A system for detecting unusual sounds from sound environment observed by microphone arrays

Mitsuru Kawamoto\textsuperscript{1,2}, Futoshi Asano\textsuperscript{3}, Koichi Kurumatani\textsuperscript{1,2}  
National Institute of Advanced Industrial Science and Technology (AIST)  
\textsuperscript{1,2}2-41-6 Koto-ku, Tokyo 135-0064, JAPAN  
\textsuperscript{3}Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki, 305-8568, JAPAN  
Emails: \{m.kawamoto, f.asano, k.kurumatani\}@aist.go.jp  
\textsuperscript{2}CREST, JST

Yingbo Hua  
Department of Electrical Engineering  
University of California  
Riverside, CA, 92521, USA  
Email: yhua@ee.ucr.edu

Abstract—In this paper, we propose a system that can detect unusual sounds and directions by observing sound environment with microphone arrays. One of the attractive features of the system is to detect the unusual information through daily environmental sound measurement. Therefore the system does not require such troublesome processes that the detected sounds must be predefined, the predefined sounds must be collected, and using the collected sounds, their features must be modeled, where the conventional systems have such troublesomeness. Moreover, unlike conventional systems using video cameras, our system is not limited by video camera angles. A simple experimental result shows the validity of the proposed system.

Keywords—Security and safety system; Unusual sounds; Sound direction estimation; Sound classification; Microphone array

I. INTRODUCTION

Sound localization (e.g., [1], [10]), sound detection (e.g., [2], [5]), sound classification (e.g., [6], [11]), and so on, are attractive tools for understanding and analyzing sound environments.

In this paper we combine these signal processing techniques to develop a system that provides "security and safety" for public spaces such as department stores, hospitals, and schools. Security and safety are provided by detecting unusual sounds.

In the conventional researches on security and surveillance systems, methods for detecting unusual sounds can be roughly divided into two types: One is such a method that unusual sounds defined in advance are detected (e.g., [9], [12]). The other is such a method that unusual sounds measuring a distance to a model of normal situations are detected (e.g., [3], [4]). The former type of methods need to model the features of the predefined unusual sounds using, e.g., GMM, and hence the method has such a troublesomeness that training data corresponding to unusual sounds defined in advance must be collected. In the latter types of methods, normal situations must be grasped and must be modeled using training data for the normal situation. However, it is not simple task to understand what is the normal situation in a given environment.

In contrast to the conventional works, the objective of our system is to detect unusual sounds, which are not predefined, from among the sounds measured in a given environment. Namely, the proposed system can detect the unusual information without passing through training process. Exploring this new approaches is the main contribution of this paper.

\begin{figure}[h]  
\centering  
\includegraphics[width=\textwidth]{diagram.png}  
\caption{Proposed security monitoring instrument}  
\end{figure}

Fig. 1 shows a block diagram of the proposed approach. To detect the unusual sound, first of all, a sound localization technique with a microphone array is used. In this paper, we propose a new algorithm of using a likelihood function as the sound localization method. Because a decoupling mechanism is embedded in the localization process of the proposed algorithm, we expect it to be simpler to localize

\textcopyright{} 2009 IEEE  
DOI 10.1109/IAS.2009.200  
978-0-7695-3744-3/09 $25.00
sources from mixtures of sounds generated from different sources, compared to other localization methods such as MUSIC [10]. Next, based on the localized information, that is, the maximum value of the likelihood function \( \max_{ij} L_{st}( \theta_{ij} ) \) (see Fig. 1), we detect environment sounds and then classify the sounds using their powers. Moreover, feature extraction techniques, e.g., MFCC, are applied to analyze the sounds categorized for each class. Finally, for each class, we determine whether the sounds are normal or unusual, using a method of measuring the similarity of the extracted features.

The proposed system combines existing sound processing techniques to create a new functionality which has not been available previously and which provides security and safety by audio surveillance. In addition, in contrast to video surveillance systems, the proposed system can be used to provide security and safety even when sound sources (and thus the sources of hazards) are not included in the video camera angles.

We report the results of experiments using data recorded in one environment.

II. PROPOSED METHOD

A. Sound measurement

![Microphone array](image)

The microphone array shown in Fig. 2 is used to measure the sounds generated in public spaces, where the microphone array is made by referring to the vector sensor [7]. There are four cubes in the microphone array, and hence it is called Quad-Cube Microphone Array (QCMA). Each cube has four microphones (fig. 2 (b)), and hence the total number of microphones is 16.

Let \( y_i(t) \) denote the observed signal measured by each microphone of the QCMA, where the subscripts \( i = 1,2,3,4 \) of \( y_i(t) \) denote the observations measured by, respectively, \( o, x, y, \) and \( z \) shown in Fig. 2 (b), and the superscripts \( c = 1,2,3,4 \) denote the cube number. Then for each cube, the signals used as the observations are defined as

\[
\begin{align*}
y_1^c(t) &= v_1^c(t) \quad (1) \\
y_i^c(t) &= v_i^c(t) - v_1^c(t), i = 2,3,4. \quad (2)
\end{align*}
\]

That is, we use the differences between \( v_i^c(t) \) (\( i = 2,3,4 \)) and \( v_1^c(t) \) to detect sounds.

B. Sound Direction Estimation Algorithm

We detect environment sounds using the QCMA. To this end, at first, the sound localization is implemented based on the outputs \( y_i^c(t) \) (\( i,c = 1,2,3,4 \)) using the following process.

Hereafter, when the sound localization and detection are implemented, we analyze all signals in the frequency domain. The short-time Fourier transform (STFT) of the output is denoted as \( y(t, \omega) = [Y_1^c(t, \omega), Y_2^c(t, \omega), \ldots, Y_4^c(t, \omega)]^T \), where \( Y_i^c(t, \omega) \) is the STFT of the \( i \) th microphone input in the \( c \) th cube at time \( t \) and frequency \( \omega \). In the remainder of this paper we omit the index of frequency \( \omega \) for simplicity. The output vector can be modeled as

\[
y(t) = As(t) + n(t), \quad (3)
\]

where \( A \) is a location vector matrix defined as \( A = [a_1, \ldots, a_L] \), \( s(t) = [S_1(t), \ldots, S_L(t)]^T \) is a source spectrum vector, and \( n(t) = [N_1(t), \ldots, N_M(t)]^T \) is a background noise spectrum vector. Here, \( L \) is the number of active sound sources and \( M \) is the number of microphones. The noise is assumed to be zero mean Gaussian noise. We assume that the covariance matrices of \( s(t) \) and \( n(t) \) are defined as, respectively, \( E[s(t)s^H(t)] = K_s = \text{diag}(\gamma_1, \ldots, \gamma_L) \) and \( E[n(t)n^H(t)] = \sigma I \), where \( \text{diag}\{ \cdot \cdot \cdot \} \) denotes a diagonal matrix with the diagonal element \( \gamma_l (l = 1,2, \ldots, L) \) denote the power spectrums of \( S_l(t) \) (\( l = 1,2, \ldots, L \)), \( \sigma \) denotes the power of the noise \( n(t) \), and \( I \) denotes the identity matrix.

The output vector can be decomposed into the following equation:

\[
y(t) = \sum_{l=1}^{L} x_l(t) = H x(t), \quad (4)
\]

where \( x_l(t) = a(\theta_l)S_l(t) + n_l(t), x(t) = [x_1^T(t), \ldots, x_L^T(t)]^T \) is a \( ML \)-column vector, and \( H = [I, \ldots, I] \) is an \( M \times ML \) matrix. The symbol \( n_l(t) \) is an arbitrary decomposition of the noise vector satisfying \( \sum_{l=1}^{L} n_l(t) = n(t) \) and \( E[n_l(t)n_l^H(t)] = \frac{\sigma}{L} I \). In this paper, the localization of \( S_l(t) \)'s is implemented using the covariance matrix of \( x_l(t) \). Namely, we expect that since \( x_l(t) \) includes only one source \( S_l(t) \), the localization for each source might be...
simpler than the one obtained by using the covariance matrix of \( y(t) \), e.g., MUSIC. However, \( C_{xt} = \frac{1}{N} \sum_{n=1}^{N} x_i(n) x_i^T(n) \) is not calculated directly. Hence \( C_{xt} \) can be estimated by the following conditional expectation.

\[
C_{xt} := E[C_{xt}|C_y, \hat{K}_y] = \hat{K}_{xt} - \hat{K}_{xt}(\hat{K}_y)^{-1}\hat{K}_{xt}
\]

\[
\hat{K}_{xt} = \hat{\gamma}_t a_i a_i^H + \frac{1}{L} I,
\]

\[
\hat{K}_y = \sum_{t=1}^{L} \hat{K}_{xt},
\]

where \( \hat{\gamma}_t \) can be estimated by \( (a_i^H C_y a_i)|a_i|^4 \), \( C_y \) is a covariance matrix of \( y(t) \). Then, we calculate the following likelihood function for \( x_i(t) \) with respect to \( \theta_{ij} \), \( i, j = 1, \cdots, P \), where \( \theta_{ij} \) represents the 2D direction of the sound source and \( P \) is the number of virtual sources.

\[
L_{xt}(\theta_{ij}) = \exp \left(-\frac{1}{2} \text{tr} \left[ C_{xt}(\theta_{ij}) \hat{K}_{xt}^{-1}(\theta_{ij}) \right]\right).
\]

Using (8), the position \( \theta_{ij} \) can be found so that the likelihood function is maximized, and then if the value of \( L_{xt}(\hat{\theta}_{ij}) \) for the estimate position \( \hat{\theta}_{ij} \) is greater than a threshold, we determine that there is a source at the position \( \hat{\theta}_{ij} \), where the threshold is suitably chosen.

\section{Sound Classification}

Moreover, we classify the sound generated from a localized position, using its power, that is,

\[
P_y(t) = \sum_{\omega} 10 \log_{10} |Y_i(t, \omega)|^2,
\]

where \( |*|^2 \) denotes the square norm of *, and \( Y_i(t, \omega) \) is STFT of \( y_i(t) \) in (1). Fig. 3 shows the value of \( P_y(t) \) with time \( t \). We detect the absolute values of the differences between local minimums and local maximums, e.g., \( d_m \) in Fig. 3 and count the differences which belong to the following categories:

\[
\{ C1 : d_m > l_1 \}, \quad \{ C2 : l_1 \geq d_m > l_2 \}, \quad \{ C3 : l_2 \geq d_m > l_3 \}.
\]

In this paper, three types of sounds are classified as follows, depending on the counted number for each category.

<table>
<thead>
<tr>
<th>Class</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Class 2a</td>
<td>\geq 2</td>
<td>\geq 0</td>
<td>\geq 0</td>
</tr>
<tr>
<td>Class 2b</td>
<td>1</td>
<td>1</td>
<td>0, 1, 2</td>
</tr>
<tr>
<td>Class 3</td>
<td>0</td>
<td>2</td>
<td>\geq 3</td>
</tr>
</tbody>
</table>

Class 1, Class 2, and Class 3 are basically related to bangs. Class 2 and Class 3, however, include different sounds with Class 1, for example, speech, footsteps, and so on.

\section{Unusual Sound Detection}

Unusual sounds are detected for among each class, that is, Class 1, Class 2, and Class 3. In order to detect the unusual sound, the MFCC and Fluctuation Pattern (FP) analysis are applied to the sounds categorized for each class. The FP analysis describes the amplitude modulation of the loudness per frequency band and describes characteristics of the audio signal which are not described by the spectral similarity measure. Therefore, the FP analysis, in addition to MFCC, is necessary to classify bangs. On details of the FP analysis, see [8].

Using the MFCC and the FP, we compare the similarity of the sounds for among each class. To this end, we calculate the symmetric Kullback-Leibler divergence between the pairs of the MFCCs obtained from the sounds, where the value obtained by the symmetric Kullback-Leibler divergence is denoted by \( d_{KL} \). Moreover, the Euclidean distance between the pairs of the FPs is calculated, where the value is denoted by \( d_E \). Then the similarity between a pair of sounds is judged using the following value.

\[
d_{KL} + d_E,
\]

That is, if the value \( (d_{KL} + d_E) \) is greater than a threshold, the pair is similar, if not, the pair is not similar, where the threshold is suitably chosen. The unusual sound for each class is judged by using the similarity.

---

**Figure 3.** The power of a detected sound obtained by (9)

**Figure 4.** Microphone array used for observing sounds
A. A Simple Experiment

1) Setup: Experiments were implemented in a normal room. The QCMA was positioned near to the ceiling of the room (see Fig. 4). The sound was generated from appropriate positions which were about 1.5m from the tripod on which the QCMA was mounted. There were four kinds of sounds, that is, a sound of dropping a plate (Sound A), a sound of hitting something (Sound B), a sound of dropping a magazine (Sound C), and a sound of hitting something (Sound D). Their sequential generation is shown in Fig. 5. The parameters \( l_1, l_2, \) and \( l_3 \) in (10) were set to be 600, 250, and 50.

![Image of sound detection](image.png)

Figure 5. One of the sounds observed by the QCMA.

2) Results: As result of the sound detection using the method shown in subsection II-B, 5 sounds were obtained as a bang, where the parameter \( l \) in (8) was set to be 1. It should be noted that every sounds were categorized into Class 1. Sound B was not categorized into the three classes.

![Table of classification results](table.png)

<table>
<thead>
<tr>
<th>Detected sounds</th>
<th>Similar sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det1 (Sound A)</td>
<td>First bang</td>
</tr>
<tr>
<td>Det2 (Sound A)</td>
<td>Det1 (Similarity:62%)</td>
</tr>
<tr>
<td>Det3 (Sound C)</td>
<td>Warning!!</td>
</tr>
<tr>
<td>Det4 (Sound D)</td>
<td>Warning!!</td>
</tr>
<tr>
<td>Det5 (Sound D)</td>
<td>Det4 (Similarity:61%)</td>
</tr>
</tbody>
</table>

We numbered these detected sounds as Det1, Det2, \cdots, Det5. Table 1 shows the classification results obtained using the value of (11) for the detected sounds (Det1–Det5). The left column shows the detected sounds. The right column shows the classification results. Namely, the sound which is the most similar to each detected sound shown in the left column is shown in the right column, where the similarity is represented with the percentage which is based on the value of \( d_{KL} \). We should note that if the similarity is more than 60\%, the pair is similar. Since there are no similar sounds before Det3 and Det4 are generated, those sounds were judged, respectively, as an unusual sound, where in the table, "Warning!!" means the unusual sound, that is, in our surveillance system, if unusual information are detected, the "Warning!!" will be indicated in order to call a guard’s attention. The sounds Det2 and Det5 were judged, respectively, as an usual sound, because there are similar sounds before they are generated. From the results we conclude that our proposed method of classifying sounds has an ability of detecting usual and unusual sounds.

IV. Conclusion

We have introduced the techniques of measuring, detecting, and classifying sounds for a security monitoring system. The main contribution of our work is that we presented the techniques of detecting sounds which are usually not audible in sound environment. We confirmed that the proposed techniques can be used to detect the unusual sound in a simple experiment.

REFERENCES