Home Environmental Sound Alert System for Deaf and Hard-of-Hearing Users

Hye-Seung Cho, Hyoung-Gook Kim

Kwangwoon University, Seoul, Rep. of Korea
{hye_seung401; hkim}@kw.ac.kr

Abstract

Sound signals provide a great deal of information about their sound sources. However, deaf and hard-of-hearing people are not able to access this important information. Therefore, an assistive technology is required that automatically recognizes sound information and converts it into usable information for hearing-impaired people. In this paper, a home environmental sound alert system for deaf and hard-of-hearing users is presented. The system detects the sound generated in the home environment, converts the sound into text, and provides this text to the user. The core component of the environmental sound alert system is an accurate sound event detection mechanism. For precise sound event detection, we proposed an improvement method including signal estimation, channel selection, and a bidirectional gated recurrent neural network.

Index Terms: environmental sound alert, sound event detection, wireless sensor network, gated recurrent neural network

1. Introduction

Sensory abilities such as vision and hearing are very important in human life. In particular, certain information in our society usually depends on communication via sound. This social practice renders important information inaccessible to many deaf and hard-of-hearing people [1]. Therefore, to provide hearing-impaired people with information about sound, different assistive technologies have been developed. For instance, the light system that flash the light when somebody rings the doorbell is one of the representative assistant infrastructures. However, such approaches are only targeted to specific events and they are ineffective when multiple sound events are generated at the same time.

In this paper, we propose a home environmental sound alert for deaf and hard-of-hearing people based on sound event detection (SED). The proposed system is composed of a wireless sensor network (WSN) and the user's smart device.

The environmental sound alert system needs to detect and recognize sound events accurately in various situations in real life. Therefore, we also proposed a method to improve the performance of the SED of the proposed system. The proposed method is comprised of signal estimation, channel selection, and a bidirectional gated recurrent neural network (GRNN) [2]. However, even if a high-performance SED is applied, detection errors can occur due to various external factors. In order to minimize the risk caused by these errors, the proposed system provides both the event detection results and their probabilities.

The outline of this paper is as follows. In Section 2, the proposed system and the detail of the SED method are explained. Experimental results are presented in Section 3 and conclusions are given in Section 4.

2. Proposed System

Fig. 1 schematically illustrates the proposed environmental sound alert system based on SED.

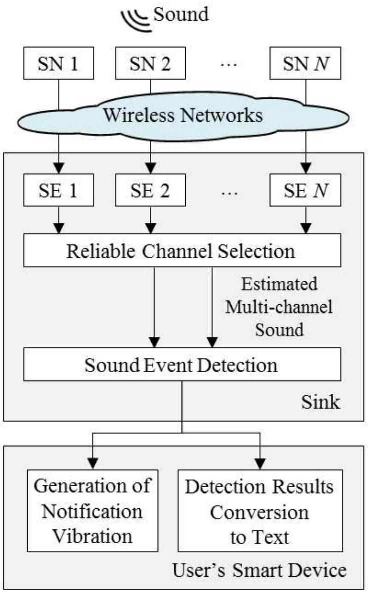
Sound
SN 1
SN 2
···
SN *N*
Wireless Networks
SE 1
SE 2
···
SE *N*
Reliable Channel Selection
Estimated Multi-channel Sound
Sound Event Detection
Sink
Generation of Notification Vibration
Detection Results Conversion to Text
User's Smart Device

Figure 1: *Schematic illustration of the proposed system.*

The wireless sensor nodes (SNs) can simultaneously capture sounds generated in a room. Each SN is equipped with a single microphone. The microphone at each SN receives mixed or polyphonic sounds. The recorded mixed sounds are encoded and transmitted via sound packets to the sink through networks with wireless links that are associated with packet loss. When each microphone packet arrives at the sink via the wireless networks, it is decoded into a signal frame. Lost packets are recovered by packet loss concealment in signal estimation (SE).

Then, a set of microphones of which the signals are the most highly correlated with each other are chosen among the multi-channel microphones to increase computational efficiency and achieve better performance. In this paper, we used a signal-based channel selection (CS) method using a multi-channel cross-correlation coefficient (MCCC) [3]. The basic concept of this approach is to treat the channel that is uncorrelated with the other channels as being unreliable and to select only a subset of microphones with the most correlated signals.

After CS, the signals of the selected two-channels are then used for environmental SED. Motivated by the human auditory system (using two ears), we extracted a noise-reduced spectrogram and time difference of arrival (TDOA) [4] from the two-channel audio information. The features were used as high-resolution spectral inputs to train the bidirectional GRNN (BGRNN). The BGRNN is one of the most recent neural networks, and demonstrates good performance in sequence modeling. It provides a fast and stable convergence rate compared to the long short-term memory recurrent neural networks (LSTM-RNN).

Detected sound event labels and probabilities are sent to the user device. This information is sorted in the order of highest to lowest accuracy and converted to text. The vibration for notification is also generated. The user is then notified with a vibration and text notification letting them know which sound event has occurred.

3. Experimental Results

In this section, the performance of the proposed SED method of the proposed system is evaluated using real life audio.

The living area used in our experiments was a 30 m^2 apartment. The rooms were equipped with sound sensors. Real sounds were recorded with sound sensors from various everyday environments. The sound corpus contained 10 sound classes. The polyphony percentages of the test set among the annotated frames were as follows: 83.5%, 10.6%, 4.3%, and 1.6% at polyphonic levels of 1, 2, 3, and 4, respectively. These samples for each class were distributed randomly (60% in training set, 20% in validation set, and 20% in test set). The performance of the proposed method was compared to that of different classifiers in combination with different features as follows, where NR, ST, and 2 denote the noise reduction, spectrogram, and two-channels, respectively:

Baseline (BL): The baseline system used MFCC coefficients (20 static, 20 delta, and 20 acceleration) extracted from one-channel audio. Before feature extraction, SE and CS were performed, although NR was not applied. A Gaussian mixture model (GMM) was used for SED.

Proposed Method (PM): The PM was composed of SE, CS, NR2, ST2, TDOA, and BGRNN. The input layer of the BGRNN comprised 40 units and three hidden layers with 200 GRU units. 3-layer BGRNNs were initialized with orthogonal weights and rectifier activation functions. The network was trained by binary cross-entropy as loss function.

Method 1 (M1): M1 was composed of SE, CS, NR, ST, TDOA, and BGRNN. One-channel features per frame were applied to the BGRNN classifier.

Method 2 (M2): M2 was composed of CS, ST, TDOA, and BGRNN. One-channel spectrogram features without SE or NR per frame were extracted and applied with 1 TDOA to the BGRNN classifier.

Method 3 (M3): M3 was composed of SE, CS, NR2, ST2, TDOA, and GRNN. Instead of using the BGRNN classifier of M1, a GRNN was used as the classifier with two-channel features.

Method 4 (M4): M4 was composed of SE, CS, NR2, ST2, TDOA, and LSTM-RNN. The LSTM-RNN was used as a classifier with two-channel features. The input layer of the LSTM-RNN comprised 40 units and two hidden layers with 200 LSTM units. The network was trained by binary cross-entropy as loss function.

For the evaluation metrics of system performance for SED, we used error rate (ER) and F-scores calculated in one second segments [5]. Experimental results show the baseline system (BL) has an event average error rate (ER) of 1.1 and F-score of 64.5%. The PM significantly outperforms the baseline system in terms of ER and F. These results confirm that SE, CS, and NR significantly contributed to the detection of overlapping sound events. In addition, BGRNN achieves better classification results than GMM, GRNN, and LSTM-RNN, which were trained on the same audio features. Spatial and noise-reduced spectrogram features from the multi-channel audio show considerable improvements over those when only mono-channel audio was used with the same classifier.

Table 1: *Performance Comparison for different combinations of classifiers and feature.*

Methods	**ER**	**F(%)**
BL	1.1	64.5
PM	0.63	86.7
M1	0.71	83.2
M2	0.98	74.9
M3	0.65	86.3
M4	0.67	85.9

4. Conclusions

In this paper, we proposed a home environmental sound alert system for deaf and hard-of-hearing users. The proposed system provides the user with information of all of the detected sound events and their probabilities. We also presented an improved sound event detection scheme for a reliable and effective alert system. Experimental results demonstrate the potential effective use of the proposed alert system in practical situations. Future work will focus on extending the BGRNN to improve detection accuracy for various sound event detection based systems.

5. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2015R1D1A1A01059804).

6. References

[1] D. Bragg, N. Huynh, and R. E. Ladner, “A personalizable mobile sound detector app design for deaf and hard-of-Hearing users,” *Proceedings of the ACM SIGACCESS, October 24–26, Reno, NV, USA*, 2016, pp. 3–13.

[2] M. Zoehrer and F. Pernkopf, “Gated recurrent networks applied to acoustic scene classification and acoustic event detection,” *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.

[3] M. Wölfel, “Channel selection by class separability measures for automatic transcriptions on distant microphones,” *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 582–585.

[4] D. Pavlidi et al., “Real-time multiple sound source localization and counting using a circular microphone array,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[5] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.