Acoustic Scene Classification. Room Identification. **MUS-15**

Yuliya Sergiyenko

July 10, 2015

Keywords/ phrases

auditory scene analysis (ASA), Computational auditory scene analysis (CASA), Acoustic scene classification, Equal error rate (EER), Room impulse response (RIR), Gaussian Mixture Model (GMM), Mel-frequency Cepstral Coefficient (MFCC) Roomprints.

Introduction

Every day humans encounter multiple complex audio scenes and have the ability to understand what they are. Out hearing system can perform such tasks as easily following conversation with one speaker in a crowded room and closing our eyes and still being able to distinguish which scene we are at. Since the emergence of computers researchers have been trying to answer the questions of whether we can reproduce this remarkable human ability with a computer, whether we can create an algorithm that will be just as good as humans themselves. There have been a lot of research done and algorithms developed in Computational auditory scene analysis (CASA) and most of them try to reproduce human auditory system in order to identify and classify sounds just like humans. Many of the existing algorithms already do perform on the level of human expert listeners. But acoustic scene classification is one area where there hasnt been a lot of research or definite success yet. Acoustic scene classification aims at identifying the location at which the audio or video recording was made. In the field of acoustic scene classification there is a branch that focuses their research on indoor scene classification or also called room classification. In this paper I will present the state of the art in the field of Room classification, the results obtained in the research and also present my conclusion on the topic.

Background

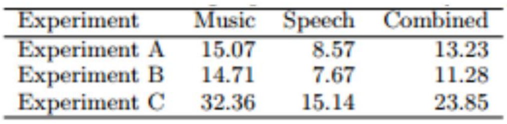
People started to investigate the science of speech perception already in 1930s in Bell Labs [1]. In 1953 a scientist Colin Cherry first formulated a phenomenon of human hearing system. One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. We may call it the cocktail party problem wrote Cherry in his work [2]. In the 1990 Albert Bregman has defined a field of Auditory scene analysis and Computational auditory scene analysis. (CASA) is the study of auditory scene analysis (ASA) by computational means. [3] The goal of the CASA system is to be able to separate sound mixtures in the same way that humans are able to do. In the field of auditory scene classification the research was conducted mostly as a background or additional study for other field. The first method that focused solely on auditory scene classification was published in 1997 by Sawhney and Maes [4] in a report from the MIT Media Lab. They have described a simple classification of five predefined classes of environmental sounds: people, voices, subway, traffic, and other, via extraction of several discriminating features. Their results showed overall classification accuracy of 68 percent. Since 1997 there have been multiple algorithms developed in this field. The most recent findings were presented as results of the challenge on detection and classification of acoustic scenes and events (DCASE) in 2013 [5]. The new dataset was developed for this challenge, the 11 existing algorithms were tested and results were compared with baseline method and a human expertise. The best results showed accuracy of 78 percent. But in all of the before mentioned methods the researchers generally define a set of categories, record samples from these environments, and treat ASC as a supervised classification problem within a closed universe of possible classes [5]. The classification includes both indoor and outdoor classes. As humans, we spend most of our times indoors so there arose the need to research the field of indoor acoustic scene classification or room classification. Identifying a room is a very young field of research.In 2010 group of researches proposed that classification of a room volume from reverberant speech signals can be useful in acoustic scene analysis applications. An Equal error rate (EER) of 22.38 percent was achieved [6].

Room Identification: Definition and Application

Indoor acoustic scene classification or Room identification is given the audio or video recording, being able to identify the particular room in which the recording was made. Room identification has multiple applications in different fields of research such as location estimation, music recommendation, speech recognition indoors, law-enforcement and forensics. Right now we can see a lot of location-based multimedia applications being developed and used. For example, automatic tagging of uploaded user pictures. So the room location is and important information to have. Currently applications use GPS data to identify the location but the GPS estimations do not work well indoors. There have been attempts to increase accuracy with the help of using the strenght of the WiFi signal. But in case GPS and/or WiFi coverage is insufficient, or devices do not support the technology, then the location cannot be estimated. There have been another approach to estimate location based on visual similarities in a video recording. But this approach does not take into consideration changes in spatial configuration like moving of the furniture. Room identification can help music recommendation systems to create playlists based on the particular location. In the speech recognition field the automated systems are affected by unknown room reverberance. Knowing the room can help adapt recognition system and reach better results. And in the field of forensics the room identification in the emergency phone calls can be helpful for speeding up the process and even filtering the fake calls [7].

Room Identification Using Acoustic Features in a Recording

In 2012, researches N. Peters, H. Lei, and G. Friedand published a paper Name That Room: Room Identification Using Acoustic Features in a Recording [7]. In this paper they have proposed to identify the room in an audio or video recording through the analysis of acoustical properties using machine learning techniques. This is the first research in the room identification field that uses both speech and musical material. Since there exists no standardized dataset for room identification, they have created their own set from anechoic audio recordings, each filtered with a variety of impulse responses from a number of rooms. Room impule response (RIR) is like a "fingerprint" og the specific room, it is a recording of what it would sound like if an extremely loud and short click was played in the room like a gun shot. RIR depends on the location of sender and receiver, therefore no RIR within a room is completely similar to another. The researches have collected RIRs from different public databeses and they composed 7 room. For each selected room, 24 RIRs are available. The rooms were classified as: "Bedroom", "Studio", "Classroom", "Church1", "Church2", "Great Hall" and "Library". After evaluating previous research these scientists decided on using for their approach is the Mel-frequency Cepstral Coefficients (MFCCs) since it performed well in previous cases. MFCCs are based on the Mel-scale." The Mel scale is defined such that a tone of 1000Hz equals 1000mel. If tone A is perceived twice as high as tone B then its mel value is twice as high "[8]. "MFCC features C0-C19 (with 25 ms window lengths and 10 ms frame intervals), along with deltas and double-deltas (60 dimensions total), are extracted. For each audio recording, one room-dependent Gaussian Mixture Model (GMM) is trained for each room using MFCC features from all audio recordings associated with that room. The system performance is based on the equal error rate (EER)", - researches describe the system in the paper. They have carried out 4 sets of experiments. They have first tested speech and music samples separately and then tested them combined. Based on their results, the researches have made the following observations [7]: 1) In all experiments the EER of the speech material is about twice better than the EER of the musical material (Table1).



Experiment	Music	Speech	Combined
Experiment A	15.07	8.57	13.23
Experiment B	14.71	7.67	11.28
Experiment C	32.36	15.14	23.85

Table1. Resulting Equal Error Rates (EER)[7].

2) The EER of the combined materials, where testing dataset contained both music and speech content, is about the average of the EER for music and speech in separation. 3) MFCC window size is the most prominent parameter that can influence the feature extraction process and the resulting EER. For room identification short-term MFCC features are more suitable than-long term MFCC features. 4) The estimation error is not randomly distributed but depends on the (acoustical) similarities of the tested rooms. 5) The rooms formed clusters according to the databases they originally came from, which suggests differences in measurement systems or technique may have caused some of the between-room variability. 6) The system achieved overall accuracy of 61 percent for music and 85 percent for speech signals. In the future the researches will focus on improving the accuracy for music materials by exploring different additional features.

Roomprints For Forensic Audio Applications

In the field of forensics the state of the art is the work presented in 2013 by scientists A.H. Moore, M. Brookes and P.A. Naylor called "Roomprints For Forensic Audio Applications"[8]. In their research they have proposed the concept of a "Roomprint". "Roomprint" of a room is similar to "fingerprint" of a person. The researches propose collecting a database of "Roomprints" for room against which all of the other rooms can be compared. The comparison will be aimed to answer the following questions [8]: 1) Verification - if the claim is madethat a recording was made in a particular room, is there sufficient evidence to reject the claim? 2) Identification - if we have a knowledge that a recording was made in one of a number of rooms, can we determine which one is most likely? According to the researches, the "Roomprint" must meet a number of requirements [8]: 1) A "Roomprint" must only include features of a room which allow it to be distinguished from other potentially similar rooms. 2) A "Roomprint" should ideally be independent of the location of the speaker and microphone in the room. 3) A "Roomprint" should ideally be independent of time. In this research we are given examples of features that can be used in a "Roomprint". Geometric features such as a size and a shape of a room can be promising sing they do not depend on time or on the speaker and microphone in the room. But this features are quite hard to infer from an audio recording. There are multiple researches made on the topic of extracting partial information. According to this research this imformation can be included in the "Roomprint" but will probably not be sifficient by itself. Room acoustic parameters are another promising features to include in a "Roompront". They include Early decay time (EDT) and Reverberation time (T60). They do not require speciial arragement of the microphone and do not depend on direction or orientation of audio source. Environmental sounds such as fans are not useful in the ”Roomprint” by themselves since they are not directly related to the room and may wary over time. But their presence or absence could be used in the verification process. In this research the scientists decided to evaluate the idea of a ”Roomprint” using ”a frequency dependent measure of reverberation time under a number of alternative transformations”. The dataset was comprised of 22 rooms in total. For each room 22 RIRs were selected.The dataset of 484 observations were transformed according toeach of 6 different already existing methods [8] to give six alternative representations. For each representation, a 14-dimensional Gaussian distributions was estimated for each room. The results have shown that: 1) The overall error rate was 32.6 percent (Table2).

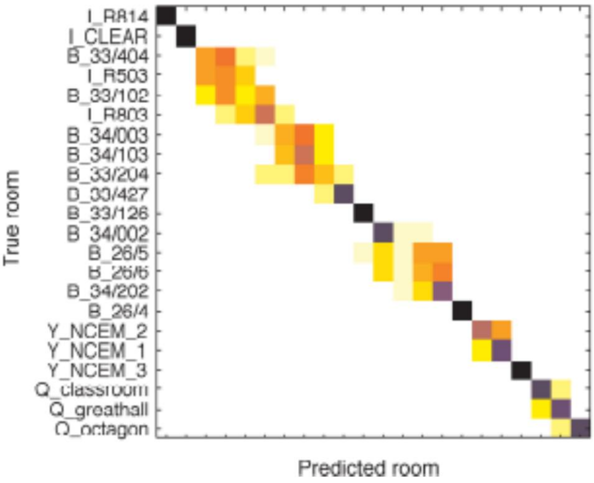


Table2. Confusion matrix for room identification experiment using T60 [8].

2) The best result with error rate of 3.9 percent was achieved by using the logarithm of frequency dependent reverberation time as a ”Roomprint” feature. 3) If the two rooms are in the same building and are built to thesame plan, they have almost identical distributions in each frequency. Despite this, the classifier was able to distinguish even these rooms correctly 70 percent of the time. 4) The is no universal datasets so it makes the concept limited. The researches plan to continue exploring this ”Roomprinting” method and experimenting with different features.

Conclusion

In this paper I have presented a brief overview of the topic of Acoustic scene classification with main focus on Room Identification and the state of the art in that field to this date. This is a very young field and the research is somewhat limited. Also, despite showing very high accuracy results, aforementioned two methods still have a lot of limitations and do not surpass human experts. This suggests that there is still a lot of room for improvement before any algorithm can reach and outperform the human ability to classify the acoustic scene based on sounds alone.

References

[1] B. H. Juang, L. R. Rabiner. *Automatic Speech Recognition A Brief History of the Technology Development* . Elsevier Encyclopedia of Language and Linguistics (2005), pp. 1,2.

[2] E. C. Cherry. *Some experiments on the recognition of speech, with one and two ears.* Journal of the Acoustical Society of America, 25 (1953), pp. 975–979.

[3] A. S. Bregman. *Auditory scene analysis: the perceptual organization of sound.* The MIT Press, Cambridge, MA (1990).

[4] N. Sawhney and P. Maes. *Situational awareness from environmental sounds.* Technical report, Massachussets Institute of Technology (1997).

[5] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley. *Acoustic Scene Classification.* IEEE. School of Electronic Engineering and Computer Science (2014).

[6] N. Shabtai, B. Rafaely, Y. Zigel. *Room volume classification from reverberant speech.* In Proc. of intl Workshop on

Acoustics Signal Enhancement, Tel Aviv, Israel (2010).

[7] N. Peters, H. Lei, G. Friedland. *Name that room: Room identification using acoustic features in a recording*. In Proceedings of the 20th ACM international conference on Multimedia (2012), pp. 841–844.

[8] A. H. Moore, M. Brookes, P. A. Naylor. *Roomprints for forensic audio applications*. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY (2013).