Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 215-226, Ağustos 2020
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 215-226, August 2020
Copyright © 2020 EJOSAT
Research Article

Real-Time Auditory Scene Analysis using Continual Learning in Real Environments

Barış Bayram[1], Gökhan İnce[2*]

[1] İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-5588-577X)
[2] İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-0034-030X)

(Bu yayın 26-27 Haziran 2020 tarihinde HORA-2020 kongresinde sözlü olarak sunulmuştur.)

(**DOI:** 10.31590/ejosat.779710)

ATIF/REFERENCE: Bayram, B. & İnce, G. (2020). Real-Time Auditory Scene Analysis using Continual Learning in Real Environments. *Avrupa Bilim ve Teknoloji Dergisi*, (Special Issue), 215-226.

Abstract

Continual learning for scene analysis is a continuous process to incrementally learn distinct events, actions, and even noise models from past experiences using different sensory modalities. In this paper, an Auditory Scene Analysis (ASA) approach based on a continual learning system is developed to incrementally learn the acoustic events in a dynamically-changing domestic environment. The events being salient sound sources are localized by a Sound Source Localization (SSL) method to robustly process the signals of the localized sound source in the domestic scene where multiple sources can co-exist. For real-time ASA, audio patterns are segmented from the acoustic signal stream of the localized source for extraction of the audio features, and construction of a feature set for each pattern. The continual learning is employed via a time-series algorithm, Hidden Markov Model (HMM), on these feature sets from acoustic signals stemming from the sources. The learning process is investigated by conducting a variety of experiments to evaluate the performance of Unknown Event Detection (UED), Acoustic Event Recognition (AER), and continual learning using a Hierarchical HMM algorithm. The Hierarchical HMM consists of two layers: 1) a lower layer in which AER is performed using an HMM for each event and *the event-wise likelihood thresholds;* and 2) an upper layer in which UED is achieved by one HMM with *a suspicion threshold* through the audio features with their proto symbols stemming from the lower layer HMMs. We verified the effectiveness of the proposed system capable of continual learning, AER and UED in terms of False-Positive Rates, True-Positive Rates, recognition accuracy and computational time to meet the demands in a learning task of multiple events in real-time. The effectiveness of the AER system has been verified with high accuracy, and a short retraining time in real-time ASA having nine different sounds.

Keywords: Continual learning, auditory scene analysis, acoustic event recognition, unknown event detection, hierarchical HMM

Gerçek Ortamlarda Artımlı Öğrenme ile Gerçek Zamanlı İşitsel Sahne Analizi

Öz

Artımlı öğrenme ile sahne analizi, farklı duyusal modaliteler kullanarak geçmiş deneyimlerden daha önce bilgi sahibi olunmayan olayları, eylemleri ve hatta gürültü modellerini aşamalı olarak öğrenmek için durmaksızın gerçekleşen bir süreçtir. Bu çalışmada, dinamik olarak değişen gerçek bir ev ortamında akustik olayları aşamalı olarak öğrenmek için artımlı bir öğrenme sistemine dayanan İşitsel Sahne Analizi (ASA) yaklaşımı sunulmuştur. Ortamdaki en baskın ses kaynakları olan olaylar, birden fazla kaynağın bulunduğu işitsel sahnede bu kaynaktan elde edilen sinyalleri verimli ve kesintisiz bir şekilde işlemek için bir Ses Kaynağı Yerelleştirme (SSL) yöntemi ile yer tespiti yapılmaktadır. Gerçek zamanlı sahne analizinde, ses örüntüleri, bu örüntülerden ses özniteliklerin çıkarılması ve öznitelik setinin oluşturulması için bu kaynağın akustik sinyal akışından segmente edilir. Artımlı öğrenme, kaynaklardan elde edilen akustik sinyallerden bu öznitelik kümelerinde zaman serisi algoritması tabanlı olan Gizli Markov Modeli (HMM) kullanılmıştır. Öğrenme süreci, Bilinmeyen Olay Algılama (UED), Akustik Olay Tanıma (AER) ve Hiyerarşik HMM yöntemi kullanarak sürekli öğrenmenin performansını değerlendirmek için çeşitli deneyler yapılarak geliştirilmiştir. Hiyerarşik HMM iki katmandan oluşur: 1) AER'nin her bir olay için HMM ve olay bazlı eşik değerleri kullanılarak gerçekleştirildiği bir alt katman; ve 2) bir ses önitelik seti için ilgili alt katman HMM'inden çıkartılan proto sembolleri ile ses özniteliklerinin birleştirilip bir HMM ile bir şüphe eşk değeri kullanılarak UED'nin gerçeklştirildiği bir üst katman. Artımlı öğrenme, AER ve UED'e sahip bu sistemin, Yanlış-Olumlu Oranlar, Doğru-Olumlu Oranlar, tanıma doğruluğu ve hesaplama süresi gözetilerek birden fazla olayın söz konusu olduğu gerçek zamanlı öğrenme için gereken gereksinimleri karşılayacak seviyede olduğunu doğruladık. AER sisteminin etkinliği, yüksek doğruluk ve dokuz farklı ses içeren gerçek zamanlı ASA'da kısa bir yeniden eğitim süresi ile doğrulanmıştır.

Anahtar Kelimeler: Sürekli öğrenme, işitsel sahne analizi, akustik olay tanıma, bilinmeyen olay algılama, hiyerarşik HMM

1. Introduction

In the last decade, event recognition for a variety of problems has been carried out using acoustic signals. The recognition ability is developed to detect and classify urban and natural events [1, 2], household events [3], interaction with objects [4], animal life monitoring [5], surveillance [6, 7, 8], etc. However, learning is a continuous process, and also the information collected from the real, dynamically-changing environments expands during the lifelong deployment of the system. To accomplish continual learning for scene analysis, one has to deal with important challenges on perceiving the environments with a high level of background noise, and collecting and processing a large amount of data. Needless to say that the ability requires advanced hardware- and software-based solutions for practical uses in real-time.

Analyzing acoustic scenes in a real environment is one of the major applications of audio perception [9, 10], since various target or irrelevant sources are included in the real environments, causing a mixture of sounds obtained from multiple events, background noise, etc. Also, in the Auditory Scene Analysis (ASA) approach, recognition ability requires preprocessing techniques such as Sound Source Localization (SSL) and Sound Source Separation (SSS) in complex environments with overlapping sounds to extract important temporal and spatial information [11]. To analyze the scenes for understanding the relationships of entities such as events, objects, and humans, and efficiently labeling them is known as the cocktail party problem [12].

Besides, in real environments, when a novel object is recently detected, a few knowledge exists about it. Thus, an object-wise model is generated using this knowledge. Such learning with few examples, and novelty detection, mostly deep learning-based approaches have been proposed that require too high computational time at the model generation for this object. For object recognition and novelty detection, compared to humans, machine learning models not based on deep networks require large-scale datasets for training. Furthermore, there are rare studies on the combination of few-shot learning and novelty detection. The learning ability required for the cognitive aspects of scene analysis is recently considered as a continual learning ability [13]. In the learning process, new knowledge of a known class is recognized, and a new object is detected, and the relative knowledge of these classes inserted into the model; and so this process continues in a lifelong manner [14]. However, there is an important problem for the incremental learning ability, which is the catastrophic forgetting in which the recently obtained knowledge dominates the learning, and the previously learned knowledge is forgotten. In our real-time ASA study, we investigate the continual learning by retraining the data to meet the demands of real-time processing.

In this work, the real-time ASA is composed of five main stages that are 1) feature extraction including Sound Source Localization (SSL), audio pattern segmentation and audio feature extraction, 2) Acoustic Event Recognition (AER), 3) Unknown Event Detection (UED), 4) model generation after the detection of an unknown event, and 5) continual learning by fast retraining to avoid the catastrophic forgetting problem is presented to incrementally learn the acoustic knowledge from the sound sources, to distinguish and then to also accumulate new knowledge of a known source, and to detect the unknown sources in the scenes. Thus, using the real-time ASA framework, it is possible to analyze an auditory scene, and consequently to infer the source of a sound. We aim to build signal processing stages in the framework of ASA using the real-time audio streams acquired as given in the order below:

- the most salient sound source in the scene is localized to separate signals of the other sources and noises,
- audio patterns are segmented from the signals of the source,
- audio features are extracted from the patterns, and a feature set is constructed for each patterm, and handled as time-series data,
- a time-series method based on Hierarchical HMM is exploited, in which the lower HMMs are used for AER to recognize the feature sets of sound sources, and upper HMMs are used to detect unknown sources through the proto symbols of the lower HMMs.
- the proposed method incrementally learns the feature set of known or unknown sources.

The continual learning module mainly covers AER to recognize a known event for retraining of new knowledge of this event, and UED responsible of the detection of unknown events to generate a model using a few samples of the event. Therefore, the generation or retraining of a model is designed and developed considering the requirements of real-time ASA. For this purpose, an HMM in the lower layer of the Hierarchical HMM is utilized for each known sound event to recognize audio patterns. Also, an HMM in the upper layer for all the events is employed to decide whether the pattern belongs to a known event class, or none of the existing classes. After the recognition of a known event or the detection of an unknown event, using the feature set the HMM of the recognized event is updated by retraining or a HMM is generated for the unknown event. The performance of audio feature extraction, UED, and continual learning with retraining is assessed by various experiments in a domestic environment in terms of prediction performance and computational time.

The rest of the paper is organized as follows: in Section 2 the related works are reviewed. The stages of the proposed continual learning for ASA are described in Section 3. In Section 4, the offline and real-time experiments carried out in a domestic environment, the evaluation metrics and the results of the experiments are explained, followed by the conclusion and the future work.

2. Related Work

The types of features used in Acoustic Event Recognition (AER) can be typically divided into time-domain and frequency-domain features. The Zero-Crossing Rate (ZCR) [15] is an important time-domain feature mostly used for the classification of vocal and non-vocal sounds by indicating the most dominant frequency of a frame. It is a measure of the number of times when the sign changes between successive samples along with a signal. Also, the time feature is mostly utilized to discriminate voices from unvoiced speech [16], and to model the music. Moreover, it is utilized by being combined with other features like Mel-Frequency Cepstral Coefficient (MFCC) and discrete wavelet transform [17].

MPEG-7 low-level descriptors are frequency-domain features such as the spectral centroid, the spectral spread, and the spectral flatness [18]. These descriptors are also used to extract features from the acoustic events: the spectral centroid is a measure of the shape of the power spectrum; the spectral spread gives the information about the spectral shape by taking the root-mean-square deviation of the spectrum from its centroid; and the spectral flatness measures the flatness of the power spectrum. Another frequency-domain feature is spectral roll-off [19] used to measure the frequency, which takes 95% of the power spectrum and indicates the skewness of the spectral shape, and pitch ratio calculated on the pitches estimated in each frame to discriminate the speech from non-speech signals [20].

For AER, several works have used MFCC features as discriminative features to describe the spectral shape of the signal. These coefficients are computed by applying the discrete cosine transform to the log-energy outputs of Mel-scaling filter-bank; and its first and second-time derivatives ΔMFCCs, Δ^2MFCCs [21] are widely-used cepstrum features. Also, MFCC has been combined the other audio features such as spectral centroid, spectral spread, and spectral flatness to model events by HMM for surveillance [7]. In [22], MFCC features are combined with an MPEG-7 low-level descriptors, and wavelet features to generate an HMM. Furthermore, many AER studies have focused on feature selection based on Kullback-Leibler (KL) distance-based AdaBoost to analyze the discriminative audio features in order to select the most distinctive feature set [23]. In our study, we also employed the KL distance for feature selection to determine the discriminant feature vectors with the least number of dimensions to accomplish the fast retraining of HMMs in both of the layers for continual learning in real-time.

For acoustic object/event recognition, various generative classification algorithms such as HMM, long short-term memory, etc., and discriminative classification algorithms such as Support Vector Machine (SVM), multilayer perceptron, random forest, k-nearest neighbor, etc. have been utilized on different kinds of features such as spectral, temporal or perceptual features [24, 25]. In [26], non-speech event sounds are modeled for time-series classification using HMM with MFCCs. Also, two-dimensional time-frequency representations are utilized for time-series based AER like Stabilized Auditory Image (SAI) [27], Spectrogram Image Feature (SIF) with SVM and with a deep neural network studied in [28], with convolutional neural network studied in [29], etc.

In the work [30], an acoustic event detection approach used in the scenes such as busy street, restaurant, kid play area, pool hall, etc., was developed by applying multiple instance learning classifier, maximum margin early event detector both operating in the feature space and the landmarking space. The similarity measuring method is proposed to provide a mapping between the temporal sequences of MFCC features to a landmarking space. An AER approach proposed [22] for automatic recognition of different environmental sounds at home, which are speech, music, dog and cat sounds, doorbell sound, baby crying, explosion, gunshot, and laughter. Moreover, an ASA approach [31] utilized for robotic tasks has been proposed to recognize the physical and functional properties of the objects, that are the material type and the event occurred while the robot interacts with the objects. Also, in another ASA study [32], five manipulation behaviors performed by the robot on 36 objects, that are grasping, shaking, dropping, pushing, and tapping are recognized. In this work [33], an AER platform for monitoring elderly people to improve their life quality is proposed which is able to localize and recognize the human voice and activities in a domestic environment.

Various algorithms have been developed for Sound Source Localization (SSL) to estimate the Direction Of Arrival (DOA) on the emitted signals from the microphone array [34]. The spatial feature gives much information about the time, the location of the event, and the location of the noise sources interfering with the target sound source, which may be useful for AER. The well-known approaches for SSL are based on beamforming [35] and Multiple SIgnal Classification (MUSIC) [36] used in different types of environments. MUSIC, also used in our work, is a subspace-based method used by deriving a steering vector to detect reliable peaks with low computational cost, allowing to work in real-time. Another SSL approach for dynamic environments is Generalized Eigen-Value Decomposition based MUSIC (GEVD-MUSIC) [37] which performs noise-robust localization. It uses a noise correlation matrix to suppress the noises. Also, Bian et al. worked on an SSL approach to monitor and understand domestic activities [38]. Moreover, the localization can be utilized for tracking of the sound sources in a living space after recognizing the sources. Thus, it is based on employing sound recognition and SSL based on MUSIC together by using Gaussian Mixture Model (GMM) with outlier rejection [39]. Their tracker on moving sound sources is applied in real-time.

Only in few studies, the development of lifelong learning abilities has been investigated in the audio domain. Most of the studies whose main aim is object learning with a few examples, novel object detection, and incremental learning, have used deep learning methods. Due to the limited number of labeled data, most of the works conducted in the field of computer vision have targeted the learning model based on a few examples, which is also known as few-shot learning. However, in a few approaches, acoustic data have been utilized to generate few-shot learning models. In the work of Wang [40], a metric-based few-shot learning method has been proposed for AER due to high cost of listening to a mixed sound to label each location of an event. Another few-shot learning approach based on the acoustic data has used an Attentional Graph Neural Network [41]. However, none of these existing deep learning works are appropriate for few-shot learning particularly in real-time.

In addition, for few-shot learning and novelty detection, some studies have utilized acoustic data to generalize an audio class with few number of examples, and detect novel audio classes. In [42], few-shot method based on meta-learning has been presented for acoustic event detection to detect the unknown acoustic classes. Also, only a few works exist using unsupervised algorithms for novelty detection. The acoustic novelty detection studies have utilized unsupervised deep networks such as Recurrent Neural Network based Autoencoders on a benchmark speech dataset [43], and Convolutional Long Short Term Memory (LSTM) Autoencoders and Convolutional Autoencoders on the sounds of the different types of manufacturing processes [44]. However, using these deep networks, the generation of the model for a novel class is not possible in real-time due to the massive computational resources.

In various studies, the lifelong learning problem have been extensively tackled using various types of machine learning methods. However, only a few works have achieved to combine the novelty detection and few-shot learning. An incremental approach has been proposed for novelty detection with few-shot learning, based on a Parzen window kernel density estimator. The method is applied to real-time data streams regarding gestures under the problems, concept-drift, and existence of novel classes. Also, a class-based incremental learning approach [45] has been proposed, in which unknown classes are detected and incrementally adapted into a model by preserving the knowledge of the known classes. An incremental few-shot learning approach based on Attention Attractor Network, a meta-learning method has been presented to incrementally achieve few-shot learning without retraining the data [46]. In general, most existing approaches suffer from a high computational cost; so there is apparently a lack of studies on continual learning in real-time.

3. Real-time Auditory Scene Analysis System

3.1. Overview of System

The proposed real-time Auditory Scene Analysis (ASA) approach shown in Figure 1 includes five main modules: (1) **feature extraction** including Sound Source Localization (SSL) used for detection and location monitoring of a sound source in a scene, audio pattern segmentation, extraction of audio features and construction of a feature set of this pattern, (2) **model generation** of a lower HMM for each event, and an upper HMM for all the known events, (3) **Acoustic Event Recognition** (AER), (4) **Unknown Event Detection** (UED) regarding the outputs of AER module, and (5) **continual learning** by retraining of the lower-layer and upper-layer of Hierarchical HMM. All the stages are applied to the signal captured from a microphone. However, for real-time ASA, fast and efficient techniques for these modules are developed. The ASA begins with the localization of a salient sound source in a real environment. The acoustic signal stream from this source is processed to recognize acoustic known events and to detect unknown events, or undefined events. Subsequently, continual learning is achieved to learn knowledge about a known event by retraining, or unknown event by adapting its knowledge into the model. In the offline initial model generation module being a separate pipeline, however, at least one initial model may be generated using one, a few, or several audio patterns.

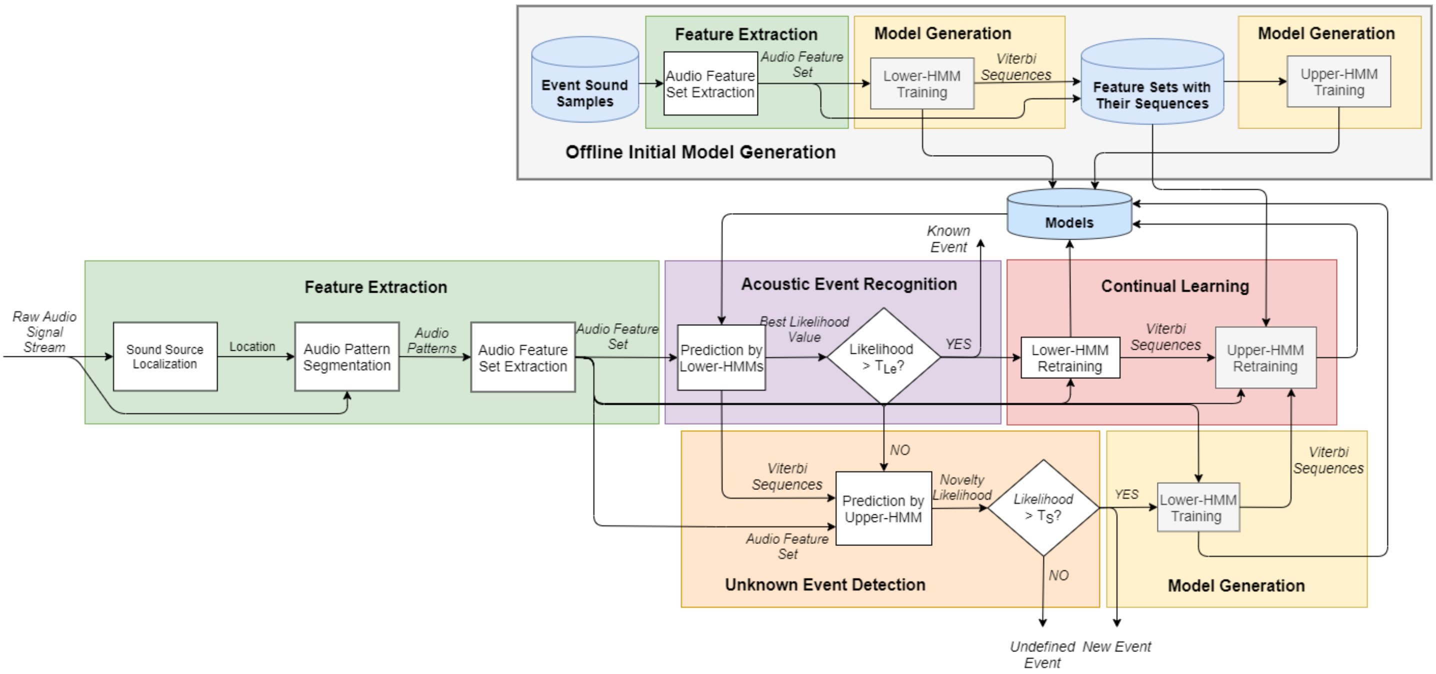

Figure 1. The overview of the proposed auditory scene analysis framework.

In real-time, an audio feature set utilized in the AER, UED, and continual learning stages is composed of a number of audio feature vectors extracted from the patterns segmented on the signals of the localized salient sound source. The ASA approach requires two types of thresholds for the recognition of known events and the detection of unknown events. These are an event-wise likelihood threshold, T_{Le} for each event e, and a suspicion threshold, T_S for all events to be compared with the outputs of the AER and UED models in order to obtain the likelihood value and the novelty rate. The threshold values are dynamically updated by taking the median of all the likelihood values computed by the upper HMM, or each lower HMM.

Details of each module of the system will be elaborated in the following sections.

3.2. Sound Source Localization

The location of the sound can be important to discriminate useful sound sources from the background noises, or non-target sources in a dynamically-changing environment. Thus, the location of a sound is utilized to ignore the sound sources not in the scene. Besides, the spatial information may carry essential cues, depending heavily on the characteristics of the scene. The spatial feature of the source can be utilized as an additional feature during pattern segmentation, event recognition, and unknown event detection, since it may provide important information about the event. Moreover, non-target sounds such as footsteps of a person, keystrokes, opening and closing of the door, the human voice, etc., can be ignored.

In this work, Generalized Eigen-Value Decomposition based Multiple SIgnal Classification (GEVD-MUSIC) [37] method which performs noise-robust localization is employed, since it is quickly able to localize the azimuth angle of multiple sound sources. Using this method, a sound source was monitored over its location in order to process its signal in an ASA pipeline. Thus, each stage of ASA is performed on the signal of a source. If a new source is localized, its signal will be analyzed by a separate ASA process. To monitor a detected source, a power threshold and an interval threshold in angle are utilized. If the source has higher power than the threshold, and the difference between its current position and the previous position is smaller than the interval threshold, the same source id is given, namely, the location of the source did not change over time.

3.3. Audio Pattern Segmentation

The aim of the segmentation is not only to distinguish the audio pattern of an event from silence, but also from another overlapped sound event. Therefore, the patterns with variable lengths are segmented from the signal stream of a localized source. A pattern has a boundary with the onset and offset frames on the stream. Thus, these frames need to be detected to segment an audio pattern in real-time. For the detection of the frames, the energy feature is extracted, and then a peak is picked by a sliding window on these features. If the energy value of the peak frame is higher than a threshold dynamically updated, this frame is tagged as the onset frame. Thereafter, we need to estimate the ending point of the pattern. The last frame before another peak frame with energy higher than the recent threshold value, or a silent frame with around zero energy is the offset frame. All the frames between the onset and offset frames are assumed as a useful audio pattern. For good performance in the AER, a fast and robust real-time segmentation technique is crucial.

3.4. Audio Feature Set Extraction

After the segmentation of the audio pattern, audio features are extracted, and a feature set for each audio pattern, including a number of feature vectors consequently as many as the length of the pattern is constructed. Thus, different cepstral, spectral, and temporal types of audio features are used as in Table 1 (to be later benchmarked in terms of AER performance and computational time of continual learning).

Table 1: The audio features evaluated for AER.

Cepstral	Spectral	Temporal
MFCC	*Flatness*	*ZCR*
MFCC with Derivatives	*Centroid*	*Energy*
LPCC	*Roll-off*	
LSP	*Flux*	

Furthermore, to reduce the computational time of retraining process in the continual learning, Kullback-Leibler distance based feature selection is applied to evaluate the discrimination capability of each feature unit in the vector of the estimated best audio feature as follows:

$$D(p_{ij}||q_i) \ = \int p_{ij} log \frac{p_{ij}}{q_i},$$

where p is the distribution of the i^{th} feature unit of the j^{th} audio event class, and q is the distribution of the i^{th} feature units of all the known event classes. Thus, the discriminant capability of a unit is;

$$D(p_i) = \sum_j b_j D(p_{ij}||q_i),$$

where b_j is the prior probability of the j^{th} event class.

The feature units with high distance to all the others are selected to be used in audio feature extraction stage for continual learning. The feature selection is required to remove redundant and similar feature units of different classes. It aims to attain similar AER performance using lower number of features for the decrease of the computational time, because after a new knowledge is obtained the retraining process will be employed.

3.5. Unknown Event Detection

Initially, HMMs in the lower layer of Hierarchical HMM are generated using the feature sets of audio patterns segmented from the sound samples of the existing events, and then the sets with their Viterbi sequences are utilized to generate an HMM in the upper layer. Using the event-wise likelihood thresholds for AER, and a suspicion threshold for UED, the abnormal feature sets are detected as illustrated in Figure 1.

The UED functions as follows; (1) the salient sound source is detected and localized, (2) audio patterns segmented from new incoming signal from this source are predicted by existing lower-layer HMMs, (3) the likelihood value of the most likely event is compared with the event's likelihood threshold estimated during the training of the existing sound samples of this event, (4) if it is less than the threshold, its Viterbi sequence is extracted, and combined with the corresponding feature set, (5) the combined feature set is predicted by the upper-layer HMM, (6) if the likelihood computed by the HMM is lower than the novelty rate, the pattern is predicted as highly anomaly, an unknown event class is detected.

In real environments, a novel sound source may emerge in the domestic scenes. Therefore, the real-time ASA with continual learning system needs to detect the abnormal audio features from the unknown sound sources. Firstly, the audio feature set of an audio pattern extracted from the acoustic data stream is recognized in the AER stage by the lower layer HMMs and the event-wise likelihood thresholds. If the best log likelihood value is less than the relevant threshold, the feature set with its proto Viterbi symbols belonging to the event with the best likelihood is predicted by the upper layer HMM. Then, the likelihood value computed by this HMM used for UED is compared with a suspicion score for the UED to decide whether the feature set belongs to 1) a new event, 2) to a known one with the highest value, or 3) to none of them because of the low precision.

3.6. Continual Learning

The methods for modelling time-series data suffer from various problems affecting the performance, such as large-scale and high-dimensional data, series with variable-length, missing values and patterns of the series, modelling of occluded multiple series, high computational time required for real-time tasks and so on. Under these challenges we developed a Hidden Markov Model (HMM) based lifelong learning approach with UED and retraining. Therefore, the computation time for the stages of ASA is as important as the AER performance. Obtaining a robust and fast performance from HMM is taken into consideration while developing the model for continual learning.

The HMM has been utilized for modeling time-series data like in different signal processing tasks such as analysis and recognition of speech, voice, acoustic events, etc, and modeling sequential data like in natural language processing and bioinformatics tasks such as analysis of text data, biosignal, etc. In our study, we have used the HMM algorithm for analysis of auditory scenes on the audio feature sets composed of a number of audio feature vectors. For each event, an HMM is generated on all the feature sets belonging to this event regarding a number of iterations, and a distinct number of states. At the training time, the HMM aims to maximize an objective function computing the likelihood value of a Viterbi sequence most likely representing these sets. In the prediction time of an audio pattern, the likelihood value estimated by the HMM of an event represents the probability of belonging to this event. The AER is accomplished by comparing the highest likelihood value with a likelihood threshold of the most likely event. If the likelihood is less than the threshold, the combination of the feature set and its Viterbi sequence is exploited in UED. However, UED is applied to only one HMM generated on all the feature sets of all known acoustic events. Therefore, only one suspicion threshold is utilized to decide on whether the combination may belong to an unknown or undefined event.

3.6.1 HMM Formulation Adapted for AER

In this problem, the ASA process may begin with an initial HMM $H_e(X_e, \lambda_e)$ of an event, e to maximize the objective function:

$$H_e(X_e, \lambda_e) = \underset{\lambda_e}{\operatorname{argmax}}\, P(I|X_e, \lambda_e),$$

in which X denotes a matrix consisting of a consequent set of the feature vectors from an audio pattern segment of the event, I represents the feature vector (state) sequence, and λ is the set of HMM parameters different for each event, updated during the initial training and also retraining. A number of states is determined based on the acoustic event. The states of the HMM are connected by transitions, and the likelihood value is computed by this object function that aims to estimate the most likely state sequence. Each audio feature vector of the set belonging to an audio pattern is represented by a state, and the sequence is composed of a number of state symbols as many as the size of the set.

One of the parameters of HMM is the transition probability matrix including the probability values of the transition from the state, s_i to the state, s_j, is computed as $a_{ij} = P\big(q_{t+1} = s_j | q_t = s_i\big)$ where s_i is the actual state at time, t. Thus, the probabilistic model is generated regarding a Markov chain induced by these transition probabilities. In other words, an audio feature set of each segment of an event is modeled by the transition probabilities, and characterized in terms of the state symbols at the training and prediction times.

At the training time, the parameters of HMM are optimized by the Expectation-Maximization algorithm that consists of two steps, M-step and E-step. The expectation of the parameters is achieved in the E-step. In M-step, it is basically aimed to maximize the likelihood over a number of iterations selected by the user, or convergence criteria. The main parameters are the state transition probabilities and the initial state probabilities optimized by the forward and backward processes. At each step, the new parameter set, λ'_e of an HMM belonging to an event is estimated, and this re-estimation loop continues over the forward and backward variables until the stopping criterium is satisfied;

$$\log P(O|\lambda'_e) - \log P(O|\lambda_e) < \epsilon,$$

where ϵ denotes a small constant value. The optimization process is applied to each event's model, H_e in the lower layer of the Hierarachical HMM, but for UED, only one HMM is generated using all the existing feature sets with the Viterbi sequences estimated by its own model, H_e.

Besides, the computational time of the training of an HMM scales linearly with the length of the feature sets of an event, and quadratically with the number of states. Therefore, it is aimed to find the optimal number of iterations to decrease the computational time while maintaining the prediction performance high. Also, the space complexity scales with the characteristics of the data, and an

HMM. There are a few works proposed to achieve online HMM based on the online Baum-Welch and Expectation-Maximization techniques [47], but none of them provides AER performance as much as retraining itself.

4. Experiments and Results

4.1. Software and Hardware Settings

The audio preprocessing operations for real-time ASA are implemented in *HARK* [48], which is an open-source robot audition software. For the time-series based AER, the evaluation and the visualization *hmmlearn*, *scikit-learn*, and *matplotlib* libraries are used, respectively. For the audio feature extraction stage, the audio python library, *Librosa*, and *pymir* are utilized. The experiments were performed by a computer with Intel quad Core i7-3700K with 16 GB RAM on the recordings taken by Kinect microphone array with 4-channel.

4.2. Experimental Procedure

In the first part of the experiments, namely the offline experiments, the audio feature set and parameters for the models learned for UED and AER were estimated so that they can be later used in the real-world, real-time experiments. For these offline experiments, the recordings captured in a domestic environment have been utilized. In these offline experiments, a bunch of domestic sounds such as being recorded from opening and closing doors, footsteps, kettle, vacuum cleaner, washing hands, dishwasher, toilet flush, and washing machine have been generated in the near vicinity of a microphone array. There are 436 event sounds manually labeled (41 door opening and 50 door closing, 36 footsteps, 65 kettle, 55 vacuum cleaner, 24 washing hands, 85 dishwasher, 32 toilet flush, and 48 washing machine sounds). Using the pattern segmentation approach, 31905 audio patterns are obtained from the sound instances, consisting of 1981 door opening and 2711 door closing, 2416 footsteps, 2449 kettle, 3972 vacuum cleaner, 6008 washing hands, 5112 dishwasher, 3184 toilet flush, and 4072 washing machine audio patterns. Sound files are divided into a training set (70%) with which the model generation and retraining are performed, and a test set (30%) only used for performance evaluation in offline experiments. The recordings in the training set were utilized as training and validation. Thus, continual learning begins using one or a few audio patterns segmented from a recording in the training set and learns the rest in time.

The audio features are extracted with the sampling rate selected as 16kHz, Fast Fourier Transform size as 512, and the window size as 256. Thus, a feature vector includes the features extracted from a window. The number of states and iterations used at the training time of HMM are selected for each event considering the length of the audio pattern, and the number of feature sets.

In the second part of the experiments, the audio feature set and parameters for the models learned in the offline experiments were utilized for the generation of AER and UED models to be used in the real-world, real-time experiments. For these experiments, the same domestic sounds have been generated in real-time.

4.3. Evaluation Metrics

This section presents the evaluation criteria used in the framework of Hierarchical HMM based AER, continual learning with retraining, and UED. The accuracy of AER, f1-score and time for retraining is computed to determine the most appropriate type of audio features. We also evaluated the performance of the UED in terms of precision, recall, and the detection time including the prediction time of the lower HMMs and upper HMM. The calculation of the computational time is performed for each stage to evaluate the real-time performance of the proposed framework.

4.4. Results of Offline Experiments on Auditory Scene Analysis

In these experiments, the sounds of events recorded from a real domestic environment were utilized to evaluate the audio features, continual learning with retraining, and UED. Subsequently, the outcomes of the experiments were exploited to develop the models for the real-time ASA experiments in the domestic environment.

4.4.1 Results on Audio Feature Extraction

In this experiment, different types of audio features listed in Table 1 are evaluated in terms of prediction performance and retraining time. We need to estimate most appropriate cepstral feature set that will comprise the major part of the set, and the promotive features in the spectral and temporal features. Firstly, the cepstral features (Table 2) and then other variations of features (Table 3) are evaluated by performing the experiment 10 times and taking the average accuracy, f1-score and retraining time. During this continual learning, the retraining time is measured under the worst scenario in which most of the feature sets is retrained.

Table 2: The AER performance using the cepstral features.

Features	Dimension	Avg. Accuracy	Avg. F1-Score	Avg. Retraining Time (sec.)
MFCC	13	0.881	0.716	2.524
MFCC with Derivatives	39	0.914	0.772	5.110

LPCC	20	0.711	0.389	4.171
LSP	255	0.761	0.471	12.820

Table 3: The performance of AER using different feature combinations including MFCCs with its derivatives.

Features	Dimension	Accuracy	F1-Score	Retraining Time (sec.)
MFCC	13	0.881	0.716	2.524
MFCC with Spectral Features	16	0.889	0.734	2.828
MFCC with Energy	14	0.879	0.708	2.580
MFCC with ZCR	14	0.883	0.721	2.644

In Table 2, the results are listed, in which MFCCs and MFCCs with their derivatives have the best prediction performances. It is decided to prefer only MFCCs by considering the dimension of the features, and the computational time required with respect to the dimension. As table 3 illustrates, using the variations of features with 1-dimension, the performance either deteriorated, or not improved remarkably.

Furthermore, to decrese the computational time in the feature domain, the Kullback-Leibler distance based feature selection is applied on the 13-dimensional MFCC features. Through the realization of five experiments, for MFCC and selected coefficients of MFCC, the average AER accuracy and f1-score of lower HMMs, and the average computation time of the retraining of an HMM in worst case are listed in Table 4.

Table 4: AER performance and retraining time of both MFCC and selected coefficients of MFCC features.

Features	Accuracy	F1-Score	Prediction Time (sec.)	Retraining Time (sec.)
MFCC with 13 coefficients	0.881	0.716	0.0060	2.524
MFCC with selected 8 coefficients	0.849	0.670	0.0043	1.986
MFCC with selected 4 coefficients	0.811	0.577	0.0038	1.180

4.4.2 Results on Unknown Event Detection

In the UED experiments, different scenarios are considered to evaluate the performance of the detection of an unknown sound source. First of all, we need to demonstrate the benefit of the upper HMMs for UED. For UED, only lower HMM with predefined thresholds is compared with the Hierarchical HMM in terms of precision and recall (Table 5). The single layer HMM based solution for UED especially suffers when multiple HMMs provide lower likelihood values when little differences between them exist, which eventually deteriorates the recall of the detection of the known events to unknown. Therefore, analyzing the novelty also on the Viterbi sequences by an upper HMM increases the precision and recall due to the decrease in the false-positive, and false-negative rates.

Table 5: UED performance and detection time of only HMM and Hierarchical HMM.

Number of Known Event	Avg. Precision	Avg. Recall	Detection Time(sec.)
Only HMM	0.799	0.494	0.0178
Hierarchical HMM	0.889	0.818	0.0271

Furthermore, the performance of UED using Hierarchical HMM is investigated considering; (1) if only one class is known, (2) if three event classes are known, (3) if five event classes are known and (4) if only one event class is unknown (eight are known), in which HMMs in the lower and upper layers of the Hierarchical HMMs have been trained using few audio patterns segmented from only one isolated sound sample. The performances of UED in terms of precision, recall, the computational time of generation of a lower HMM, and retraining of the upper HMM are shown in Table 6. Also, the time for the detection of one unknown class, in which the predictions of all existing lower HMMs and the upper HMM is evaluated.

Table 6: UED performance and detection time of Hierarchical HMM in different scenarios.

Number of Known Event	Avg. Precision	Avg. Recall	Detection Time (sec.)
1(only one event is known)	0.884	0.818	0.0103
3	0.859	0.792	0.0179
5	0.830	0.761	0.0210
8 (only one event is unknown)	0.819	0.745	0.0218

In Table 6, the results of Hierarchical HMM in the scenarios in which only one, three, or five events are known, or only one event is unknown. The selected three and five events are not similar to each other, therefore, the performance of UED is evaluated when at least one event similar to the unknown one is known. A sample of UED experiment is demonstrated in Fig. 2, in which the suspicion threshold (the red point), and the likelihood values computed by the upper HMM of the patterns belonging to unknown events (dish machine, door closing, and vacuum cleaner) are shown. The *correctly detected* patterns are represented by the blue points, and the patterns detected as *undefined* and *known* are represented by the green and yellow points, respectively.

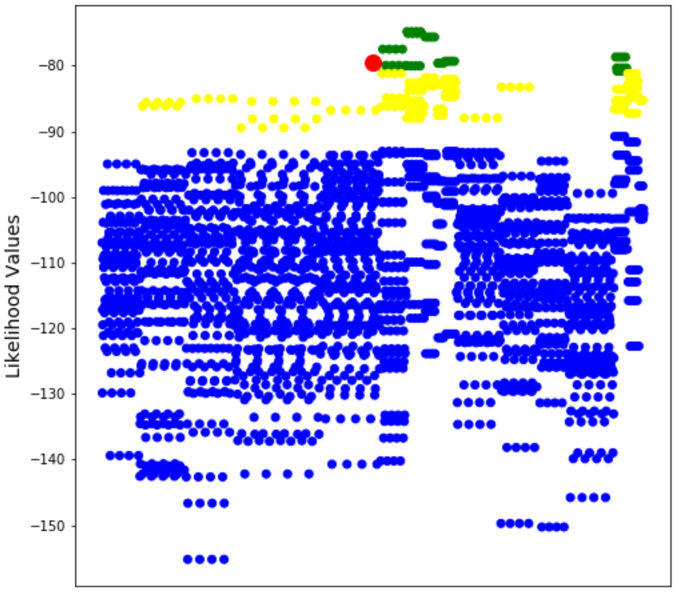

Figure.2: The likelihood values of the unknown events that are the true-positive samples (blue points), the false-positive samples (green points), the patterns detected as undefined (yellow points), and the suspicion threshold (red point) over the whole test data represented with the indices along the x-axis.

In the offline UED experiments, it is observed that the number of sound samples used for training HMMs directly affects the threshold values. In real-time experiments, an initial model is generated with the parameter values obtained in these offline experiments by using few patterns of a randomly selected event.

4.5. Results of Real-Time Experiment on Auditory Scene Analysis

In this experiment, using a few sound samples of a number of domestic events, we have trained an initial lower HMM for each known event and an initial upper HMM for all the events. In real-time, the sounds of absolutely unknown events including footsteps, toilet flush, washing hands and opening and closing the doors are generated. Also, the sounds of the known events including kettle, vacuum cleaner, washing machine and dishwasher are generated in real-time to evaluate the AER performance on the known classes of the initial model.

The results of the two real-time ASA experiments are shown in Table 7, including the precision and recall values of AER and UED. The lowest precision is belonging to the kettle due to the similarity between the audio patterns of kettle boton with the footstep, and vice versa. Also, several patterns of toilet flush sound are mixed with the patterns of the dishwasher and vacuum cleaner. Most distinctive sound is the vacuum cleaner and closing the outdoor.

Table 7: AER and UED performances and retraining time of lower and upper HMMs for different sounds.

Events	Precision for AER	Recall for AER	Precision for UED	Recall for UED
Door Openning	0.811	0.770	0.888	0.804

Door Closing	0.874	0.841	0.910	0.864
Footsteps	0.790	0.731	0.764	0.681
Kettle	0.757	0.689	0.784	0.599
Vacuum Cleaner	0.889	0.814	0.929	0.871
Toilet Flush	0.803	0.782	0.861	0.802
Washing Machine	0.819	0.800	0.844	0.814
Washing Hands	0.821	0.804	0.809	0.821
Dishwasher	0.796	0.764	0.840	0.790

The AER performances of two different continual learning experiments are shown in Figure 3. Also, the number of detected events are indicated on the AER accuracy line that represents the number of patterns being detected as unknown and adapted into the Hierarchical HMM at this point. While the number of unknown classes increase, the AER accuracy also improves. An average of 20-35 different classes are detected in various experiments considering 101 event sounds with 12901 audio patterns are introduced.

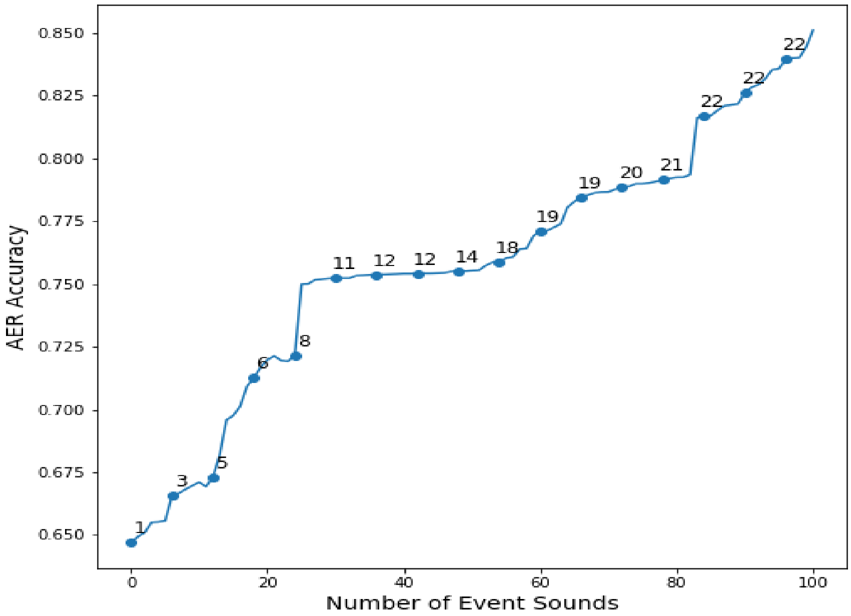

Figure 3: The performances of AER in 4 different sessions of lifelong learning experiments.

5. Conclusion

In this study, we proposed an approach for auditory scene analysis with continual learning and unknown event detection (UED) to recognize the acoustic events in real domestic environments. For this purpose, we developed a system comprising audio feature extraction, audio pattern segmentation, unknown event detection, and time-series based continual learning with retraining. In the learning module, we utilized Hierarchical HMMs in which the lower HMMs are utilized for AER to recognize the feature sets, and the upper HMM is employed for UED to predict the sounds being either "known", "unknown" or "undefined" events. The new knowledge regarding the event is used to retrain the previous knowledge, or brand new lower and upper HMMs are generated for this event satisfying the continual learning constraints. However, the retraining process may require a long time after a while. Thus, in this work, not only learning performance but also the complexity of prediction and retraining were investigated. We assessed the effectiveness of the system using scenes in the offline and the real-time ASA experiments. As a result, it is observed that the time-series based method on the Mel-frequency cepstral coefficients (MFCCs) can incrementally learn and recognize the acoustic events in real-time ASA tasks.

As future work, the number of everyday sounds to be recognized will be increased to the order of hundreds and the performance of the system will be verified. Besides, the real-time ASA will be utilized with a real robot where intelligent and autonomous behaviors such as getting closer to the events and tracking the moving targets in the scenes will be performed to improve the performance of AER.

References

[1] Salamon, J., Jacoby, C. and Bello, J. P. (2014). A dataset and taxonomy for urban sound research, in Proc. ACM Int. Conf. Multimedia, pp. 1041- 1044.

[2] Young, S. H. and Scanlon, M. V. (2001). Robotic vehicle uses acoustic array for detection and localization in urban environments, SPIE Proc. Mobile Robot Perception, vol. 4364, pp. 264-273.

[3] Wang, J. C., Lee, H. P., Wang, J. F. and Lin, C. B. (2008). Robust environmental sound recognition for home automation, IEEE

Trans. Autom. Sci. Eng., vol. 5, no. 1, pp. 25-31.

[4] Sinapov, J., Weimer, M. and Stoytchev, A. (2008). Interactive learning of the acoustic properties of objects by a robot, in Proceedings of the RSS Workshop on Robot Manipulation: Intelligence in Human Environments.

[5] Lee, C. H., Han, C. C. and Chuang, C. C. (2008). Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients, IEEE Transactions on Audio, Speech and Language Processing, 16(8), pp. 1541-1550.

[6] Carletti, V., Foggia, P., Percannella, G., Saggese, A., Strisciuglio, N. and Vento, M. (2013). Audio surveillance using a bag of aural words classifier, in Proc. IEEE Int. Conf. Adv. Video Signal Based Surveillance, pp. 81-86.

[7] Vozarikova E., Pleva M., Juhar J., Cizmar A. (2011). Surveillance system based on the acoustic events detection. Journal of Electrical and Electronics Engineering; 4 (1):255-258.

[8] Chandrakala, S., and Jayalakshmi, S. L. (2019). Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Computing Surveys (CSUR)*, *52*(3):1-34.

[9] Imoto, K. (2018). Introduction to acoustic event and scene analysis. *Acoustical Science and Technology*, *39*(3):182-188.

[10] Zeremdini, J., Messaoud, M. A. B., and Bouzid, A. (2015). A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation. *Brain informatics*, *2*(3), 155-166.

[11] Okuno, H. G., Ogata, T., Komatani, K., & Nakadai, K. (2004). Computational auditory scene analysis and its application to robot audition. In *International Conference on Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004.* (pp. 73-80). IEEE.

[12] Okuno, H. G., Nakatani, T., and Kawabata, T. (1995). Cocktail-Party Effect with Computational Auditory Scene Analysis – Preliminary Report –, In Proceedings of the Sixth International Conference on Human-Computer Interaction.

[13] Ade R. R and Deshmukh P. R. (2013). Methods for incremental learning: a survey. International Journal of Data Mining & Knowledge Management Process; 3(4):119-125.

[14] Hulley, G. and Marwala, T. (2007). Evolving classifiers: Methods for incremental learning. arXiv preprint arXiv:0709.3965.

[15] Atrey, P. K., Maddage, N. C., and Kankanhalli, M. S. (2006). Audio based event detection for multimedia surveillance. In Proc. of ICASSP.

[16] Cachu, R., Kopparthi, S., Adapa, B. and Barkana, B. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal, ASEE.

[17] Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., & Hamzaoui, R. (2017). Audio-based event recognition system for smart homes. In 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pp. 1-8.

[18] Wellhausen, J., and Höynck, M. (2003). Audio thumbnailing using MPEG-7 low-level audio descriptors. In *Internet Multimedia Management Systems IV* (Vol. 5242, pp. 65-73). International Society for Optics and Photonics.

[19] Rabaoui, A., Lachiri, Z. and Ellouze, N. (2009). Using hmm-based classifier adapted to background noises with improved sounds features for audio surveillance application. International Journal of Signal Processing, 5(1):46–55.

[20] Abu-El-Quran, A., Goubran, R. and Chan, A. (2006). Security monitoring using microphone arrays and audio classification. Instrumentation and Measurement, IEEE Transactions on, 55(4):1025 –1032.

[21] Logan, B. (2000). Mel frequency cepstral coefficients for music modelling, Proc. Int. Symp. Music Info. Retrieval (ISMIR).

[22] Ntalampiras, S., Potamitis, I. and Fakotakis, N. (2010). A Multidomain Approach for Automatic Home Environmental Sound Classification, Proc. 11th Annual Conference of the International Speech Communication Association, pp. 2210-2213.

[23] Zhou, X., Zhuang, X., Liu, M., Tang, H., Hasegawa-Johnson, M., and Huang, T. (2007). HMM-based acoustic event detection with AdaBoost feature selection. In Multimodal technologies for perception of humans, pp. 345-353.

[24] Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. Pattern Recogn Lett 24:2895–2907.

[25] Rabaoui, A., Kadri, H., Lachiri, Z. and Ellouze, N. (2008). Using robust features with multi-class SVMs to classify noisy sounds, International Symposium on Communications, Control and Signal Processing.

[26] Dong, R., Hermann, D., Cornu, E., and Chau, E. (2007). Low-power implementation of an HMM-based sound environment classification algorithm for hearing aid application. In *2007 15th European Signal Processing Conference*, IEEE, pp. 1635-1638.

[27] Walter, T. C. (2011). Auditory-based processing of communication sounds, Ph.D. dissertation, University of Cambridge.

[28] McLoughlin, I., Zhang, H. M., Xie, Z.P., Song, Y. and Xiao, W.. (2015). Robust Sound Event Classification using Deep Neural Networks. IEEE Transactions on Audio, Speech, and Language Processing. 23, pp. 540– 552.

[29] Zhang, H., McLoughlin, I. and Song, Y. (2015). Robust Sound Event Recognition using Convolutional Neural Networks. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. 2635. IEEE; 2015. pp. 559–563.

[30] Sangnier, M., Gauthier, J. and Rakotomamonjy, A. (2015). Early frame-based detection of acoustic scenes, In IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics.

[31] Saltalı İ, Sariel S, and İnce G. (2016). Scene analysis through auditory event monitoring. In: Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents (DAA) 2016, pp. 1-6. doi: 10.1145/3005338.3005343

[32] Sinapov, J., and Stoytchev, A. (2009). From Acoustic Object Recognition to Object Categorization by a Humanoid Robot, In Proceedings of the RSS Workshop: Mobile Manipulation in Human Environments.

[33] Do, H.M., Sheng, W. and Liu, M. (2015). An open platform of auditory perception for home service robots, in PRoceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS ’15), pp. 6161–6166.

[34] Asano, F., Goto, M., Itou, K., and Asoh, H. (2001). Real-time sound source localization and separation system and its application to automatic speech recognition, in Seventh European Conference on Speech Communication and Technology. Citeseer.

[35] Kushwaha, M., Koutsoukos, X., Volgyesi, P., and Ledeczi, A. (2009). Acoustic source localization and discrimination in urban environments. In 2009 12th International Conference on Information Fusion, IEEE, pp. 1859-1866.

[36] Schmidt, R. (1986). Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas and Propagation, vol. 34, no. 3, pp. 276–280.

[37] Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., and Tsujino, H. (2009). Intelligent sound source localization for dynamic environments, in IROS, pp. 664–669.

[38] Bian, X., Abowd, G. D., and Rehg, J. M. (2005). Using Sound Source Localization to Monitor and Infer Activities in the Home, In the Proceedings of the Third International Conference on Pervasive Computing.

[39]Liu, Y. W., Liang, H. M., Lao, S. Y., Wu, C. W., Hao, H. K., Kung, F. J., Ho, Y. T., Lee, P. Y., and Kang, S. C. (2014). Developing ‘voice care’: real-time methods for event recognition and localization based on acoustic cues, Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW ‘14).

[40]Wang, Y., Salamon, J., Bryan, N. J., and Bello, J. P. (2020). Few-Shot Sound Event Detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 81-85.

[41]Zhang, S., Qin, Y., Sun, K., and Lin, Y. (2019). Few-Shot Audio Classification with Attentional Graph Neural Networks. In *INTERSPEECH*, pp. 3649-3653.

[42]Shi, B., Sun, M., Puvvada, K. C., Kao, C. C., Matsoukas, S., and Wang, C. (2020). Few-Shot Acoustic Event Detection Via Meta Learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 76-80.

[43]Nguyen, D., Kirsebom, O. S., Frazão, F., Fablet, R., and Matwin, S. (2019). Recurrent Neural Networks with Stochastic Layers for Acoustic Novelty Detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 765-769.

[44]Bayram, B., Duman, T. B., and Ince, G. (2020). Real time detection of acoustic anomalies in industrial processes using sequential autoencoders. *Expert Systems*, e12564.

[45]Shmelkov, K., Schmid, C., and Alahari, K. (2017). Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3400-3409.

[46]Ren, M., Liao, R., Fetaya, E., and Zemel, R. (2019). Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, pp. 5275-5285.

[47]Khreich W., Granger E., Miri A., and Sabourin R. (2012) A survey of techniques for incremental learning of HMM parameters. Information Sciences 2012; 197: 105-130. doi: 10.1016/j.ins.2012.02.017

[48]Nakadai, K., et al (2008). An open source software system for robot audition HARK and its evaluation. In: *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*. IEEE. pp. 561-566.