Wake Word Based Room Identification with Personal Voice Assistants

Mohammadreza Azimi
School of Computer Science and Information Technology (CSIT)
University College Cork
m.azimi@cs.ucc.ie

Utz Roedig
School of Computer Science and Information Technology (CSIT)
University College Cork
u.roedig@ucc.ie

Abstract

Personal Voice Assistants (PVAs) are used to interact with digital environments and computer systems using speech. A wake word such as 'Alexa' is spoken by the user to initiate interaction with the PVA. We use the audio recording of the wake word to determine the room in which user - PVA interaction takes place. We collected data from 10 different rooms in which a user speaks the wake word at different locations. This dataset is used to evaluate three different neural network based algorithms for room identification. Our evaluation shows that rooms can be identified with 90% accuracy. The impact is twofold: (i) PVA audio recordings leak private information about the user environment; (ii) Acoustic room identification is an option for augmenting user - PVA interaction.

1 Introduction

PVAs such as Amazon Alexa or Google Home are now commonplace. We use these systems to interact with our environment and computer systems. A PVA records a user's voice and converts speech to text using Automated Speech Recognition (ASR). The obtained transcript is then interpreted by the system and actions are carried out. The system may then generate an audio response which is played back to the user via the PVA's integrated speakers.

The audio signals a PVAs records can be analysed to extract more information than speech. It is possible to identify gender [9], emotional state [14] or health condition [2]. It has also been shown that information about a user's environment can be extracted from the acoustic channel. Consequently, users are increasingly concerned about their privacy when using PVAs. In this work we describe how speech signals recorded with a standard PVA can be analysed in order to determine the specific room in which the user PVA interaction took place. The audio channel can be used passively to identify rooms. On the one hand, our research demonstrates that room information may be retrieved from audio signals, which may jeopardize users' privacy. On the other side, the capacity to recognize a room might be employed as a feature. The user's engagement with the PVA can be adjusted to the precise room in which the conversation takes place. Room identification can also be utilized as an added security feature to limit user interaction to specified rooms.

We assume in this work that either the PVA is mobile (a mobile Phone) or that it is a smart speaker that can be easily carried from room to room. We further assume that a wake word such as 'Alexa' is spoken by the user to initiate interaction with the PVA. We use recordings of the wake word as reference signal which is used to determine the room. The work presented in this paper builds on our previous work [1] in which we investigated room identification based on the existing database of the ACE challenge [3], containing speech samples (so called babble noise) collected in different rooms. However, this work differs regarding dataset and used algorithms used for room identification. For this work we collected a dataset specifically tailored to the outlined PVA context. A PVA was placed in 10 different rooms and a user spoke the wake word 'Alexa' multiple times at different locations.

The specific contributions of this work are:

- *Room Identification*: We describe three different methods for room identification based on a spoken wake word.
- *Wake Word Dataset*: We collected a dataset providing samples of the spoken wake word 'Alexa' at different positions in 10 rooms.
- *Evaluation of Room Identification*: We evaluate the proposed methods using the collected dataset. Results show that a total accuracy of up to 90 percent is achievable under specific conditions.

In the next section we discuss related work. Section 3 describes on a system level how room identification is used in a PVA context. In Section 4 we detail our dataset. Section 5 describes our methods for room identification and in Section 6 we present our evaluation results. The last Section concludes the paper and includes the final remarks.

2 Related Work

In 2012, the first system for performing room identification was proposed and introduced by Peters et al. [13].

The proposed system is based on a Gaussian Mixture Model (GMM) system that employs Mel-Frequency Cepstral Coefficient (MFCC) features. To train the model, a dataset of audio samples extracted from a video clip were collected and offered. Both speech and non-speech (music) samples were included. Our study differs as we employ neural network-based models and use a wake word dataset collected in a practical setting.

Moore et al. [5] proposed in 2014 to utilize a Gaussian Naive Bayes Classifier (GNBC) with Frequency Dependent Reverberation Time (FDRT) traits for room identification. A database of 484 Room Impulse Responses (RIRs) for 22 rooms varying in capacity from 29 to 9500 cubic meters was employed. The FDRTs served as the classifier's input feature. An Equal Error Rate (EER) of 3.9 % is achievable. The FDRTs features must be measured using specialized equipment. In our work, we use an ordinary phone to capture sound.

A novel method for identifying the place in which the capturing device is located, is proposed by Moore et al. [6]. Regarding the used database, the authors artificially created a dataset containing 400 samples from five different rooms. The original samples used for creation are taken from the evaluation dataset of the ACE challenge [4]. The main hand-crafted features are sub-band negative-side variance features that were extracted from the artificially created database. For classification of the extracted features Naive Bayes classifier were utilized. In our work we are using a real instead of generated data set. Moreover, we use different analysis techniques (i.e. neural network based techniques).

Hand crafted feature extraction heavily depends on human expertise and experience, while Deep Neural Networks (DNNs) seeks to circumvent these constraints by feature learning for model training automatically. Papayiannis et al. [12] proposed deep learning based techniques including a Convolutional Recurrent Neural Network (CRNN) with an attention-mechanism for performing room identification. In their studies, the CRNN classifier's accuracy of the classification is 78% with 5 hours of training data and 90% percent when using 10 hours. They have used an artificially constructed dataset to train the chosen models. In our work, a newly collected database is used. We explore also the possibility of transfer learning techniques instead of training for scratch.

In another research by Papayiannis et al. [11] a novel method for data augmentation for the training of DNN based room classifiers is proposed. The method relies on the training of Generative Adversarial Networks (GANs), which are used to generate artificially Acoustic Impulse Responses (AIRs) that increase the training data available for the classifiers. This is a novel idea, but still the database is artificially created. In our paper, we have collected a real database by the use of an ordinary phone. Also, this work is focused on generating data and does not explore classification.

3 System Overview

Figure 1 shows the PVA system overview. A device such as a mobile phone or dedicated Internet of Things (IoT) device is used to capture sound. When the device (front end) recognises a wake word (e.g. 'Alexa') it records the subsequent sound and sends this recording to a cloud-based back end for analysis. The back end uses powerful ASR to transcribe speech to text. Thereafter the text is analysed to extract user commands and, if required, an action is carried out. Sometimes user feedback is provided following a command, for example, via an audio signal emitted via the front end.

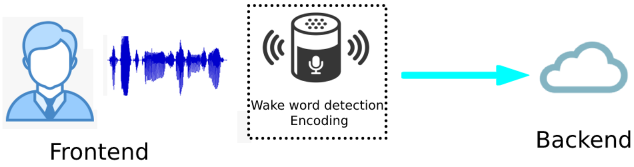

Figure 1. Personal Voice Assistant (PVA) system overview. The device (front end) is activated by recognising a wake word (e.g. 'Alexa'). The sound recorded after the wake word is transported to a back end for further analysis (e.g. Automated Speech Recognition (ASR))

In this work we assume that the recorded sound transported to the back end undergoes an additional analysis to identify the room. Such analysis may be carried out by the operator of the PVA infrastructure or by a third party that is able to get hold of the recorded audio. In both cases a user may not be aware that the sound is analysed in this particular way and a user may also not be aware of privacy issues arising from such analysis.

As the PVA user always initiates communication with the PVA by using the wake word (e.g. 'Alexa' in case of an Amazon PVA) it is possible to use this part of the communication as reference signal. Usually the same user is speaking the same wake word in potentially different locations. Using a reference signal simplifies the task of identifying a room compared to a situation where different users and arbitrary spoken words or sentences have to be used.

As a user is always starting the communication with the wake word a large pool of wake word samples spoken in different rooms can be acquired quickly. We assume it is possible to collect wake word samples for which the room is known. These samples are then used to train a system which is then subsequently used to classify rooms based on a recorded wake word.

4 Dataset

Our investigation aims to assess whether it is possible to detect the room in which the device is placed using wake word samples recorded via a PVA (i.e. by an ordinary smartphone).

We created a dataset using the following procedures. A single user (male) speaks the wake word 'Alexa' in a room at 3 different positions multiple times (20 to 30 times). A standard mobile phone (a Galaxy A5) is used to record the spoken wake words. The phone is placed in the middle of the room; the 3 speaker locations are approximately 1 meter away from the phone and the three three different positions are also approximately one meter apart. This setup is used to record the wake word in 10 different rooms. 7 rooms are

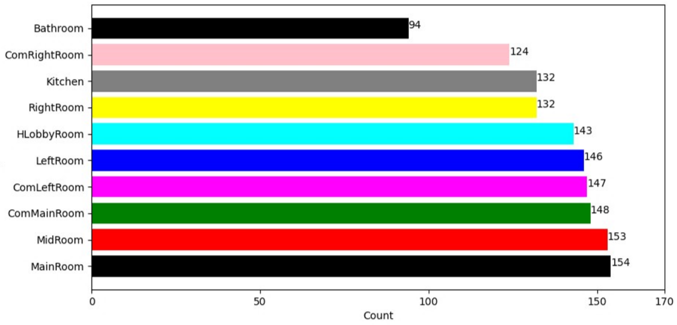

Figure 2. Number of samples recorded in 10 different rooms.

located in an office building and 3 in a residential building. The whole procedure is repeated after a day. Data collected at the first day is labeled Session1 and data collected the second day is labeled Session2. It has to be noted that the phone is placed approximately at the same position on the table and the speaker is taking approximately the same position in the second session. However, positions are not exactly replicated between both sessions. The number of wake word samples collected in the different rooms is shown in Figure 2. We collected 1371 wake word samples in ten different rooms.

Using a single speaker is realistic as in a practical PVA context the same user would usually interact with the device.

A standard phone is used for recording sound as this would also be the case in a practical setting.

We use 10 rooms in two different locations. The office environment creates a particular challenge as the rooms are acoustically very similar. Some offices have the exact identical shape and furniture. In a real setting this would not be the case; in a house it is to be expected that rooms are usually different.

We use one place for the PVA and multiple speaker locations. We believe that users tend to place a phone in a room at a similar spot (e.g. place it on the table). Users may vary their position when interacting with the device.

We use two session to ensure that the dataset represents the fact that a user entering a room and interacting with the PVA will not always place the device in the exact same spot and also will not always be in the exact same position when interacting with the PVA

The experiment was approved by the Social Research Ethics Committee (SREC) at University College Cork (Log 2021-139).

5 Algorithms for Room Identification

We assume it is possible in our scenario to obtain labeled wake word samples in a set of rooms. For example, a database may be built over time, collecting the wake word every time the user interacts with the PVA in a known context. The obtained samples are then used to train a model which is subsequently used to identify the room based on a recorded wake word in a situation where the location is not known.

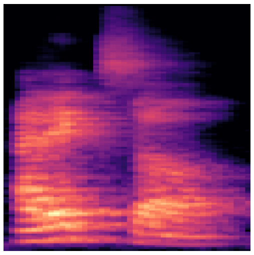

Figure 3. Spectogram of the Wake Word 'Alexa'.

We chose to design and explore three different algorithmic approaches to perform room classification. Our three approaches are based on successful methods for sound classification reported in literature. Our three methods make use of Convolutional Neural Network (CNN) as they have a track record of producing good results in the context of sound classification and categorization [10].

In our first approach we split the sound signal into two parts: Harmonic and Percussive parts. These features are then used as input for a CNN based classifier. Mu et al. [7] proposed this Harmonic Percussive Source Separation (HPSS) approach for environmental sound classification tasks.

In our second approach we directly use the sound spectrogram (see Figure 3) as input instead of pre-processing the acoustic signal to extract features such harmonic and percussive components. The spectrograms are then used as input for a CNN based classifier. Nanni et al. [8] have used such approach successfully to identify sources of noise in environments.

For the third approach we chose a transfer learning technique. The wake word spectrograms (similar to the aforementioned approach) are used as input for a pretrained model for sound classification (Imagenet pretrained VGG16). The extracted feature vectors are then processed and classified by adding new layers on the top of the existing pretrained models. This approach was used to deal with a situation where less data (wake word samples) is available for training. Next we discuss each approach in and chosen techniques in more detail.

5.1 First Approach - HPSS

For this approach the Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the the obtained wake words. Thereafter the HPSS technique is applied to split the input signal into two sub signals. The mean values of the MFCCs were then calculated. We chose to use 120 mels and, thus, the output of this prepossessing stage is a two dimensional array with size (120,2). The extracted feature vectors are then used as input to a CNN. The architecture of our chosen CNN is tabulated in Table 1; one dimensional convolutional layers are stacked on top of each other.

Table 1. First Approach - HPSS: configuration of the chosen CNN.

Layers	Neurons	Activation Func.	Kernel Size
Conv1D	64	RELU	2
MaxPooling	64	N/A	2
Conv1D	64	RELU	2
MaxPooling	64	N/A	2
Conv1D	64	RELU	2
MaxPooling	64	N/A	2
Conv1D	64	RELU	2
MaxPooling	64	N/A	2
Dense	64	SIGMOID	N/A
Output	N/A	SOFTMAX	N/A

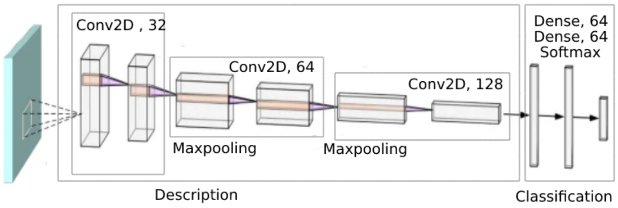

Figure 4. Architecture outline of approach 2 - CNN shallow.

It is worth mentioning that for performing our multi-class classification task we have chosen a categorial crossentropy loss function and an Adam optimizer. The learning rate is 5e-3 in this case. This approach is very cost-efficient and this is the advantage of using this approach for performing the room identification task.

5.2 Second Approach - Shallow CNN

For the second approach, instead of extracting MFCCs from the samples and working with voice signals, we generated spectrograms (see Figure 3). The *opencv* function *imread* is used to process the spectrograms and images are resized to an array of size (224,224,3). Figure 4 shows the chosen CNN architecture processing the spectrograms. We have used three blocks of convolutional layers with 32, 64 and 128 filters of size 3×3. For each block, one 2×2 Maxpooling layer is added to obtain image features with lower learning parameters. In order to connect the feature extraction stage with the classification stage one flattening layer is used. In the classification stage two dense layers with 64 neurons followed by a soft-max layer for prediction of the classes are used. The loss function is categorical crossentropy and the chosen optimizer is Adam and the resulting learning rate is 5e-5. It is worth mentioning that all the CNNs are followed by a rectified linear unit activation function. Only for the one layer before the last layer, a sigmoid function is used to have the output between 0 and 1.

5.3 Third Approach - Transfer Learning

Similar to the second approach we first resize the obtained mel spectrogram images representing the wake word. Thereafter we use *Vgg16*, one of the most powerful pretrained image classification systems available in order to process the obtained images. To tune the network from the task of object classification to the task of room identification we eliminate the classification block of the network (FC6, FC7 and the softmax layer) then we add two convolutional layers (number of neurons = 1024), one dense layer (number of neurons = 64) and one softmax layer on the top. Hence, we only train the added layers and the existing convolutional layers are frozen when training the system using the collected wake word dataset.

6 Evaluation

We use the aforementioned three approaches together with our collected wake word dataset to evaluate to which degree it is possible to determine a room based on the wake word spoken by a user.

We perform two different evaluations. In the first evaluation we use the data collected in Session1 and Session2 for training and evaluation. We use this first evaluation to judge general performance of our three different classification methods. In the second evaluation we only use data from Session1 for training and then use data from Session2 for testing. This was done to assess if it is possible to determine the room if speaker and PVA are not placed at the exact same position as during training of the system. In a practical setting, user and PVA may be at similar locations but not necessarily the exact same position.

In our first evaluation we use all three of our classification approaches. In the second evaluation we focus only on the third approach (transfer learning) as it is the most suitable approach for these conditions.

6.1 Evaluation 1

The test and training data are randomly picked form the sample pool (Session1 and Session2). We use 80% of the samples for training the networks and the rest for testing.

Figure 5, Figure 6 and Figure 7 show the classification results obtained with our three approaches.

Using the first approach (HPSS) we can obtain a very good result of 99% of accuracy for the chosen scenario. As it is shown in the confusion matrix, the classifier can discriminate between the samples recorded in different rooms with a very high accuracy.

Using the second approach and by training a shallow network from scratch the overall accuracy of 100% is achieved.

With the third approach, using transfer learning, an overall accuracy of 96% is achieved.

All methods provide a useful high accuracy. However, the second approach is here in this scenario the best choice.

6.2 Evaluation 2

We use Session1 as training data and Session2 as test data. We use the third approach of transfer learning to classify rooms. Figure 8 shows the results.

The overall accuracy drops substantially from 96% to 68% by changing the training-testing scenario. The major reason for this is that throughout a session, successive samples can be considered as identical samples. Samples from Session2 used for testing are not included in the training data obtained from Session1. Furthermore, in terms of size, geometry, and decoration, the office rooms included in the evaluation are remarkably similar. Thus, any differentiation between these very similar rooms is challenging. Another as-

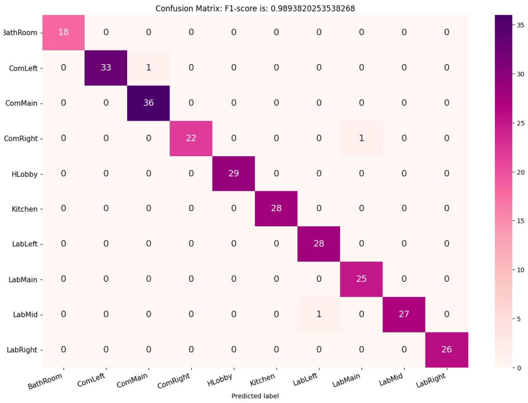

Figure 5. First Approach: HPSS, Ten Rooms.

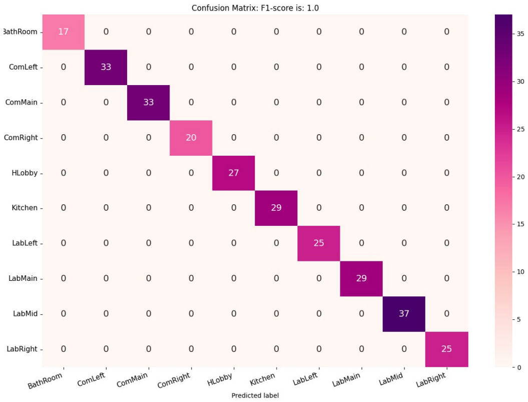

Figure 6. Second Approach: Shallow CNN, Ten Rooms.

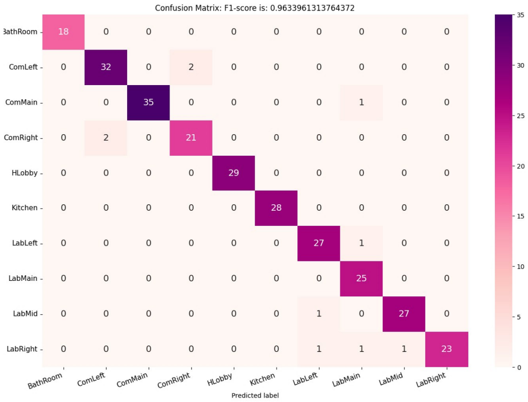

Figure 7. Third Approach: Transfer Learning, Ten Rooms.

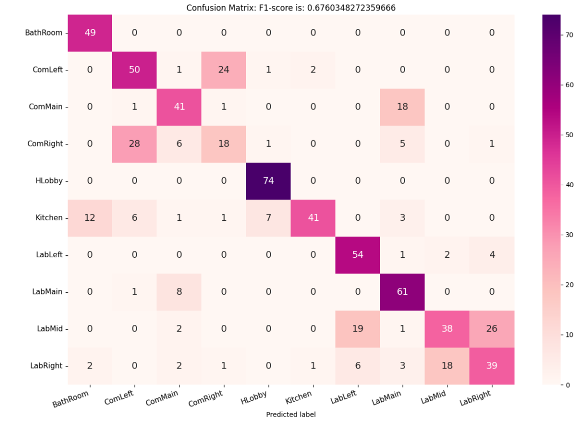

Figure 8. Cross-Session Comparison. Third Approach: Transfer Learning, Ten Rooms. Results: Ten Rooms.

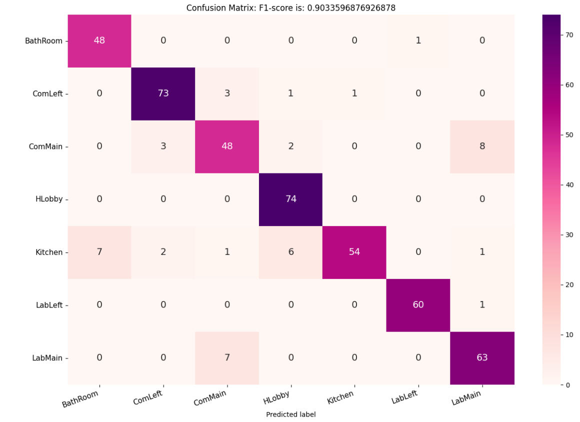

Figure 9. Cross-Session Comparison Results. Third Approach: Transfer Learning, Seven Rooms.

pect to consider is that in this evaluation scenario less data is available for training (as only taken from Session1).

In most real-world scenarios a PVA would be used in environments (e.g. residential buildings) where rooms have significant different acoustic properties. For example, a kitchen is different to a living room which is different to a bedroom.

To represent this situation we choose to remove three rooms (the most similar office rooms) and run the classification algorithm again. The results are shown in Figure 9.

The results improve under these conditions significantly and an overall accuracy of 90% is achieved. This shows that if rooms are sufficiently different an acoustic sample is sufficient to determine the room in which the interaction between PVA and user takes place.

7 Conclusion

We have shown that it is possible to build a system that is capable to identify a room in which a user interacts with a PVA based on recorded wake words. We created a dataset specifically for the purpose of evaluating wake word based room identification in a PVA context.

Our results show that it is possible to identify the room with a high accuracy. It is necessary to collect training data in the target environment. However, as the required training data consists of the wake word that must be spoken whenever the user interacts with the PVA we believe it is possible to gather the required information over time. The evaluation shows limitations; if rooms are very similar in shape, size and materials it is difficult to clearly identify the room using our proposed methods.

The results of our work can be interpreted in two ways. First, PVA audio recordings leak private information about the user environment. Clearly a user may not want to reveal to a PVA provider any information about their environment. Second, room identification is an option for augmenting user - PVA interaction. It may be possible to use this approach to tailor PVA interaction to the room in which it takes place.

It also has to be pointed out that the reported findings are transferrable to other settings. For example, in video conferencing calls users obfuscate the background to conceal the room they are in. This may be ineffective if the sound is analysed to reveal which room it is.

Acknowlegement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 19/FFP/6775. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

8 References

[1] M. Azimi and U. Roedig. Room identification with personal voice assistants (extended abstract). In *Computer Security. ESORICS 2021 International Workshops - CyberICPS, SECPRE, ADIoT, SPOSE, CPS4CIP, and CDT&SECOMANE, Darmstadt, Germany, October 4-8, 2021, Revised Selected Papers*, volume 13106 of *Lecture Notes in Computer Science*, pages 317–327. Springer, 2021.

[2] S. Deb, S. Dandapat, and J. Krajewski. Analysis and classification of cold speech using variational mode decomposition. *IEEE Transactions on Affective Computing*, 11(2):296–307, 2020.

[3] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor. Estimation of room acoustic parameters: The ace challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693, 2016.

[4] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor. Estimation of room acoustic parameters: The ace challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693, 2016.

[5] A. H. Moore, M. Brookes, and P. A. Naylor. room identification using roomprints. *journal of the audio engineering society*, june 2014.

[6] A. H. Moore, P. A. Naylor, and M. Brookes. Room identification using frequency dependence of spectral decay statistics. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6902–6906, 2018.

[7] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1):1–14, 2021.

[8] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci. An ensemble of convolutional neural networks for audio classification. *Applied Sciences*, 11(13):5796, 2021.

[9] A. Nediyanchath, P. Paramasivam, and P. Yenigalla. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7179–7183, 2020.

[10] Y. R. Pandeya, D. Kim, and J. Lee. Domestic cat sound classification using learned features from deep neural nets. *Applied Sciences*, 8(10):1949, 2018.

[11] C. Papayiannis, C. Evers, and P. A. Naylor. Data augmentation of room classifiers using generative adversarial networks. *CoRR*, abs/1901.03257, 2019.

[12] C. Papayiannis, C. Evers, and P. A. Naylor. End-to-end classification of reverberant rooms using dnns. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3010–3017, 2020.

[13] N. Peters, H. Lei, and G. Friedland. Name that room: Room identification using acoustic features in a recording. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, page 841–844, New York, NY, USA, 2012. Association for Computing Machinery.

[14] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee. A dialogical emotion decoder for speech emotion recognition in spoken dialog. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6479–6483, 2020.