SURROUNDING AUDIO SCENE AND EVENT IDENTIFICATION

[1]Ankush Soni , [2] Shubham Thombare, [3]Kunal Sonar, [4] Vishal Kajale, [5]Pradnyesh Bhisikar

Department of computer engineering,
SITRC, Nashik, India.

Abstract : Audio analysis is useful for audio monitoring, multimedia retrieval, audio tagging, and criminal investigations. Due to the presence of multiple sound sources, background noises, and the presence of overlapping or polyphonic contexts, environmental audio scene recognition (EASR) and sound event recognition (SER) for audio monitoring are difficult tasks. We concentrate on developing strong and efficient descriptions for essential aspects of ambient audio scenarios, which have shown to be successful in language and audio-related activities For the SER task, the Environmental Sound Classification (ESC-50) dataset is used. The exclusionary character of model-driven presentations improves EASR and SER setting a good example. The existing solutions are better suited to duties with small datasets.

Index Terms—

***IndexTerms* - EASR (Environmental Audio Scene Recognition), SER (Sound Event Recognition), STFT (Short-Time Fourier Transform), FFT (Fast Fourier Transform), Frequency, Amplitude, Timbre(Pitch).**

I. Introduction

Monitoring human and social activities is becoming increasingly important in our day to day living. Automated surveillance systems, that use systems techniques employing both video and audio information, have recently gained importance[1-3].When visual cues cannot reliably recognize the activities (events) and environments/contexts, audio cues are complementary to visual cues. The information collected from a semantic audio analysis can be beneficial for audio surveillance and related applications such as analyzing and forecasting patterns of events, classifying/searching audio records, customer alerts, and robot It can be instrumental for various applications such as audio forensics and audio surveillance in different environments such as child-care centres, smart-homes, elder-homes, orphanages, residential areas, office rooms, roads, elevators, cities, and intelligent conference rooms. navigation[4-5]. The existence of sound events and audio scenes indicates human and socially related actions as normal or abnormal activities. Besides speech and music signals, environmental sound signal analysis is becoming more important, and it has a wide range of applications in different domains. It can be instrumental for various applications such as audio forensics[6] and audio surveillance[7] in different environments such as child-care centres, smart-homes, elder homes, orphanages, residential areas, office rooms, roads, elevators, cities, and intelligent conference rooms. It can also be tailored with environmental audio scene recognition(EASR) [8], sound event recognition(SER) [7], audio tagging [9] , ambient assisting living[10], and fall detection [11] applications. Automated surveillance is one of the major applications where recognition of acoustics events and scenes play a vital role. Environmental audio scene recognition refers to the process of recognizing the context or environment of an audio stream, with applications in devices requiring contextual awareness. Some of the indoor audio scenes include the following: grocery shop, cafe/restaurant, and home. Outdoor audio scenes include busy streets, open-air market, and park. Sound event recognition aims to locate and recognize each occurrence of a monophonic event or polyphonic event in a specific environment. For example, a car horn, riding a motorcycle, and footsteps sound are specific events of a recording in a busy street. Some common characteristics of environmental audio scenes and sound events include the following: (i) the signal to noise ratio (SNR) is typically very small in an audio signal, particularly if the microphone is not very near to the acoustic source; (ii) Discriminative information existing in low-frequency ranges; (iii) no specific structures like phonemes or prosody; (iv) inability to identify a dictionary of basic units.

Compared with speech and music signals, these audio events/scenes do not have a grammar-like structure and so it is hard to model the characteristics of such events/scenes[14- 15]. All these characteristics increase the difficulty of the problem in an audio event/scene recognition task. Some of the issues related to SER and EASR include the following: presence of multiple sound sources[12]; the existence of overlapping or polyphonic events[16] ; recognition of confusable events/scenes, e.g., street traffic vs. restaurant [17]; both the system recognition and human recognition tend to have similar confusions mainly within the high-level categories (vehicles, outdoors, and public).

II. Related work

Data in the acoustical event categorization can complement a sound event detection system with additional contextual information. Sound events in a context of real life [17] are equally different from the categorization in a still environment of isolated occurrences [10, 12]. A. Later studies on the recognition of an environmental audio scene (EASR) We examined three strategy categories:

(1) The use of the developed/generic characteristics such as time frames, spectrum region and conceptually triggered characteristics;

(2) to use the learnt characteristics of algorithms for machine learning;

The overlapping sceneries and noise resulted in most of the strategies addressed in the literature. The functional learning approaches were employed to overcome the difficulties caused by overlapping sights and sounds. The effectiveness of machine learning algorithms depends mainly on data display. The learnt features (high-quality audio description) were possibly built on innovative features, providing perhaps powered or precision.

III. DATASET

Associated Detailed Dataset

Nearly all of these publicly accessible detailed data sets are aimed towards identifying people’s actions or gestures. They collected the existing datasets. Our collection also facilitates the development of audio-based categorization algorithms. We also covered a greater number of actions that help to categorise the many scenes that assist health systems recognise. Some activities in the dataset may overlap with current dataset activities, while we have created a new, complicated collection of everyday activities in our data collection.

Related Audio Dataset

Our data collection is a substantial dataset in terms of the number of actions and the quantity of audio streams per action compared to current audio datasets.

TABLE I

Dataset	Provider	No. of classes	No. of sources	Scene category
ESC-50	Karol picz	5	1	32

Our dataset is a full data set in relation to the number of actions and the amount of audio streams per activity as compared to previous audio data sets.

Dataset Statistics

Inspired by studies carried out under Daily Living Activities[18] by health specialists. In our dataset, we conceived 24 actions. These include: calling, tinkle, drink, eat, enter from the door, go out, lie down, open a pill box, pick up an object, read, sit still, sit, sit, sleep, stand up, sweep, use a laptop, use a phone, wake up, walk, wash the hand, watch television, pour water and writing.

IV. ACTIVITY RECOGNITION

Acoustics (Audio) is one of our most important sensory inputs for us. Virtually every activity or occurrence has a distinctive sound in our environment. Audio contains three primary features that help us differentiate between multiple sounds.

- Amplitude : Sound Loudness
- Frequency : The pitch of the sound
- Timbre — Quality of the sound or Audio character

Say, a wav file is an audio event created by an activity. The activity can be talk, buzzing, cracking, strolling, throwing water, etc. Since we detect events with very little audio, we have been training as people. Someone can state that people can learn new sound events quite effectively and can distinguish sound events. Hearing a podcast utilises the capacity to acoustically detect the audio solely. We may employ sound recognition to perceive the environment with other sensory information. However, it is quite tough to recognise audio events consistently (ideally with a computer or algorithm). This is mostly due of :

- The noisiness of recorded sound clips — transducer noise and background noise.
- An event can be occurring at various loudness levels and various time durations.
- Having a limited number of examples to feed into an algorithm.

Pre-Processing of audio

First, a way to represent audio snippets must be developed (.wav files). The voice signal should therefore be used as inputs to the learning algorithms of the system. The library Librosa offers some important features for the processing of python audio. In a numpy array, the audio files are loaded with Librosa. The arrays would be composed of the intensities of the audio recording at a rate named ’Sampling rate.’

The first difficulty, noisiness, should be solved after loading an audio clip onto a tablet. The Program employs the algorithm for ’spectral gating’ to reduce sound distortion. The noisereduce python package provides an excellent implementation of the method. You may find the algorithm in the noisereduce library documentation.

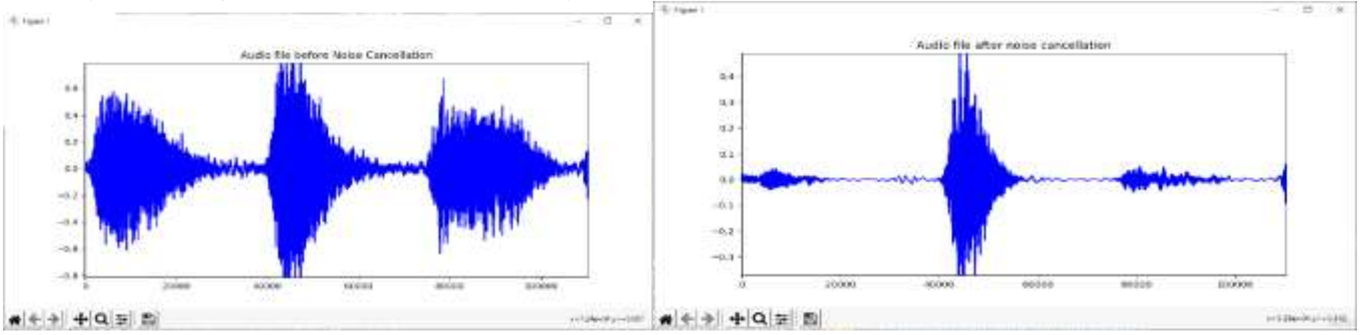

Fig. 1.

There is an empty length in the produced audio clip (unnecessary quiet). Let us reduce the leading and following sections of quiet above the loudness of a threshold level.

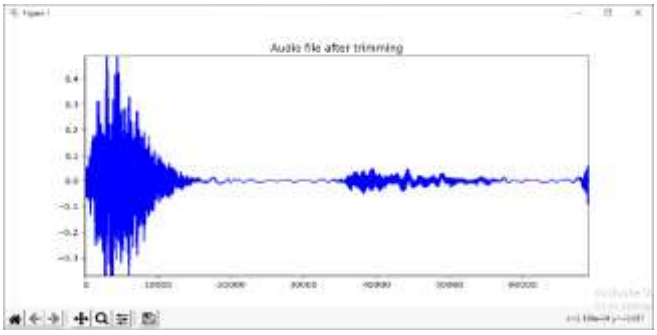

Fig.2.

Extraction of Feature

The pre - processed recordings can sometimes be utilised as sound events themselves. To make the categorization process more effective and precise, we must extract features out of the audio segments. Extract from each audio clip the absolute Fourier Transform Short-Time values (STFT). Fast Fourier transform size(n fft) of the window is utilised to compute STFT as 512. The n stft = n fft/2+1 is determined by means of a 512-size frame. 257 [16]frequency bins(n stft) will be calculated. In computing the STFT[19] the window is relocated by 256 hops in length to properly superpose the windows.

V. Experimentation And Evaluation

All of the below assessments are completed with 5-time cross-validation against cross subjects. Cross-subject tests were used to get independent responses from the subject. We have achieved an average accuracy of 70.06 percent with the described CNN-LSTM model in a profound sequence classification model.

TABLE II

Phases	Accuracy
Training	94.00
Testing	43.13
Real time	61.69

The efficiency of adopting a neural 3D-convolution network was also assessed to categorise depth sequences, resulting in 65.67 percent accuracy. Due to LSTM's ability to model sequences over CNN, the CNN-LSTM model was favoured.

VI. Conclusion And Future WorK

The efficiency of identified daily life activities using depth sequences and audio streams for the prospective application of environmental aided living systems was assessed in this article. A multi-modal dataset was prepared for assessment and a multi-modal fusion model was established. In comparison with the different stream classification, the efficacy of multimodal classification may be demonstrated by superior results.

We want to construct a real-time recognition system as part of the future work. We think that it is possible to further improve the precision of the multi module activity categorization. We have combined findings from individually trained models for various streams in this article. Additional techniques such as early fusion of audio and depth data.

VII. References

1. N. Takahashi, M. Gygli, and L. Van Gool, "AENET: Learning deep audio features for video analysis," IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 513–524, 2018.
2. Classification of Activities of Daily Living Based on2019 IEEE 14th International Conference on Industrial and Information Systems (ICIIS), 18-20 Dec., Peradeniya, Sri Lanka.

3. Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, “Sound event detection and time–frequency segmentation from weakly labeled data,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 27, no. 4, pp. 777–787, 2019.
4. 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, 2017.
5. L. Jing, B. Liu, J. Choi, A. Janin, J. Bernd, M. W. Mahoney, and G. Friedland, “DCAR: A discriminative and compact audio representation for audio processing,” IEEE Transactions on Multimedia, vol. 19, no. 12, pp. 2637–2650, 2017.
6. H. Malik, “Acoustic environment identification and its applications to audio forensics,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 11, pp. 1827–1837, 2013.
7. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746, 2015.
8. G. Mafra, N. Duong, A. Ozerov, and P. P´erez, “Acoustic scene classification: an evaluation of an extremely compact feature representation,” in Detection and Classification of Acoustic Scenes and Events, 2016.
9. Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. Jackson, and M. D. Plumbley, “Unsupervised feature learning based on deep models for environmental audio tagging,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1230–1241, 2017.
10. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, “Robust sound event classification using deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 540–552, 2015.
11. M. Cheffena, “Fall detection using smartphone audio features,” IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1073– 1080, 2016.
12. M. Crocco, M. Cristani, A. Trucco, and V. Murino, “Audio surveillance: a systematic review,” ACM Computing Surveys (CSUR), vol. 48, no. 4, p. 52, 2016.
13. S. Chachada and C.-C. J. Kuo, “Environmental sound recognition: A survey,” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013, pp. 1–9.
14. H. Phan, L. Hertel, M. Maass, R. Mazur, and A. Mertins, “Learning representations for nonspeech audio events through their similarities to speech patterns,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 4, pp. 807–822, 2016.
15. M. Lagrange, G. Lafay, B. Defreville, and J.-J. Aucouturier, “The bag of- frames approach: a not so sufficient model for urban soundscapes,” The Journal of the Acoustical Society of America, vol. 138, no. 5, pp. EL487–EL492, 2015.
16. J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste et al., “An exemplar-based nmf approach to audio event detection,” in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013, pp. 1–4.
17. S. Ntalampiras, I. Potamitis, and N. Fakotakis, “Probabilistic novelty detection for acoustic surveillance under real-world conditions,” IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 713–719, 2011.
18. "Activities of Daily Living Evaluation I • Encyclopedia.com." [Online]. Available: • https://www.encyclopedia.com/caregiving/encyclopediasalmanacs-transcripts-and-maps/activities-daily-living-evaluation. [Accessed: 10-Aug-2019].
19. Sound Event Classification: A to Z,,. Chathuranga Siriwardhana Computer Science & Engineering passionate Engineer: https://towardsdatascience.com/sound-event-classification-using-machine-learning-8768092beafc.