Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

A Design of Acoustic Security System in Near Field based on Paired Microphones and Automatic Video Camera

Kohei Hayashida (1), Masanori Morise (2) and Takanobu Nishiura (2)

(1) Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Japan
(2) College of Information and Science, Ritsumeikan University, Kusatsu, Japan

PACS: 43.60.JN

ABSTRACT

Many conventional security systems use visual information from a video camera. However, these systems may not be able to acquire the important scenes in the blind area of the video camera. Acoustic security systems can support conventional visual security systems with acoustic events detection. In our research, we focused on acoustic security systems in the near field, and we designed a prototype of a three-dimensional acoustic security system using paired microphones and an automatic video camera. This system detects a sound event and automatically focuses the video camera on the detected sound source in real time. Conventional automatic video camera steering systems localize a sound source assuming it is in the far field. Moreover, they detect only speech because this type of system was developed for videoconferencing. However, detection errors increase in the near field. Thus, we designed a prototype acoustic security system that robustly detects a sound event, localizes the sound source in the near field, and then automatically steers the video camera to the detected sound event in real time. We carried out evaluation experiment in a real office environment. An evaluation experiment in a real office environment showed that the proposed security system could robustly detect a sound event and quickly steer a video camera toward it.

INTRODUCTION

Many conventional security systems use visual information from video cameras. However, these systems may not be able to acquire important scenes in the blind area of the video camera. On the other hand, acoustic security systems can detect sound events with sensing microphones. These systems can be used to support conventional visual security systems with acoustic events detection. In our research, we focused on acoustic security systems in the near field, and we designed a prototype of a three-dimensional acoustic security system in the near field based on paired microphones and an automatic video camera. This system detects a sound event and automatically focuses the video camera on the detected sound source in real time.

An automatic video camera system steered using real-time sound source localization for videoconferencing has been developed by H. Wang [1]. This system localizes a sound source based on the direction estimation method [2, 3], and utilizes different microphone pairs to estimate a horizontal angle, an elevation angle, and a distance in far field assumption. However, the localization error increases with it in the near field. In addition, the target signal in this system is speech, because the system was developed for videoconferencing. In this research, we designed a prototype acoustic security system in the near field, which robustly detects a sound event, localizes the sound source, and then automatically steers the video camera to the detected sound event in real time. The proposed acoustic security system detects the sound event based on a hidden Markov model (HMM) [4]. Moreover, a sound source localization method under the near-field assumption has previously been developed [5]. In this research, we introduce this localization method to a proposed system for robust localization in the near field.

CONSTRUCTION OF PROPOSED ACOUSTIC SECURITY SYSTEM

The proposed acoustic security system consists of a microphone array and an automatic video camera. Figure 1 shows a picture of the proposed system, and Fig. 2 shows the placement of the microphone array. The proposed system uses a three-microphone array 0.18 [m] wide and 0.11 [m] tall. Figure 3 shows a processing flow of the proposed system. First, the proposed system captures audio signals with the microphone array and identifies the captured sound. We used a hidden Markov model (HMM) [4] to identify the sound source. Next, if the captured signal is the target sound, the sound source is localized by three-dimensional cross-power spectrum phase analysis with multiple paired microphones (Multiple paired 3D-CSP) [5]. The three-dimensional sound source location, which consists of a horizon, an elevation, and a distance between the sound source and the microphone array, is estimated from the time delays of the three paired microphones. Finally, the video camera is automatically steered toward the detected sound source.

Figure 1. Proposed system.

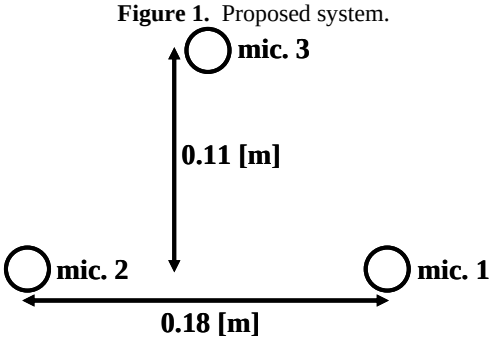

Figure 2. Microphone array configuration.

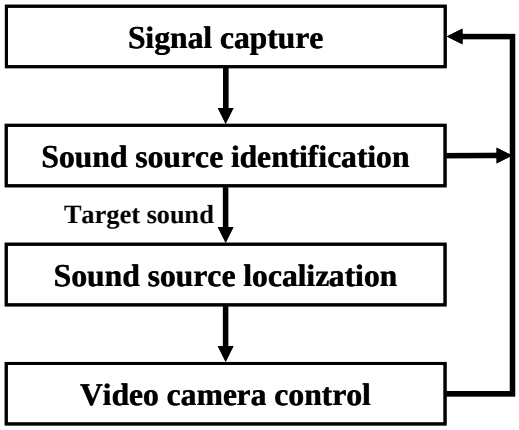

Figure 3. Processing flow of the proposed system.

Sound source identification based on HMM

The hidden Markov model (HMM) [4] is a time sequence signal probability model. The HMM represents a non-stationary time sequence signal by changing between plural stationary time sequence signals. It is generally used as an acoustic model. We used the HMM for sound source identification. Moreover, we utilized several features for sound source identification: mel-frequency cepstrum coefficient (MFCC), delta MFCC, and delta power. In this research, we defined a clapping sound as the target sound and speech as the unnecessary sound. We used the sound scene database in a real acoustic environment that is recorded by Real World Computing Partnership (RWCP-DB) [6] to build the model of the clapping sound and the Advanced Telecommunications Research Institute International (ATR) phoneme balance of 216 words [7] to build the model of speech. We built the HMM with the Hidden Markov Model Toolkit (HTK) [8].

Localization method based on CSP in the near field

In this research, we used three-dimensional cross-power spectrum phase analysis with multiple paired microphones (Multiple paired 3D-CSP) [5] for robust near-field sound source localization.

CSP for sound source direction estimation in far field assumption

Cross-power spectrum phase analysis (CSP) [2, 3] is a powerful direction of arrival (DOA) estimation technique under the far field assumption that does not depend on the frequency characteristics of the sound sources. It estimates the CSP coefficients and the time delay of arrival (TDOA) on the basis of captured signals of a paired microphone. The CSP coefficients $CSP(k)$ and the TDOA $\hat{\tau}$ are derived from Eqs. (1) and (2).

$$CSP(n) = DFT^{-1}\left(\frac{DFT(x_i(t))DFT(x_j(t))^*}{|DFT(x_i(t))||DFT(x_j(t))|}\right), \quad (1)$$

$$\hat{\tau} = \arg\max_n CSP(n). \quad (2)$$

The symbols $x_i(t)$ and $x_j(t)$ denote the captured signals with a paired microphone, DFT the discrete Fourier transform, and DFT^{-1} the inversed discrete Fourier transform. The CSP method first calculates the discrete Fourier transform of captured signals $x_i(t)$ and $x_j(t)$ with a paired microphone, then calculates phase differences on the basis of an amplitude normalization, and it finally acquires the CSP coefficients with inversed discrete Fourier transform, as shown in Eq. (1). On the other hand, the TDOA is acquired by utilizing the time lag on the basis of the maximum CSP coefficient, as shown in Eq. (2). The DOA is derived from Eq. (3).

$$\theta = \cos^{-1}\left(\frac{\hat{\tau}c}{d}\right). \quad (3)$$

The symbol d denotes the interval between two microphones, and c the sound propagation velocity. The CSP has a maximum value of the CSP coefficient in a DOA with the dominant power. The CSP method calculates the DOAs on the basis of the plane wave.

2D-CSP with multiple paired microphones for sound source localization under the near field assumption

On the basis of CSP, two-dimensional cross-power spectrum phase analysis with multiple paired microphones (multiple paired 2D-CSP) [5], an expanded version of conventional CSP, has previously been developed. It estimates two-dimensional sound source locations which consist of a horizon and a distance under the near-field assumption. The multiple paired 2D-CSP first estimates the CSP coefficients and the TDOA $\hat{\tau}_l$ derived from Eqs. (1) and (2). These are exactly the same steps as those of the CSP in the previous section. In addition, the TDOA $\tau_l(S_x, S_y)$ from sound source (S_x, S_y) to paired microphone l is derived from Eq. (4) on the basis of the spherical wave, with a sound propagation time $\tau_{M_i}(S_x, S_y)$ based on Eq. (5).

$$\tau_l(S_x, S_y) = \tau_{M_i}(S_x, S_y) - \tau_{M_j}(S_x, S_y), \quad (4)$$

$$\tau_{M_i}(S_x, S_y) = \sqrt{(S_x - S_{M_{ix}})^2 + (S_y - S_{M_{yi}})^2}\,/c. \quad (5)$$

The symbol l denotes a paired microphone that consists of microphone M_i and M_j, (M_{ix}, M_{iy}) denotes a coordinate of the microphone M_i. The multiple paired 2D-CSP localizes a sound source utilizing the TDOA $\hat{\tau}_l$ derived from Eqs. (1) and (2) and the TDOA $\tau_l(S_x, S_y)$ derived from Eq. (4). The

spatial coefficient $P(S_x, S_y)$ for multiple paired 2D-CSP is derived from Eq. (6).

$$P(S_x, S_y) = \sum_{l=1}^{N} \left(\tau_l(S_x, S_y) - \hat{\tau}_l\right)^2. \quad (6)$$

The symbol N denotes the number of paired microphones. If $\hat{\tau}_l$ accords with $\tau_l(S_x, S_y)$ in Eq. (6), $P(S_x, S_y)$ goes to 0. When $P(S_x, S_y)$ is at its least value, (S_x, S_y) is estimated as the sound source location.

3D-CSP with multiple paired microphones for sound source localization under the near-field assumption

On the basis of CSP, three-dimensional cross-power spectrum phase analysis with multiple paired microphones (multiple paired 3D-CSP) [5], an expanded version of multiple paired 2D-CSP, has previously been developed. It estimates three-dimensional sound source locations, which consist of a horizon, an elevation and a distance under the near-field assumption. The multiple paired 3D-CSP first estimates the CSP coefficients and the TDOA $\hat{\tau}_l$ derived from Eqs. (1) and (2). In addition, the TDOA $\tau_l(S_x, S_y, S_z)$ from sound source (S_x, S_y, S_z) to paired microphone l is derived from Eq. (7) on the basis of the spherical wave, with a sound propagation time $\tau_{M_i}(S_x, S_y, S_z)$ based on Eq. (8).

$$\tau_l(S_x, S_y, S_z) = \tau_{M_i}(S_x, S_y, S_z) - \tau_{M_j}(S_x, S_y, S_z), \quad (7)$$

$$\tau_{M_i}(S_x, S_y, S_z) = \sqrt{(S_x - S_{M_{ix}})^2 + (S_y - S_{M_{yi}})^2 + (S_z - S_{M_{iz}})^2} \Big/ c. \quad (8)$$

The symbol (M_{ix}, M_{iy}, M_{iz}) denotes a coordinate of the microphone M_i. The multiple paired 3D-CSP localizes a sound source utilizing the TDOA $\hat{\tau}_l$ derived from Eqs. (1) and (2) and the TDOA $\tau_l(S_x, S_y, S_z)$ derived from Eq. (7). The spatial coefficient $P(S_x, S_y, S_z)$ for multiple paired 3D-CSP is derived from Eq. (9).

$$P(S_x, S_y, S_z) = \sum_{l=1}^{N} \left(\tau_l(S_x, S_y, S_z) - \hat{\tau}_l\right)^2. \quad (9)$$

If $\hat{\tau}_l$ accords with $\tau_l(S_x, S_y, S_z)$ in Eq. (9), $P(S_x, S_y, S_z)$ goes to 0. When $P(S_x, S_y, S_z)$ is at its least value, (S_x, S_y, S_z) is estimated as a sound source location.

EVALUATION EXPERIMENT

Experimental conditions

We carried out the evaluation experiment in a real office environment. The microphone array used for sound source localization is shown in Fig. 2. Table 1 shows the experimental conditions. We captured the audio signals by 16 [kHz] sampling frequency and 16 [bit] quantization. Figure 4 shows the placement of the talker and the proposed system. The talkers were positioned at horizontal angles of 0 and 45 [deg.], elevation angles of 0 and 20 [deg.], and distances of 1.0 and 2.0 [m]. A total of six talkers (one female and five males) participated in this experiment. The participants said Japanese words or clapped their hands at each position. The fan noise of personal computer (PC) servers was used as the stationary noise. In this research, we defined a clapping sound as the target sound and speech as the unnecessary sound. Thus, the proposed system detects a clapping sound and steers the video camera to it. Moreover, the proposed system ignores speech and does not steer the video camera toward speech sounds. The localization performance was evaluated on whether the sound source (participant's hands) appears in the video camera image or not. The sound identification performance was evaluated by the detection rate of the target sound (clapping sound) and the rejection rate of the unnecessary sound (speech).

Experimental results

Figure 5 shows the experimental results that involve the localization rate, the detection rate of the target sound (clap sound) and the rejection rate of the unnecessary sound (speech). In Fig. 5, the localization rate was 83.3%. The detection rate of the target sound was 91.7%, and the rejection rate of the unnecessary sound was 85.4%. In addition, the sound source identification and the sound source localization were finished in real time using a commercially available laptop PC. The experiment demonstrated that the proposed system could detect a sound event and quickly steer the video camera to it. In addition, Fig. 6 shows a camera image before the camera is controlled to respond to a sound. Figure 7 shows a camera image after the camera has been steered toward a sound and Fig. 8 shows a camera image after the camera has been steered and focused. In Figs. 6-8, we confirmed that the video camera focused on the participant's hand after the participant clapped his/her hands. Therefore, the prototype acoustic security system in the near field using paired microphones and automatic video camera was demonstrated to work.

DISCUSSIONS

The sound source localization mostly failed in Positions 2 and 4 (Fig. 4) where the distance between the microphone array and the sound source was 2.0 [m]. We consider reduced SNR due to room reverberation to be a cause of the localization error. In addition, multiple paired 3D-CSP localized the sound source with the intersection of TDOAs on the basis of each paired microphone. Thus, the localization accuracy depends on the estimation accuracy for the time delay between paired microphones. In other words, the localization accuracy of multiple paired 3D-CSP depends on

Table 1. Experimental conditions.

Audio conditions	
Environment	Office room
Sampling frequency	16 [kHz]
Quantization	16 [bit]
Room reverberation ($T_{[60]}$)	0.45 [sec]
Ambient noise	47.0 [dBA]
Stationary noise	PC server noise
Sound source identification conditions	
Feature vector	33 orders 16 orders MFCC + 16 orders ΔMFCC + 1 order ΔPower
Number of states	8
Number of mixtures	128
Decoder	Julius [9]
Localization conditions	
Horizontal angle	0, 45 [deg.]
Elevation angle	0, 20 [deg.]
Distance	1.0, 2.0 [m]

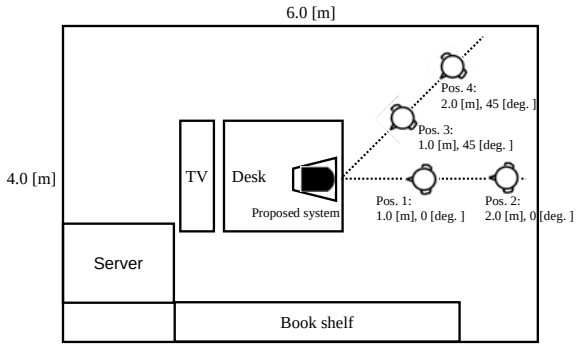

Figure 4. Overview of experimental environment.

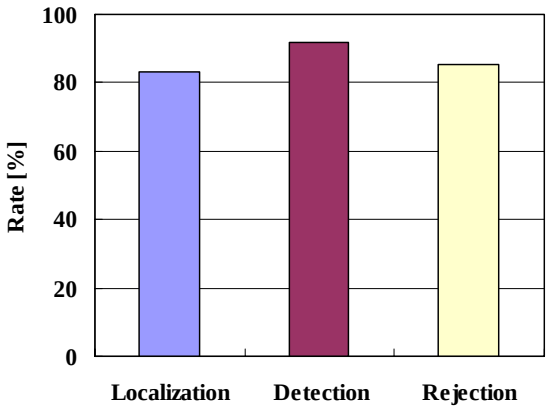

Figure 5. Results of experiment.

the sampling frequency and the interval between microphones. Therefore, we need to consider the arrangement of microphones and make the localization method more robust to improve the localization accuracy in the near field. The proposed system can identify the target sound and reject the unnecessary sound with over 85% accuracy. However, this performance is not sufficient for sound event detection. Moreover, reduced sound identification accuracy is expected when there are more target sounds. Therefore, we need to consider the optimum parameters for the detection of sound events.

CONCLUSIONS

In this research, we designed a prototype of a three-dimensional acoustic security system in the near field using paired microphones and an automatic video camera. Our evaluation experiment demonstrated that the system could respond to a sound event and automatically focus the video camera on a sound source in real time. In future work, we will try to consider the arrangement of the microphone array, improve localization accuracy, and make the system capable of identifying various sound events.

ACKNOWLEDGEMENTS

This work was partly supported by Global-COE and Grants-in-Aid for Scientific Research funded by The Japanese Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

1 H. Wang, *et. al.*, "Voice source localization for automatic camera pointing system in videoconferencing," *ICASSP '97*, pp. 187-190, 1997.

Figure 6. The video camera image before the control.

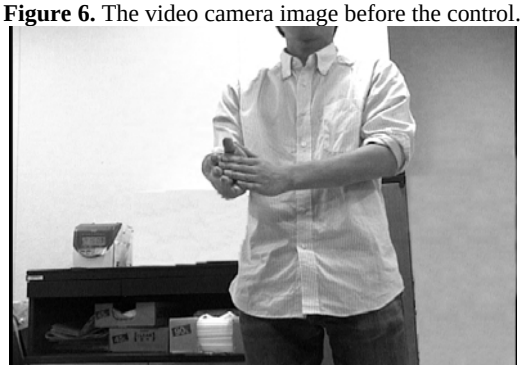

Figure 7. Video camera image after steering.

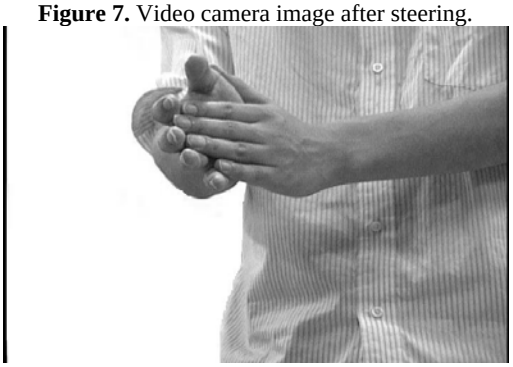

Figure 8. The video camera image after steering and focusing.

2 C. H. Knapp, *et. al.*, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust.* , vol. ASSP-24, no. 4, pp. 320-327, 1976.
3 M. Omologo and P.Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. ICASSP94*, pp. 273-276, 1994.
4 X. D. Huang, *et. al.*, "Hidden markov models for speech recognition," *Edinburgh Univ. Press*, 1990.
5 D. Rabinkin, *et. al.*, "A DSP implementation of source location using microphone arrays," *Proceedings of the SPIE, Vol. 2846*, pp. 88-99, 1996.
6 Satoshi Nakamura, *et. al.*, "Design and status of sound scene database in real acoustical environments," *Proc. of Acoustical Society of Japan*, pp. 137-138, Sep. 1998.
7 K. Takeda, *et. al.*, "Acoustic Phonetic Labels in a Japaese Speech Database," *Proc. European Conference on Speech Technology, vol. 2*, pp. 13-16, 1987.
8 HTK Web site, http://htk.eng.cam.ac.uk/
9 Akinobu Lee, *et. al.*, "Julius --- an Open Source Real-Time Large Vocabulary Recognition Engine," *Proc. EUROSPEECH*, pp. 1691-1694 , 2001.