POSITION ESTIMATION USING A MICROPHONE AND STEREO LOUDSPEAKER WITH AN AUDIO-EMBEDDED HIDDEN TIME SYNCHRONIZATION SIGNAL

TaeJin Park and Kyeong Ok Kang

Electronics and Telecommunications Research Institute, Daejeon, Korea

ABSTRACT

We proposed a method for estimating the position of mono channel microphones using the sound from loudspeakers with hidden time synchronization waveforms in the audio signal. Unlike previous studies based on the SS (Spread Spectrum) technique, we used OFDM (Orthogonal Frequency Division Multiplexing) time synchronization signal embedded in the sound, maintaining the inaudibility of the embedded signal. While the microphone detects the time synchronization waveform, users remain unaware of the hidden time synchronization signal. With this hidden signal, we can successfully obtain the location of the microphone while the music is playing. We have anticipated various applications of our proposed method such as a configuration of multichannel audio systems or spotting target locations for private listening zones generated by a speaker array.

Index Terms— position estimation, microphone position**,** OFDM, time synchronization

1. INTRODUCTION

Various types of newly developed audio systems such as multichannel compatible sound-bars or private listening zone systems with multiple loud speakers [1] have recently been released. Since these sound systems have a targeted sweet spot, to set up the best sound playback environment, the playback system should control the delay and gain of each loudspeaker channel according to the listener's position. These demands for a listener positioning estimation in multichannel audio systems have been solved by placing a microphone in the listening position while the audio system is playing a particular measuring signal for the time delay estimation. While these position-estimating schemes work quite well, they usually employ a configuration mode for estimating the position, which means the user should consciously turn on the tuning mode of the speaker system. However, a number of general consumers might view this type of particular position estimating process as a complicated or troublesome chore. We therefore propose a scheme for unconsciously estimating the position of the listener rather than operating a particular configuration mode with a particular measuring signal. To accomplish this, the inaudibility of the additional signal used for position estimation is an important issue of our research. In this paper, we assumed that the microphone-equipped controllers are synchronized with the main audio system using two loudspeakers connected with a cable or wireless network. The overall signal flow of our proposed method is described in Fig. 1.

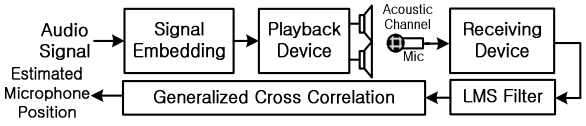


Fig. 1. Proposed position estimation system

Only a few attempts have thus far been made for position estimation using sound from the loudspeakers and a mono microphone. An attempt to apply an SS data-hiding method to an indoor navigation system was made [2]. However, the position estimation focused solely on distinguishing the nearest loudspeaker based on the strength of the watermark used. A similar study for finding the position of a camcorder used to record a movie in a theater was conducted by Nakashima et al. [3] based on the use of an SS watermarking soundtrack. As this work is focused on a movie theater location, the channel domination effect caused by the nearest loud speaker is not a crucial issue since the microphone position is far from each loud speaker. However, when it comes to a smaller room with a home theater system, the user may occasionally approach the loudspeaker, and one speaker signal will dominate another. To overcome this problem, we applied a totally different approach to solve the channel domination problem. Without using the SS technique, the OFDM time synchronization signal investigated in [4] was engaged to assure the orthogonality of the channel signal. In addition, our proposed scheme assures not only a shorter detection time but also an accurate mean estimation error compared to previous methods.

2. GENERATING A TIME SYNCHRONIZATION SIGNAL

2.1. Properties of a time synchronization signal

Since we applied the TDOA (Time Difference of Arrival) method for position estimation, the time synchronization issue is crucial for our problem. By applying a time synchronization signal in the pilot subcarrier, we can build a time synchronization signal by manipulating the phase of the audio signal. To determine the exact synchronization point, the received microphone signal should be cross-correlated by the original time synchronization signal. A greater gap between the maximum and second-largest cross-correlation values assures more robustness during the time synchronization in a noisy environment. The most essential property for time synchronization is therefore the PSPR (Peak to Side-Peak Ratio) of the time synchronization signal. We thus set the PSPR value as the key factor in optimizing the performance of the time synchronization signal. The PSPR and cross correlation are described in (1).

$$r(\tau) = \mathrm{Re}\left[\sum_{n=0}^{N-1} s(n)s^*(n+\tau)\right]$$
$$\mathrm{PSPR} = \frac{r(0)}{\max_{\tau \neq 0} |r(\tau)|} \tag{1}$$

2.2. Optimization scheme for a time synchronization signal

As mentioned in the introduction, if a microphone approaches a loudspeaker, the signal from the nearest speaker will dominate the signals from the other speakers. To mitigate this effect, two time synchronization signals, $U_L(k)$ and $U_R(k)$, are exclusively placed in the frequency domain to maintain orthogonally, as described in (2). The orthogonality of the time synchronization signal therefore ensures a lesser effect from the dominant loudspeaker.

$$U_L(k) = \sum_{n=0}^{N-1}\left(\left(\sum_{l\in L}\sin(\omega_l n + \Theta_l)\right)e^{-j\frac{2\pi kn}{N}}\right)$$
$$U_R(k) = \sum_{n=0}^{N-1}\left(\left(\sum_{r\in R}\sin(\omega_r n + \Theta_r)\right)e^{-j\frac{2\pi kn}{N}}\right) \tag{2}$$
$$\sum_{k=0}^{N-1} U_L(k)U_R(k) = 0$$

When placing pilot subcarriers for time synchronization in the signal, the properties of the time synchronization signal should be considered. Since an analytical treatment of the pilot subcarrier location problem is impossible, we employed a previous genetic algorithm in [4] and adapted it for our problem by changing its merit function and chromosome structure. The criterion of the chromosome selection, which is called the merit function, was set as in (3), where merit value **J** for the selection of the PSPR criterion consists of a sum of the PSPR values and the inverse of the PSPR magnitude difference between two channels. And α is the weighting coefficient balancing PSPR disparity and sum of the PSPR values.

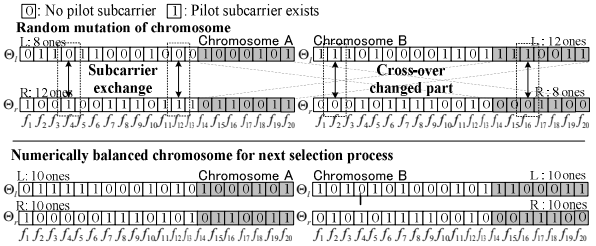


Fig. 2. Example of the proposed genetic algorithm process

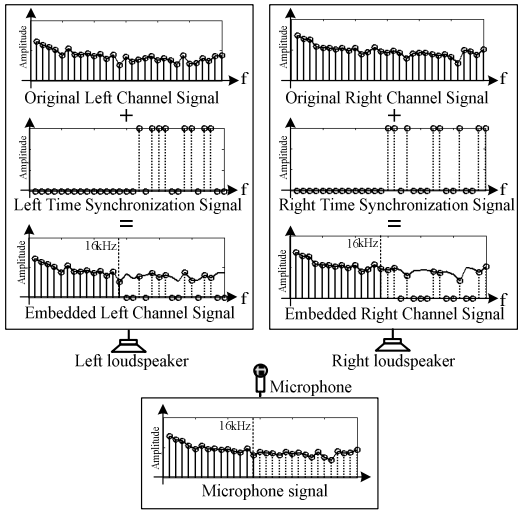


Fig. 3. Example of the proposed signal embedding scheme

$$\mathbf{J} = \alpha(PSPR_L + PSPR_R) + (1-\alpha)\frac{1}{|PSPR_L - PSPR_R|} \tag{3}$$

However, in terms of chromosome structure, the left- and right-channel chromosome set for a genetic algorithm process is coupled with the exclusive frequency, as described in Fig. 2. As shown in Fig. 2, the number of subcarriers in each left and right channel is maintained through a random mutation process. After this random mutation process, we calculate merit value of every chromosome set and preserve the chromosome sets with higher merit values. Through a repetition of these processes, we can therefore obtain the time synchronization waveform created by the pilot subcarrier using an optimized frequency placement.

2.3. Embedding of a time synchronization signal in the frequency domain

Since the time synchronization signal generated using the above method is clearly audible to the human ear, we adjusted the magnitude of the time synchronization signal in the frequency domain to the same magnitude as the original signal in the frequency domain, preserving the phase of the generated time synchronization signal to over 16 kHz, as described in Fig. 3. Also, we applied a window to every frame in the time domain to remove the frame transition noise. Although the windowing process slightly influenced the orthogonally of the two signals, it helped preserve the inaudibility of the embedded time synchronization signal since the listener might not sense the omission of the frequency component owing to cross-talk between the two channels and the audible frequency limitations of the human ear.

3. SIGNAL POSTPROCESSING AND DETECTION

3.1. Channel interference mitigation

As described in section 2, a single mono microphone receives an audio signal from both loudspeakers. Therefore, a robust channel separation scheme is needed to effectively reduce the channel interference. For this purpose, we previously mentioned the orthogonality of two signals. However, owing to the environmental noise and the characteristic response of a loudspeaker and microphone, the orthogonality of the two signals experiences some interference. To overcome this problem, we applied an adaptive filter [5] to estimate the redundant signals caused by environmental noise or the non-linearity of the loudspeaker and microphone. As described in (4) and (5), as well as in Fig. 4, we estimated the noise terms $n_L(n)$ and $n_R(n)$ of each channel from the microphone input using original time synchronization signal $\mathbf{u}_L(n)$ and $\mathbf{u}_R(n)$.

$$\begin{aligned} s_L(n) &= \frac{1}{N}\sum_{k\in L} S(k)e^{j\frac{2\pi kn}{N}} \\ s_R(n) &= \frac{1}{N}\sum_{k\in R} S(k)e^{j\frac{2\pi kn}{N}} \end{aligned} \tag{4}$$

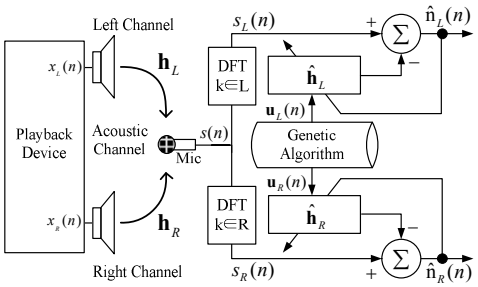


Fig. 4. Overview of the signal embedding scheme

$$\begin{aligned} s(n) &= s_L(n) + s_R(n) = \left(x_L(n) + n_L(n)\right) + \left(x_R(n) + n_R(n)\right) \\ \hat{x}_L(n) &= \hat{\mathbf{h}}_L{}^H(n)\mathbf{u}_L(n) \\ \hat{n}_L(n) &= s_L(n) - \hat{x}_L(n) \\ \hat{\mathbf{h}}_L(n+1) &= \hat{\mathbf{h}}_L(n) + \mu\mathbf{u}_L(n)\hat{n}_L{}^*(n) \\ \hat{x}_R(n) &= \hat{\mathbf{h}}_\mathbf{R}{}^H(n)\mathbf{u}_R(n) \\ \hat{n}_R(n) &= s_R(n) - \hat{x}_R(n) \\ \hat{\mathbf{h}}_R(n+1) &= \hat{\mathbf{h}}_R(n) + \mu\mathbf{u}_R(n)\hat{n}_R{}^*(n) \end{aligned} \tag{5}$$

3.2. Generalized Cross Correlation

Unlike in a typical time-domain cross-correlation approach, the cross correlation is calculated in the frequency domain, enabling selective weighting for the frequency of the two signals, which is called GCC (Generalized Cross Correlation) proposed in [6]. The GCC is described as follows, where $\Phi(k)$ is frequency domain weighting filter. Also, $X_s(k)$ and $X_u(k)$ are the DFT of the microphone signal and original time synchronization signals, respectively.

$$\hat{r}(\tau) = \frac{1}{N}\sum_{k=1}^{N-1}\Phi(k)X_s(k)X_u^*(k)e^{j\omega_k\tau} \tag{6}$$

For weighting filters $\Phi(k)$ in GCC, we applied a spectral Wiener filter with PSD(Power Spectral Density) $P(k)$ since we estimated the residual noise from the adaptive filter, allowing us to use the power spectrum of residual noise as *a priori* knowledge for the Wiener filter. If we assume that this estimated residual noise $\hat{n}(n)$ from (5) is uncorrelated with the estimated audio embedded time synchronization signal $\hat{x}(n)$, the Wiener filter can be derived as follows.

$$\Phi(k) = \frac{P_{\hat{x}s}(k)}{P_s(k)} \approx \frac{P_{\hat{x}}(k)}{P_{\hat{x}}(k) + P_{\hat{n}}(k)} \tag{7}$$

Using spectral Wiener filter weighting, we also whiten the microphone signal to relieve the effect of the dominant frequency. Thus, the final generalized cross-correlation value of the time synchronization and microphone signal for each channel is derived as below.

$$\begin{aligned} \hat{r}_L(\tau) &= \frac{1}{N}\sum_{k\in L}\frac{P_{\hat{x}_L}(k)}{P_{\hat{x}_L}(k) + P_{\hat{n}_L}(k)}\frac{X_S(k)}{|X_S(k)|}X_{u_L}^*(k)e^{j\omega_k\tau} \\ \hat{r}_R(\tau) &= \frac{1}{N}\sum_{k\in R}\frac{P_{\hat{x}_R}(k)}{P_{\hat{x}_R}(k) + P_{\hat{n}_R}(k)}\frac{X_S(k)}{|X_S(k)|}X_{u_R}^*(k)e^{j\omega_k\tau} \end{aligned} \tag{8}$$

We therefore calculate the argument of the maximum GCC value as below.

$$\begin{aligned} n_\mathrm{L} &= \arg\max_\tau\left(\hat{r}_L(\tau)\right) \\ n_\mathrm{R} &= \arg\max_\tau\left(\hat{r}_R(\tau)\right) \end{aligned} \tag{9}$$

Using the sample shifting n of both signals, we can derive the distance from microphone to each loud speaker, which leads to the microphone position.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Position estimation experiment

A position estimation experiment was conducted in an ordinary room with a slight echo. We placed a microphone at the nine different positions shown in Fig. 5. Six different music files with three different genres (Classical, Pop, and Rock) were tested at three different volumes. The loudspeaker volumes were measured as 60, 70, and 80 dB, respectively, using an SPL (Sound Pressure Level) meter at position 8, shown in Fig 5, while playing amplitude-maximized white noise. The position estimation results are indicated in Table 1 based on the detection probability. All estimations were counted as successful for each frame when the detected microphone position was within 8 cm (ten sample errors at 44.1 kHz sampling rate) of the actual test position. The overall mean distance estimation error of our proposed work, 0.23 m.

4.2. Inaudibility test of an embedded signal

An inaudibility test was conducted using a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening test. The listening position was placed at position 8 in Fig. 5. Ten test subjects, both male and female, participated. The test volume was 70 dB, which was measured using an SPL meter at position 8 in Fig. 5, while playing amplitude-maximized white noise. We tested eight different music files from four genres (Classical, Pop, Rock, and Jazz) with mean dBFS(Decibel in Full Scale) in Fig.5 using both our proposed method and the realizable SS method proposed in [2]. The listening test results are described in Fig. 6. The test results show that the embedded signal using the proposed method and that using the SS method proposed in [2] with a gain parameter of 0.2 have a similar score with the hidden reference signal. In addition, neither the lower cutoff frequency (12 kHz) nor the higher gain parameter of the proposed method than those used in the SS data hiding method can guarantee inaudibility.

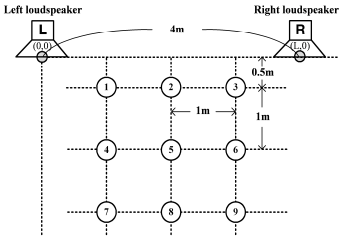
Table 1. Probability of successful distance estimation

Loc	60dB	70dB	80dB
1	0.85	0.98	1.00
2	0.07	0.54	0.96
3	0.15	0.53	0.76
4	0.97	0.99	1.00
5	0.66	0.97	1.00
6	0.50	0.87	0.99
7	0.91	0.99	1.00
8	0.73	0.97	1.00
9	0.80	0.95	0.97

Left Channel Probability

Loc	60dB	70dB	80dB
1	0.16	0.42	0.80
2	0.35	0.48	0.78
3	0.83	0.97	1.00
4	0.46	0.84	0.95
5	0.71	0.85	0.99
6	0.93	0.99	1.00
7	0.75	0.92	0.98
8	0.80	0.91	1.00
9	0.88	0.98	1.00

Right Channel Probability



Music genre	Mean dBFS
Classic	-22.8dB
Pop	-17.8dB
Rock	-17.2dB
Jazz	-18.4dB

Fig. 5 Test area dimensions and test sample specification

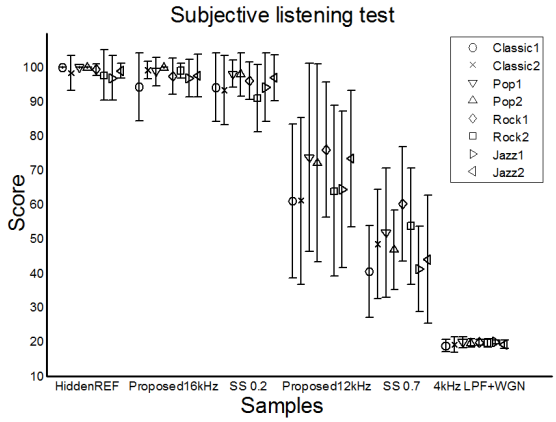


Fig. 6. Listening test results

5. CONCLUSION

In this paper, we described our proposed position estimation method using a mono microphone and loudspeaker signal. The position estimation results shown in chapter 5.1 indicate that the proposed position estimation scheme can estimate a microphone position accurately with a moderate loudspeaker volume. In addition, the MUSHRA listening test results showed the reasonable inaudibility of our embedded signal. Thus, if our proposed scheme is applied to a position estimating scheme for a multichannel audio system, the users are expected to remain unaware of the time synchronization signal while the embedded signal is emitted through the air along with an ordinary audio signal. Since the method proposed in this paper is available using two loudspeakers, further investigation using three loudspeakers or a speaker array will ensure more practical usage for various applications.

ACKNOWLEDGEMENT

This work was supported by the Broadcasting Technology R&D program of KCC/KCA. [11921-02001, Development of Multiview 3D Compatible UHDTV Broadcasting Technology]

REFERENCES

[1] Jung-Min Lee, Tae-Woong Lee, Jin-Young Park, and Yang-Hann Kim, "Generation of a private listening zone; acoustic parasol," *Proceedings of 20th, International Congress on Acoustics*, ICA 2010, Sydney, Australia, 23-27 August 2010.

[2] Nevena Lazic, and Parham Aarabi, "Communication Over an Acoustic Channel Using Data Hiding Techniques," *Multimedia, IEEE Transactions on*, vol. 8, no. 5, pp. 918-924, October 2006.

[3] Yuta Nakashima, Ryuki Tachibana, and Noboru Babaguchi, "Watermarked Movie Soundtrack Finds the Position of the Camcorder in a Theater," *Multimedia, IEEE Transactions on*, vol. 11, no. 3, pp. 443-454, April 2009.

[4] Oktay Ureten and Selcuk Tascıoglu, "Autocorrelation Properties of OFDM Timing Synchronization Waveforms Employing Pilot Subcarriers," *EURASIP Journal on Wireless Communications and Networking*, Article No. 10, January 2009.

[5] Simon Haykin, *"Adaptive Filter Theory,"* 4th Ed., Prentice Hall: Upper Saddle River NJ, 2002, p. 236.

[6] Charles H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320-327, August 1976.