IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 January 2024, accepted 7 March 2024, date of publication 19 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3378096

RESEARCH ARTICLE

Real-Time Sound Recognition System for Human Care Robot Considering Custom Sound Events

SEONG-HU KIM[1], HYEONUK NAM[1], SANG-MIN CHOI[2], AND YONG-HWA PARK[1], (Member, IEEE)

[1]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
[2]Samsung Electronics Company Ltd., Gyeonggi-do 16677, Republic of Korea

Corresponding author: Yong-Hwa Park (yhpark@kaist.ac.kr)

This work was supported in part by the Institute of Civil Military Technology Cooperation funded by the Defense Acquisition Program Administration and the Ministry of Trade, Industry and Energy of Korean Government, under Grant UM22409RD4; and in part the BrainKorea21 (BK21) FOUR Program of the National Research Foundation Korea (NRF) Grant funded by the Ministry of Education (MOE).

ABSTRACT In real-life situations where human care robots are deployed, there are custom sound events whose acoustic characteristics change depending on the user's choice unlike general sound events so that the human care robots cannot recognize custom sound events correctly in a conventional way. To solve this critical problem, a real-time sound event recognition system with customization process is proposed. The human care robot collects custom sound samples of a specific user and customizes a sound event recognition model. The overfitting-based customized model shows the best recognition performance by improving F-scores by 66.4% on average compared to the conventional recognition model. After the customization process, the human care robot performs a real-time sound recognition by consistently streaming robot's real-time microphone signals into the overfitting-based customized SER model. In this process, an optimized overlap is applied on subsequent audio inputs on SER to achieve sufficiently fast response and robust performance. As a pilot test of the human care robot implemented in actual environment, the real-time sound recognition system shows the best average F-score of 0.982 with 75% overlap for sound events including custom sounds. This pilot test result confirms that the real-time sound recognition system with customization process can be successfully applied to human care robots to respond to the custom sounds.

INDEX TERMS Sound event recognition, human care robot, custom sound event, real-time system.

I. INTRODUCTION

Elderlies need meticulous and continuous care because they are vulnerable to various physical and mental dangers. As a global phenomenon, elderlies' population is rapidly overtaking the world population and number of people who can look after elderlies is decreasing, as result total cost for society to take care of elderlies is increasing exponentially. These yield growth in burdens for caring elderlies and decline in elderlies' life qualities. Therefore, human care robot that assists and takes care of the elderlies is emerging as a solution to the aging society [1].

Thanks to the rapid development in deep learning technology, various types of artificial intelligence have been implemented to human care robot in order to make them comprehend situations of users and take appropriate actions corresponding to the situations automatically [2], [3]. For example, we could implement object detection algorithm on a robot with cameras to observe and understand the environment and circumstance [4], [5], [6]. However, visual intelligence provides limited understanding on circumstances around the robot, as it can only recognizes events within visible regions and fails to recognize events in the limited vision of the camera. On the contrary, auditory intelligence is advantageous over visual intelligence as sound can be heard even when the sound source is invisible. It also enables the robot to understand circumstances more comprehensively by recognizing urgent events where alarms ringing, people scream or moan, etc. Therefore, we come up sound recognition system for human care robot that recognizes events from given

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

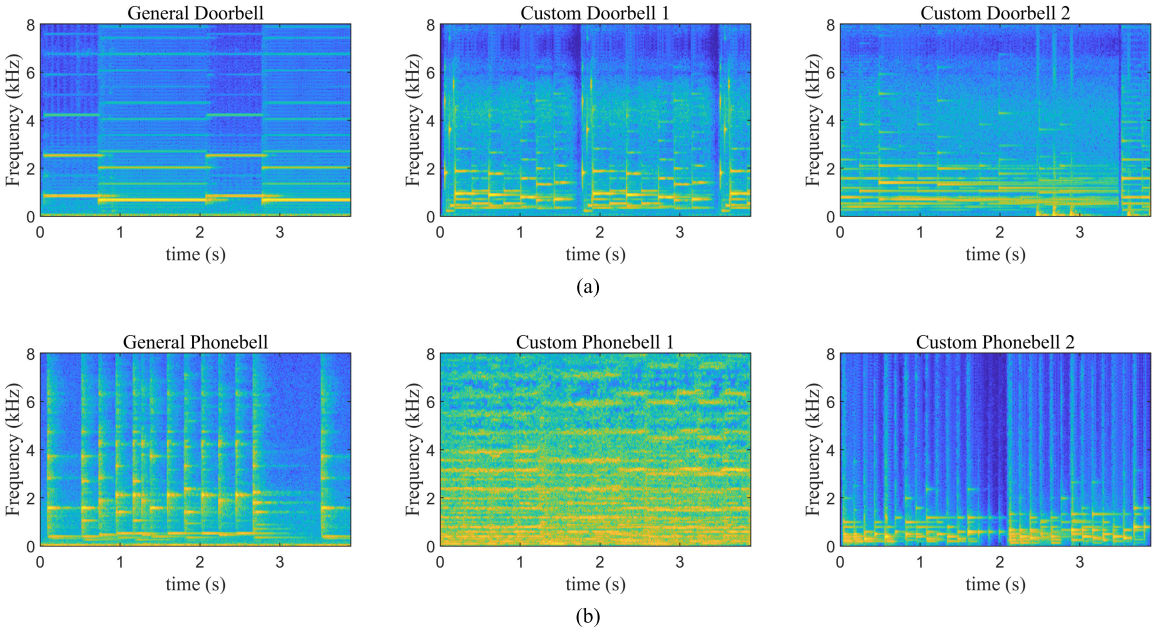


FIGURE 1. Spectrogram of example sounds of (a) general and custom doorbells and (b) general and custom phonebells.

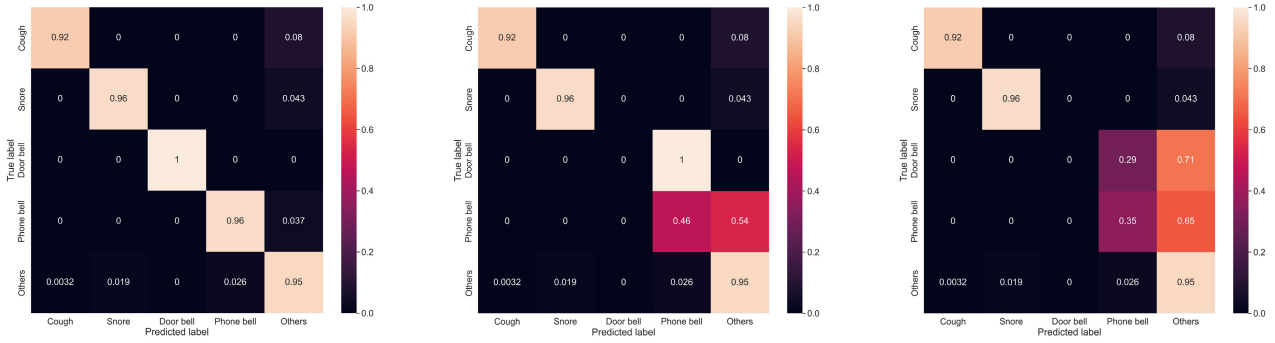
audio input to provide auditory intelligence on a human care robot.

Sound recognition field is recently led by works on sound event detection (SED), which performs classification and time localization simultaneously. In general, convolutional recurrent neural network (CRNN) [7], which combines convolutional layers and recurrent layers, is mainly used as the SED model [8], [9]. CRNN with transformer [10], [11] and conformer [12] widely used in automatic speech recognition achieved state-of-the-art performance in SED [13], [14], [15], [16], [17]. CRNN with frequency dynamic convolution, which is the content-adaptive model [18], [19], [20], improved SED performance by considering frequency dependencies as well as temporal dependencies [21]. In addition, data augmentation methods [22], [23], [24] improved not only performance but also robustness of SED model. There are other studies on applying sound recognition methods to human care robots. Previous studies for robot application have been mainly conducted to secure robust recognition performance in noisy situations by applying sound event recognition (SER) models [25], [26], [27], [28], [29]. Since human care robots do not require very precise time localization, SER instead on SED has been usually applied to the robots. In this work, we build sound recognition system for human care robot using SER as SED requires unnecessarily precise time localization. When applying SER on human care robot, we focus on identifying and resolving the effects of sound events rather than noise.

Recently, various sound event recognition systems or frameworks for human care robots and smart homes have been proposed based on the previous academic research on sound event recognition [30], [31]. The challenge human care robots and smart homes is the automatic sound recognition system and its ability to respond to problems that occur in real-life situations. Human care robots and smart homes need to recognize sounds occurring in the users' private space in real time and at all times without any off time to provide services to the users. So, human care robots and smart homes store the incoming sound data in real time as audio files with a certain duration using microphones, and they send the audio files to the server for data pre-processing and event recognition [32], [33], [34], [35]. This system enables the robots and the smart homes to recognize sound events in real time and automatically. On the other hand, there are various factors that degrade sound event recognition performance in indoor environments where robots and smart homes are applied. Different users have different room sizes and noises, and the SNR varies depending on the relative position between the microphone and the sound source. These performance degradation factors have been analyzed, and sound event recognition systems that are robust to these factors have been proposed [36]. In this study, we propose an automatic sound event system for application of human care robots in real-life situations.

A. RESEARCH QUESTIONS FOR REAL-LIFE APPLICATIONS OF HUMAN CARE ROBOTS

The ultimate goal of SER system for human care robots is to accurately recognize sound events related to users in real indoor environment without directly communicating with users. However, there is a big gap between general sound events assumed in conventional SER studies and sound events occurring in real-life situations. The conventional SER studies assume that each sound event has a similar acoustic



(a) (b) (c)

FIGURE 2. Confusion matrices of SER model tested with dataset containing (a) general bell sound set with average F-score of 0.946, (b) custom bell sound 1 set with average F-score of 0.577, and (c) custom bell sound 2 set with average F-score of 0.608.

characteristic pattern based on a common mechanical or biological mechanism. For example, blender or vacuum cleaner sound could be different by each device. But they share common mechanical structures thus produce sounds those share similar acoustic characteristics. Cough and snoring sound also could be different by each individual, but they share common biological mechanism thus produce sounds those share similar acoustic characteristics. These sound events are recognized through frequency pattern classification using deep neural networks. However, several sound events do not satisfy this assumption in real-life situations where human care robots are utilized. Due to the recent proliferation of various digital devices, users can customize sound events with any kind of pre-recorded sounds they want such as music, animal sound, sound of nature, speech, etc. When we hear these pre-recorded sounds, it could easily confuse us if those are from video clips from Youtube, or they are phonebell, or alarm sound. Thus, in real-life situations, there are custom sound events that have different acoustic characteristics from those we already know, or that have similar acoustic characteristics to other sound events.

Among various sound events that can assist elderlies, doorbell and phonebell can be easily customized by users. Generally known doorbells and phonebells tend to repeat simple melody or frequency patterns, and these general sounds for doorbells and phonebells are dominated in various sound event datasets utilized in DCASE challenges. However, in real-life situation, each individual has a different type of smartphone, and the phonebell sounds provided by each smartphone are more like music rather than monotonous sounds. In addition, in the case of doorbells, the doorbell sounds in apartments, which are the primary residence of Korean, have changed to simple music forms with melodies rather than simple patterns in recent years. Figure 1 shows example spectrograms of general sounds and custom sounds for doorbells and phonebells. In details, the general phonebell is the default ringtones on Apple smartphone, and the custom phonebells are music-like ringtones on Samsung and Google smartphones. As can be seen in the spectrograms, general doorbell and phonebell repeat simple frequencies and harmonics. On the other hand, custom sounds appear as a combination of random and complex frequencies, similar to music. Rather, custom doorbells might look similar in frequency patterns to the general phonebell. This makes it difficult to find common acoustic features between custom sounds within the same class.

Moreover, it remains to be seen how conventional SER models recognize custom sounds with different acoustic characteristics from those trained. In this study, in order to focus on the influence of custom sound events, the SER model is implemented of ResNet-34-based model, which is mainly used in the deep learning-based recognition tasks. We select four sound events consisting of custom sound events (doorbell and phonebell) and sound events with consistent frequency characteristics (cough and snore). The SER model trains with Audioset [37], FSD50K [38], and other datasets utilized in DCASE challenges. The test dataset to verify the model performance is directly collected indoor environment. The detailed information about training and evaluation processes are described in Section III. For trained SER model, a total of three tests are performed for each of general bell sound set, custom bell sound 1 set, and custom bell sound 2 set using the sounds shown in Figure 1. Cough, snore, and other sounds are the same for all three test sets. The test data is unseen data during model training, and the confusion matrices of event recognition results are shown in Figure 2. The SER model achieves high average F-score of 0.946 for the test dataset with general bell sounds. However, the model shows average F-scores of 0.580 and 0.576 for the test datasets with custom bell sounds 1 and 2, respectively. While cough and snore result in high and consistent F-scores, the SER model shows F-scores of zero for the custom doorbell 1 and 2, 0.450 for the custom phonebell 1, and 0.432 for the custom phonebell 2. The reason for performance degradation is that the conventional SER models are *overgeneralized* to general sounds, and the custom sounds do not share the similarity of acoustic characteristics between general sound samples within the same event. Cough and

snore have common acoustic characteristics regardless of who makes those sounds, so the SER model is generalized to recognize common characteristics and shows good recognition performance. These results indicate that the recognition performance is low for custom sound events that are typically encountered in real-life situations. Due to inaccurate recognition performance of custom sounds, human care robots cannot provide accurate services to users who customize sound events. Therefore, the research question is how to enable human care robots to recognize any custom sound accurately in real time.

B. RESEARCH OBJECTIVES

Prior to this study, there have been reports on custom sound events [31], but there has been no research on sound event recognition systems that can automatically respond to the custom sound events. Thus, we propose an automatic sound event recognition system that can recognize custom sounds in terms of practical applications for human care robots. The custom sound events can be set to any pre-recorded sound the user wants, so it is not feasible to generalize the acoustic characteristics of custom sounds using data-driven methods. So, we focus on the specificity of the environment in which human care robots are utilized. A human care robot usually interacts with one specific user in single specific space: user's home. The specific user may customize the sounds for several sound events, but from the robot's perspective, it only needs to be aware of the custom sounds set by that specific user, not all possible custom sounds. Thus, we aim to customize the sound event recognition system of robot for the specific user to accurately recognize custom sounds. However, this customization process should be automatic. The reason is that users of human care robots are not experts in robotics or sound event recognition, so it is difficult to customize the robots by themselves. Even if experts are involved, it is difficult to access the robots due to issues such as personal privacy. Therefore, in this paper, we propose a real-time sound recognition system with customization process for human care robot. The main contributions of this work are as follows:

- Recognition of custom sound events which composed of pre-recorded sound events upon user's preference is defined as research question in this paper; and corresponding recognition algorithm, human care robot implementation and living room validation are presented.
- The proposed sound recognition algorithm with customization process using specific user's custom sounds improves F-scores by average of 64.2% compared to the conventional recognition model.
- Real-time sound recognition system with customization process is proposed and implemented in human care robot. Predefined data-acquisition, signal processing, and inference actions are executed by the robot in a real time fashion.
- Through a pilot test in real domestic environments (Korean elderly apartments), proposed real-time custom sound recognition system showed 0.982 of average F-score, confirming feasibility of the proposed sound recognition system.

The remainder of the paper is organized as follows. Section II introduces customization methods of the SER model mounted on a human care robot and the real-time sound recognition system with customization process. Section III describes dataset, training details with evaluation metrics, and pilot test setting. Section IV shows the experiment results, and Section V discusses the results. Finally, Section VI presents conclusion and future works.

II. REAL-TIME SOUND RECOGNITION SYSTEM WITH CUSTOMIZATION PROCESS FOR HUMAN CARE ROBOT

In this section, we propose two customization methods of SER model and real-time sound recognition system with customization process for human care robot. Main target of human care robot in this work is to assist elderlies using auditory cues, so we select four sound events that are difficult to recognize using visual cues: cough, snore, doorbell and phonebell.

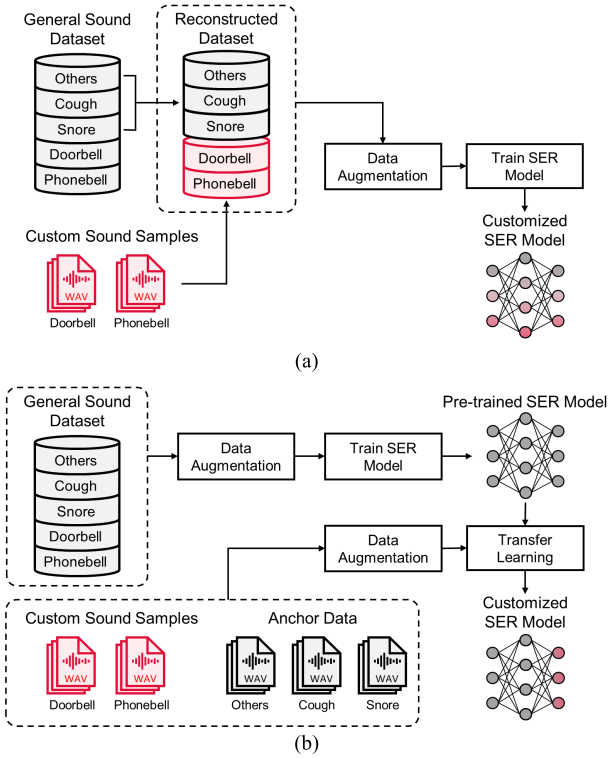


FIGURE 3. Flowcharts of (a) overfitting-based customization and (b) transfer learning-based customization methods. Custom sound samples are recorded by a human care robot in a specific user's environment, and these customization processes assume the situation after recording.

A. CUSTOMIZATION METHODS FOR SOUND EVENT RECOGNITION MODEL

1) OVERFITTING-BASED CUSTOMIZATION

As mentioned before, the human care robot need only provide accurate event recognition and services for a specific user.

The pre-recorded sound of a custom sound event doesn't change gradually like a speech changes with age or disease, but rather drastically with user preferences or electronic device replacement. The custom sounds are usually fixed for long time, and custom sound events do not change over time, so recognition performance is also independent of time. Thus, it is sufficient for the robot to recognize specific sounds for events selected by the user, and human care robot does not usually need to consider other sounds events for the same events. The first customization method starts from the idea that accurate recognition of custom sounds can be performed if the SER model is overfitted to custom sounds of a specific user. In detail, the SER model is trained using only custom sound samples for the specific user, excluding general sounds for custom sound events. The flow chart of this overfitting-based customization process is shown in Figure 3-(a). The overall training process is same as the conventional SER training process, but the composition of training dataset is different. For cough, snoring, and others, which are non-custom sound events, datasets are constructed using general sound samples. For doorbell and phonebell, which are custom sound events, datasets use sound samples recorded up to 10 seconds of sounds used by a specific user. In this case, since the training data imbalance between custom sound events and non-custom sound events is severe, the custom sounds are replicated by the number of samples in each non-custom sound event. For robustness of the network to external noise, on-line data augmentation is applied to these data, which convolves the room impulse responses and adds various external noises at every iteration, and the SER model is trained on the reconstructed dataset with on-line data augmentation (augmentation details in Section III). Since the model is trained using only specific custom sounds, it is overfitted to these sounds, resulting in accurate recognition and eventually customized to a specific speaker.

2) TRANSFER LEARNING-BASED CUSTOMIZATION

The overfitting-based customization method is intuitive, but it differs from how humans perceive and respond to custom sound events. Humans know the general acoustic features for each sound event like the conventional SER models. When humans hear a sound for a custom sound event, the sound is recognized as another sound event, such as music. Unlike conventional SER models, after misrecognition, humans learn these specific sounds as doorbell and phonebell classes, and recognize the same sound as these events for the subsequent occurrence. Humans adjust their SER models based on information about custom sound events, which is similar to the process of fine tuning in deep learning-related tasks. We propose transfer learning-based customization method that mimics human reaction to custom sound events, and the flow chart is shown in Figure 3-(b). There is a pre-trained SER model using a dataset composed of general sound samples. This pre-trained model does not yet have the ability to recognize custom sounds. The structure of SER model consists of convolution-based layers that extract a sound event-related feature vector from the input spectrogram, and a linear layer-based classifier that classifies events from the feature vector. The convolution-based layers serve as a general feature extractor that allows acoustic event information to be revealed well, so the parameters of the convolution-based layers are fixed to prevent overfitting and training instability during transfer learning process [39]. The parameters of linear layer-based classifier are tuned with pre-recorded custom sounds of the specific user. However, if the model is tuned with only custom sound samples, the model will be trained to improve performance on custom sound events without gaining recognition performance on non-custom sound events (cough and snore). To maintain performance for non-custom sound events after transfer learning, we utilize anchor data of non-custom events for tuning together. The anchor data is selected as data with high probability score that pre-training model correctly recognizes among the training data of non-custom events. For a balanced dataset, custom sound samples are recorded with the same anchor data length as the non-custom events. During the training of the network, on-line data augmentation is applied to the dataset, and the linear layer-based classifier is tuned with pre-recorded custom sound data and anchor data of non-custom sound events. Since the transfer learning-based customization method is not to train the entire model, but to tune the model with a small dataset, it is possible to customize the SER model faster than overfitting-based customization method.

B. REAL-TIME SOUND RECOGNITION SYSTEM WITH CUSTOMIZATION FOR HUMAN CARE ROBOT

Human care robot should recognize sound events in real-time to immediately provide appropriate services for various situations. We propose a real-time sound recognition system for human care robot using the customized SER model. The SER model is designed to recognize sound events for 2-second-long audio input in order to secure high recognition performance and less computation cost considering real-time processing. Since SER model is trained to perform SER on 2-second-long audio input, continuously received microphone input should be made to 2-second-long segments. A simple and intuitive method is to execute following loop: collect microphone input until its length reaches 2 seconds, then feed the 2-second-long segment to SER, then discard the received audio and collect microphone input again from scratch. However, this method causes two problems. First, it might fail to recognize short sound event happened just around the boundaries of two adjacent audio segments. Second, the maximum possible delay between sound event happening and the prediction notification is 2 seconds, which is quite long for a real-time operation. To mitigate these problems, we utilize a method to overlap adjacent input segments fed to SER model. When we apply $ol\%$ overlap, it means last $ol\%$ of i-th input segment is reused as the first $ol\%$ of $(i+1)$-th input segment. For example, when we apply 75% overlap, an input segment is obtained from $0 \sim 2$ seconds

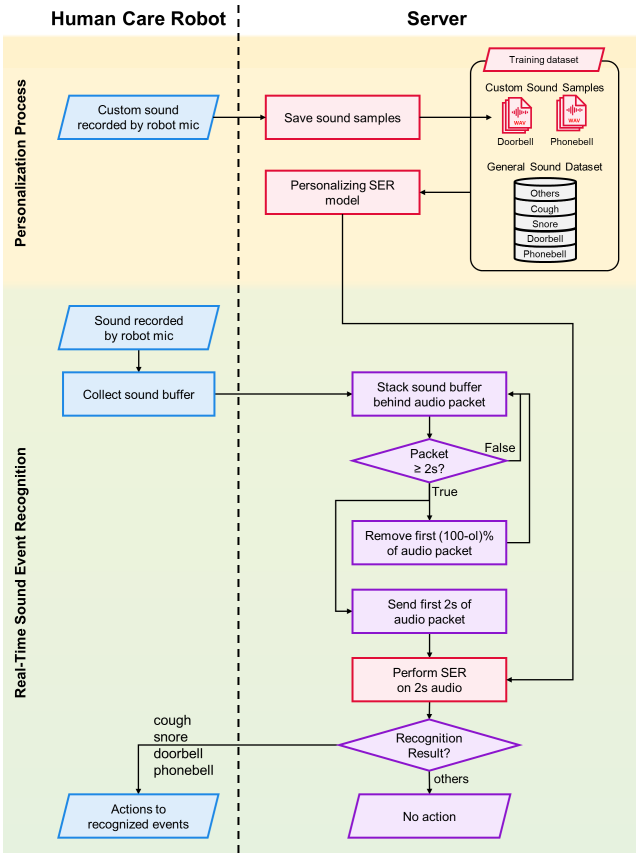


FIGURE 4. Flowchart of real-time sound recognition system for human care robot.

of microphone input and then the next input segment is obtained from 0.5 ∼ 2.5 seconds of microphone input so that the adjacent inputs are overlapped by 75% corresponding to 0.5 ∼ 2 seconds of microphone input. The overlap proportion between subsequent inputs, *ol,* can be set between 0% and 100%. This method could effectively enhance robustness of SER and recognize events faster and more frequent.

The overall process of the real-time sound recognition system with customization is described in Figure 4. We aim to develop a human care robot system under the assumption; the human behaves as natural as possible in the domestic environment using their ordinary devices. So, we proposed the system that provides appropriate services through only human care robots without any additional devices in the user's living environment. The system is composed of a human care robot and a server with GPU processor. The human care robot records sound around it using microphone and executes actions depending on SER results. The server receives audio data from the human care robot and runs SER on the received audio data. First, when the human care bot enters a user's living space, it begins the process of customizing the SER model. The human care robot records custom sounds of the user and sends them to the server. This study does not address how the user and the human care bot communicate naturally to record the custom sound samples. Instead, the user simply plays the doorbell and phonebell to the human care robot, and the robot would record the sounds. The length of recording depends on the customization methods. The robot records each custom sound event up to 10 seconds for overfitting-based customization method and the same length as the anchor data for transfer learning-based customization method. The server uses these recorded custom sound samples as a dataset to customize the SER model. Customizing SER model adopts either overfitting-based or transfer learning-based method. After the customization process, the human care robot starts real-time sound event recognition, and audio data is recorded through the robot's microphone. The robot stacks audio buffer with pre-defined length, then send it to the server continuously. Then, to match the length of SER input audio data to be 2-second-long, the server stacks audio buffers until it obtains an audio packet longer than 2 seconds. When the audio packet exceeds 2 seconds, the server gets first 2 seconds from the audio packet and feed it to SER. Meanwhile, the server removes first (100 - *ol*)% of the audio packet and then stacks sound buffer from the robot again for next SER inference. The server performs SER on the 2-second-long audio segment using stored customized SER model. If the recognition result is one of the four events (cough, snore, doorbell, phonebell), the proposed system generated pre-programed intelligent actions. When the robot recognizes the bell sounds, the robot reminds the user that a guest or phone call is coming using generated voices in the user's language. When the robot recognizes a cough sound continuously, the robot asks the elderly person whether he or she is sick using hand gestures and language expressions. On the other hand, if the result is ''others'', then no action is executed. All these corresponding actions are programed in the Pepper using dedicated program on robot operating system (ROS). The proposed human care robot system automatically performs the customization process and real-time recognition to minimize intervention between expert and user. In addition, the collection of custom sound samples is performed by the human care robot without experts. Since custom sound events have less variation in acoustic characteristics for a fixed user, customization can be enabled with only a small number of custom sound samples. So that, this data collection process does not require a lot of time and expert intervention compared to conventional SER systems without customization.

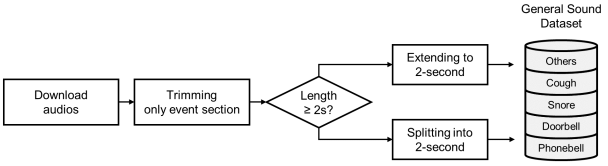


FIGURE 5. Flowchart of training dataset preparation.

III. EXPERIMENTAL SETUP

A. DATA PREPARATION

For general sound dataset with four sound events (cough, snore, doorbell and phonebell), we collect and refine

audio data from various large-scale audio datasets. Google AudioSet [37] is a large-scale dataset with manually annotated 10-second sound event clips gathered from Youtube videos. FSD50K [38] is the largest fully-open dataset of human-labeled sound clips gathered from Freesound. The SINS dataset [40] contains a continuous recording of one person living in a vacation home over a period of one week. Sound clips belonging to four events and others classes are collected from these datasets in this work. Dishwashing, glass breaking, knock, door slam, toilet flush, TV sound, cooking, breathe, and speech occur frequently in domestic environments, so we include them in 'others' class of the dataset so that SER models could learn that they are not target sound events. The overall process for collecting data and constructing training dataset is shown in Figure 5. First, we download audio data for five sound events from the sound event datasets. However, the audio data from Audioset are weakly labeled, and the time locations of sound events in sound clip are unavailable. As we aim to train SER model with 2-second inputs for real-time operation, we need time location of sound events to make sure the 2-second inputs, obtained by trimming 10-second audio clips, include target sound events. For cough, we use time locations labeled in [41]. This time location labelling is qualified to be appropriate as it showed reasonable cough recognition performance in previous studies [41], [42]. For snore, we find time location by ourselves by manually locating vibrating sound and excluding exhaling sound between vibrating sound which has too low energy. For doorbell and phonebell, we trim the data by manually listening to and locating the time location, and additionally collect from the DCASE challenge [43], [44]. All audio data is resampled to sampling rate of 16 kHz, and the length of trimmed sound event dataset precisely containing target sound events over all time span is shown in the 2nd column of Table 1. The SER model is trained with a spectrogram of 2-second inputs, so it is necessary to modify each sound clips in the training dataset into 2-second-long sound clips. Audio clips shorter than 2 seconds are extended to 2 seconds by padding random noise with -10 dB signal-to-noise ratio (SNR) from the start and end of audio clips, and audio clips longer than 2 seconds are shortened by randomly selecting 2-second-long interval within the clips. As this process is executed for each training epoch, the randomness is differently applied each epoch thus this process enhances robustness of the trained SER model and ensures consistent performance.

For validation of trained SER model's performance in the actual environment where human care robot is operated, we make a test dataset by recording cough, snore, phonebell, doorbell, and others sound event data in the real domestic environment (apartment rooms: living room, small and large bedrooms, kitchen) which is shown in Figure 6. The microphone is placed at six different points and the sound sources are within 2 meters around the microphone points to generalize dataset with various recording environments. In addition, the human care robot should recognize sound events regardless of position of the robot and the user, so we generate and record sound sources at various locations to make a test dataset that mimics this situation. Human sounds such as cough and snore are directly made by three people. Doorbell and phonebell sounds are each recorded with for two types of custom sounds to realize the actual environment where specific individual uses the human care robot. For others class, sounds of various events other than the four specific classes are recorded. Each sound event is recorded in 2-second-long audio data with sampling rate of 16kHz. The examples of recorded custom sounds are shown in Figure 1. Total length of test dataset composed of the recoded sounds is in 3rd column of Table 1, and they are used for test of trained SER model. In particular, the doorbell and phonebell are recorded for 48 seconds and 108 seconds for each sound type, so a total of 96 seconds and 216 seconds were recorded.

TABLE 1. Length in seconds of training and test dataset in each event class.

Event class	Training dataset (trimmed)	Test dataset
Cough	9,344	194
Snore	5,198	94
Doorbell	1,744	96
Phonebell	5,859	216
Others	62,083	600

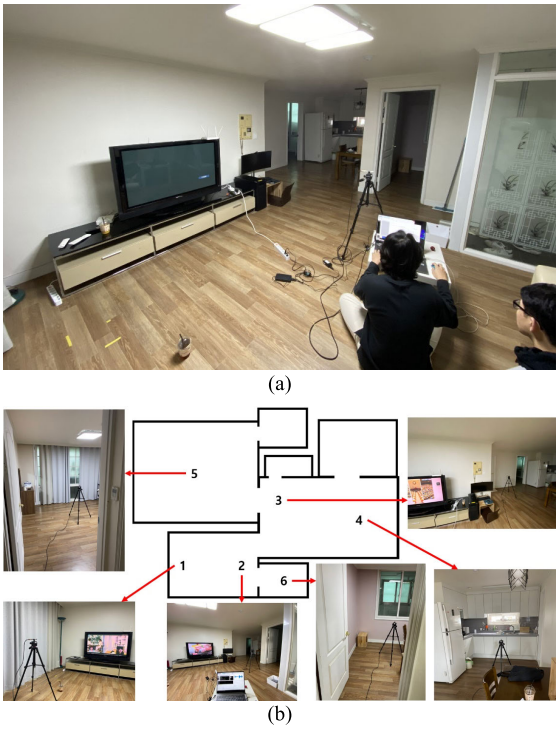


FIGURE 6. (a) An apartment used as real domestic environment where test dataset for SER model is recorded. (b) 6 locations where microphone and sound sources were placed during recording.

B. DATA AUGMENTATION AND FEATURE EXTRACTION

Since our goal is to test practicality for application of sound recognition system on human care robot, we need to test SER model that is implemented on a robot. However, there are various other noise sources interfering sound recognition system. The robot tends to produces noises from mechanical and electronic parts such as fans, servomotors, etc. In addition, it intentionally makes various sounds such as alarm and artificial speech through an installed speaker for interaction with users. Other than that, the domestic environments where the robots are used involve various background noises. These noises cause recognition error in the trained SER model. Also, domestic environment induces various types of reverberation due to different room structures shown in Figure 1. To solve these problems, various data augmentations are applied to give training data the acoustic environment similar to actual robot-running environment. We utilize a total of 30 minutes of domestic and office noise audio data from the DEMAND dataset [45] as background noise, and the noise is added to training data with a random SNR between 5 and 15. Reverberation is applied to training data by convolving sound event data with one of a total 60,600 simulated room impulse responses for small, medium, and large rooms [46]. In each epoch, additive noises and room impulse responses are randomly selected for each data. The examples of background noise and room impulse response are shown in (a) and (b) of Figure 7. By convolving the room impulse response and adding the background noise to the cough sound data as shown in (c) of Figure 7, the result is audio that is similar to acoustics in a real-world environment, as shown in (d) of Figure 7. Thus, all data is applied with different noise and reverberation every epoch, increasing the model's robustness.

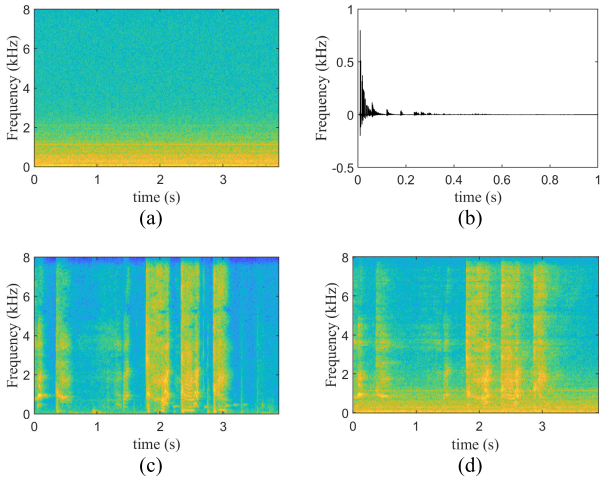


FIGURE 7. Examples of data augmentation process. (a) is spectrogram of background noise in conference room, (b) is an impulse response in large room, (c) is a spectrogram of cough sound, and (d) is a spectrogram of cough sound with data augmentation by convolving room impulse response and adding background noise with SNR of 6dB.

After data augmentation, we extract 64-dimensional log Mel-spectrograms from the 2-second segment using hamming window of width 25ms with step 10ms and number of fast Fourier transform 512. Mean and variance normalization is applied on every frequency bin of the Mel-spectrogram. Normalized Mel-spectrograms are utilized as input features of SER model.

C. SOUND EVENT RECOGNITION NETWORK ARCHITECTURE

In this work, we select ResNet-34 [47] with Squeeze-and-Excitation blocks [48], which is state-of-the-art network in sound event recognition tasks [21], [49], [50] and speaker recognition tasks [20], [51], [52], [53], in order to focus only on the custom sound events and customization methods. The detailed structure is described in Table 2. The channel sizes are reduced to half from the original model for the computational efficiency. It also has temporal average pooling layer to aggregate frame-level features. The last layer is a fully connected layer with four nodes corresponding to four sound classes (cough, snore, doorbell and phonebell). After the last layer, sigmoid function is applied to make result nodes' value to be between 0 and 1, and each node represents presence probability of each sound class. ''Others'' class refers to background sound and other sound events, and we classify output as ''others'' when none of four nodes are active.

D. DETAILS ON SER MODEL TRAINING AND EVALUATION METRIC

SER models are trained on NVIDIA TITAN RTX using PyTorch [54] in the server computer. Adam optimizer [55] is used with weight decay of 5×10^{-5} and initial learning rate of 0.001 which is decreased by a factor of 0.95 every epoch. Binary cross-entropy is used to train SER models to make them recognize sound events independently. Batch size is fixed as 200, and the models are trained for 50 epochs. These hyperparameters are optimized to achieve the best recognition performance of the conventional SER model for general sounds. For fine tuning the model, parameters of other layers except linear layers (Linear layer 1 and Linear layer 2) are frozen. The linear layers are tuned for 100 iterations using binary cross-entropy loss with Adam optimizer. The initial learning rate is 0.001, and it is decreased by a factor of 0.95 every iteration. During one iteration, the batch includes all anchor data and all custom sound samples.

For evaluation of the SER models, a sound event class is predicted to exist in the input audio clip when the node corresponding to that event yields value exceeding predetermined threshold value. When none of four result nodes' values exceeds the threshold, none of target sound event class exists thus the result becomes ''others'' class. We set the threshold as 0.5 which is most commonly used values. Trained and tuned SER models are evaluated with the test dataset, and the performance is measured using precision, recall and F-score. The precision, recall, and F-score are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

TABLE 2. Configuration of ResNet-34 with squeeze-and-excitation for sound event recognition.

Layer	Structure	Output size (C×H×W)
Input	-	1×64×200
Conv1	Conv(3 × 3, 32)	32×64×200
Conv2	$\begin{bmatrix}\text{Conv}(3\times3,32)\\ \text{Conv}(3\times3,32)\\ \text{SE Layer}\end{bmatrix}\times3$	32×64×200
Conv3	$\begin{bmatrix}\text{Conv}(3\times3,64)\\ \text{Conv}(3\times3,64)\\ \text{SE Layer}\end{bmatrix}\times4$	64×32×100
Conv4	$\begin{bmatrix}\text{Conv}(3\times3,128)\\ \text{Conv}(3\times3,128)\\ \text{SE Layer}\end{bmatrix}\times6$	128×16×50
Conv5	$\begin{bmatrix}\text{Conv}(3\times3,256)\\ \text{Conv}(3\times3,256)\\ \text{SE Layer}\end{bmatrix}\times3$	256×8×25
Flatten	-	2048×25
Temporal Pooling	Average pooling	2048
Linear layer 1	FC(256)	256
Linear layer 2	FC(4)	4

Conv denotes convolution, and numbers inside parentheses of Conv refers to (size × size of kernel, number of channels). SE denotes squeeze-and-excitation module. FC denotes the fully connected layer with the number of output nodes. Nonlinear function ReLU and batch normalization are applied after every convolution. ASP denotes attentive statistical pooling. Nonlinear function ReLU and batch normalization are applied after every convolution.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

where TP, FP, and FN denote the total counts of true-positives, false-positives, and false-negatives, respectively [56]. The result is true-positive when the label and the system output both indicate the same event to be active. The result is false-positive when the system output indicates an event class as active while the label indicates as inactive. The result is false-negative when the system output indicates an event class as inactive while the label indicates as active. The precision, recall, and F-score are calculated for four event classes only because others class refers to situations other than the specific four events thus true-positive does not exist for others class. We compare the SER performance for each event with precision, recall, and F-score, and the overall performance of SER model with class-averaged F-score.

E. HUMAN CARE ROBOT SYSTEM IN INDOOR ENVIRONMENT

In order to verify the proposed system for human care robot, we implement the real-time sound recognition system with customization process in a real domestic environment using humanoid robot 'Pepper' as shown in Figure 8. Pepper has four microphones with frequency range from 100Hz to 10kHz and collects audio data from one center microphone for fast operation. The server is equipped with Intel® Xeon® Silver 4116 CPU and NVIDIA® TITAN RTX GPUs is responsible for overall data processing. The human care robot and the server exchange data with each other through wireless communication. Based on the proposed system, the human care robot recognizes sound events including custom sound events. The pilot test environment is constructed similarly to real environment by adding speech and other real-life noises. We make sound events in-situ instead of playing recorded event sounds to ensure that the test reflects a real situation as much as possible.

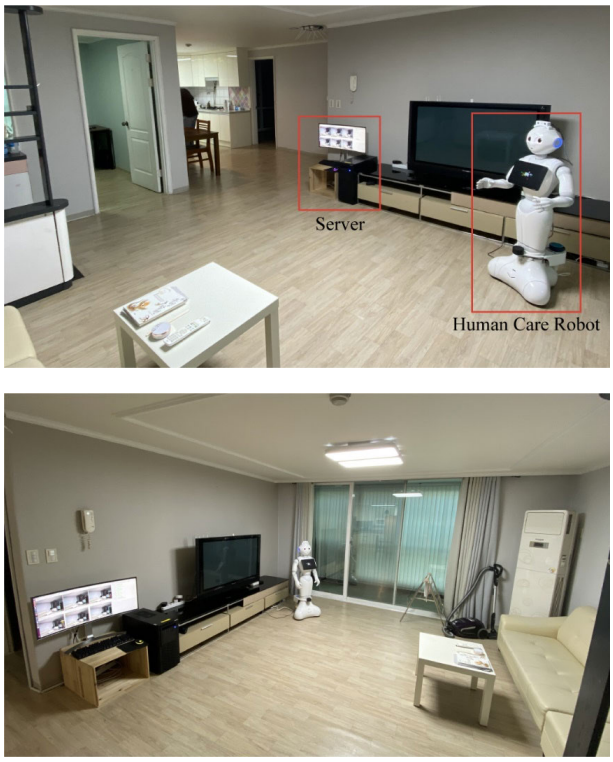


FIGURE 8. Real-time sound event recognition system in domestic environment. Human care robot and server are shown in the pictures. The customized SER model was mounted on the server.

IV. EXPERIMENTAL RESULTS

A. COMPARISON OF PERFORMANCE AND COMPUTATIONAL TIME BETWEEN CUSTOMIZED SER MODELS

We compare the sound event recognition performance of conventional model and customized models based on overfitting and transfer learning. The SER models are tested on a total of two test datasets. The first test dataset consists of the same types of sounds as custom doorbell 1 and custom phonebell 1 in, and the second test dataset consists of the same types of sounds as custom doorbell 2 and custom phonebell 2 in Figure 1. The data for cough, snore, and others classes are the same. Precision, recall, and F-score of the SER models

TABLE 3. Sound event recognition performance of conventional SER model and customized SER models for test dataset containing custom bell sound 1 and 2 sets.

Class	Metric type	Conventional model	Overfitting-based customized model	Transfer learning-based customized model			
				Number of anchor data for each class			
				10	20	30	40
Cough	Precision	0.988	0.933	1.000	0.964	0.953	0.953
	Recall	0.920	0.955	0.852	0.920	0.932	0.932
	F-score	0.953	0.974	0.920	0.942	0.943	0.943
Snore	Precision	0.882	0.956	0.878	0.872	0.911	0.909
	Recall	0.957	0.915	0.766	0.872	0.872	0.851
	F-score	0.918	0.935	0.818	0.872	0.891	0.879
Custom Doorbell 1	Precision	0.000	1.000	1.000	1.000	1.000	1.000
	Recall	0.000	0.926	0.875	0.833	0.875	0.875
	F-score	0.000	0.962	0.933	0.909	0.933	0.933
Custom Doorbell 2	Precision	0.439	0.959	0.836	0.981	0.964	0.981
	Recall	0.463	0.974	0.944	0.963	0.981	0.981
	F-score	0.450	0.967	0.887	0.972	0.972	0.981
Custom Doorbell 2	Precision	0.000	1.000	0.952	0.955	1.000	1.000
	Recall	0.000	1.000	0.933	0.875	0.875	0.917
	F-score	0.000	1.000	0.889	0.913	0.933	0.957
Custom Phonebell 2	Precision	0.559	1.000	0.846	1.000	1.000	1.000
	Recall	0.352	1.000	0.815	0.815	0.852	0.852
	F-score	0.432	1.000	0.830	0.898	0.920	0.920

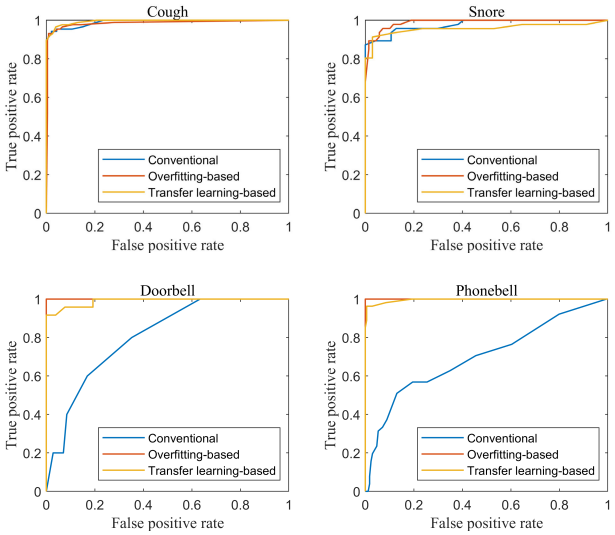


FIGURE 9. Receiver operating characteristic (ROC) curve of sound event recognition models for cough, snore, doorbell, and phonebell. The models tested on the dataset containing custom bell sound 2 set.

for overall dataset are listed in Table 3. The conventional SER model correctly recognizes events with high F-scores of 0.953 and 0.918 for cough and snore, but it shows relatively poor performances with low F-scores for doorbell and phonebell. This result indicates that the conventional SER model rarely recognize custom sounds. In contrast, the overfitting-based customized SER models outperforms the conventional SER model by achieving precision, recall, and F-score of 1.000 or close to 1.000 for doorbell and phonebell. For coughing and snoring, the scores are about 0.9 or higher, showing high performance level similar to the conventional model. Meanwhile, we compare the performance of the transfer learning-based customized model depending on the number of anchor data. The transfer learning-based customized model is a retrained version of the conventional model's classifier layers that improves precision and recall of conventional model for doorbell and phonebell, achieving overall F-scores of 0.95. However, when the number of anchor data is reduced to 10, the recognition performances for doorbell and phonebell decreases. In addition, the transfer learning method actually decreases the precision of cough and recall of snore for the conventional model that is originally good at recognizing cough and snore, resulting in a lower F-score. As the number of anchor data increase, the recall of cough and the precision of snoring improve, and these metrics converge from over 30 anchor samples. Finally, the transfer learning-based customized model with 40 anchor data achieves a good average F-scores of 0.934 for custom bell sound 1 set and 0.938 for custom bell sound 1 set. Thus, both overfitting-based and transfer learning-based customized models show better recognition performance for doorbell and phonebell than the conventional SER model. The overfitting-based customized model outperforms the transfer learning-based customized model using 40 anchor data for coughing and snoring, so it achieves the best performance with average F-scores of 0.955 and 0.969 for test datasets, respectively. Moreover, we also compare the receiver operating characteristic (ROC) curves of SER models to confirm how sensitive the model is to recognizing each event. The ROC curve shows the correlation between the true-positive rate and the false-positive rate when the threshold to determine whether or not each event has occurred

TABLE 4. Computation time comparison of different customization methods.

Methods		Computation time (s)
Overfitting-based customization		9,169.39
Transfer learning-based customization	Anchor = 10	43.55
	Anchor = 20	53.10
	Anchor = 30	63.84
	Anchor = 40	74.69

TABLE 5. F-scores with different overlap proportion in pilot test.

Class	Overlap Proportion		
	0%	50%	75%
Cough	0.909	0.966	1.000
Snore	0.723	0.868	0.929
Doorbell	1.000	1.000	1.000
Phonebell	1.000	1.000	1.000
Average	0.908	0.959	0.982

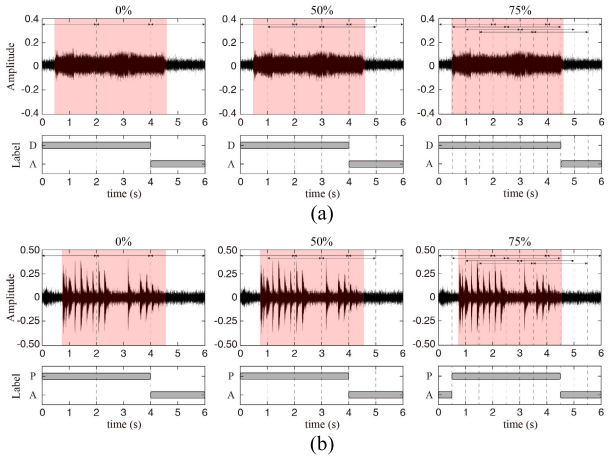


FIGURE 10. An illustration of input audio signals and real-time sound event recognition results of (a) doorbell and (b) phonebell with various overlap proportions of 0%, 50%, and 75%. The red area indicates the sound event occurrence section, and the dotted lines indicate the boundary of sound buffers stacked in audio packets. Every one buffer in the case of 0% overlap, every two buffers in the case of 50% overlap, and every four buffers in the case of 75% overlap are defined as 2-second-long packets indicated by solid arrows. Label 'P' denotes phonebell, label 'D' denotes doorbell, and label 'A' denotes absence.

is between 0 and 1. The curves of the conventional model and the customized models tested on dataset containing custom bell sound 2 set are shown in Figure 9. The customized models have area under curves (AUCs) almost 1.0 for all events, which means they are close to perfect classifiers that correctly distinguish the probability of each event being present into only 1 or 0. However, the conventional model has AUCs of 0.806 and 0.701 for doorbell and phonebell, while AUCs almost 1.0 for cough and snore. These results indicate that the customized methods improve classification ability for doorbell and phonebell while maintaining perfect classification ability for coughs and snoring with AUCs of almost 1.0. These results of ROC curves and AUCs are the same as the comparison with the previous three precision, recall, and F-score.

The customization process takes place after the human care robot enters into a specific environment and collects custom sound samples, so it is necessary to compare not only sound event recognition performance but the computation time required for the customization process. We derive the computation time of the overfitting-based and the transfer learning-based customization methods depending on the number of anchor data, as shown in Table 4. For transfer learning-based customization, the computation time increases as the number of anchor samples increases, and it took 74.69 seconds using the 40 anchor samples. Unlike this, the overfitting-based customization method trains the model with the reconstructed dataset instead of just a few samples, resulting in a computational time that is about 123 times as long. The size of reconstructed training dataset is the same as the dataset for training the conventional SER model, so the computation time is almost the same as about 9,213 seconds required to train the conventional SER model. In summary, the overfitting-based method has the best performance, but it takes a long computation time to customize, while the transfer learning-based method has the trade-off of being less performant but can be customized in less computation time. We focus on the sound recognition performance and utilize the overfitting-based method for the real-time acoustic event recognition system, with a detailed discussion of the reasons for this choice in Section V.

B. REAL-TIME SOUND RECOGNITION SYSTEM DEPENDING ON OVERLAP PROPORTION

We examine how SER performance changes with different overlap proportions and determined the optimal value of the overlap proportion. The pilot test is conducted to evaluate the performance of the proposed system in a real domestic environment. The test is performed by making 30 sounds for each sound event with the real-time sound recognition system in the domestic environment shown in Figure 8. The system is tested with different overlap proportions (0%, 50%, and 75%), and corresponding F-scores are listed in Table 5. For doorbell and phonebell sound events, the proposed real-time SER system shows F-score of 1.000 regardless of the overlap settings. We display the pilot test results of doorbell and phonebell depending on the overlap proportions, and the results are shown in Figure 10. There are 2-second-long packets completely consisting of doorbell or phonebell, and the events are accurately recognized for these sections. In general, doorbell and phonebell are longer than 2 seconds because they are mainly melodic sounds. When doorbell or phonebell rings, SER model would receive at least one 2-second segment that fully contains the acoustic characteristics of bell sound. So that, the proposed system

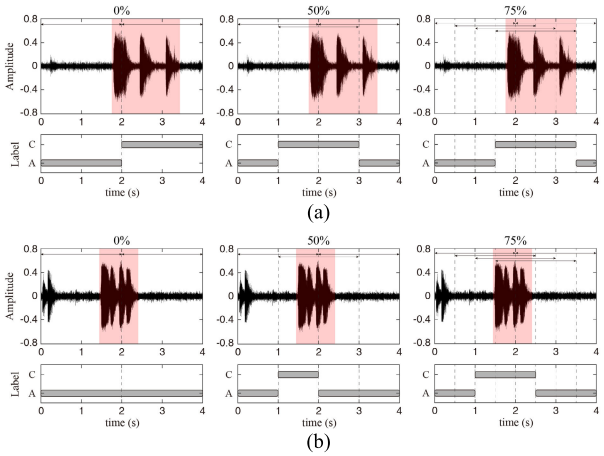


FIGURE 11. An illustration of input audio signals and real-time sound event recognition results of cough with various overlap proportions of 0%, 50%, and 75% in (a) well-recognized case and (b) poorly-recognized case. The red area indicates the sound event occurrence section, and the dotted lines indicate the boundary of sound buffers stacked in audio packets. Every one buffer in the case of 0% overlap, every two buffers in the case of 50% overlap, and every four buffers in the case of 75% overlap are defined as 2-second-long packets indicated by solid arrows. Label 'C' denotes cough and label 'A' denotes absence.

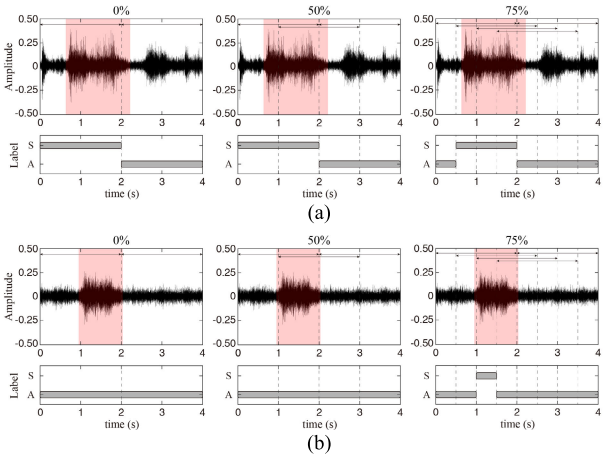


FIGURE 12. An illustration of input audio signals and real-time sound event recognition results of snore with various overlap proportions of 0%, 50%, and 75% in (a) well-recognized case and (b) poorly-recognized case. The red area indicates the sound event occurrence section, and the dotted lines indicate the boundary of sound buffers stacked in audio packets. Every one buffer in the case of 0% overlap, every two buffers in the case of 50% overlap, and every four buffers in the case of 75% overlap are defined as 2-second-long packets indicated by solid arrows. Label 'S' denotes snore and label 'A' denotes absence.

recognizes doorbell and phonebell regardless of overlap proportions. However, the accuracy of onset and offset, which are directly related to appropriate service provision time, is different depending on overlap proportions. The result of 75% overlap proportion determine the onset and offset within a smaller margin of error compared to the actual onset and offset than the results of 0% and 50% overlap proportions. Since the system with 75% overlap proportion recognizes sound events every 0.5 second, the error of onset and offset is within 0.5 second, which is consistent with the result. Thus, 75% overlap proportion is suitable for real-time recognition of doorbell and phonebell in terms of accurate recognition and onset/offset determination.

On the other hand, cough and snore sound events show relatively low performances when overlap is not applied. For snore sounds, we define the inhalation region, where the sound is caused by nasal friction occurs in the entire sound, as the snore event, and perform event recognition and time localization. As overlap proportion increases, the F-score increases drastically from 0.909 to 1.000 and from 0.723 to 0.929. We display the well-recognized and poorly-recognized cases among the pilot test results depending on the overlap proportions, and the results are shown in Figure 11 for cough and Figure 12 for snore sound events. In the well-recognized case, whole sound events of cough and snore appear within 2-second-long packets, and the SER model accurately recognizes the events. However, in the poorly-recognized case, the events are truncated at the end of 2-second-long packets, and these packets are recognized as 'absence'. Actually, cough is an impulse-like sound, and it has a very short duration. Snore sound cannot be longer than time taken to breathe in, which is usually shorter than 2 seconds. Thus, there is a high probability that 2-second-long packets will contain short part of cough or snore events, rather than whole events. This leads to low real-time recognition performance in short overlap proportions. In addition, the systems with 0% and 50% overlap proportions recognize every 2 seconds and 1 second, so even if the events are recognized, the errors of onset and offset are large. For this reason, long overlap proportion is required for accurate real-time SER of cough and snore sounds, and we choose 75% overlap proportion which shows the best performance with average F-score of 0.982 and accurate onset/offset in this experiment.

C. PILOT TEST OF REAL-TIME SOUND RECOGNITION SYSTEM IN HUMAN CARE ROBOT

Finally, we test the proposed real-time sound recognition system, and the waveform of sound captured by the human care robot and the result of real-time SER using proposed real-time sound recognition system are shown in Figure 13. Illustrated example is recorded in the same test setting. A total of eight sound events are made in real-time, two of each for four sound event classes (cough, snore, doorbell, and phonebell) in single take. It can be seen that not only accurate SER but also reasonable onset and offset are indicated. Therefore, through the pilot test, we confirm that the proposed real-time sound recognition system for human care robot with customization process provides accurate real-time SER performance in real environment.

V. DISCUSSION

A. CUSTOMIZING SOUND EVENT RECOGNITION MODEL

The conventional sound recognition model shows poor recognition performance for custom sounds that differ from general

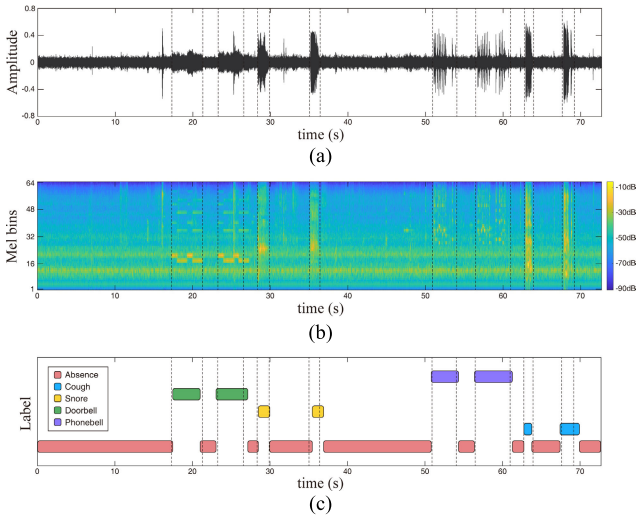


FIGURE 13. An illustration of the SER results over time. (a) a waveform plot of sound recorded by human care robot, (b) Mel-spectrogram of recorded audio data, and (c) sound event recognition results using proposed real-time sound recognition system for human care robot. The solid line indicates onset and offset points at which the event occurred and ended.

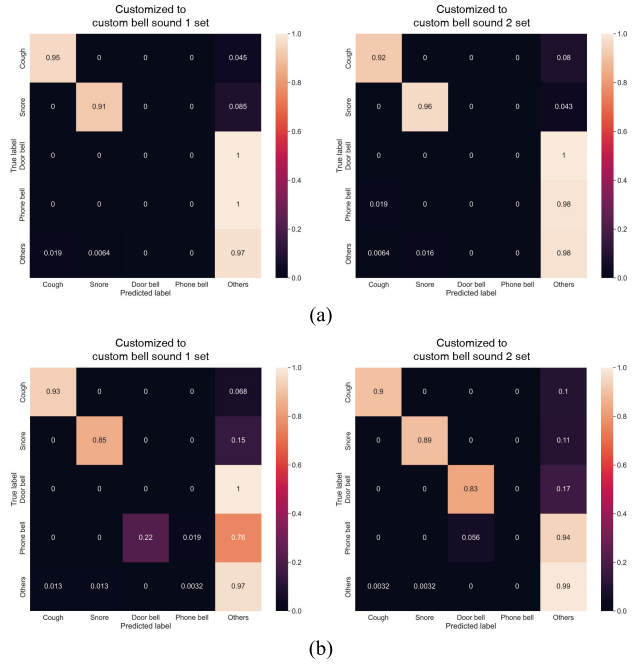


FIGURE 14. Confusion matrices of (a) overfitting-based customized models and (b) transfer learning-based customized models tested with dataset containing general doorbell and phonebell sounds.

acoustic characteristics of events. To improve the recognition performance for custom sounds, we propose two methods to customize the sound event recognition model depending on the user: overfitting-based customization and transfer learning-based customization. These two customization methods accurately recognize custom sound events while maintaining performance for general sound events. This result indicates that the proposed customization methods can help human care robots recognize custom sound events. For high quality human care robot, not only high recognition performance but also short customization time are required to provide accurate services quickly from the first time the robot enters the user's environment. The overfitting-based customized model showed the best recognition performance, but it requires much more customization time compared to the transfer learning-based customized models. The performance of transfer learning-based customized model can be improved by utilizing more anchor data, but over 30 anchor data, the performance increasement is minimal and rather only the computation time and custom sound samples collection time are constantly increased. There are advantages and disadvantages to each of these methods, making it difficult to choose the best one for human care robot. So, we focus on the exact meaning and purpose of customization. An important point of customization is not only to quickly adapt to a specific user, but also to respond only to the customized user. If the robot recognizes other sound samples containing general sounds as user's event, it risks providing services that the user doesn't want. We perform an additional recognition test with sounds other than those of specific user to determine whether the customized model has the ability to respond only to the customized user. The models customized to custom bell sound 1 and 2 sets are tested with general doorbell and phonebell as shown in Figure 14. The overfitting-based customized models recognize doorbell and phonebell sounds as others class, but the transfer learning-based customized models recognize the general bell sounds as custom sounds of specific user or other class. This means that the recall (related to false-positive) for the others class is close to 1 for overfitting-based customized models and lower for transfer learning-based customized models. While the overfitting-based customization method trains all layers of SER network on custom sound events, the transfer learning-based customization method trains only clarifier layers on custom sound events. The sound event feature extractor of transfer learning-based customized model is specialized for general sounds, so it lags behind the feature extractor of overfitting-based customized model in its ability to extract discriminative features for custom sound events. Thus, the transfer learning-based models have the potential to recognize other bell sounds as specific user's bell sounds. We believe that recognizing only custom sound events is important for the human robot to provide accurate services, so the overfitting-based customization method is applied to the real-time sound recognition system of human care robot. If a fast customization process is important, then the transfer learning-based customization method could be utilized in this case.

B. REAL-TIME SOUND EVENT RECOGNITION

For human care robots to perform accurate sound event recognition and time localization, we design the real-time sound recognition system using the segments overlapping method. Sound event recognition is performed by stacking sound buffers collected from the human care robot for

2 seconds and using the stacked packets as input to the customized SER model. In this process, recognition and time localization performance are compared when receiving segments every 2 seconds, 1 second, and 0.5 seconds with 0%, 50, and 75% of overlap proportions between packets, respectively. The results show that the recognition performances for doorbell and phonebell remain the best with an F-score of 1.0 regardless of overlap proportions. However, the performance of time-localization deteriorates because smaller overlap proportion results in longer intervals between recognizing events. Similarly, for cough and snore, the performance of time-localization degrades with a smaller overlap proportion, and furthermore, the recognition performance also degrades. The reason for this result is that cough and snore are events that appear for a shorter duration than 2 seconds, which is the length of recognition segment, and if the recognition cycle is longer, the events are present at the end of the segments and cannot be recognized. Doorbell and phonebell have sufficiently long event lengths that recognition performance is not affected by overlap proportion. Therefore, a high overlap proportion is required to improve the performance of time-localization for all events and to improve performance for short events. However, if the recognition cycle becomes shorter than the recognition time due to too high overlap proportion, real-time recognition cannot be achieved, so it is necessary to consider the specifications of the server computer together. In our system, it is confirmed through the pilot test that 75% of overlap proportion is sufficient to secure accurate event recognition and time-localization performance while maintaining a real-time process. These results indicate that human care robots can accurately recognize custom sound events and provide accurate services using the proposed real-time sound recognition system with customization.

C. OVERALL SOUND EVENT RECOGNITION SYSTEM WITH CUSTOMIZATION PROCESS FOR HUMAN CARE ROBOT

The proposed system objectively has the limitation that it requires additional time and effort in the processes of custom sound samples collection and customization compared to the conventional SER system. For customized sound data collection, we find that the SER models customized with a small number of custom sound samples achieved high performance due to custom sound events' acoustic characteristic, which have less variation for a fixed user. The robot only needs to collect a small number of users' custom sound samples. For customization process, we minimize user intervention by automating the customization process and the real-time sound event recognition process. So that, the minimum effort to collect users' custom sound samples is performed initially by the robot, and most processes are automated and rarely require an expert, which compensates for the limitations of the proposed method. Furthermore, the proposed system has the significant advantage of performing accurate recognition for custom sound events compared to the conventional SER systems. From the user's usability perspective, it is reasonable to provide more accurate services with less effort. Therefore, accurate custom sound recognition of the proposed system can fully compensate for its limitation.

VI. CONCLUSION

In this paper, a real-time sound recognition system with customization process for human care robot is proposed. There are gaps in acoustic characteristics between general sound events utilized in the conventional SER studies and custom sounds occurring in real-life environments where human care robots are applied. The custom sounds are events that users can change to various sounds such as music according to their preferences, and the acoustic characteristics are different to the general sound events. However, the conventional SER models were *overgeneralized* on the sound events, so the performance of conventional SER model is degraded for custom sound events. To address this problem, we propose the real-time sound event recognition system with customization process for human care robot. The overfitting-based and transfer learning-based customization methods utilize custom sound samples used by a specific user, taking advantage of the fact that the user and usage environment of the human care robot hardly change. Overfitting-based customized SER models significantly improve average F-scores compared to the conventional SER model from 0.580 and 0.576 to 0.955 and 0.969 for two test dataset containing custom bell sounds. Transfer learning-based customization SER models also improve recognition performance on custom sounds, but they lag slightly behind the overfitting-based customized models and have issues recognizing sounds as events that do not belong to the specific user. Since it is reasonable to recognize only the events of a specific user to provide accurate services in precise situation, we design a real-time sound recognition system with overfitting-based customization process. After the automatic customization process, the human care robot performs real-time event recognition using customized SER model. The robot sends audio buffers to server which stacks audio buffers into 2-second-long segments with 75% overlap, thus yielding audio segments every 0.5 second. The reason is that the system with 75% overlap showed high performance with F-scores of 1.000 and 0.929 for cough and snore which have impulsive acoustic characteristics with short-duration compared to 2-second. The server computer recognizes if sound events exist from audio segments using customized SER model. Proposed real-time sound recognition system show the best real-time recognition performance with average F-score of 0.982 in the pilot test. In addition, the pilot test verifies that proposed system accurately recognizes sound events including custom sound events in actual domestic environment. Therefore, the human care robot using the proposed real-time sound recognition system with customization process can understand situations of the user in the domestic environments and efficiently assist elderlies by providing appropriate services at the right time.

As future works, the proposed customization process and the real-time sound recognition system will be verified to

perform accurate recognition for a larger number of events. In addition, a big assumption of the proposed system is that a user and the custom sounds are rarely unchanged from the moment the human care robot is deployed, but in real-world application, problems may arise where custom sound events change or are added that deviate from this assumption. In this case, new custom sound samples need to be collected and customize the SER system anew. Thus, it is necessary to develop an automated process of re-customization through communication between the user and the human care robot in order for human care robot to address these issues by itself without external help.

REFERENCES

[1] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: A review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, Apr. 2009.

[2] Y. A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M. I. Oros-Flores, and D.-L. Almanza-Ojeda, "Human activity recognition using temporal convolutional neural network architecture," *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116287.

[3] S. Cebollada, L. Payá, M. Flores, A. Peidró, and O. Reinoso, "A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114195.

[4] D. Barry, M. Shah, M. Keijsers, H. Khan, and B. Hopman, "XYOLO: A model for real-time object detection in humanoid soccer on low-end hardware," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, New Zealand, Dec. 2019, pp. 1–6.

[5] E. Martinez-Martin and A. P. del Pobil, "Object detection and recognition for assistive robots: Experimentation and implementation," *IEEE Robot. Autom. Mag.*, vol. 24, no. 3, pp. 123–138, Sep. 2017.

[6] S. Chatterjee, F. H. Zunjani, and G. C. Nandi, "Real-time object detection and recognition on low-compute humanoid robots using deep learning," in *Proc. 6th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2020, pp. 202–208.

[7] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[8] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," 2017, *arXiv:1710.02997*.

[9] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, Sep. 2021.

[10] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–8.

[11] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036–5040.

[13] Y.-W. Wang, C.-P. Chen, C.-L. Lu, and B.-C. Chan, "Cht+ nsysu sound event detection system with multiscale channel attention and multiple consistency training for dcase 2021 task 4," DCASE Community, Tech. Rep., DCASE2021 Challenge, 2021.

[14] T. Na and Q. Zhang, "Convolutional network with conformer for semi-supervised sound event detection," DCASE Community, DCASE2021 Challenge, Tech. Rep., 2021.

[15] R. Lu, W. Hu, D. Zhiyao, and J. Liu, "Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios," DCASE Community, Tech. Rep., DCASE2021 Challenge, 2021.

[16] Y. Chen, "Convolution-augmented conformer for sound event detection," DCASE Community, Tech. Rep., DCASE2021 Challenge, 2021.

[17] H. Koo, H.-M. Park, J. Park, and M. Oh, "Sound event detection based on self-supervised learning of wav2vec 2.0," DCASE Community, Tech. Rep., DCASE2021 Challenge, 2021.

[18] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–8.

[19] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11027–11036.

[20] S.-H. Kim, H. Nam, and Y.-H. Park, "Temporal dynamic convolutional neural network for text-independent speaker verification and phonemic analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6742–6746.

[21] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," in *Proc. Interspeech*, Sep. 2022, pp. 2763–2767, doi: 10.21437/interspeech.2022-10127.

[22] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," 2021, *arXiv:2107.03649*.

[23] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617.

[24] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4308–4312.

[25] M. Janvier, X. Alameda-Pineda, L. Girinz, and R. Horaud, "Sound-event recognition with a companion humanoid," in *Proc. 12th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2012, pp. 104–111.

[26] S. Park, W. Choi, D. K. Han, and H. Ko, "Acoustic event filterbank for enabling robust event recognition by cleaning robot," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 189–196, May 2015.

[27] K. Nakamura and K. Nakadai, "Robot audition based acoustic event identification using a Bayesian model considering spectral and temporal uncertainties," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4840–4845.

[28] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 447–458, Mar. 2017.

[29] S.-H. Kim, S. Choi, H. Nam, and Y.-H. Park, "Deep learning-based personalized sound event recognition for human care robot," in *Proc. 24th Int. Congr. Acoust. (ICA)*, 2022.

[30] S. Chandrakala and S. L. Jayalakshmi, "Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 1–34, May 2020.

[31] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 335–371.

[32] R. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás, "HomeSound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring," *Sensors*, vol. 17, no. 4, p. 854, Apr. 2017.

[33] A. W. Ramadhan, A. Wijayanto, and H. Oktavianto, "Implementation of audio event recognition for the elderly home support using convolutional neural networks," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2020, pp. 91–95.

[34] S. Pandya and H. Ghayvat, "Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence," *Adv. Eng. Informat.*, vol. 47, Jan. 2021, Art. no. 101238.

[35] S. Mondal and A. D. Barman, "Human auditory model based real-time smart home acoustic event monitoring," *Multimedia Tools Appl.*, vol. 81, no. 1, pp. 887–906, Jan. 2022.

[36] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Audio content analysis for unobtrusive event detection in smart homes," *Eng. Appl. Artif. Intell.*, vol. 89, Mar. 2020, Art. no. 103226.

[37] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[38] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 829–852, 2022.

[39] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, 2018, pp. 328–339.

[40] G. Dekkers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," *Detection Classification Acoustic Scenes Events*, vol. 2017, pp. 1–5, Nov. 2017.
[41] F. Al Hossain, A. A. Lover, G. A. Corey, N. G. Reich, and T. Rahman, "FluSense: A contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–28, Mar. 2020.
[42] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, "Deep learning based cough detection camera using enhanced features," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117811.
[43] A. Mesaros, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE Workshop Detection Classification Acoustic Scenes Events*, 2017.
[44] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.
[45] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, Art. no. 035081.
[46] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digit. Signal Process.*, Jul. 2009, pp. 1–5.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[49] Y. Kiyokawa, S. Mishima, T. Toizumi, K. Sagi, R. Kondo, and T. Nomura, "Sound event detection with resnet and self-mask module for dcase 2019 task 4," DCASE Community, Tech. Rep., 2019.
[50] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1251–1264, 2023.
[51] J. S. Chung, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, Jun. 2020, pp. 2977–2981.
[52] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: Lessons from VoxSRC 2020," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5809–5813.
[53] S.-H. Kim, H. Nam, and Y.-H. Park, "Decomposed temporal dynamic CNN: Efficient time-adaptive network for text-independent speaker verification explained with speaker activation map," 2022, *arXiv:2203.15277*.
[54] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–6.
[55] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.
[56] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016.

SEONG-HU KIM received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2017, 2019, and 2024, respectively. His research interests include text-independent speaker identification, text-independent speaker verification, and sound event detection.

HYEONUK NAM received the B.S. and M.S. degrees in mechanical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include automatic speech recognition, speech dereverberation, semi-supervised sound event detection, and sound event localization and detection.

SANG-MIN CHOI received the B.S. degree in mechanical engineering from Sung Kyun Kwan University, Seoul, South Korea, in 2020, and the M.S. degree in mechanical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2022. His research interest includes sound event localization and detection.

YONG-HWA PARK (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from KAIST, in 1991, 1993, and 1999, respectively. In 2000, he joined the Aerospace Department, University of Colorado at Boulder, as a Research Associate. From 2003 to 2016, he was with Samsung Electronics, Visual Display Division, Samsung Advanced Institute of Technology (SAIT), as a Research Master in the field of micro-optical systems, with applications to imaging and display systems. In 2016, he joined KAIST as an Associate Professor of noise and vibration control plus (NOVIC+) with the Department of Mechanical Engineering, devoting to researches on vibration, acoustics, vision sensors, and recognitions for human–machine interactions. His research fields include structural vibration, event/condition recognition from sound and vibration signatures utilizing AI, blood pressure and health monitoring sensors, 3-D sensors, and LiDAR for motion measurements. He is a Board Member of KSME, KSNVE, KSPE, and SPIE. He has been the Conference Chair of MOEMS and Miniaturized Systems in SPIE Photonics West, since 2013.