Tracking daily routines of elderly users through acoustic sensing: An unsupervised learning approach

Pavlos Nicolaou
School of Engineering, University of Kent, UK
pn218@kent.ac.uk

Christos Efstratiou
School of Computing, University of Kent, UK
c.efstratiou@kent.ac.uk

***Abstract*—Assistive technologies that can passively track people's daily activities with dementia can deliver significant benefits for the patients themselves and their carers. This work investigates the feasibility of developing a system for the unsupervised tracking of daily activities at home through acoustic sensing. Motivated by the wide adoption of intelligent voice assistant devices in home environments, we developed a prototype algorithm to identify diversions from typical activities using the captured sounds, without the need for activity labeling. The system relies on sound embeddings through a pre-trained model, a novel dimensionality reduction algorithm, and the application of dynamic time warping for pattern matching. Our evaluation through synthetic activity sequences using data from our data collection in addition to public datasets shows very good performance (precision 0.99, recall 0.95).**

***Index Terms*—assisted living, unsupervised learning, dementia**

I. Introduction

There is a growing demand for the development of assistive technologies that can provide support for elderly people living with dementia. As the specific condition can progressively lead to a significant deterioration of the ability of people to function without support, there is great value in the design of a system that can monitor their condition unobtrusively and alert carers when there are signs of significant cognitive decline.

A vital sign that can indicate deterioration of cognitive abilities for people with dementia is a progressive difficulty following their typical daily routines, where specific regular tasks, such as having a meal, can be skipped or repeated in short intervals. In this work, we explore the design of a system for the passive tracking of daily activities to detect diversions from regular routines.

Considering the well documented limitations of deploying wearable technologies to support people living with dementia [1], in this work, we consider the feasibility of employing acoustic sensing technologies as the primary modality for monitoring the daily activities. Indeed, the wide adoption of smart assistant devices like Alexa and Google Home has also motivated the development of similar products tailored for elderly people (e.g. MiiCube [2]). These technologies offer the opportunity to utilise sound as a rich sensing modality that can be used to provide assistive applications. As part of an ongoing collaboration with the company MiiCare Ltd – a supplier of smart assistants for the elderly – we explore the feasibility of using acoustic sensing to identify diversions from the typical daily routines of elderly people with dementia. Considering the significant challenge of collecting datasets with appropriate labelling of the activities performed, our approach investigates the development of an unsupervised approach of tracking daily routines through sound.

In this preliminary work, we present an architecture for the unsupervised tracking of daily routines through sound and detecting changes of the regular sequence of activities or the "skipping" of particular activities as part of the daily routine. The general idea behind our approach is to consider a small set of sound sequences as representative patterns of typical activities. The detection of diversion from these patterns is performed through the combination of mapping sounds into multi-dimensional embeddings through the use of a pretrained model (VGG-ish [3]), the reduction of the dimensionality of the produced embeddings through a novel approach presented in this paper, and the application of Dynamic Time Warping as a pattern matching technique. This paper presents preliminary system performance results, evaluated through synthetic data using public datasets of sounds within domestic environments, and data collected through a controlled study with healthy participants. Using the public dataset, the system achieves 0.74 precision and 0.93 recall and using our collected dataset the system obtains 0.99 precision and 0.95 recall.

II. Related Work

A. Daily Activities with Dementia

People diagnosed with dementia, depending on the stage of their condition, may experience a significant reduction in their ability to perform daily activities or follow regular routines [4]. They may, for example, face increased difficulty getting dressed or may skip brushing their teeth as part of their morning routine. These alterations in elderly peoples' routines can introduce challenges for carers, who may need to intervene in such situations. Monitoring and recording changes in their daily routines can be a valuable tool in tracking the progress of their condition.

There has been extensive work in the use of assistive technologies for people with dementia. Riikonen et al. [5] emphasise in their work that the most valuable technologies

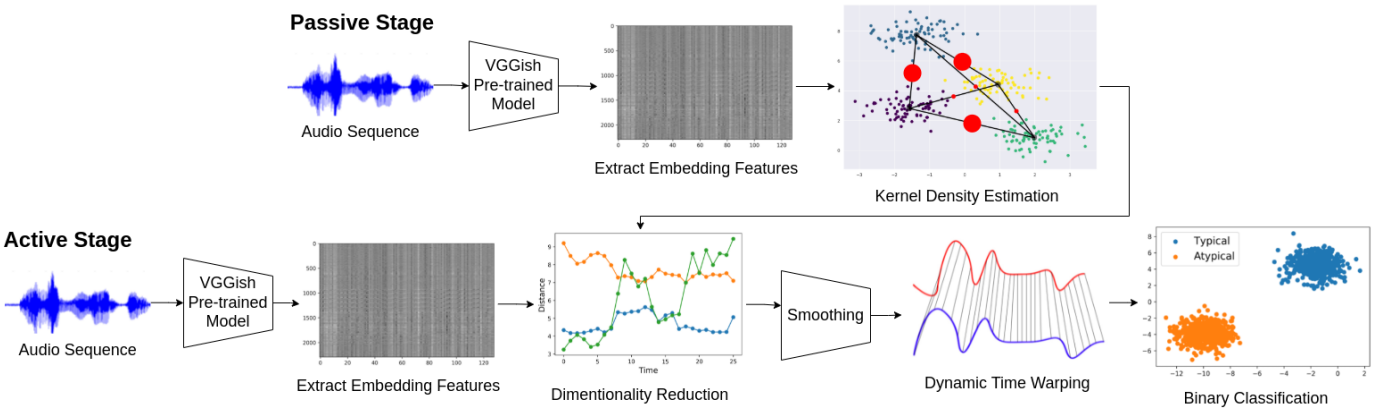


Fig. 1: Pipeline: The process aims to analyse a stream of sounds captured over a long time window where the user performs a typical routine (e.g. morning routine) and compares the sound patterns with previously captured "normal" routine for that individual. The process does not rely on the identification of the specific activities, and therefore does not require labeled data for training.

do not require any interaction by the user and can operate passively.

B. Acoustic Scene Classification

Recognising various acoustic environments from recorded acoustic signals is an active research field that has received significant attention. Acoustic scene classification (ASC) is the task of recognising the acoustic environment based on the recorded acoustic signal, for example, "office" or "park" [6]. Interestingly, recent results in this domain demonstrate that the state-of-the-art of acoustic scene classification can outperform humans on the same task [7]. ASC algorithms have matured and are already in real-world application scenarios. However, one of the main challenges in this domain is the difficulty of collecting well curated datasets. One of the most popular methodologies to apply ASC in new domains is using a pretrained model (PANNs [8]) and applying transfer learning. In this work we adopt this approach as part of the early processing of acoustic signals.

C. Unsupervised activity recognition

Supervised methodologies are prevalent in activity recognition applications although, manual labelling of large datasets is expensive and time-consuming. These challenges are particularly difficult when considering labeling of acoustic data where the particular sounds for similar activities can differ greatly between users [9].

Unsupervised learning can be used to discover recurring sequences of activities captured by human activity recognition sensors with unlabeled data. These methods have been used for behaviour pattern discovery in smart home environments before, and they can also be used for providing proactive assistance from home [10]. Moreover, unsupervised sensing where specific activities are not explicitly detected, can offer an additional layer of privacy with respect to the output of the specific system.

III. Motivation

By 2020, 50 million people were estimated to be affected by dementia worldwide [11]. Furthermore, the number of cases is rising by about 10 million every year [12]. In the face of these projections, there is an increasing need to provide technology that can observe their daily lives and the progress of their condition while reducing the need for unnecessary involvement of carers.

In this work, we consider the feasibility of using voice assistants within the homes of people with dementia to detect diversion of their daily routines, which could be a sign of cognitive decline. Considering example technologies like MiiCube, there is an opportunity to work with large acoustic datasets collected from the living environments of older users. However, the nature of these deployments makes it highly challenging to collect any ground truth information about the specific activities people perform.

The main objective of this work is to develop a system that can "learn" the typical activity patterns of users in an unsupervised way, using sound. Although the typical day of any person is not predictable, there are certain times that most people follow regular routines, for example, morning routine, mealtime, or bedtime routine. We consider the proposed system as a passive sensing tool to detect diversions within those regular routines. The system should identify diversions from the "typical" pattern by detecting skipped activities that were not performed in the typical order. Using only sound signals as input, the main idea is to transform acoustic signals into a low dimensionality time series that can indicate when the user switches from one activity to the next.

The proposed system does not need to identify the exact activities involved, and therefore does not rely on training over labeled data. Instead, with the transformation of acoustic signals into low dimension time series, we aim to apply time series pattern matching techniques like Dynamic Time Warp-

ing (DTW) [13] (commonly applied in gesture recognition [14] for example).

IV. Methodology

The overall architecture of the proposed system can be found in Figure 1. The system operates in two stages: the *passive stage* aims to collect acoustic data over a period of time in order to "learn" the sound characteristics of typical routines for each user; the *active stage* is where the system analysis acoustic data aims to detect diversions from the typical routines. The general pipeline employed involve extracting features from the acoustic signal, applying a technique for dimensionality reduction, smoothing, and a final stage of pattern matching using DTW.

A. Data Collection

The proposed system was developed and evaluated using two different datasets. The Freesound Audio Tagging 2019 dataset [9] dataset is a large public dataset of a range of sounds, typically used for ASC tasks. The dataset is labeled with ground-truth, but the labeling is only used as part of the evaluation process.

In addition to the public dataset, we evaluated our system through a small-scale collection of real-world sounds. We asked 10 participants to record their morning routine using their smartphones, over a period of 5 days. The data collection was performed through the AudioHive App [15], a purpose-built application that we developed for this study. The app allows participants to label their activities. Activity durations, and sequences varied across participants. There were activities of short duration (e.g. average 2 mins for "brushing teeth"), or much longer (e.g. average 10.6 mins for "having coffee"). Similar to the public dataset, the labels were only used as part of the system evaluation process.

Although beyond the scope of this paper we are currently in the process of collecting acoustic data from a real-world deployment of acoustic sensing within two care-homes in the United Kingdom, involving 20 participants diagnosed with dementia.

B. Feature Extraction

The approach followed for extracting features from acoustic signals is influenced by typical methods applied for acoustic scene classification [16]. Traditionally, features on sound signals are extracted by transforming the signal into the frequency domain. However, since the publication of the VGGish model in 2017 [3], the acoustic sensing community has increasingly explored the application of transfer learning to extract sound features that can be used for acoustic scene classification.

The VGGish model is a pre-trained CNN network that has been trained over the AudioSet [17] dataset. The model takes as input frames of 975ms of audio data, converts them into a spectrogram, applies the Mel-Frequency Filter Banks, and uses this as input for training for acoustic scene classification. Applying VGGish for direct activity recognition in new environments can produce poor results, as the model is trained over a fixed set of labels. Instead, the VGGish model is typically used as a feature extraction method [3] by stripping the last layer of the model and utilising the generated 128-dimensions feature vector produced by the model as a multi-dimensional embedding representing features that can be used for acoustic sensing.

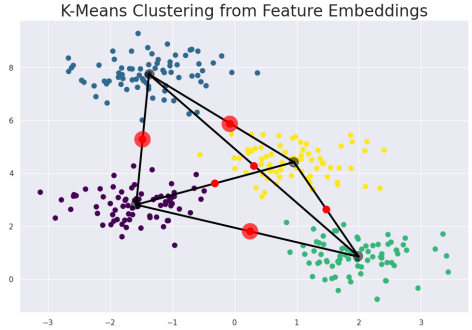


Fig. 2: Identifying Reference Points (illustration): We apply k-means over the dataset of domestic sounds. The red dots represent the mid points between cluster centres. We select the 3 points with the lowest density.

In our work, we employ a similar methodology for feature extraction. Our assumption for the VGGish produced embeddings is that the model is trained to provide high discrimination between sounds of different activities and closer similarity between sounds of the same activity. Preliminary investigation through the AudioSet dataset has validated that assumption. Indeed using Euclidean distances over the embedding vectors between audio samples of similar activities (e.g. cooking) are significantly smaller than the distances across activities (e.g. cooking vs bathroom). We also notice that for certain sound types, there are significant variations with respect to the generated embeddings. Indeed, sounds within the same activity can vary over short periods. As the VGGish model operates over frames of less than a second, these variations are manifested as changes on the embedded vectors produced for the same activity.

In order to address the issues with high variability within the sound embeddings, we incorporate a smoothing function over the generated embeddings before they are forwarded to the next stage. Specifically, as we intend to identify long activities (such as cooking or eating), we aggregate the produced embeddings over a window of 1min. Essentially the pipeline operates over a 1min sliding window (50% overlap), where the VGGish embeddings are averaged to produce a single 128-dimension embedding for each 1min window.

C. Dimensionality Reduction

DTW is a common technique used for pattern matching of time series. However, DTW is not well suited for time series with high number of dimensions, as the complexity increases exponentially with each additional dimension. Commonly DTW is applied on series with up to three dimensions.

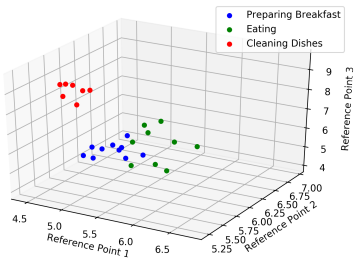


Fig. 3: Distances between cluster centers midpoints and sequence

As we intend to apply DTW over a time series of sound signals, it is essential to transform the 128-dimension embeddings into a significantly reduced dimensionality vector. Following our previous observation regarding Euclidean distances between embeddings, we consider using a small set of fixed reference points within the 128-dim embedding hyperspace and using the distances from these reference points as a new lower dimension feature vector. In the proposed system, we aimed to reduce the dimensionality of the embeddings into 3-dimensional feature vectors. Each value within the 3D vector should hold the distance of each sound sample from three identified reference points within the embedding space.

Selecting the appropriate reference points is essential to ensure that the calculated distances will still have a discriminatory effect in separating sounds of different activities. This is a key part of the "passive stage" of the overall system. Over the "passive stage" a range of domestic sounds are collected over multiple days. As the VGGish model is trained over a larger range of sounds (including a variety of outdoor acoustic scenes), we expect that the distribution of domestic sounds within a specific household would be relatively sparse when embedded into the 128-dimension space. Therefore, we aim to identify three reference points within that sparse space to help differentiate activities through their Euclidean distance.

Our approach for selecting the reference points is based on the following rationale:

- Reference points should not be within dense areas of sound activity in the embedding space: The rationale is that a dense area can contain representations of a specific activity, and a reference point within that space can have a highly discriminatory effect for sound samples of the same activity.
- Reference points should be "near" the areas in the embedding space where domestic sound activity is located: This way, estimated distances from sounds of different activities can have a more discriminatory effect.
- The three reference points should be far from each other: we want to minimise any correlations between the three dimensions produced.

The approach that satisfies these requirements is the following: We operate over a set of domestic sounds collected during the "passive" stage. Using the VGGish model, the dataset is transformed into a 128-dim data series of embeddings. We perform a sequence of k-mean clustering operations over the embeddings of all the domestic sounds of the household, with $k \in \{3, 4, 5, ...\}$. We produce k cluster centres for each clustering operation, which we assume are within dense areas in the embedding space. Using the cluster centres, we calculate the midpoint within each pair of them: a total of $k(k-1)/2$ candidate points for each clustering (Figure 2). We consider these points as candidates that can be "near" the areas of activities but potentially within low-density space. Using the Kernel Density Estimation (KDE) function fitted over the dataset of domestic sounds, we calculate the density for each candidate point. Finally, we identify the 3 candidate points with the lowest density as potential reference points. This process will generate a triplet of points for each clustering operation. The last step is to select the triplet with the maximum distance between the candidate reference points: For each triplet, we calculate the surface area of the triangle produced between the three points as an estimator of the distance between them. The final selection includes the triplet produced by one of the clusters with the highest surface area.

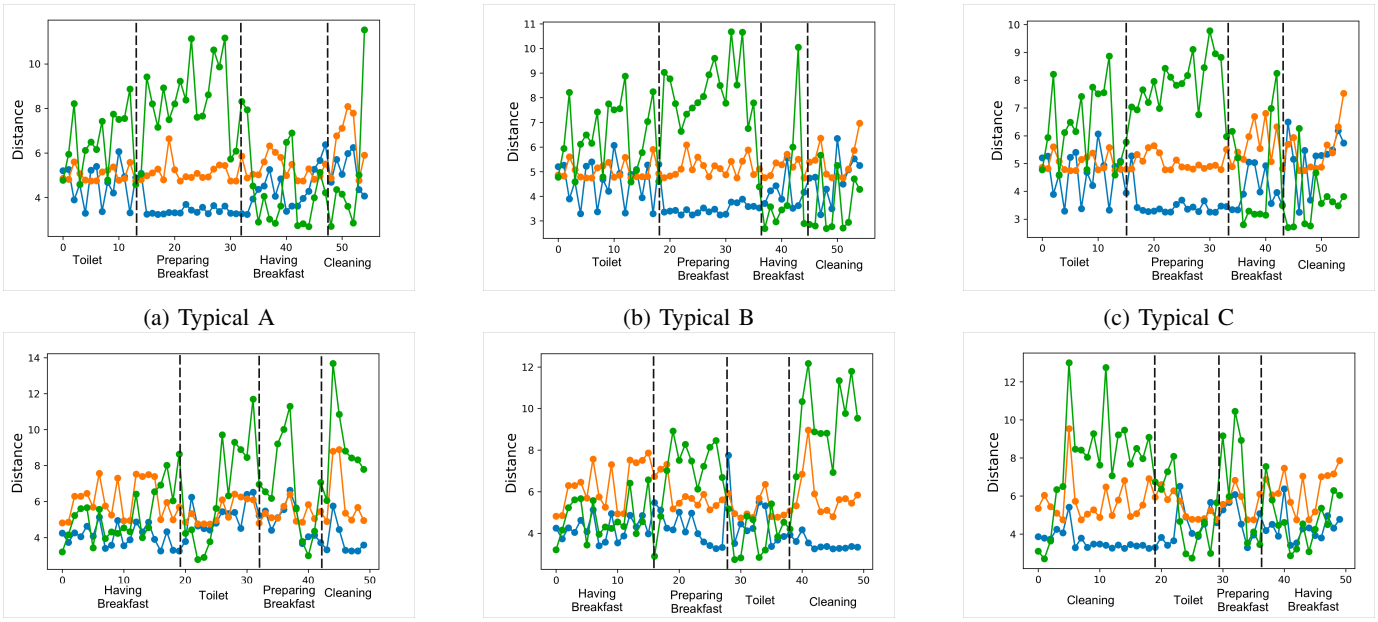
The identification of the three reference points allows us to apply dimensionality reduction over the 128-dimension embeddings. Specifically, we can calculate the Euclidean distance of each sound sample from the three identified reference points. For example, in Figure 3 you can see a sample of the distribution of three domestic activities after dimensionality reduction. Although there is a significant loss of information from the reduction to a 3D vector, the proposed approach can still maintain the discriminant characteristics of the specific features.

D. Pattern Matching with DTW

Dynamic Time Warping [18] is an algorithm that calculates the dissimilarity or distance between two time series while allowing the warping (compression or expansion) of the time axis in order to find the best alignment of the two series. Specifically, DTW calculates the distance between each possible pair of points within two time series. Through these it calculates the cumulative distance matrix and identifies the ideal warping path that minimizes the distance between the two series. When working with multi-dimensional time series the multivariate DTW [14] algorithm has been successfully used for gesture recognition using 3D-activity time series.

In this work we consider the use of DTW as a form of identifying activity sequences that are similar to the typical routine of the specific user, and flag sequences that are considered atypical. After the dimensionality reduction process, a time series of sound signals mapped into a sequence of 3D vectors, can be used for activity sequence pattern matching. Different sequences of daily activities can be used to identify the "typical" daily pattern of activities. Similarly new sound sequences can be classified using DTW based on their similarity with the typical activity patterns.

When monitoring a particular "routine" of activities (e.g. morning routine, bedtime routine), we consider a set of sound sequences collected over a few days, as the typical pattern



(a) Typical A (b) Typical B (c) Typical C

(d) Atypical A (e) Atypical B (f) Atypical C

Fig. 4: Distances from reference points and audio samples from the Data Collection. In the first row the sequences contain the same set of activities. The patterns of the three dimensions are similar although the duration can vary. Second row the sequences are different and this is reflected in changes in the overall patter compared to the "typical" sequence.

of the user. We term this the *training* set. Calculating DTW between the sample of sound sequences from the *training* set, allows us to estimate a threshold θ as the maximum DTW distance between sequences of the training set. When tracking new sound sequences, we calculate the average DTW distance of that sound sequence with all sequences of the *training* set. If the average distance is lower than θ the particular sequence is considered "typical", and when the distance is higher it is considered "atypical".

V. Results

We evaluate the performance of the proposed model using two datasets: the public Freesound Audio Tagging 2019 dataset [9], and our datasets collected through the AudiHive app [15] by 10 users over 5 days. Freesound Audio Tagging datasets contains 297,144 samples of audio data, accurately labelled with the activity they represent. For this work, we selected only a subset of the dataset domestic sounds. Similarly, the AudioHive dataset was manually labelled by the participants, with activities representing their morning routine.

A. Synthetic Sequences

In order to evaluate the model, we needed a sufficient dataset of both "typical" and "atypical" sequences of activities. We syntesised such sequences by combining data from the original datasets, stitching sound samples of different duration from various activities.

In particular, we generated three new datasets through the stitching of activities from the AudioHive dataset and three new datasets from the Freesound Audio Tagging 2019 dataset.

- Typical sequence: We defined a sequence $S_t = a_1, a_2, ..., a_n$ of n activities selected from the complete set of activities within the dataset, to represent a typical set of activities performed in a household, e.g. preparing breakfast, cleaning dishes, etc. For each activity a_i within the sequence, we estimated the distribution of each duration as recorded by our participants. Based on this distribution, we generated samples of variable length within the range of two standard deviations from the mean duration for each activity.
- Reordered sequence: Using the same S_t we produced random re-orderings of the set of activities. These sequences contain the same activities as those in the typical sequence, but in randomly mixed order.
- Missing activity sequence: Using the S_t sequence generated a set of activities where one of the a_i activity is removed from the sequence.

Through this process, we generated 30 acoustic signals of "typical" sequences, 30 re-ordered sequences which represent the "atypical" set, and 30 sequences with a missing activity using the public dataset. Next, we did the equivalent with the AudioHive dataset; we created 30 acoustic signals of "typical" sequences, 30 re-ordered sequences, and 30 missing activities using the AudioHive dataset.

For the training of the DTW algorithm, we selected a random subset of 10 typical activities to estimate the acceptable range of DTW distances to classify a sequence as "typical". The validation set consists of the remaining 20 "typical" sequences and 60 "atypical" sequences.

The "passive" stage of the system involves the identification

TABLE I: Results of the selection of reference points using clustering. k=7 generates points with the widest distance between them.

Clusters (k)	Avg. Density 3 midpoints	Surface area
k = 3	20.50	1.47
k = 4	48.25	1.71
k = 5	11.24	2.52
k = 6	34.17	1.43
k = 7	8.027	2.73
k = 8	-11.57	1.10
k = 9	-12.78	1.22
k = 10	20.50	1.08

TABLE II: Performance of proposed clustering method, and comparison with baseline (PCA)

Method	Precision	Recall
Clustering (AudioHive Dataset)	0.99	0.95
Clustering (Public Dataset)	0.74	0.93
PCA (Public Dataset)	0.59	0.8

of appropriate reference points for dimensionality reduction. These are selected for each environment and the set of domestic sounds contained in each dataset. Table I illustrates the outputs of that process over a series of clustering steps. For each of these, we calculated the midpoint between all cluster centres, selected the three midpoints with the lowest density, and calculated the surface area between them. In this case, midpoints from $k = 7$ cover the largest surface area.

During the "active" stage, the system uses these reference points to reduce the dimensions of any 1min sound sample. Figure 4 shows samples of the produced 3-d vectors for different sequences. It can be observed visually that sequences of the "typical" pattern demonstrate similarly shaped time series, whereas those of "atypical" behaviour do not. This indicates that the time series transformation into a 3D model produces suitable results for DTW pattern matching.

We evaluated the algorithm using the *validation* set, consisting of 20 correct, and 60 wrong activity sequences for each dataset. The results are shown in Table II. The results are also compared with the effects of a similar system that relies on a more traditional dimensionality reduction technique using pricipal component analysis (PCA). As shown the proposed algorithm can achieve very high performance. We note that the performance is particularly high for the real-world collection through participants. The reason for this high performance is that the "passive" stage analyses the patterns of the sounds that are generated within each household. This leads to a transformation that is tailored to the sound patterns produced by each participant. Instead, the public dataset consists of activity sounds from a range of different environments group together.

VI. Conclusions

This paper presents a novel technique for the unsupervised tracking of changes in daily routines using acoustic sensing. This work aims to develop a system that can detect significant changes in the daily routines of people with dementia.

The proposed system relies upon the VGGish model to generate embeddings of sound samples. A novel dimensionality reduction technique transforms the acoustic signal into a 3D time series of features. The application of DTW is then applied to match different patterns of activity sequences. The evaluation of the system through synthetic data achieves a precision of 0.99 and recall of 0.95. We are currently in the process of collecting acoustic data through a real-world deployment involving two care-homes in the UK. In our future work we intent to evaluate the performance of this technique under real-world conditions.

References

[1] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2012.

[2] K. Summoogum, J. Wall, O. Levy, S. Mantri, A. Smith, D. Das, H. Phan, I. McLoughlin, T. Whittaker, and D. Parsons, "Miicare," 2020, https://www.miicare.co.uk/.

[3] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[4] C. M. Giebel, C. Sutcliffe, and D. Challis, "Activities of daily living and quality of life across different stages of dementia: a uk study," *Aging & Mental Health*, vol. 19, no. 1, pp. 63–71, 2015.

[5] M. Riikonen, K. Mäkelä, and S. Perälä, "Safety and monitoring technologies for the homes of people with dementia," *Gerontechnology*, vol. 9, no. 1, pp. 32–45, 2010.

[6] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes acolornd events*. Springer, 2018.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "Assessment of human and machine performance in acoustic scene classification: Dcase 2016 case study," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 319–323.

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[9] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[10] D. J. Cook and N. Krishnan, "Mining the home environment," *Journal of intelligent information systems*, vol. 43, no. 3, pp. 503–519, 2014.

[11] W. H. Organisation, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[12] ——, 2021. [Online]. Available: https://www.who.int/publications/i/item/risk-reduction-of-cognitive-decline-and-dementia

[13] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[14] G. A. Ten Holt, M. J. Reinders, and E. A. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Thirteenth annual conference of the Advanced School for Computing and Imaging*, vol. 300, 2007, p. 1.

[15] P. Nicolaou, 2020. [Online]. Available: https://ubicomp-kent.org/projects/audiohive/

[16] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] J. B. Kruskall, "The symmetric time warping algorithm: From continuous to discrete," *Time warps, string edits and macromolecules*, 1983.