Room Identification with Personal Voice Assistants

Mohammadreza Azimi and Utz Roedig

School of Computer Science and Information Technology, University College Cork, Cork, Ireland
{m.azimi,u.roedig}@cs.ucc.ie

Abstract. Personal Voice Assistants (PVAs) are used to interact with digital environments and computer systems using speech. In this work we describe how to identify the room in which the speaker is located. Only the audio signal is used for identification without using any other sensor input. We use the output of existing trained models for speaker identification in combination with a Support Vector Machine (SVM) to perform room identification. This method allows us to re-use existing elements of PVA eco-systems and an intensive training phase is not required. In our evaluation rooms can be identified with almost 90 percent accuracy. Room identification might be used as additional security mechanism and the work shows that speech signals recorded by PVAs can also leak additional information.

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.

A PVA may also use other techniques in addition to ASR to analyse recorded speech samples. For example, speaker identification may be carried out. In this case the speech signal is analysed in order to identify who the speaker is that is supplying a voice command. Such method may be useful in order to tailor a PVA action to the interacting user. For example, if a user requests to play their favourite music it is necessary for the system to identify the correct user. Also, such feature can be used to improve security and is used to implement user specific PVA access control. Other features that can be extracted from speech signals are the user's gender [9], emotional state [14] or health condition [2].

In this work we investigate how to extract features from audio samples captured by a PVA that allow us to determine the room in which the sample has been recorded. Such *room identification* feature is useful to further tailor PVA usage to the user environment. For example, if the user requests to play their favourite music the system can recognise in which room the command was issued

and play music via the correct speaker system. We assume here that either the PVA is mobile (a mobile Phone) or that it is a smart speaker that can be easily carried into another room. Room identification is also important from a security perspective. Room identification can be used as additional security feature. A PVA could be configured to only accept commands that are placed in specific rooms. For example, a doctor may only interact with patient data via a PVA in specific environments such as the consultation room but not the hospital's cafeteria. It has also to be noted that audio based room identification represents a privacy issue. Users that interact with a PVA do not necessarily want to sacrifice location privacy.

Existing work has shown that a Deep Neural Network (DNN) can be trained to identify the room in which a sound was recorded. However, a large data set is required and training of the DNN takes considerable effort. Also, this new capability requires additional processing capabilities. To overcome these issues we investigate in this work a different approach. We propose to use existing trained models used for speaker recognition to perform the additional task of room identification. Specifically we evaluate this approach using two trained speaker recognition systems that we call *thinResnet* [13] and *VGGVox* [8]. We use the output vectors of the speaker recognition system as input for an SVM which we then use for room identification. The SVM can be configured using a relatively small number of sound samples and complex training of a specialised DNN is not necessary. In a PVA eco-system sophisticated trained models for ASR and speaker recognition are available and the effort to implement room identification can be reduced.

The specific contributions of this work are:

- *Room Identification via Trained Models*: We describe a method for room identification using existing trained models; specifically trained speaker recognition models.
- *Evaluation of Room Identification*: We evaluate the proposed method using the two well known speaker recognition systems that we call *thinResnet* [13] and *VGGVox* [8]. We use a public available data set from the Acoustic Characterisation of Environments (ACE) challenge. We show that rooms are identified with 89 % accuracy.

In the next section we discuss related work. Section 3 describes on a system level how room identification is used in a PVA context. Section 4 describes our method for room identification using existing speaker recognition models. In Section 5 we detail our evaluation; evaluation setup, data sets and results are described. Section 6 concludes the paper.

2 Related Work

A number of techniques are available to characterise a room. Some of techniques have been used to perform room characterization and/or room identification. Here we detail work closest to ours and highlight differences. The main difference

to existing work is that i) we use unprocessed original audio files recorded in different rooms and ii) we use preexisting NN-based models trained for a different purpose than room identification or verification.

Peters et al. [11] introduced in 2012 a system for room identification by analysing audio in a video clip. Mel-Frequency Cepstral Coefficient (MFCC) features are used for analysis. An accuracy of 61% for music and 85% for voice signals is achieved with no shared data between training testing phase. The term "Room Identification" was first coined by the authors [12]. Our work differs as we re-use existing speaker identification models.

Moore et al. [5,6] proposed in 2013 the use of Gaussian Naive Bayes Classifier (GNBC) using Frequency Dependent Reverberation Times (FDRTs) features for room identification. A database consisting of 484 Room Impulse Responses (RIRs) for 22 rooms, with volumes ranging from 29 to 9500 cubic meters, were used. The FDRTs was used as input feature to the classifier. According to the obtained results, in the best case scenario an Equal Error Rate (EER) of 3.9% can be achieved. Special equipment is required to measure the FDRTs. In our work we use recorded speech directly for room identification instead of dedicated acoustic measurements.

Murgai et al. [4] conducted research to see if blind estimation of the reverberation fingerprint of an unknown room could be performed by monitoring recorded speech signals. Despite the fact that the cited paper's main research goal was room volume classification, the obtained reverberation fingerprints can also be used for room identification. In this work we look at how to extract characteristics from audio samples to specifically identify a room using existing speech identification models.

In 2018 Moore et al. [7] proposed a new method for room identification using sub-band negative-side variance features. A GNBC is used to classify the features. The evaluation used recording samples taken from the evaluation dataset of the ACE challenge [3]. Voice recordings in five rooms were used. For the best-case scenario where the training data includes utterances spoken from the same position as the test data, a 90.5% accuracy is obtained. While our work uses the same dataset for evaluation we use a different analysis method. We use existing trained speaker identification models and their output as features to identify rooms using an SVM.

Papayiannis et al. [10] explored room identification based on the influence of reverberation on speech. The authors propose Convolutional Recurrent Neural Networks (CRNNs) to identify the room. For evaluation, Acoustic Impulse Responses (AIRs) are used from the ACE challenge dataset, measured in 7 rooms. The AIRs are used to artificially add reverbaration to speech samples; then these artificial samples are used to identify the rooms. According to the achieved results, the classification accuracy of the CRNN is 78%. In our work we do not use generated samples and we use existing trained speaker identification models to identify rooms.

3 System Overview

A PVA is a system comprised of two major components: the front end and the back end (see Figure 1). The front end is either implemented as a dedicated device, often called a *smart speaker*, or realised as an app on the user's smartphone or other system such as a TV.

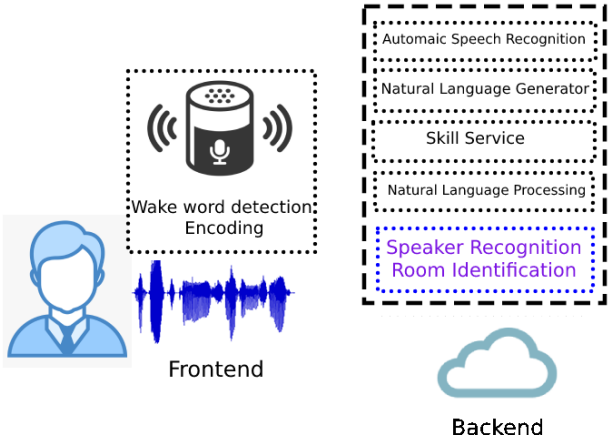

Fig. 1. PVA system overview comprising front end and back end infrastructure.

The front is able to record and play audio. It also comprises a wake word detection; once a wake word such as *Alexa* is recognised the following speech signal is recorded and transmitted to the back end.

The following are the key components usually found in the back end: ASR module, Natural Language Processing (NLP) module, skills management module (skill service) and natural language generator module (text to speech generator).

The process of turning the recorded speech into text is implemented by the ASR. This process is carried out using acoustic and language models. An NLP module is needed for intent recognition. The meaning of the speech and the user's expectations are expressed by the *intent* which results in a structured codified user request. The natural language generator module may be used to generate a speech response message played to the user by the front end.

A speaker recognition module may also be used in the back in order to identify the speaker. The obtained user identification might be used to prevent execution of a command. For example, when a specific user is deemed not to be allowed to issue a specific command.

To implement room identification it would be possible to include an additional module specifically for this purpose in the back end. This module would be supplied with the recorded voice, similar to the Speech Recognition (SR) to

perform the task of room identification. As discussed in the related work, models are existing that could be used for this purpose.

However, this approach introduces two challenges. Firstly, the additional back end module would require resources to execute; all currently processed speech samples submitted by front ends require this additional resource. Thus, the back end infrastructure would need to be scaled up which is costly and also not energy efficient (Energy consumption is a significant cost factor of data centers). Secondly, the additional back end module would require to be trained which requires effort. Significant amount of data from user homes would need to be available. While it is possible to do to it is an additional overhead.

To overcome these challenges we propose therefore another approach. We propose to use the output (the feature vectors after processing) of the existing speaker recognition module to perform room identification. The output of the speech recognition module is used within a simple SVM to classify the rooms.

4 Room Identification

For speaker identification a neural network can be used. These take an acoustic signal (the speech signal) as input and then classify the speaker based on the features extracted from the input signal. Here we make use of a trained neural network for speaker identification, however, we take the output feature vector of the neural network to feed an Support Vector Machine (SVM) which is then used to classify the different rooms.

We take a number of acoustic signals collected in the rooms of interest and feed these to the trained neural network for speaker recognition. Then we use the resulting feature vectors to train an SVM. The training of the SVM can be performed with relatively few samples from rooms and training to classify rooms is much simpler than training a full end-to-end DNN for this purpose.

The SVM input data is mapped to a higher dimensional feature space via a kernel function. The feature space is derived using the kernel function, instead of being strictly defined. In this way, the selection of the kernel is the key to determine the feature space. We chose the Gaussian Radial Basis Function (RBF) for our SVM. We use in our system two well known speaker recognition systems that we call *thinResnet* [13] and *VGGVox* [8].

thinResNet (512 dimensional feature vector): In this case we use the 'thinResNet' [13] trunk architecture with a dictionary-based NetVLAD layer for aggregating extracted features across time. This neural network model was trained end-to-end. It is also worth mentioning that here, voice activity detection (or automatic silence removal) is not applied. The output of the fully connected layer is used here as the extracted feature vector of 512 elements that is used as input for our SVM.

VGGVox (1024 dimensional feature vector): VGGVox was proposed by Nagrani et al. [8] and this architecture is based on the VGG-M [1] Convolutional Neural

Network (CNN), which is noted for its great efficiency and image classification ability. Using the 1024 dimension FC7 vectors, feature vectors from the classification network can be obtained. Here, we use the extracted feature vectors for training our SVM classifier.

5 Evaluation

5.1 Dataset

The ACE Challenge database is used. The ACE Challenge was set up to encourage research on blind estimation of acoustic parameters from noisy speech using newly collected reverberant speech samples under different conditions [3]. The database contains so called *babble noise* recorded in seven different rooms (Two offices, two lecture rooms, two meeting rooms and lobby).

The babel noise is created by four to seven persons sitting in close proximity and chat constantly for the duration of the audio recording. The files were recorded on two separate occasions using the same microphones, with the microphones moved to the new position between the two occasions. For each of the rooms, two babble noise samples were obtained (for two different microphone positions).

We have seperated the babble noise samples into 2.5 second length audio samples. This way we we obtained 1352 samples in total distributed across the 7 rooms as follows: No.1. First Living Room (FLRoom) 200 samples, No.2. First Meeting Room(FMRoom) 167 samples, No.3. First Office (FOffice) 153 samples, No.4. Second Living Room (SLRoom) 243 samples, No.5. Second Meeting Room(SMRoom) 178 samples, No.6. Second Office (SOffice) 205 samples, and No.7.Lobby (Lobby) 206 samples.

5.2 thinResNet

In order to train our SVM we used 502 voice samples, with 850 samples being used to test the chosen model. The training and test data-set samples were chosen randomly. Table 1 shows the summary of results obtained. Figure 2 shows the obtained confusion matrix.

As seen in Figure 2, the multiclass-classifier accurately identified all samples as belonging to the first office, with the exception of three that were wrongly identified as samples being recorded in the first living room (two samples) and the second meeting room (one sample). It can be illustrated in the same figure that, for the second meeting room, the number of mistakenly rejected samples was zero and the false negative rate is zero.

According to Table 1, the best F1-score of 99% is achieved for the second living room (there were only two incorrectly classified samples and there was no incorrectly rejected sample in this class), while the worst F1-score of 77% is achieved when we want to recognize the second meeting room.

We can conclude from Table 1 that, the overall accuracy of 89% can be obtained when we train our model using the thinResnet feature vectors (512 dimensional extracted feature vectors).

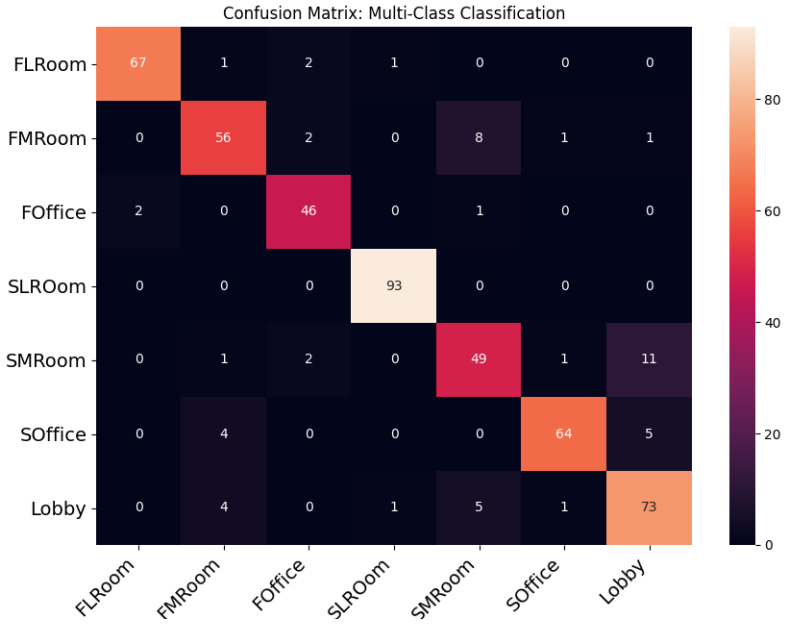

Fig. 2. thinResnet: The obtained confusion matrix using 850 voice samples.

Table 1. SVM Classification Results - thinResnet

Type of Rooms	Precision	Recall	F1-score	Support
FLRoom	0.97	0.94	0.96	71
FMRoom	0.85	0.82	0.84	68
FOffice	0.88	0.94	0.91	49
SLRoom	0.98	1.00	0.99	93
SMRoom	0.78	0.77	0.77	64
SOffice	0.96	0.88	0.91	73
Lobby	0.81	0.87	0.84	84
accuracy			0.89	502

5.3 VGGVox

Figure 3 shows the resulting confusion matrix using VGGVox. As it can be seen in Figure 3 for the second living room, 94 samples were classified and identified correctly, while two samples were mistakenly and incorrectly rejected by the system. As it is shown in Figure 3, the worst case scenario is when we want to identify Room no.2 (the first meeting room) using the obtained samples. As it is mentioned before, the training data set is completely separate from the test data-set and the samples were randomly chosen for these two separate data sets.

Table 2 summarises the results obtained using VGGVox in vombination with the SVM. The best f1-score is 98% for Room no.4 and the lowest f1-score is 67% for Room no.2. The overall accuracy is 86%.

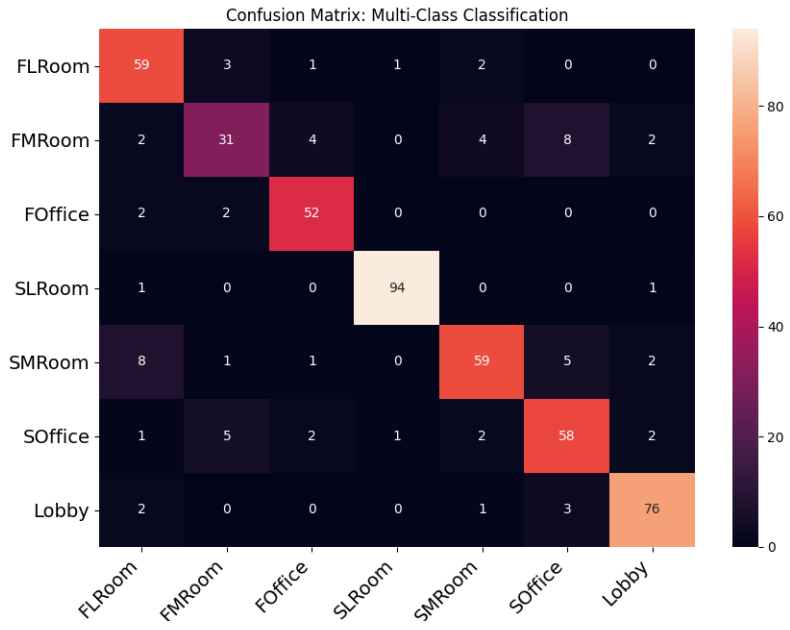

Fig. 3. VGGVox: The obtained confusion matrix using 850 voice samples.

6 Conclusion

This work has shown that room identification based on voice samples are feasible and that existing neural networks used for other tasks such as speaker identification can be re-purposed for this task. By doing so it can be avoided to train

Table 2. SVM Classification Results - VGGVox

Type of Rooms	Precision	Recall	F1-score	support
FLRoom	0.79	0.89	0.84	66
FMRoom	0.74	0.61	0.67	51
FOffice	0.87	0.93	0.90	56
SLRoom	0.98	0.98	0.98	96
SMRoom	0.87	0.78	0.82	76
SOffice	0.74	0.82	0.80	71
Lobby	0.92	0.93	0.92	82
accuracy			0.86	498

complex networks just for this task and existing elements of a PVA infrastructure can be re-used.

In this work we used the public available dataset collected by the Acoustic Characterisation of Environments (ACE) Challenge. This data was not specifically collected for a PVA context. Thus, in our next steps we plan to collect our own data set issuing voice commands to a PVA. We will then repeat the experiments detailed in this paper using this more specific dataset.

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.

References

1. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets (2014)
2. Deb, S., Dandapat, S., Krajewski, J.: Analysis and classification of cold speech using variational mode decomposition. IEEE Transactions on Affective Computing **11**(2), 296–307 (2020). https://doi.org/10.1109/TAFFC.2017.2761750
3. Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A.: Estimation of room acoustic parameters: The ace challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(10), 1681–1693 (2016). https://doi.org/10.1109/TASLP.2016.2577502
4. M. Murgai, M.P., Rau, J.: blind estimation of the reverberation fingerprint of unknown acoustic environments. journal of the audio engineering society (october 2017)
5. Moore, A.H., Brookes, M., Naylor, P.A.: Room geometry estimation from a single channel acoustic impulse response. In: 21st European Signal Processing Conference (EUSIPCO 2013). pp. 1–5 (2013)

6. Moore, A.H., Brookes, M., Naylor, P.A.: room identification using roomprints. journal of the audio engineering society (june 2014)
7. Moore, A.H., Naylor, P.A., Brookes, M.: Room identification using frequency dependence of spectral decay statistics. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6902–6906 (2018). https://doi.org/10.1109/ICASSP.2018.8462008
8. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: A large-scale speaker identification dataset. Interspeech 2017 (Aug 2017). https://doi.org/10.21437/interspeech.2017-950, http://dx.doi.org/10.21437/Interspeech.2017-950
9. Nediyanchath, A., Paramasivam, P., Yenigalla, P.: 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). pp. 7179–7183 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054073
10. Papayiannis, C., Evers, C., Naylor, P.A.: End-to-end classification of reverberant rooms using dnns. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 3010–3017 (2020). https://doi.org/10.1109/TASLP.2020.3033628
11. Peters, N., Lei, H., Friedland, G.: Name that room: Room identification using acoustic features in a recording. In: Proceedings of the 20th ACM International Conference on Multimedia. p. 841–844. MM '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2393347.2396326, https://doi.org/10.1145/2393347.2396326
12. Peters, N., Lei, H., Friedland, G.: Room identification using acoustic features in a recording (US patent, US 9,449,613 B2, Sep 20, 2016)
13. Xie, W., Nagrani, A., Chung, J.S., Zisserman, A.: Utterance-level aggregation for speaker recognition in the wild. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5791–5795 (2019). https://doi.org/10.1109/ICASSP.2019.8683120
14. Yeh, S.L., Lin, Y.S., Lee, C.C.: 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). pp. 6479–6483 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053561