Evaluating Spectral Magnitude Representation and Spectral Energy for Audio-based Activity Detection

Anastasios Vafeiadis, Ioannis Papadimitriou, Anastasis Papanagnou, Dimitrios Giakoumis,
Konstantinos Votis and Dimitrios Tzovaras
Information Technologies Institute - Center for Research & Technology Hellas - Thessaloniki, Greece
E-mail: {anasvaf, i.papadimitriou, pkanastas, dgiakoum, kvotis, tzovaras}@iti.gr

***Abstract*—Acoustics has received a great research interest for human activity detection in indoor and outdoor environments. Compared to vision-based approaches, microphones can achieve a high percentage of recognition accuracy in a variety of activities, while not being affected by lighting conditions. Furthermore, audio-based activity detection can be considered an unobtrusive method, as long as the data is not related to speech or other sensitive information and no data is sent on cloud. Selecting the appropriate audio features that can achieve a high recognition accuracy and generalize in multiple domestic environments is a challenging task. In this work, three of the most commonly used spectrogram representations are evaluated, based on their spectral magnitude and the spectral energies. Specifically, using multi-channel audio data, the Short-time Fourier Transform (STFT), the Mel and the Gammatone spectrograms are extracted and trained on a 2D Convolutional Neural Network (CNN). The F1-Score of each feature representation are computed, while the McNemar tests and the Receiver Operating Characteristic (ROC) curves ensure the statistical independence between the magnitude and the energy representations. Extensive experimental results on a public database for detection of daily activities in a home environment, show that the overall highest recognition accuracy is achieved by the STFT magnitude representations.**

***Index Terms*—Sound event detection, activity recognition, convolutional neural networks**

I. Introduction

Human activity is one of the most important aspects of context information, which can be used for a plethora of ubiquitous applications [1]. Activity recognition is a tool that can provide real-life benefits in human-centered applications and sectors, such as healthcare and eldercare. It is one of the most promising topics for a variety of research areas, namely mobile and pervasive computing [2], [3], context-aware computing [4]–[6] and ambient assistive living [7]–[10].

Human activity recognition (HAR) mainly consists of four basic tasks; to choose and deploy appropriate sensors to objects and environment to monitor and capture relevant behaviour; to collect, store and process the data; to create models that learn the mapping between the aforementioned data and the set of activities of interest; and finally to develop algorithms to infer the activities from sensor data [11].

The sensor selection and deployment is crucial, as it needs to satisfy and comply with the nuances of the problem at hand, including the location, the ambient environment and the user profile and requirements. With regard to the type of sensor that is used for monitoring, activity recognition can be classified into two categories; vision based and sensor based [11]. The former is based on the use of video sensing facilities (e.g., video cameras) and exploits well-known computer vision techniques. Although playing a prominent role in the industry, image and video-based techniques are not always sufficient. For example, body pose can be used to classify a large range of human activities, but posture information cannot always provide unique evidence about the actions a human is engaged in, as quite different activities can be carried out in resembling body poses [12]. Thus, the second category has emerged, the sensor-based HAR. This usually involves the processing of time series data of state changes and various parameter values. Audio techniques can supplement this set of sensors or even be used on their own as a standalone solution for specific cases [13].

Until recently, the main focus of sound analysis research was on speech recognition [14], music classification [15] and speaker identification [16]. These methods' applicability on environmental audio analysis is limited, as there is a fundamental difference with regard to speech, in that there is no underlying phoneme-like structure. Furthermore, another important difference from speech recognition or speaker identification is that typically there is close proximity between the sound source (human speech) and the microphone, so as to ensure that the background sound energy is lower than the foreground one, not impairing the recognition system. This is not always true in the case of environmental audio classification [17], where the Signal-to-noise ratio (SNR) has been shown that can significantly affect the recognition accuracy [13].

To date, deep neural networks (DNNs) piqued research interest, owing to the fact that they outperform traditional classification techniques in several application domains, as well as that the cost of modern graphics processing units is relatively small. In the computer vision domain, CNNs have been widely adopted directly over raw video and image data in several application fields [18]–[20]. However, recently, deep learning methods are revealing their effectiveness also in applications of audio analysis; although a taxonomy of the published papers is still far from realized, two main emerging trends can be distinguished. The first consists of methods that analyze directly the raw audio data in the time domain by exploiting Deep Belief Networks or Restricted Boltzmann Machines [21], [22]. These approaches are related to the use of temporal-based features, which are not generally handcrafted but extracted

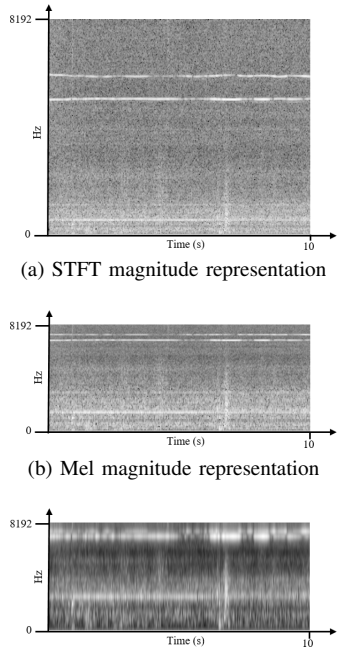
with the help of deep networks. The second trend consists of methods that use precomputed representations obtained by CNNs, starting from raw data. A good example is offered by the various time-frequency representations of the input signal, such as the STFT spectrogram or the Mel-Frequency Cepstral Coefficients spectrogram [23], [24]. AENet [25], SoReNet [26] and AReN [27] are recent contributions to this field and outstanding examples of a CNN fed by spectrogram images achieving very promising results for the problem of sound event recognition. Hence, it can be easily concluded that the representation automatically extracted by means of deep networks is definitively better in finding a high level representation of the data and is confirmed by various studies [28], [29].

Research using STFT, mel and gammatone spectrograms is abundant throughout the literature. The studies are either focused the magnitude representation [30] of spectrograms, or the corresponding spectral energies [31]. The selection of the spectrograms for environmental sound classification was based on their high recognition accuracy in various tasks such as music genre classification or speech recognition. Furthermore, the main focus of past research is typically on the CNN architectures for classification or novel data augmentation techniques. The scope of the present work is the extensive evaluation of three different spectrogram types (STFT, mel and gammatone) based on their magnitude representation, where the input features are pixel values from 0 to 255, and their spectral energies, where the input features are the frequency energies at each time frame. To the best of the authors' knowledge, this is the first attempt where the spectral magnitude representations and the spectral energies of the three aforementioned spectrograms are studied for the task of domestic audio-based human event detection.

The paper is organized as follows. In Section II the feature extraction methods and the network architecture used are presented. Section III demonstrates the experimental results, while the final conclusions are drawn in Section IV.

II. Methodology

Since the main objective of this study was the audio-based human activity detection in a domestic environment, it was important to use an annotated real-world dataset. Based on this, the dataset that was selected was the development set of a derivative of the SINS database [32], used for the Task 5 of the DCASE 2018 challenge. The particular part of the dataset consists of continuous audio recordings of approximately 200 hours of data from 4 sensor nodes of one person living in a vacation home over a period of one week. Each microphone array consisted of four linearly arranged microphones. The continuous recordings were split into audio segments of 10 s. Segments containing more then one active class (e.g., a transition of two activities) were left out, so each segment represented one activity. Data was labeled on daily activity level, ranging within nine different activities (absence, cooking, dishwashing, eating, other, social activity, vacuum cleaner, watching TV and working). For evaluation the cross-validation folds provided by the challenge organizers were used



(a) STFT magnitude representation

(b) Mel magnitude representation

(c) Gammatone magnitude representation

Fig. 1. 2D magnitude representations for the vacuum cleaner class. The x-axis represents the time (10 s) and the y-axis the frequencies (up to 8 kHz based on the Nyquist theorem)

The most common representation encountered in the literature for environmental sound classification, is the spectrogram. A spectrogram could present the linear frequencies (STFT), mel frequencies, or gammatone frequencies. The mel frequency spectrogram is a time-frequency visualization, but it is adapted on how sound is perceived by the human auditory system; most significantly, the ear's frequency sub-bands get wider for higher frequencies, whereas the spectrogram has a constant bandwidth across all frequency channels. A Gammatone spectrogram or gammatonegram is a time-frequency visualization based on a Fast Fourier Transform (FFT)-based approximation to gammatone sub-band filters, for which the bandwidth increases with increasing central frequency. The aforementioned three different-magnitude representation types, extracted from the raw audio using the LibROSA [33] library, were studied; STFT, mel (the same representation, with the only difference that the frequency axis is scaled to the mel scale using overlapping triangular filters) and gammatone spectrograms (using overlapping gamma distribution filters).

The parameters used throughout all the experiments were a sampling rate of 16 kHz, which was the original sampling frequency of the dataset. The four audio channels were averaged into a monophonic audio and an FFT size of 512 with 512 samples between successive frames (hop length). The FFT size and hop length were selected so as to result in an array that was as close to a rectangular shape, preserving

the same information in the frequency and the time axes. For the mel and the gammatone spectrograms, 128 mel bins and 128 gammatone bins were selected, respectively, since no significant information, except for weak harmonics, was noticed in higher frequencies. The colour mode for all representations was grayscale and the resulting shape for the STFT shape was 257×313×1, and 128×313×1 for the mel and gammatone spectrograms (Figure 1). Two ways of providing the input spectrograms for the models in the present study were used. The first was by feeding the model with image input (magnitude representation) and array input (spectral energy representation). The process of producing the former type of representation was done by using the *specshow* function of the LibROSA package, while the conversion from RGB to grayscale was carried out with the Rec.601 standard as described in [13].

For the training and testing processes, Python libraries such as TensorFlow [34], SciPy [35] and Spafe [36] were utilized. The deep learning network architecture that was selected for the features of this study was the B0 variant of EfficientNet (Table I) [37] and was trained on a Nvidia RTX 3090 graphics card.

TABLE I
EFFICIENTNET-B0 STRUCTURE FOR THE STFT REPRESENTATIONS

Stage	Operator	Resolution	#Channels	#Layers
i	$\mathcal{F}_i$	$H_i \times W_i$	C_i	L_i
1	Conv 3×3	257×313	32	1
2	MBConv1, k 3×3	129×157	16	1
3	MBConv6, k 3×3	129×157	24	2
4	MBConv6, k 5×5	65×79	40	2
5	MBConv6, k 3×3	33×40	80	3
6	MBConv6, k 5×5	17×20	112	3
7	MBConv6, k 5×5	17×20	192	4
8	MBConv6, k 3×3	9×10	320	1
9	Conv 1×1 & Pooling & FC	9×10	1280	1

As the goal of this study was to solely evaluate the effect of the selected audio features on the event-based detection of human activity, no augmentations in the audio or the image domain (e.g., pitch shift, dynamic range compression and image rotate), were applied.

III. RESULTS AND DISCUSSION

The models were initially set to train for 200 epochs. The Adam optimizer was used, with an initial learning rate of 0.001, β_1 of 0.9 and β_2 of 0.999. In order to avoid overfitting during the training phase, an early stopping criterion was applied, with a patience of eight consecutive epochs. Additionally, there was an extra reduction of the learning rate of the optimizer when there was no improvement in the validation macro F1-Score for 5 consecutive epochs.

In Table II the F1-Scores for all spectrogram approaches (STFT, mel-spectrograms and gammatonegrams) and input types (2D magnitudes and 2D energies) are shown. The proposed 4-fold validation split was used. The classes were unbalanced, skewed towards the absence, working and watching TV classes, which was the reason to use the macro F1-Score as the evaluation metric. An example (taken from [38] of the class balance can be seen in Figure 2. Fold 3 proved to

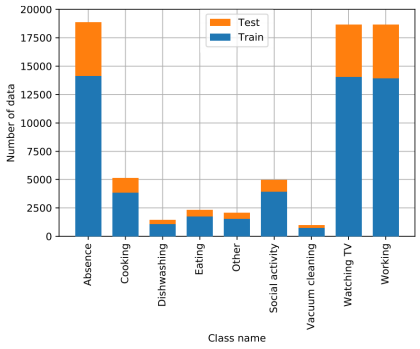


Fig. 2. Data distribution of each activity in Fold 1 of the development dataset (borrowed from [38]). Folds 2, 3 and 4 follow a similar distribution.

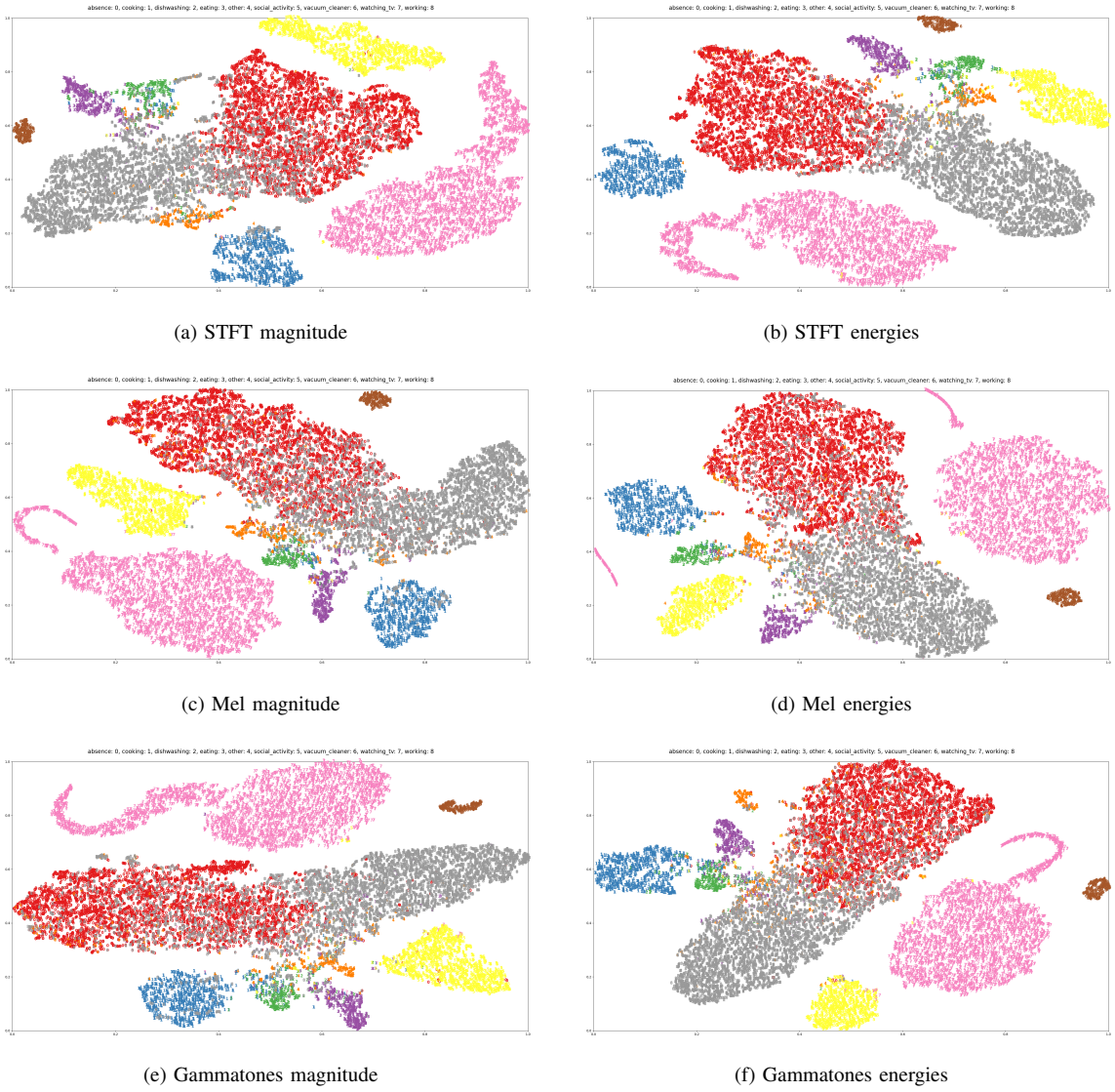
be the most challenging one with the F1-Scores ranging from 83.4% for gammatone magnitude representation to 87.92% for the respective STFT representation. On the other hand, the best performance was exhibited in Fold 4 where the F1-Score varied from 91.76% (mel magnitude representation) to 93.35% (STFT magnitude representation). On average, the magnitude representation showed better performance than the spectral energies in the case of STFT features, while the opposite was evident for the mel and gammatone features. McNemar testing was carried out to determine the proportion of errors between the selected features for the 4-fold evaluation setup. The selected p-value was set to 0.05 and comparing the spectral energies with the magnitude representations for the three audio features, the null hypothesis was rejected for each pair, ensuring they were statistically significant.

The albeit slightly better performance of the STFT representations compared to the other two, most probably owes to the fact that the STFT can map more abstract information. On the other hand, mel and gammmatone representations are more suited to speech as they both try to mimic the human auditory system, in the way of boosting lower frequencies and reducing higher ones. The former attempt that using triangular filters that abruptly cut specific frequencies, while the latter do that with gamma distribution filters which provide smoother filtering of the frequencies. It must be noted that the model trained on the STFT magnitude representation outperformed the baseline model in [39] in terms of F1-Score, which used mel spectral energies as input to a 1D CNN classifier, by 7.04%.

In Figure 3 the t-distributed Stochastic Neighbor Embedding (t-SNE) plots are shown. for all representations with the False Positive rate on x-axis and True Positive rate on y-axis. It can be seen that in all cases, vacuum cleaning (brown), social activity (yellow), watching TV (pink) and cooking (blue) are easier to distinguish from the rest of the classes in the dataset. Eating (purple) and dishwashing (green) are found neighbouring in all representations, while working (grey) and absence (red) are difficult to distinguish between, most probably owing to the background noise sharing the same features. The 'other' class (orange) is found neighbouring

TABLE II
F1-SCORE OF THE STFT SPECTROGRAMS, MEL-SPECTROGRAMS AND GAMMATONEGRAMS ON THE DEVELOPMENT SET OF THE SINS DATABASE

SINS dataset features		**Development Dataset F1-Scores**				
		Fold 1	Fold 2	Fold 3	Fold 4	**Average**
STFT	spectral energies	90.48%	89.66%	85.37%	92.69%	89.55%
	magnitude representation	90.99%	90.41%	87.92%	93.35%	**90.67%**
Mel	spectral energies	90.56%	89.82%	85.46%	93.02%	89.71%
	magnitude representation	89.89%	90.56%	86.05%	91.76%	89.57%
Gammatones	spectral energies	90.24%	90.21%	84.96%	92.43%	89.46%
	magnitude representation	89.32%	89.08%	83.4%	91.83%	88.4%



(a) STFT magnitude

(b) STFT energies

(c) Mel magnitude

(d) Mel energies

(e) Gammatones magnitude

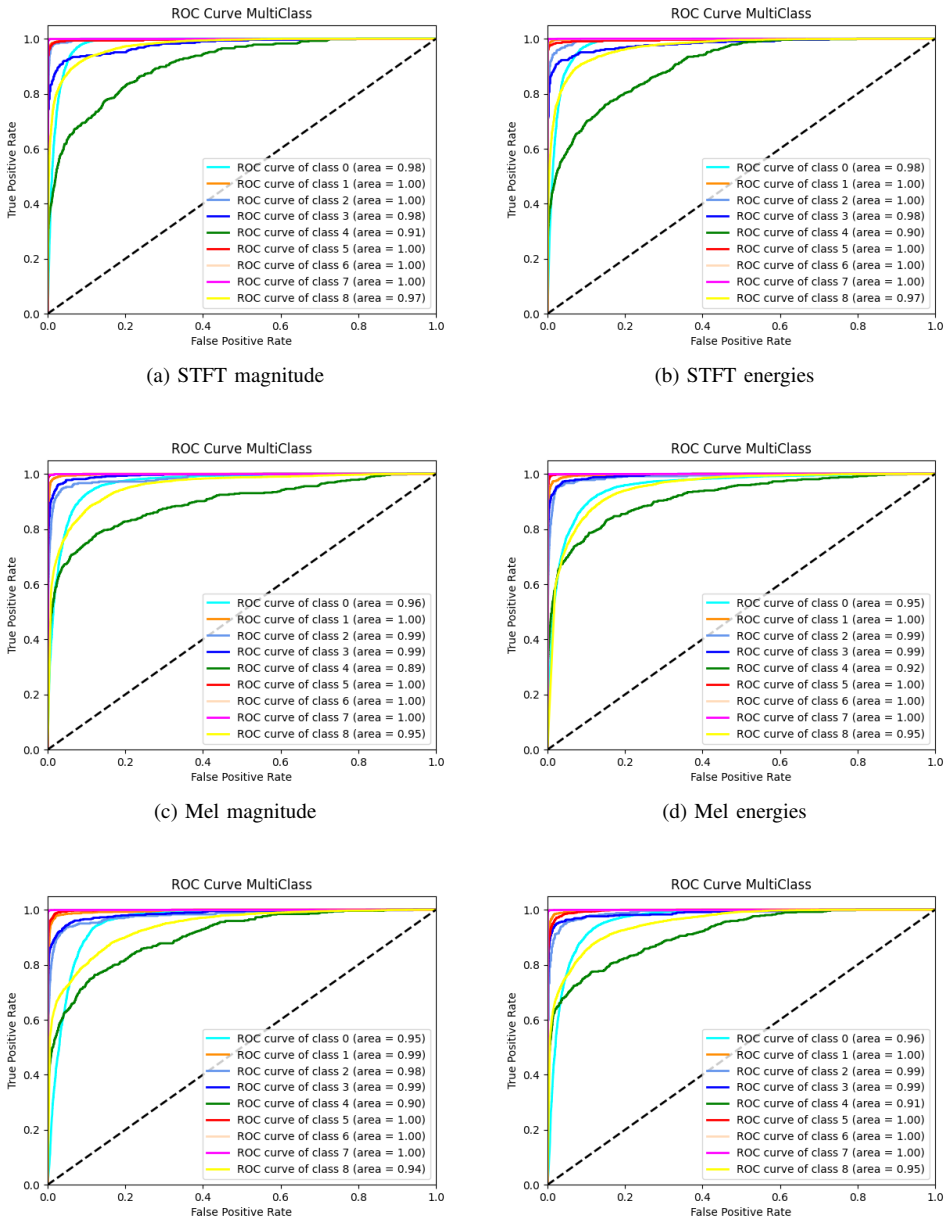
(f) Gammatones energies

Fig. 3. t-SNE plots for the nine classes. The mapping between the displayed colors and classes are as follows. Absence is the red color, cooking is in blue, dishwashing in green, eating in purple, other in orange, social activity in yellow, vacuum cleaning in brown, watching TV in pink and working in grey.

with different activity classes, depending on the representation, which was expected as it shares features with most of the classes. This is also evident in Figure 4, where the Receiver Operating Characteristic (ROC) curves for all classes and each representation are shown. In all cases the area under curve of the 'other' class was the smallest among all classes, ranging from 0.89 for the mel magnitude representation to 0.92 for the mel spectral energies.

IV. CONCLUSIONS

Recognition of human activity is typically associated with computer vision, but recently audio analysis has received a great research interest due to the fact that it can achieve



(a) STFT magnitude

(b) STFT energies

(c) Mel magnitude

(d) Mel energies

(e) Gammatones magnitude

(f) Gammatones energies

Fig. 4. ROC curves for the nine classes. Classes 0-8 are mapped alphabetically as follows. Absence, cooking, dishwashing, eating, other, social activity, vacuum cleaning, watching TV and working.

high recognition accuracy in a range of activities, while simultaneously being relatively unobtrusive and not affected by lighting conditions. The present study shows that simple audio based solutions can be applied to HAR. A comparison between STFT, mel and gammatone, magnitude and energies representations were compared, with STFT magnitude representation exhibiting the best performance in terms of F1-Score. STFTs are known to be better than mel and gammatone spectrograms due to the fact that they can map abstract information more accurately than mel and gammatone spectrograms that try to emulate human hearing and are more suited to speech rather than environmental audio. Environmental audio signals are non-stationary, which renders them not generalizable. Moreover, McNemar tests and ROC curves analysis showed the statistical independence between every pair of this study. Future work includes the use of simpler CNN architectures and inference on other public datasets for domestic human activity detection (e.g., DASEE [40]).

ACKNOWLEDGMENT

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First

Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Name: ACTIVE, Project Number: HFRI-FM17-2271).

REFERENCES

[1] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *International symposium on handheld and ubiquitous computing*. Springer, 1999, pp. 304–307.

[2] M. Weiser, "The computer for the twenty-first century scientific american," *September Elsevier Ltd*, 1991.

[3] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. Pedja, K. Koscher, A. LaMarca *et al.*, "The mobile sensing platform: An embedded activity recognition system," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 32–41, 2008.

[4] K. Van Laerhoven and K. Aidoo, "Teaching context to applications," *Personal and Ubiquitous Computing*, vol. 5, no. 1, pp. 46–49, 2001.

[5] C. R. Wren and E. M. Tapia, "Toward scalable activity recognition for sensor networks," in *International Symposium on Location-and Context-Awareness*. Springer, 2006, pp. 168–185.

[6] M. Stikic and B. Schiele, "Activity recognition from sparsely labeled data using multi-instance learning," in *International Symposium on Location-and Context-Awareness*. Springer, 2009, pp. 156–173.

[7] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel, "Inferring activities from interactions with objects," *IEEE pervasive computing*, vol. 3, no. 4, pp. 50–57, 2004.

[8] D. J. Cook and M. Schmitter-Edgecombe, "Assessing the quality of activities in a smart environment," *Methods of information in medicine*, vol. 48, no. 5, p. 480, 2009.

[9] L. Chen and C. Nugent, "Ontology-based activity recognition in intelligent pervasive environments," *International Journal of Web Information Systems*, 2009.

[10] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, "Recognizing independent and joint activities among multiple residents in smart environments," *Journal of ambient intelligence and humanized computing*, vol. 1, no. 1, pp. 57–63, 2010.

[11] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.

[12] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 509–514.

[13] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations," *Electronics*, vol. 9, no. 10, p. 1593, 2020.

[14] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[15] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE transactions on multimedia*, vol. 13, no. 2, pp. 303–319, 2010.

[16] A. Roy, M. M. Doss, and S. Marcel, "A fast parts-based approach to speaker verification using boosted slice classifiers," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 241–254, 2011.

[17] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Audio content analysis for unobtrusive event detection in smart homes," *Engineering Applications of Artificial Intelligence*, vol. 89, p. 103226, 2020.

[18] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 476–483.

[19] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-based object recognition using multimodal convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 43 110–43 136, 2019.

[20] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[21] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on affective computing*, vol. 8, no. 1, pp. 3–14, 2015.

[22] F. Guo, D. Yang, and X. Chen, "Using deep belief network to capture temporal information for audio event classification," in *2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. IEEE, 2015, pp. 421–424.

[23] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[24] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 559–563.

[25] 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, 2017.

[26] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "Sorenet: a novel deep network for audio surveillance applications," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 546–551.

[27] A. Greco, N. Petkov, A. Saggese, and M. Vento, "Aren: A deep learning approach for sound event recognition using a brain inspired representation," *IEEE Transactions on Information Forensics and Security*, 2020.

[28] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3407–3411.

[29] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.

[30] Z. Mushtaq and S.-F. Su, "Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images," *Symmetry*, vol. 12, no. 11, p. 1822, 2020.

[31] R. N. Tak, D. M. Agrawal, and H. A. Patil, "Novel phase encoded mel filterbank energies for environmental sound classification," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2017, pp. 317–325.

[32] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The sins database for detection of daily activities in a home environment using an acoustic sensor network," *Detection and Classification of Acoustic Scenes and Events 2017*, 2017.

[33] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[34] M. Abadi, A. Agarwal *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[35] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[36] "Welcome to spafe documentation." [Online]. Available: https://spafe.readthedocs.io/en/latest/

[37] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[38] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, and R. Tachibana, "Domestic activities classification based on cnn using shuffling and mixing data augmentation," *DCASE 2018 Challenge*, 2018.

[39] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "Dcase 2018 challenge - task 5: Monitoring of domestic activities based on multi-channel acoustics," 2018.

[40] A. Copiaco, C. Ritz, S. Fasciani, and N. Abdulaziz, "Dasee a synthetic database of domestic acoustic scenes and events in dementia patients environment," *arXiv preprint arXiv:2104.13423*, 2021.