Multimedia Tools and Applications (2020) 79:16127–16139
https://doi.org/10.1007/s11042-019-7547-y

Sound learning–based event detection for acoustic surveillance sensors

Jeong-Sik Park [1] · Seok-Hoon Kim [2]

Received: 28 March 2018 / Revised: 25 July 2018 / Accepted: 25 March 2019 /
Published online: 6 April 2019

Abstract
This study proposes an event detection technique for acoustic surveillance that detects emergency situations by using acoustic sensors. Most surveillance systems have widely depended on visual data recorded by closed-circuit television (CCTV) cameras, but more intelligent systems are now beginning to use audio information for more reliable detection of emergency situations. Most of the conventional studies on acoustic event detection adopt limited types of acoustic data and are based on simple algorithms, such as energy-based determination. Thus, these approaches are easily realized, but may induce serious detection errors in real-world applications. In this study, we propose an event detection technique based on a sound-learning algorithm to be adopted by real-time acoustic surveillance systems. One main process of this technique is to construct acoustic models via learning algorithms from sound data collected according to types of acoustic events. The models are used to determine whether audio streams entering an acoustic sensor refer to the events or not. In event detection experiments performed in an outdoor environment, the proposed approach outperformed conventional approaches in the real-time detection of acoustic events.

Keywords Acoustic surveillance sensor · Surveillance system · Sound learning · Acoustic event detection

1 Introduction

Surveillance is the monitoring of human behavior, activities, or other changing information for the purpose of protecting people from hazardous and emergency situations [10]. In general,

✉ Seok-Hoon Kim
kimshn@pcu.ac.kr

Jeong-Sik Park
parkjs@hufs.ac.kr

[1] Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies, Seoul, Republic of Korea

[2] Department of Electronic Commerce, Paichai University, Daejeon, Republic of Korea

this includes observation from a distance by means of electronic equipment like closed-circuit television (CCTV) cameras. Over the past few years, governments of the world have increased the amount of equipment and have attempted to advance technical features to maintain social control, recognize and monitor threats, and prevent criminal activity.

Many types of electronic devices are utilized to collect sensory information in danger zones and deliver the information to a main surveillance system. Cameras representatively work for the purpose of observing an area [4, 17, 19]. A set of cameras, including wide field-of-view (FOV) passive cameras and pan-tilt-zoom (PTZ) active cameras, autonomously capture high-resolution videos of pedestrians. In particular, they are often connected to a recording device or IP network to be maintained in the form of a visual sensor network.

Although most surveillance systems highly depend on visual sensory information captured by CCTV cameras, they occasionally may fail to attain reliable detection performance. First of all, the quality of captured images and videos is greatly affected by luminance. Thus, the data captured under dark conditions or in bad weather may be not amenable to visual analysis for gesture or facial recognition. Occlusion, or an invisible area created by the observation angle of a CCTV camera, is also a main factor that interferes with accurate recognition of visual sensory data. In addition, camera-based surveillance systems have difficulty in protecting blind spots that cameras cannot cover. Be that as it may, the cost of a high-resolution camera makes it difficult to infinitely deploy surveillance cameras.

To complement the limitations of camera-dependent surveillance systems, people began to pay attention to other sensory information, including audio, thermal, and vibration [20]. Among them, acoustic sensory information offers distinct advantages [12]. First, acoustic data are rarely affected by illumination or occlusion. And acoustic sensors, such as a microphone, can cover wider areas than cameras. Finally, a relatively low cost is required to set up the devices.

While acoustic surveillance in underwater environments has been widely investigated for military purposes or for exploring marine resources [5, 9], many challenging issues still remain in outdoor places with a variety of acoustic noises. In this paper, we propose an efficient acoustic surveillance technique to be adopted in outdoor environments.

The remainder of this paper is organized as follows. Section 2 introduces conventional studies of acoustic surveillance. Section 3 proposes an efficient acoustic detection approach for a surveillance system. Section 4 explains the experimental setup and results. Finally, conclusions are presented in Section 5.

2 Conventional studies on acoustic surveillance

An acoustic surveillance system aims at identifying hazardous situations by analyzing auditory sensory information recorded through an acoustic sensor to detect an acoustic event, thus protecting pedestrians from physical and psychological damage. For example, detection of a scream from a victim flusters the perpetrator and protects the victim by creating a warning sound and informing a security guard of the emergency situation.

In comparison with video-based surveillance, acoustic surveillance is still an emerging technique and has challenging research issues. Although several approaches for acoustic event detection have been introduced in recent years, they seem to require further verification in order to be adopted in real-world surveillance environments. Most of the early studies

concentrated on the detection of peculiar and highlight events, such as an explosion and gunshots [3]. They reported reliable detection performance, because the highlight events generally have acoustic characteristics easily distinguished from normal acoustic sounds. Later, several research works extended the sorts of acoustic sources to include footsteps, ringing phones, applause, or knocking [7, 21]. They still targeted object-induced sounds, however, and excluded human voices.

In recent years, a few studies eventually began to recognize human voices as acoustic events [11, 12]. However, most of them targeted normal speech irrelevant to emergency situations. The other studies collected abnormal voices, such as a scream from movie clips, but it provided offline verification without consideration of outdoor environments where a variety of noises exist.

Most of the conventional studies have provided reasonable research directions for acoustic event detection. Nevertheless, the use of limited kinds of acoustic sounds or a restricted verification environment creates uncertainty about whether the approaches would operate properly in a real-world emergency situation, in particular, in outdoor environments. A critical point is that virtually none focus on crimes committed against women and children, even though, for a long while, that has been recognized as one of the most malicious of criminal activities. For this reason, this study presents an acoustic event detection approach for outdoor environments. In particular, we carefully concentrate on the detection of abnormal human voices, such as a scream, asserting that the abnormal voice provides the most essential and certain cue of a crime, more so than any other kind of acoustic sound.

3 Acoustic event detection based on learning of abnormal voices

Surveillance sensors are required to continuously collect environmental information and submit it to a main surveillance system. A surveillance sensor can be a microphone sensor having a function of transmitting the recorded signals to the recognition system (surveillance system). The main surveillance system should also keep operating in real time while processing the information to detect emergency situations. Thus, several issues should be considered for real-time operation of acoustic surveillance in outdoor environments.

First of all, the system frequently encounters various background noises, such as cars, a subway, or the sounds of animals, as well as weather-related noises, such as wind and thunder. Normal voices of pedestrians also occasionally interrupt event detection and are misrecognized as an abnormal voice. In particular, a loud voice similar to screaming mainly disturbs the acoustic surveillance system.

The distance between the sound source and acoustic sensors should be also considered. In standard multimodal surveillance, multimodal sensors (including visual sensors and acoustic sensors) are maintained together to work in conjunction with each other [20]. Hence, the acoustic sensors are placed alongside other sensors, such as CCTV cameras. Due to this property, the acoustic surveillance system may occasionally fail to detect sound sources induced far from the acoustic sensor.

In this section, we investigate acoustic characteristics of abnormal voices picked up by an acoustic sensor. Then, an effective acoustic event detection approach is proposed to cope with the above issues.

3.1 Characteristics of the abnormal voice

A scream is a representative abnormal voice that people naturally manifest in case of emergency, and it can be used as an essential acoustic event in an emergency. The scream has acoustic characteristics distinguished from normal speech, as shown in Fig. 1. It reveals a distinct distribution of spectral energy in overall formant frequencies ranging from low to high, and the energy distribution is retained during a certain time, whereas a normal voice has intensive energy in low-frequency regions, and the energy distribution significantly varies over time.

This evident difference in spectral characteristics is likely to make easy detection of screaming by utilizing only temporal acoustic features, such as energy, duration and pitch. However, these features may not be suitable for discriminating between a screaming voice and normal voices such as a loud voice or an expression of the joy mode. Figure 2 represents a spectrogram of a male speaker's scream and a voice recorded while a male is calling someone's name loudly. As demonstrated in this figure, these two kinds of voices have similar distribution of spectral energy and duration. A task of discriminating between screaming in emergency and a loud voice in the joy mode has a similar ambiguity problem.

The conventional studies have reported that a temporal feature-directed acoustic event detection technique is very useful for detecting highlight events such as a gunshot. But, this approach is not evidently applicable to detecting the abnormal voice of a human. In this study, we propose a voice learning-based event detection to carefully handle these problems. Supervised learning–based approach is expected to provide a greater possibility of discriminating an abnormal voice from a normal, loud voice as well as object-induced sounds, based on the basic property that labeled training data provide a reliable decision boundary between two different classes.

3.2 Procedure of learning-based acoustic event detection

In various pattern recognition tasks using acoustic information, statistical model–based learning algorithms provide an effective criterion to discriminate and determine recognition units [2, 13, 14, 18]. A representative statistical model is the hidden Markov model (HMM), which can generalize temporal acoustic characteristics by using Gaussian mixture models (GMMs). In recent years, an advanced neural network-based learning technique called deep neural network (DNN) is widely adopted for various pattern recognition tasks [6, 8]. Although the technique provides reliable performance, it requires a tremendous amount of training data and high capacity hardware resources for intensive computation. For this reason, DNN-based learning approach may be not appropriate for sensor-based real-time monitoring system.

Many studies have already employed HMMs for speech recognition tasks. A general paradigm of conventional HMM-based speech recognition is composed of two phases:

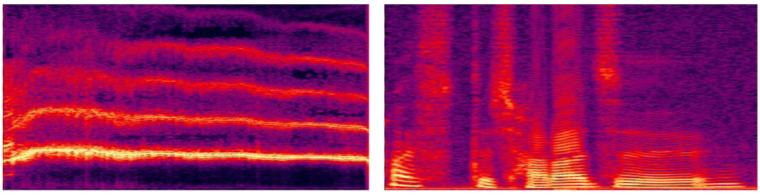

Fig. 1 Spectrogram of a scream (left) and a normal voice (right) uttered by a female

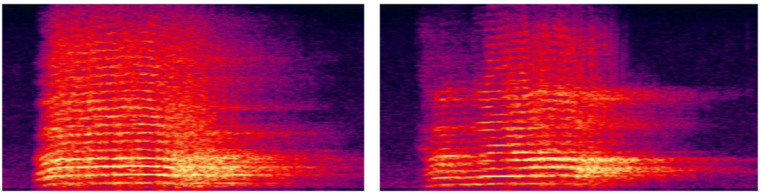

Fig. 2 Spectrogram of a scream (left) and a normal loud voice (right) uttered by a male

acoustic model training plus recognition. In the training phase, acoustic models of recognition units are constructed according to the HMM learning procedure. Then, the models are used to determine the results of the given input data during the recognition phase. Several conventional studies of acoustic event detection focused on the general HMM paradigm [7, 11, 12], but fewer considered the necessary issues addressed in the beginning of this section.

For real application of acoustic surveillance, we organize a new paradigm for acoustic event detection. The proposed procedure is demonstrated in Fig. 3. Acoustic signals picked up by an acoustic sensor are submitted to a main surveillance system, and then the system performs event detection procedures that are composed of three main modules: pre-processing, recognition, and post-processing.

The main roles of pre-processing include the detection of voice regions, the reduction of background noise components, and the extraction of acoustic features. The pre-processing module first detects voice regions from acoustic signals picked up by an acoustic sensor that is remotely located from the sound sources. The detected voice signals are then processed by noise reduction algorithms to exclude noise components that contaminate the original human voice. This process can provide a solution to improve the recognition accuracy against adverse conditions like weather or background noises. The conventional noise reduction approaches from a standard spectral subtraction technique to filter-based techniques such as a Wiener filter can be adopted for this task.

Finally, acoustic features are extracted from the noise-reduced voice signals. The types of acoustic features obtained during pre-processing should correspond to the features of acoustic models. During the recognition process, the features are compared with acoustic models that are constructed in the training phase. Next, the recognition module finds a model for which the features indicate the highest observation probability. If the model is relevant to one of the emergency sound models, the corresponding voice region is designated as a candidate stream.

In real-world environments, a significant number of detected events are eventually confirmed as false detection. Therefore, disregarding normal situations is also a necessary task in acoustic event detection. One main goal of the post-processing module is to correctly determine whether the candidate stream certainly is an emergency. This re-verification process plays a role in reducing false detection errors.

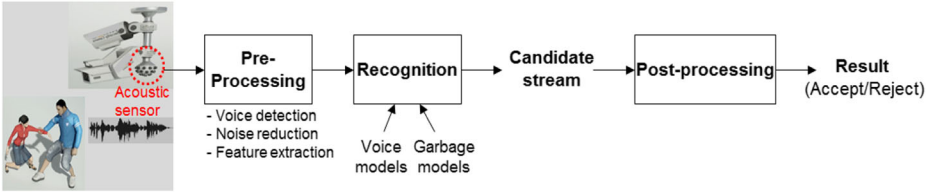

Fig. 3 The proposed procedure of acoustic event detection

3.3 Acoustic model training

In learning-based pattern recognition tasks, the correctness of acoustic models (e.g., HMMs) directly affects recognition performance. For construction of reliable acoustic models, the training process of the acoustic surveillance system requires a sufficient amount of speech data that is clearly relevant to the emergency. As addressed in Section 2, most of the conventional studies dealt with limited types of speech data. In this study, we used two methods of data collection to acquire a sufficient amount of speech data that well reflects real-world emergency situations.

In the first method, we searched for emergency situations from video and audio materials, such as films. Although the sound recordings belong to abnormal speech simulated by actors, they have clear characteristics that express the situation. For more general cases, we collected recordings divided according to gender and age.

In the second method, we attempted to consider real-world environments. Most of the data, including sound recordings in films, do not consider the distance between the acoustic sensor and the sound sources. In addition, most of the original sources are recorded in a silent recording room. The most accurate way of considering distance is to directly record the sound by using a remote acoustic sensor in an outdoor environment. For this, we installed an acoustic sensor at the height of general CCTV cameras and recorded screams from male and female participants who stayed approximately 30 m away from the sensor. Figure 4 represents a picture of such a recording.

Configuration of acoustic models also highly affects recognition performance. A general way in the conventional studies is to construct a common acoustic model for each acoustic event, using full speech data corresponding to the event. These types of model can be compatible with object-induced sounds, such as gunshots, because such sounds have common and typical acoustic characteristics. On the other hand, this approach is not suitable for human voices, which have different characteristics in accordance with gender, age, or even culture.

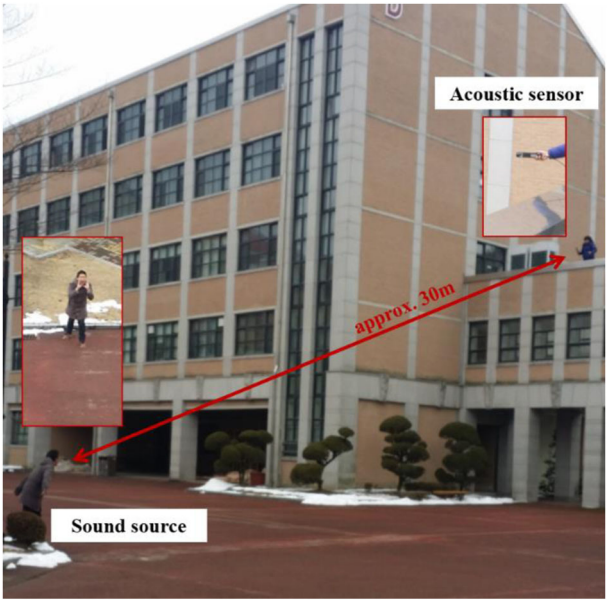

Fig. 4 Recording screams in an outdoor environment

The proposed model construction approach carefully considers the characteristics of the human voice. First of all, we maintain two kinds of models: voice models and garbage models. The garbage models are constructed by using several types of object-oriented sounds to play a role in disregarding those sounds. Voice models are divided into several kinds in order to discriminate normal speech and loud voices from abnormal voices, such as a scream. Each kind of model is constructed using corresponding speech data. Among the models, abnormal voice models are relevant to emergency situations, whereas other voice models are used for the purpose of disregarding the normal voice. There is a general tendency in the abnormal voice whereby males and females have distinct acoustic characteristics when screaming. As compared in Figs. 1 and 2, screaming of a female presents more distinct spectral energy at high frequencies than a male speaker. In order to consider this general difference between males and females, we construct two gender models using abnormal voice data divided by gender. Figure 5 summarizes our proposed configuration of acoustic models for abnormal voice-based event detection.

After determining the types of acoustic models, acoustic features are extracted from the training data that were divided according to the model type. Then, model parameters (the mean and the variance of the Gaussian distribution) are estimated using the iterative expectation–maximization (EM) algorithm [1].

3.4 Recognition of input signal utterances

The acoustic models constructed during the training phase are used to recognize given input utterances. The constructed HMMs try to identify a type of input utterance according to the following procedures. For a sequence of feature vectors $X = \{\overrightarrow{x_1}, ..., \overrightarrow{x_T}\}$ that are extracted from an input utterance, the log-likelihood of each HMM model λ_i ($i = 1, ..., E$ if there are E acoustic models) is computed with (1).

$$\log P(X|\lambda_i) = \sum_{t=1}^{T} \log P\left(\overrightarrow{x_t}|\lambda_i\right) \tag{1}$$

Then, a model that has the maximum log-likelihood with a given input utterance is determined to be a recognition result, as stated in (2).

$$\hat{e} = \underset{e}{\operatorname{argmax}} \log P(D|\lambda_e), e = 1, ..., E \tag{2}$$

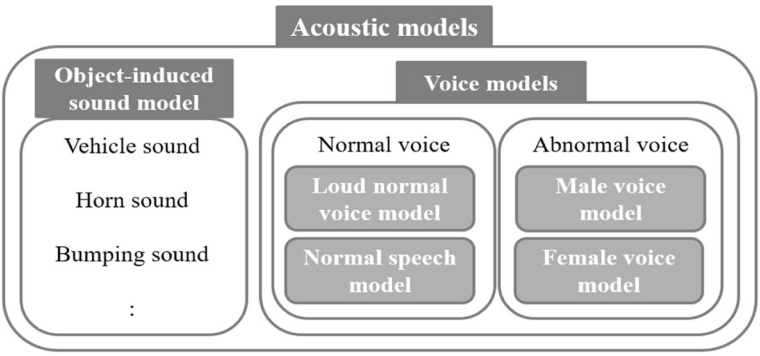

Fig. 5 Proposed configuration of acoustic models for abnormal voice-based event detection

The acoustic surveillance system determines whether the input utterance corresponds to an abnormal voice or not, according to the recognition result, $\hat{e}$. If $\hat{e}$ belongs to one of two gender models constructed from abnormal voice data, the utterance is finally categorized as an abnormal voice.

3.5 Post-processing for verification of recognition results

The recognition results of signals recorded in an online manner may contain a tremendous amount of errors, mainly due to adverse environmental conditions including weather and background sounds. A representative error type is false alarm that identifies a noise sound as an abnormal voice. False alarm frequencies significantly increase in an outdoor environment, thus degrading the reliability of a surveillance system. For this reason, it is very necessary to verify the correctness of each recognition result to determine if the result can be finally accepted or ignored. In the proposed procedure of acoustic event detection shown in Fig. 3, the post-processing module operates for the verification.

HMM-based modelling approach supports efficient verification ways [15]. Among them, the log-likelihood-based method provides reliable verification performance without intensive computation. Thus, this method is very suitable for the online acoustic event detection task in which a number of recognition results are continuously generated.

In accordance with (1) and (2), E log-likelihood results are obtained with a given input utterance when there are E acoustic models. The correctness of the recognition result, $\hat{e}$, can be verified using a criterion measure, as stated in (3).

$$M(x) = \frac{\log P\left(x|\lambda_{R_1(x)}\right)}{\sum_{r=2}^{E} \log P\left(x|\lambda_{R_r(x)}\right)} \tag{3}$$

in which $R_r(x)$ refers to the r-th-ranked model index (in a range from 1 to E) in E log-likelihood results obtained from all E models with a given utterance x. $\lambda_{R_r(x)}$ and $\log P\left(x|\lambda_{R_r(x)}\right)$ indicate the model corresponding to the index and the log-likelihood result at the r-th rank, respectively. This measure computes the relative distance between the log-likelihood at the first rank, that is, the recognition result, and the overall log-likelihood results, based on the assumption that the distance increases more when an utterance retains distinctive properties of an abnormal voice. After finishing the recognition phase, the log-likelihood results are submitted to the post-processing step and then the criterion measure is estimated according to (3). If the estimation value is regarded to be sufficiently high, the result is finally accepted. Otherwise, it is disregarded.

4 Experimental setups and results

In order to validate the proposed approach, we performed acoustic event detection experiments using two types of evaluation task: offline and online. The offline task refers to detection of isolated events, whereas the online task detects a set of events in continuous sequences. In this study, we concentrate on verifying the efficiency of two main approaches proposed herein: the method of data collection and the configuration of acoustic models. Considering the features of these two approaches, the method of data collection and model configuration were validated by offline and online evaluation tasks, respectively.

4.1 Experimental setup

There are no publicly available natural corpora consisting of atypical sound events for surveillance applications, because the data have private characteristics and are manifested only scarcely. For this reason, we collected sound data for the experiments. The data consist of sound events collected in two ways: from recordings of video materials and by direct recording in outdoor environments. The sound events include object-oriented sounds, such as vehicle sounds, and normal/abnormal voices. Fourteen male and female participants uttered sounds in both normal and abnormal voices for the direct recordings. They stayed approximately 30 m away from an acoustic sensor, as demonstrated in Fig. 4. The acoustic sensor has one channel and omnidirectional microphone. About 500 sound events were collected.

We used 12-dimensional Mel-frequency cepstral coefficients (MFCCs) as feature vectors. All vectors were computed within frames of 30 ms with a Hamming window shifted by 10 ms. Acoustic models were represented by three-state, left-to-right HMMs with 32 Gaussian mixtures. The number of mixtures providing the best condition was empirically determined.

4.2 Experimental results

First, we carried out acoustic event detection experiments using two kinds of abnormal voice data: recordings from video materials and data recorded directly in an outdoor environment. Thus, two abnormal voice models were trained from each set of data and used for verification. In addition, an object-induced sound model and a normal voice model were consecutively trained using relevant data recorded in outdoor environments. Finally, each type of data that was not included in the model training was evaluated. For fair validation, we used leave-one-out cross-validation, dividing each type of data into three sets.

We investigated the recognition ratio of test data on each acoustic model for which test data indicated maximum log-likelihood. Table 1 represents the results. As expected, most of the test data indicated maximum log-likelihood on their corresponding acoustic model. The two kinds of abnormal voice data demonstrated similar recognition performance on the relevant model. But from the recognition results, it is difficult to explain which type of abnormal voice data provides better detection performance in surveillance applications.

So, we further investigated the recognition ratio of two types of abnormal voice data on each acoustic model, except for the corresponding abnormal voice model. This experimental result is demonstrated in Table 2. As shown in this table, abnormal voice models from direct recording provided better performance on detection of abnormal voice data, compared to the video materials. 72% of all video material data were recognized by abnormal voice model constructed using direct recording data. On the other hand, only 64% of all direct recording data were categorized as abnormal voice on acoustic models constructed using video material

Table 1 Recognition ratio (%) of test data on each acoustic model

Model type Test set	A	B	C	D
A. Abnormal voice: video material	86	9	3	2
B. Abnormal voice: direct recording	7	89	2	2
C. Object-induced sound	11	8	76	5
D. Normal voice	8	9	3	80

Table 2 Recognition ratio (%) of abnormal voice data on each acoustic model except for the corresponding model

Model type Test set	A	B	C	D
A. Abnormal voice: video material	–	72	9	19
B. Abnormal voice: direct recording	64	–	12	24

data. This result suggests that data obtained from direct recording have advantages in constructing more reliable acoustic models for acoustic event detection.

The next experiment was carried out to verify the efficiency of the proposed model configuration described in Fig. 5. In this experiment, we performed an online evaluation task to consider event detection in a real-world application operating in an outdoor environment. Therefore, we constructed a set of acoustic models by using each type of acoustic data recorded by an acoustic sensor in an outdoor environment. Then, the models were used to recognize acoustic events in real time. In addition to models organized by the proposed set (Fig. 5), a simple set of acoustic models was also constructed according to the conventional approach. This simple set is composed of an object-induced sound model and an abnormal voice model. For further comparison with the conventional approaches, we developed a simple acoustic event detection program based on signal energy. In this system, signal energy was continuously estimated for a certain number of frames, and events were determined by comparing the energy with a pre-defined threshold. Accordingly, three detection systems operated concurrently to detect abnormal voice events (such as screaming) that are related to emergency situations.

For sophisticated validation, we conducted the experiments four times. In each trial, maintained for about 30 min, four male and female participants uttered screaming sounds several times, staying 30 m away from an acoustic sensor. At the same time, the systems operated to detect the acoustic events while recording the detection time.

After finishing the online evaluation, 12 detection results were confirmed through the experiments conducted four times. Each result is given by computing a false detection (FD)

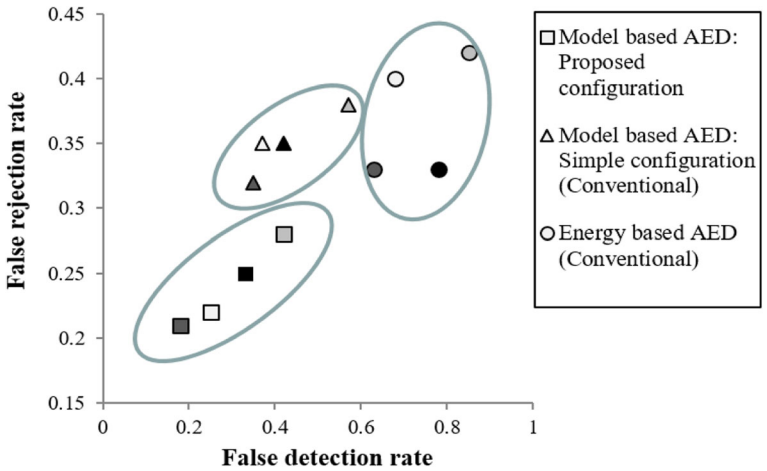

Fig. 6 Event detection results (false detection error rates and false rejection error rates) of the proposed and conventional approaches

error rate and a false rejection (FR) error rate, which are general criteria for measuring detection accuracy in pattern recognition tasks and are defined as (4) [16].

$$FD = \frac{\text{number of incorrect detections}}{\text{number of detected events}}, FR = \frac{\text{number of failed detections}}{\text{number of detected events}} \quad (4)$$

Figure 6 demonstrates the event detection results of our proposed approach and two conventional approaches. Three kinds of approaches are represented with respective marks. And the four times the experiments were conducted are discriminated by the brightness of the marks. As shown in this figure, the proposed model configuration provided significant detection accuracy in overall trials, representing low FD and FR rates. The conventional model–based approach demonstrated FR rates similar to the energy-based approach, but showed better performance in FD rates. We can see that FD rates are worse than FR rates in the overall experimental results. This was induced by wind noises that mainly caused incorrect detections. The energy-based approach frequently failed to discriminate wind noises from abnormal voices, while the model-based approaches successfully disregarded the wind noises.

5 Conclusions

In this paper, we proposed an efficient acoustic event detection technique based on a learning algorithm. To overcome the limited types of data used in conventional studies, we introduced new ways of data collection. In addition to collecting acoustic data from video materials, we recorded data directly from an acoustic sensor in an outdoor environment to consider features of real-world surveillance systems. We also proposed an effective acoustic model configuration that is organized by gender-characterized abnormal voice models and normal voice models, as well as object-induced sound models.

To validate the proposed acoustic event detection technique, we conducted several detection experiments on video recording data and directly recorded data. The results confirmed that our configuration model, adopting directly recorded data, provides better conditions for detecting acoustic events in real-world environments. For future work, we will investigate effective techniques of multimodal surveillance in outdoor environments.

Acknowledgements This research was supported by Hankuk University of Foreign Studies Research Fund, Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1A09000903), the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and the research grant of Pai Chai University in 2019.

References

1. Bilmes J (1997) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. University of Berkeley, International Computer Science Institute, Tech. Rep., ICSI-TR-97-021
2. Campbell JP (1997) Speaker recognition: a tutorial. IEEE 85(9):1437–1462
3. Clavel C, Ehrette T, Richard G (2005) Events detection for an audio-based surveillance system. In: IEEE international conference on multimedia and expo

4. Cornacchia M, Ozcan K, Zheng Y, Velipasalar S (2017) A survey on activity detection and classification using wearable sensors. IEEE Sensors J 17(2):386–403
5. Didrikas T, Kubilius R, Spegys M (2011) Surveillance of marine resources using multi-frequency hydroacoustics. Klaipeda Univ Tech Rep, NOR-LT0047
6. Hinton GE, Deng L, Yu D et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
7. Huang PS, Zhuang X, Hasegawa-Johnson M (2011) Improving acoustic event detection using generalizable visual features and multi-modality modeling. In: IEEE international conference on acoustics, speech, and signal processing
8. Kang K, Ouyang W, Li H, Wang X (2016) Object detection from video tubelets with convolutional neural networks. In: IEEE conference on computer vision and pattern recognition
9. Lurton X (2002) An introduction to underwater acoustics: principles and applications. Springer Science & Business Media
10. Lyon D (2007) Surveillance studies: an overview. Polity Press, Cambridge
11. Mesaros A, Heittola T, Eronen A, Virtanen T (2010) Acoustic event detection in real life recordings. In: European signal processing conference
12. Ntalampiras S, Potamitis I, Fakotakis N (2009) On acoustic surveillance of hazardous situations. In: IEEE international conference on acoustics, speech, and signal processing
13. Park JS, Kim JH, Oh YH (2009) Feature vector classification based speech emotion recognition for service robots. IEEE Trans Consum Electron 55(3):1590–1596
14. Park KM, Park JS, Oh YH (2010) GMM adaptation based online speaker segmentation for spoken document retrieval. IEEE Trans Consum Electron 56(2):1123–1129
15. Park JS, Jang GJ, Kim JH (2012) Multistage utterance verification for keyword recognition-based online spoken content retrieval. IEEE Trans Consum Electron 58(3):1000–1005
16. Phillips PJ, Martin A, Wilson CL, Przybocki M (2000) An introduction evaluating biometric systems. Computer 33(2):56–63
17. Qureshi FZ, Terzopoulos D (2006) Surveillance camera scheduling: a virtual vision approach. Multimed Syst 12(3):269–283
18. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. IEEE 77(2):257–286
19. Turaga P, Yuri AI (2011) Diamond sentry: integrating sensors and cameras for real-time monitoring of indoor spaces. IEEE Sensors J 11(3):593–602
20. Zhu Z, Huang TS (2009) Multimodal surveillance: sensors, algorithms and systems. Artech. House
21. Zhuang X, Zhou X, Hasegawa-Johnson M, Huang PS (2010) Real-world acoustic event detection. Pattern Recogn Lett 31(12):1543–1551

Jeong-Sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and

Technology) in 2003 and 2010, respectively. He is now an associate professor in the Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies. His research interests include speech signal processing, speech recognition, and voice interface for human-computer interaction.

Seokhun Kim was an assistant professor in Mobile Media at Suwon Women's University in 2012 and 2017. He is currently an assistant professor in the Electronic Commerce at Paichai University. His teaching and research specialties are in the fields Mobile computing, Web-App programming, Web-Database, information security.