about 4.9 kHz. The implemented system, constructed from the compact loudspeaker array, an amplifier for 16 loudspeakers including D/A (430 mm $\times$ 300 mm $\times$ 90 mm), a loudspeaker stand, a laptop, a tablet and cables, can be carried out with a single suitcase (Fig. 1(a)). The demo system is implemented with PureData (Pd) [5] and controlled by the tablet via open sound control. The previous demo system realized four- or eight-language sound spot synthesis for four or eight directions combined with multilingual neural text-to-speech (TTS) developed by NICT [6]. Additionally, we have implemented a four-language simultaneous speech-to-speech translation system[3] with multiple sound spot synthesis [4], by introducing neural network-based multilingual speech translation technologies (automatic speech recognition (ASR) [7] + simultaneous interpretation [8] + TTS [6]) developed by NICT and implemented in VoiceTra[4], which is 21-language speech-to-speech translation application for smartphones.

(a) Previous system carried out with a suitcase



(b) Proposed system carried out with a backpack

Figure 1: *(a) Previous portable multiple sound spot synthesis system carried out with a suitcase [3, 4]. (b) Proposed compact multiple sound spot synthesis system combined with multilingual simultaneous speech-to-speech translation implemented on-premise on a laptop and carried out with a backpack.*

The previous demo system has the following issues.

- Although the previous demo system is portable, a suitcase is required for transportation. This is because only the circular array is compact but the multichannel amplifier and D-sub 25 cables connecting amplifier and loudspeaker array are still bulky. However, further downsizing of the demo system is required because transporting the demo system with a suitcase is not convenient. (Meanwhile applications for speech communication technologies that can be easily demonstrated elsewhere using a smartphone.)
- The previous demo system by itself can only perform multiple sound spot synthesis using pre-prepared sound sources including synthesized speech. A demo combined with mul-

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

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

[3]https://youtu.be/uyTRd5Hu6hw

[4]https://voicetra.nict.go.jp/en/index.html

This paper is accepted at Interspeech 2025 (Show & Tell).
The preprint is published according to Interspeech Policy (https://isca-speech.org/Interspeech-Policy).

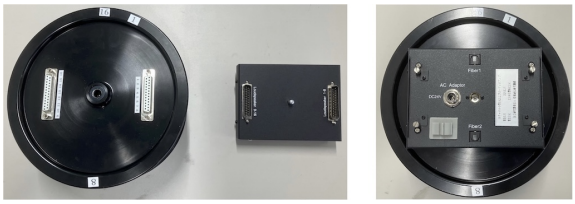


Figure 2: *Implemented very small amplifier directly mounted under a circular array of 16 loudspeakers.*

tilingual simultaneous speech-to-speech translation cannot be conducted using only the previous demo system, and another server-based multilingual simultaneous speech-to-speech translation system with network connection is required.

3. Proposed System

To further improve the mobility and convenience of the demo system, we implement a very small 16-channel amplifier (120 mm × 90 mm × 18 mm ≈ 1/60 of previous amplifier) with two optical digital audio input terminals for 16-channel audio signal inputs and an AC adopter input terminal for power supply, which can be directly mounted under the circular array and connected with two thin optical cables. Then, D-sub 25 cables connecting amplifier and loudspeaker array are not required. By the development, the demo system can be drastically downsized while keeping the synthesis quality and output power compared with the previous demo system. Additionally, we implement a 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 (Fig. 1(b)). The configuration of the implemented compact demo system is shown in Fig. 3.

4. Show & Tell Demonstration

In the Show & Tell, we conduct three types of demonstrations using the compact demo system carried out with a backpack.

1. Eight-language sound spot synthesis (Fig. 4(a)).
2. Eight different English content sound spot synthesis (Fig. 4(b)).
3. Four-language sound spot synthesis combined with multilingual simultaneous speech-to-speech translation (Japanese to English, Chinese and Korean) (Fig. 4(c)).

5. Acknowledgments

We thank NICT Universal Communication Research Institute (UCRI) members for their help in implementing multilingual simultaneous speech-to-speech translation system. We also thank NICT Innovation Design Initiative (IDI) members for the social deployment of multiple sound spot synthesis technology. This study was partly supporred by JSPS KAKENHI Grant Number JP23K11177.

6. References

[1] 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.

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

[3] ——, "Multilingual sound spot synthesis systems," in *Proc. Internoise*, Aug. 2023, pp. 5861–5865.

[4] T. Okamoto, K. Ueno, T. Okabe, K. Tani, Y. Yoshikata, M. Sudo, M. Kuwahara, and K. Hikita, "Improved portable multiple sound spot synthesis system with a baffled circular array of 16 loudspeakers," in *WASPAA 2023 Demonstrations*, Oct. 2023. [Online]. Available: https://www.okamotocamera.com/waspaa_2023_demo.pdf

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

[6] 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.

[7] P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko, and C. Hori, "The NICT ASR system for IWSLT 2014," in *Proc. IWSLT*, Dec. 2014, pp. 113–118.

[8] W. Xiaolin, M. Utiyama, and E. Sumita, "Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network," in *Proc. MTSummit*, Aug. 2019, pp. 1–11.

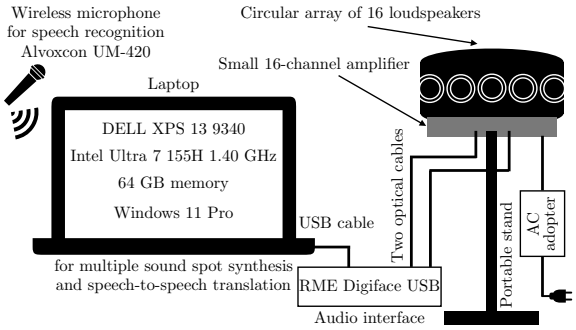


Figure 3: *Configuration of implemented compact demo system carried out with a backpack.*

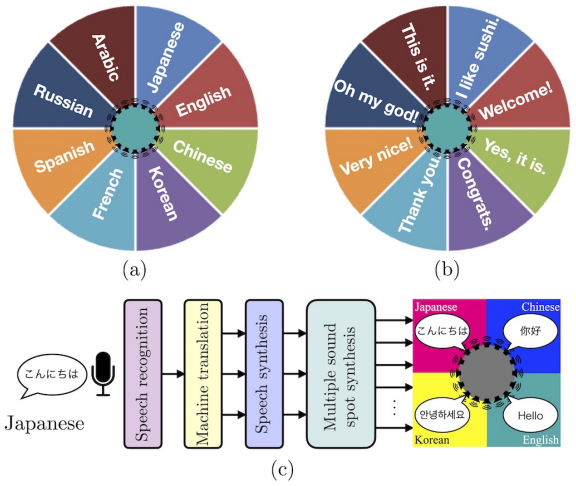


Figure 4: *(a) Eight-language sound spot synthesis. (b) Eight different English content sound spot synthesis. (c) Four-language sound spot synthesis combined with multilingual simultaneous speech-to-speech translation (Japanese to English, Chinese and Korean).*