Design of 3D microphone arrays using automatic speech recognition and psychoacoustic evaluation for immersive sounds

Hansol Lim[1], Hyung Suk Jang[1], Jin Yong Jeon[1]
Department of Architectural Engineering, Hanyang University, 133-791 Seoul, Korea[1]

Summary
A microphone array system was developed to capture the speech for the transmission to a remote space. The captured signals by the different types of arrays were synthesized into the binaural signals and evaluated for speech recognition. The automatic speech recognition algorithm, based on hidden Markov model, was applied to quantify the accuracy of speech recognition with a word list. Several noise environments were considered with different signal to noise ratios and the multi-channel sound capturing systems was found as most efficient.

PACS no. 43.60.+d, 43.72.+q,

1. Introduction

A coexistent reality (CR) is a reality in which users from remote locations can mutually communicate, share, collaborate, and feel each other's presence. To get a sense of coexistence, techniques in 3D vision, audio, movement, force feedback, haptic feedback, and other sensations need to be developed and combined. The first step in generating a feeling of acoustic coexistence is developing a teleconferencing system that transmits spatial audio speech to a real remote space.

The spatial audio in a CR needs to provide dynamic directional sound characteristics in real-time. In this study, sound capturing hardware and algorithms were developed to create a CR space. The microphone arrays and binauralization algorithms were compared using automatic speech recognition (ASR) to evaluate the performance of speech transmission.

2. Binauralization

Using a multi-channel microphone array, spatial sound was recorded to create an immersive sound landscape. Binauralization algorithms were needed to reproduce the sound in a 2-channel speaker system or headphones.

2.1. Spherical microphone array

A 32-channel spherical microphone array (Eigenmike, mh acoustics) was used to capture the spatial sound with directivity and source movement. Also, a binauralization algorithm for use with a spherical microphone was developed (Figure 1).

First, speech sounds were captured using the microphone array. A spatial Fourier transform was applied to convert the frequency domain into the spherical-wave spectrum domain, and the sound field was decomposed into spherical harmonics during sound field synthesis. In the spherical microphone array, the decomposed plane wave describes the directions of the plane waves on the sphere.

For binaural synthesis, minimum variance distortionless response (MVDR) adaptive beamforming was

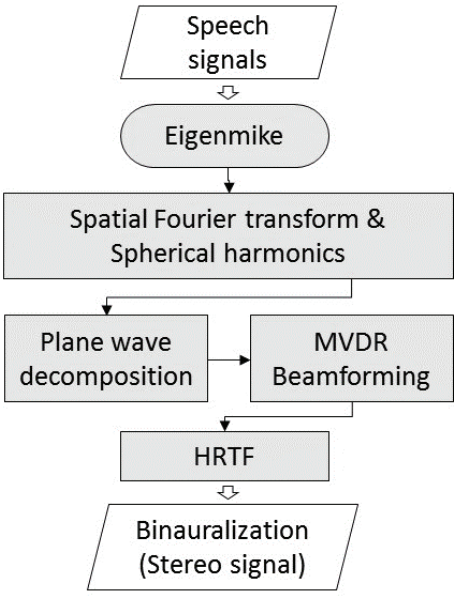

Figure 1. Binauralization of the spherical microphone array

applied with a figure-eight beam pattern to focus towards the ear. Finally, a Fabian head-related transfer function (HRTF) was used to process spatial stereo signals.

2.2. 6-ch 3D microphone array

A 6-channel 3D microphone array was developed to capture dynamic and directional sounds. Compared with the spherical microphone array, it has a lower resolution but has the advantage of real-time processing, which is important for teleconferencing. The microphone units used were electret microphones (First: WM-61A, Panasonic; Second: mke2-3, Sennheiser) and the 6 channels were aligned in pairs in the x, y, and z directions 15 cm. The distance was minimized to distinguish the directivity over the 1 kHz signals. Similar to the spherical microphone, the algorithm for binauralization was developed, and the procedure is shown in Figure 2.

The captured sounds are calculated to find the directions of the source using direction of arrival (DOA) estimation based on a cross-correlation function. On the basis of the DOA estimation results, the directivity was applied to the HRTF, which then represented the spatial stereo signals.

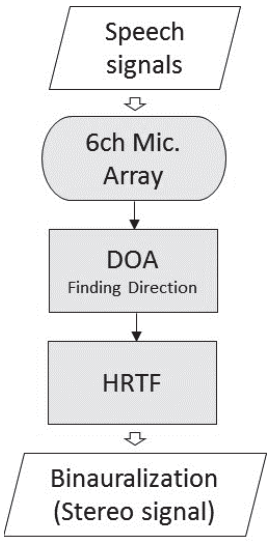

Figure 2. Binauralization of the 6-ch microphone array

3. Automatic speech recognition (ASR)

To compare microphone systems with respect to speech recognition, four types of microphones were selected, including omni (C414, AKG) and binaural (4128C, B&K) microphones. The speech source level was 60 dBA at a distance of 1 m away in the test room. Pink noise was presented at different noise levels from 45 dB to 65 to set up signal-to-noise ratios of 30, 15, 5, 0, -5 dB. In this case, 30 dB was the signal-to-noise ratio in the test room without any noise.

A configuration of 189 words in 50 sentences with 2 to 9 words per sentence was used. The total duration of each session was 2 minutes and 45 seconds. A total of 20 sessions was carried out with four types of microphone systems and five different signal-to-noise ratios.

3.1. Hidden Markov Model Toolkit

For the evaluation of speech recognition, HTK (Hidden Markov Model Toolkit, version 3.4) [1] was used. The training set used was Julius MFCC from Voxforge [2]. The word accuracy was calculated to represent the results instead of the word error rate.

As an example, if the input is “dial four three two”, and the response was “dial fours three a”, “dial” is a correctly recognized word, represented by H, and “fours” is a substitution, represented by S. In addition, an “a” was inserted (I), and “two” was deleted (D). The total number of words is represented by N. %Correct is defined as H/N, and Accurate Words (W_{Acc}) is defined as (H-I)/N. On the basis of this equation, the accurately recognized words are evaluated.

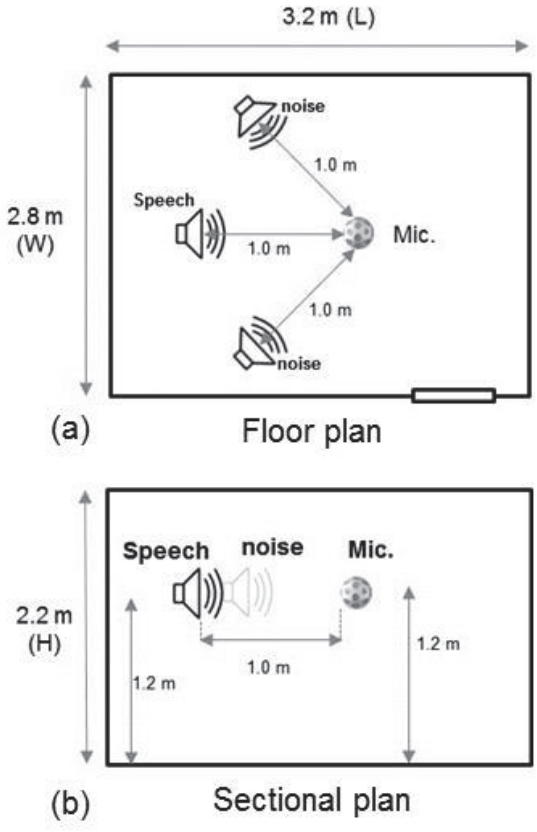

Figure 3. Measurement setup

3.2. Measurement setup

Figure 3 shows the measurement setup and conditions. The measured room volume was approximately 20 m^3 with a size of approximately 3 m length, width, and 2 m height. The room is typically used for listening tests, and the finishing materials are mainly absorptive materials; thus, the reverberation time was quite low (0.2 s).

The speech source was located at the center of the room, and the noise sources were located to the sides at the same distance to the receiver (Figure 3a). The speaker was a directional speaker (Genelec 8020 CPM), and the height was 1.2 m, at the same height as the receiver (Figure 3b). The background noise level was 30 dBA during stable conditions.

3.3. Results

The results of the word recognition rate are shown in Figure 4. Overall, the signal-to-noise ratio has a significant effect on word recognition.

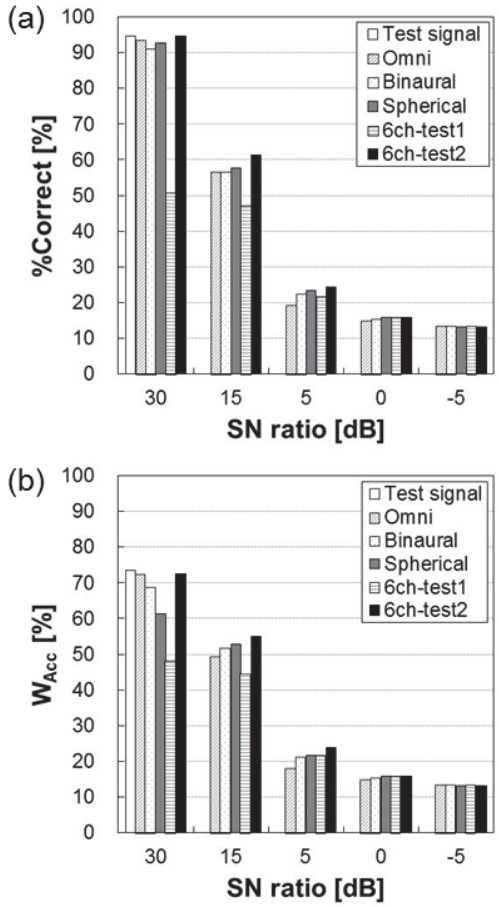

Figure 4. Results of speech recognition
(a) Correct words: H/N (b)Accurate words: (H-I)/N

Therefore, for a higher accuracy, a noise-reduction filter needs to be developed for the capturing system.

In a quiet environment, the omni, binaural, and spherical microphones have similar % of correct values as reference test signals. However, the spherical microphone showed somewhat lower values for accurate words. It seems when the algorithm resynthesized sound, there were some distortions in the speech signals, causing some inserted words to be detected.

The prototype 6-channel microphone has noise itself; thus, it exhibited lower word accuracy. This is because it showed a similar value at a signal-to-noise ratio of 15 dB. Therefore, the 6-channel microphone array needs hardware quality improvements.

In noisy conditions at a 15 dB signal-to-noise ratio, the spherical microphone displayed a higher word accuracy than the word accuracy of other microphones. Further, the omni microphone showed lower recognition rates in noisy environments. It is noted that the binauralized signal from a large number of microphones can improve the detection of speech signals.

4. Summary

In this study, ASR was applied to design a microphone array.

Our results showed that word accuracy significantly depended on the signal-to-noise ratio. Therefore, a noise-reduction filter needs to be developed for any CR teleconferencing system. The microphone array showed higher word accuracy as a speech transmission tool in noisy environments.

In the future, a 6-ch microphone array with higher word accuracy will be developed and evaluated with a moving source. In addition, it is planned to conduct the speech recognition test using Korean words.

Acknowledgement

This work was supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government (MSIP) (2013M3A6A3079356).

References

[1] S. Young et al.: The HTK book. Vol. 2. Cambridge: Entropic Cambridge Research Laboratory, 1997.

[2] VoxForge, Free Speech Recognition, www.voxforge.org