2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Acoustic Based Footstep Detection in Pervasive Healthcare

Summoogum. K, *Member IEEE, MiiCare Ltd*, Das. D, *MiiCare Ltd,* Dasgupta. S, *MiiCare Ltd*,
McLoughlin. I, *Senior Member IEEE*, Efstratiou. C, *Member IEEE*, Palaniappan. R, *Senior Member IEEE*

Abstract — **Passive detection of footsteps in domestic settings can allow the development of assistive technologies that can monitor mobility patterns of older adults in their home environment. Acoustic footstep detection is a promising approach for non-intrusive detection of footsteps. So far there has been limited work in developing robust acoustic footstep detection systems that can operate in noisy home environments. In this paper, we propose a novel application of the Attention based Recurrent Deep Neural Network to detect human footsteps in noisy overlapping audio streams. The model is trained on synthetic data which simulates the acoustic scene in a home environment. To evaluate performance, we reproduced two footstep detection models from literature and compared them using the newly developed Polyphonic Sound Detection Scores (PSDS). Our model achieved the highest PSDS and is close to the highest score achieved by generic indoor AED models in DCASE. The proposed system is designed to both detect and track footsteps within a home setting, and to enhance state-of-the-art digital health-care solutions for empowering older adults to live autonomously in their own homes.**

I. Introduction

Older adults tend to have a higher risk of physical accidents or falls as part of their daily lives [1]. The overall physical decline associated with aging can make such accidents the cause of serious complications with long term effects on an individuals' health and wellbeing. Prior work [2] has demonstrated that the gait characteristics of an individual can be linked to potential risk of accidents. Indeed, detecting the walking patterns of older adults using wearable devices has been explored in the past, relying on the use of accelerometer sensors to capture and analyse gait characteristics. However, evidence suggests that older adults generally do not find smart wearables useful [3]. A recent study [4] concluded that whether an older adult actually uses a medical wearable in the long run directly depends on their present well-being and their need to continue wearing it. This raises the question of suitability of wearable technologies for long-term continuous tracking of walking patterns for older adults.

In this work we explore the feasibility of employing acoustic sensing, through sensors embedded in the environment, to detect and analyse walking patterns of older adults. Our objective is to develop techniques that will enable acoustic sensing devices deployed within a home (i.e., coexisting with voice assistants and similar devices), to act as gait sensing devices, namely for the detection of walking activities at home, and subsequent analysis of gait patterns within those activities. We term this approach *acoustic gait analysis* (AGA), which can be considered as a special case of acoustic event detection (AED) or acoustic event analysis (AEA).

Traditional AED/AEA focus on the use of acoustic signals to identify general events that occur in specific settings. These can include events such as cooking, conversation, street traffic, etc. Our review of the literature available for AED and AEA provided three key insights:

1. Most available datasets for AEA and AED contain nonoverlapping sounds, except for a few from DCASE [5].
2. Many of the methodologies and models are developed and trained using acoustic datasets [6] recorded in controlled laboratory environments. These contain little or no background acoustic noise.
3. There are very limited publications and research into applied AEA/AED for healthcare, and more specifically, in the field of geriatrics.

The few publications [7] that do focus on AGA in indoor/domestic environments propose methodologies that predominantly assume isolated, non-overlapping audio events. However, real life recordings in a home environment are noisy and contain acoustic events that overlap with each other randomly (e.g., footsteps while the TV is on). Hence existing methodologies may not be suitable for use in real world settings.

This paper summarises our approach to (i) use noisy overlapping audio stream in home environments of older adults; (ii) develop a footstep detection model that can isolate audio windows where human footsteps are present; (iii) demonstrate a novel application of the Attention mechanism and the Bidirectional LSTM layer to improve model performance over that for simple DNN, CNN or LSTM based models; and (iv) introduce a novel post-processing algorithm for "confidence" of predictions of presence of footsteps in detected audio windows.

In Section II, we distinguish between overlapping and nonoverlapping AED and critically analyse the methodologies from two representative state-of-the-art approaches from this research area. In Section III, we explain our method and performance compared to those approaches, using a synthetically generated overlapping and noisy footstep sound dataset. In Section IV, we apply a recent evaluation criterion called the Polyphonic Sound Detection Score [8], adopted by DCASE, as the competitive metric for home environment-based AED performance. We use this criterion on the detection results from the aforementioned models to compare their effectiveness against two alternative models driven by the current state-of-the-art in detecting footsteps from a real-world audio setting containing noisy and overlapping sounds.

II. BACKGROUND

A. Overlapping vs Non-overlapping AED

AED allows the detection of multiple acoustic events in an audio signal in contrast to typical classification problems that assigns a class label to a recording of one acoustic event. The complexity of AED largely depends on the way multiple acoustic events can occur in an audio signal. A simple AED

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

task comprises detection of a sequence of temporally separated acoustic events [8, 9]. A more complex AED task involves detecting multiple overlapping acoustic events in the audio signal [10]. We use the term non-overlapping AED for the former and overlapping AED for the latter. Fig. 1 illustrates the difference between non-overlapping and overlapping AED.

In this paper, the environment we target is the indoor residential home of an older adult; sounds in this environment are typically noisy and involves overlapping acoustic sources such as television, telephone, kettle, doors, etc. as well as noises from residents and visitors.

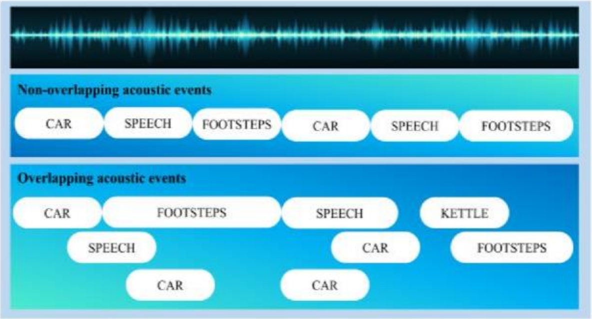

Figure 1. Non-overlapping v/s Overlapping audio [11]

B. State of the Art in AED

Haubrick et. al. [10] proposed a stacked multi-class neural network classifier using features extracted by the pretrained VGGish [12] model. This was trained on the AudioSet [13] dataset, to generalise over 635 acoustic events in different acoustic scenes. The training data distribution of the VGGish model includes a majority of *outdoor* events (which may be overlapping as well as non-overlapping) and therefore clearly differs from a residential *indoors* environment. We also note a contrast in the spectrum of typical acoustic events: Human footsteps have a frequency spectrum predominantly between 10 - 300 Hz [14] while AudioSet events (speech, music, vehicles, animal sounds, etc.) generally have dominant content at higher frequencies. Furthermore, the VGGish preprocessing pipeline computes STFTs over 960ms frames, far longer than typical single footstep durations of around 400-600ms [14]. The STFT output is then integrated into 64 mel-frequency bins. The Mel spectrum and the well-known MFCC acoustic feature are generally used for *speech* analysis because the mel scale approximates how audio loudness is perceived by the human ear. However, 64 mel-bins over the audible frequency range does not yield good features for *footstep* events because their perceived loudness is severely reduced on a mel scale and the useful information is largely concentrated into only a few bins. The four mismatches, namely frequency, frame-length, bin-spacing, and training data, strongly motivate our proposed approach.

Before the advent of Deep Neural Networks (DNNs) for audio classification, AED and AGA relied on classifiers using Support Vector Machines [7], Gaussian Mixture Models [15], Hidden Markov Models [16], Non-Negative Matrix Factorisations [17] and so on. For the related acoustic footstep detection task, we will thus compare our proposed DNN-based approach against a good example, the Nakadai et. al. [7] SVM-based footstep detector. In that work, the audio data covered four acoustic event classes (footsteps walking, footsteps running, handclaps and speech), but all with nonoverlapping and clean sound events. As mentioned in Section I, such datasets are commonly used for AED, but are not suitable for evaluating systems that aim to operate in real world home environments. We thus re-evaluate both systems [7, 10] for the noisy, overlapping footstep detection task.

III. METHODOLOGY

Our objective in this section is to demonstrate the effectiveness of a novel application of an Attention based Recurrent Deep Neural Network (RNN) model for detecting footsteps from overlapping audio over traditional machine learning and neural network-based models. We evaluate our work by comparing it with our reproductions of Nakadai et.al [7] and Haubrick et. al. [10] (which uses SVM and CNN respectively) on our synthetic footstep detection dataset. All systems are trained and evaluated using synthetic audio data. The process of synthesising appropriate data was employed to allow us to generate a range of scenarios involving overlapping sounds that consists of footsteps as well as other ambient sounds.

A. Synthetic Overlapping Audio Dataset generation

Although there is a vast collection of public acoustic data for general acoustic event classification, most of the available datasets do not contain acoustic data for footsteps. Available public sources that do, include the TUM GAID [6], ESC50 [18] and AudioSet. TUM GAID from Technische Universität München consists of 3-second 16 kHz noisy footstep clips from 305 individuals, recorded in 3 different scenarios. These include walking normally, walking with a 5kg heavy backpack and walking with coating shoes, and incorporates variations in gait. ESC50 consists of 2000 5-second clean audio clips from 50 different acoustic event classes at 44.1 kHz, out of which 40 clips are of human footsteps. ESC50 was compiled from the publicly available Freesound Project [19]. AudioSet offers machine-labelled low-quality footstep events across 1683 Youtube videos detected at a 90% confidence level.

To train and evaluate our work we relied on audio synthesis to create a sizeable dataset of sounds that contain a mix of overlapping and clean sounds, including footsteps, and other sounds that are typically present in indoor environments. The synthetic dataset was produced using “scaper” [20], a library recommended by DCASE, to generate overlapping audio files with a random number of predefined acoustic event classes and random ambient background noise. The source files used for the synthetic dataset consist of three classes:

- *Footsteps*: sources from TUM GAID and ESC50 sets, containing 3,400 and 40 footstep audio files
- *Ambient home sounds:* sources from HoME dataset, 3,440 randomly sampled files of various acoustic events in indoor scenes; DEMAND48 dataset [21] ambient sounds in Living Room, Hallway and Kitchen.
- *Noise:* An hour-long Brown Noise [22] audio and an hour-long White Noise [23] audio from YouTube.

We generated the synthetic dataset with the intention to produce a random mixture of overlapping sounds using these sources. We produced 2,000 12-second clips at 44.1 KHz with the following configuration for the synthesis: background noise levels at -3 dB; events per clip uniformly selected between 1-5; duration of events per clip $\mu = 5\ minutes, \sigma = 2$; duration of background noise uniformly selected between 2-5 seconds; SNR for events with respect to noise uniformly selected between 6-30; pitch variation for every event uniformly selected between -3 and 4 octaves; time stretching variation for each event uniformly selected between 60% and 130%. We chose a clip duration of 12 seconds as a trade-off between adequacy for realism in the study, and of storage and computing resources. The 1200-clip dataset was found to have 2494 footstep events (labelled “footsteps”) and 2510 other acoustic events (found in the home environment, labelled “others”) with an average of 3 acoustic events per clip which may or may not overlap. The number of clips in the training, validation and testing dataset are 867 (72.25%), 153 (12.75%) and 180 (15%) respectively.

B. Reproducing Nakadai et. al. on our synthetic dataset

The specific model adopts a typical shallow approach for acoustic event classification. As a pre-processing step, the system identifies a potential event directly from the input audio signal by detecting peaks separated by a pre-set minimum distance and surrounded by a pre-set background noise level. The audio segment around the peak is clipped and pre-processed to extract 6 time-domain, 7 spectral, 4 geometric and 24 MFCC features and form a 36-dimensional feature vector. The 36-dimensional feature dataset is directly fed into a multi-class SVM with Radial Basis Function (RBF) as the kernel.

C. Reproducing Haubrick et. al. on our synthetic dataset

The first component of the proposed system uses the VGGish pretrained model as a feature extractor. Pre-processing thus follows the pretrained model setup process described in the original paper [12]. The output is a 128-dimensional feature dataset. We use a three 3-layer deep binary classifier with yielding 6-dimensional vector of output probabilities which are then refined in a 2-layer, 2-node classifier, since our dataset has only two acoustic event classes: ‘footsteps’ and ‘everything else’.

D. Implementing our proposed system

Our system performs footstep detection in 3 phases. In the first phase, an audio signal is converted into a spectrogram matrix and split into 2-second windows with 50% overlap. In the second phase, the DNN classifies each spectrogram window as ‘1’ (footsteps) or ‘0’ (others). In the third phase, predictions for the overlapping windows are post-processed to infer the presence of footstep events.

Our model architecture is based on the findings of [24, 25]. The rationale is [24] reported a 15.1 % increase in $f(1)$ score for multilabel AED using Bidirectional LSTMs over DNNs while [25] found that Bidirectional LSTMs when paired with an Attention layer is better for Acoustic Scene Classification than traditional NN structures like DNN, CNN or simple LSTMs. For pre-processing, the input audio is down sampled to 16 kHz, divided into the overlapping windows of duration 2 seconds and overlap of 1 seconds, converted into a spectrogram with a 40ms frame window, NFFT window of 64 ms and overlap of 32ms. The model accepts the transpose of the spectrograms, directly as input. We paired a Bidirectional LSTM-Attention layer with another LSTM layer for feature extraction. The three layers are then connected to a DNN with 3 hidden layers and 2 output nodes for classification. Figure 2 visualises our model architecture. We then introduce our novel post-processing algorithm wherein we reinterpret the binary output of our classifier model. We then converted the generated timestamp annotations from the acoustic signal synthesiser [20] for every synthetic clip into binary signals of same length as the clip. The 1’s in the binary signals represent the footstep windows in the corresponding synthetic clips. When we create overlapping spectrograms, the binary signal windows are converted into a single binary label. The label interpretation is set during conversion: a spectrogram window will be labelled as "1" if it contains footsteps events covering more than a threshold proportion (currently set to 60%) of the window duration. A "0" label indicates that a window has fewer footstep events than threshold, implicitly indicating stronger presence of "other" acoustic events. During inference, the binary prediction vector obtained for the input signal over time is pooled in pairs. The resultant vector belongs to the domain {0, 1, 2}. This provides a good estimate of model prediction confidence. Figure 3 illustrates the concept diagrammatically.

1. "0" now means that the model has no confidence in the presence of footsteps in an audio window
2. "1" now means that the audio window contains footsteps covering less than the threshold duration
3. "2" means that the audio window contains footsteps covering more than the threshold duration

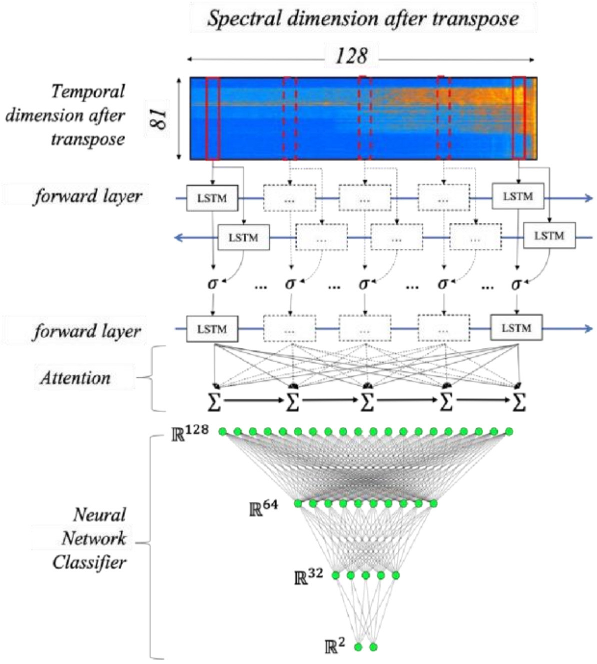

Figure 2. Proposed 3-layer DNN with bidirectional LSTM and Attention

IV. SYSTEM EVALUATION

Bilen et. al. [8] proposed a new metric of evaluation of AED models working with overlapping audio. It presents a score called Polyphonic Sound Detection Score (PSDS) which aims to deliver well-rounded insights into the performance of such AED models. Overlapping AED evaluation is prone to multiple misapplications of the $f(1)$ score and error rate miscomputations due to conventional boundary-based event annotations. We applied PSDS on the predicted boundaries of footstep and other acoustic events on the withheld dataset from Section III and compiled individual metrics for our two acoustic event classes, as depicted in Table 1.

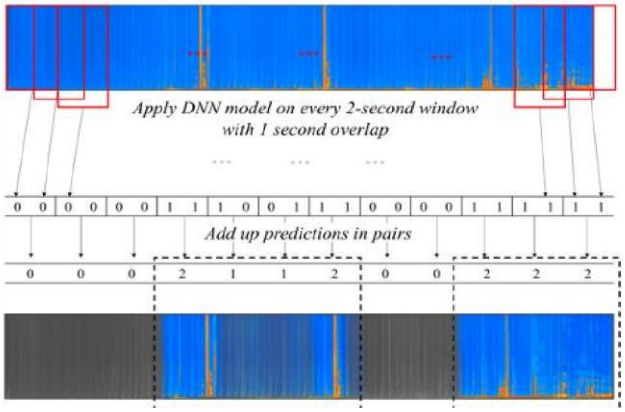

Figure 3. Our Inference Process for footstep detection

TABLE I. Evaluation of three implementations with PSDS, $f(1)$ and AUC

Model	***PSDS***	***$f(1)$ for "footsteps"***	***$f(1)$ for "others"***	***AUC***
Nakadai et. al.	0.38	0.61	0.71	0.66
Haubrick et. al.	0.41	0.86	0.83	0.86
Proposed approach	**0.65**	**0.89**	**0.94**	**0.92**

V. Conclusion

In this paper we presented a novel application of the Attention based Bidirectional LSTM DNN model with a novel post-processing algorithm to confidently isolate human footsteps in a noisy, overlapping home acoustic scene. We evaluated the proposed approach against two systems which employed SVM and VGGish respectively. The proposed architecture outperformed the other systems in $f(1)$ and AUC scores. This was confirmed when evaluated with the more recently defined PSDS metric for assessing overlapping sound detection. We believe model performance would improve further if benchmarked gait datasets from real environments become available instead of datasets which are recorded using professional equipment in a controlled environment. As future work, we therefore intend to use standard consumer equipment to record gait data from real environments and develop robust gait analysis systems that work with low quality audio containing overlapping events. In addition, we will be exploring acoustic source separation methods for acoustic events other than speech in older adult care environments and use real life data. Isolating footsteps will assist in extraction of temporal gait parameters for older adults living with or are showing symptoms of onset of dementia in commercial care homes.

References

[1] Berg, R. L., & Cassells, J. S. (1992). Falls in older persons: risk factors and prevention. In The second fifty years: Promoting health and preventing disability. National Academies Press (US).

[2] Verghese, Joe, et al. "Quantitative gait markers and incident fall risk in older adults." The Journals of Gerontology: Series A 64.8 (2009): 896-901.

[3] Yu-Huei, C., Ja-Shen, C., & Ming-Chao, W. (2019, August). Why do older adults use wearable devices: a case study adopting the Senior Technology Acceptance Model (STAM). In 2019 Portland International Conference on Management of Engineering and Technology (PICMET).

[4] Farivar, S., Abouzahra, M., & Ghasemaghaei, M. (2020). Wearable device adoption among older adults: A mixed-methods study. International Journal of Information Management, 55, 102209.

[5] Zhu, H., Ren, C., Wang, J., Li, S., Wang, L., & Yang, L. (2019). DCASE 2019 challenge task1 technical report. Tech. Rep., DCASE2019 Challenge.

[6] Hofmann, Martin, et al. "The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits." Journal of Visual Communication and Image Representation 25.1 (2014): 195-206.

[7] Nakadai, Kazuhiro, Yuta Fujii, and Shigeki Sugano. "Footstep detection and classification using distributed microphones." 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). IEEE, 2013.

[8] Bilen, Çağdaş, et al. "A framework for the robust evaluation of sound event detection." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[9] Zhang, Haomin, Ian McLoughlin, and Yan Song. "Robust sound event recognition using convolutional neural networks." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP).

[10] Haubrick, Peter, and Juan Ye. "Robust audio sensing with multi-sound classification." 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom. IEEE, 2019.

[11] Heittola, Toni, et al. "Context-dependent sound event detection." EURASIP Journal on Audio, Speech, and Music Processing 2013.1 (2013)

[12] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017.

[13] Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[14] Ekimov, A., & Sabatier, J. M. (2006). Vibration and sound signatures of human footsteps in buildings. The Journal of the Acoustical Society of America, 118(3), 2021-768.

[15] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S.Huang, "Real-world acoustic event detection," Pattern Recognition Letters, vol. 31, no. 12, pp. 1543–1551, 2010.

[16] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in Signal Processing Conference, 2010 18th European. IEEE, 2010, pp. 1267–1271

[17] Gemmeke, Jort F., et al. "An exemplar-based NMF approach to audio event detection." 2013 IEEE workshop on applications of signal processing to audio and acoustics. IEEE, 2013.

[18] Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1015-1018).

[19] Fonseca, et al. "Freesound datasets: a platform for open audio datasets."

[20] Salamon, Justin, et al. "Scaper: A library for soundscape synthesis and augmentation." 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2017.

[21] Thiemann, J., Ito, N., & Vincent, E. (2013, June). DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments. In Proc. Meetings Acoust.

[22] Audio from YouTube Video available at https://youtu.be/FcWgjCDPiP4

[23] Audio from YouTube Video available at https://youtu.be/lzmSKX5TF3g

[24] Parascandolo, G., Huttunen, H. and Virtanen, T., 2016. Recurrent neural networks for polyphonic sound event detection in real life recordings. arXiv preprint arXiv:1604.00861.

[25] Guo, J., Xu, N., Li, L. J., & Alwan, A. (2017, August). Attention Based CLDNNs for Short-Duration Acoustic Scene Classification. In Interspeech (pp. 469-473).