ACOUSTIC SCENE CLASSIFICATION USING SPATIAL FEATURES

Marc C. Green and Damian Murphy

Audio Lab
Department of Electonic Engineering
University of York
York, UK

ABSTRACT

Due to various factors, the vast majority of the research in the field of Acoustic Scene Classification has used monaural or binaural datasets. This paper introduces EigenScape - a new dataset of 4th-order Ambisonic acoustic scene recordings - and presents preliminary analysis of this dataset. The data is classified using a standard Mel-Frequency Cepstral Coefficient - Gaussian Mixture Model system, and the performance of this system is compared to that of a new system using spatial features extracted using Directional Audio Coding (DirAC) techniques. The DirAC features are shown to perform well in scene classification, with some subsets of these features outperforming the MFCC classification. The differences in label confusion between the two systems are especially interesting, as these suggest that certain scenes that are spectrally similar might not necessarily be spatially similar.

Index Terms— Acoustic scene classification, MFCC, gaussian mixture model, ambisonics, directional audio coding, multichannel, eigenmike, soundscape

1. INTRODUCTION

Since the recent increase in research into Acoustic Scene Classification (ASC) sparked by the DCASE challenges [1], the vast majority of work has focused on identifying scenes based upon mono or, at most, stereo recordings. The potential for utilising more detailed spatial properties of acoustic scenes extracted from microphone array recordings remains largely unexplored. This is partly due to inheritance of techniques from the more mature fields of Automatic Speech Recognition (ASR) and Music Information Retrieval (MIR), which often have a "perceptually motivated" approach [2] (particularly with ASR), and partly due to the common focus of ASC research on applications including use in wearable technology, smartphones and robotics, [3] where utilisation of large microphone arrays would not be practical.

Another potential application of ASC is in environmental sound research, where the focus is not on human perception or portability *per se*, but rather on obtaining a detailed understanding of the acoustic environment itself. Such detailed analysis could assist in urban planning and legislation surrounding environmental sound. The L_{Aeq} metric currently in widespread use measures the average Sound Pressure Level (SPL) over a given period of time [4], not taking into account the content of the sound. Advanced machine listening techniques could be used to provide more nuanced measures of sound to better inform acoustic surveyors and so augment the L_{Aeq} measure. This content-focused approach to acoustic assessment has been called the "soundscape approach", as opposed to the "environmental noise" approach of the majority of legislation [5].

Given this application, the limitation to low-channel-count audio is not necessary. This paper investigates the possibility of classifying acoustic scenes based upon spatial features, comparing this with the use of standard Mel-Frequency Cepstral Coefficient (MFCC) features, and is organised as follows: Section 2.1 briefly introduces the EigenScape dataset, including justification of its necessity in context with previously-released acoustic scene recordings. Section 2.2 details the methods used to extract features from the recordings, whilst Section 2.3 describes the system used to classify the data. Section 3 presents results from this study, with Section 4 providing brief additional discussion. Section 5 concludes the paper by summarising the findings of this experiment.

2. METHOD

2.1. Dataset

In order to undertake this research, a large database of spatially-recorded acoustic scenes was required, including many examples of the same kinds of locations. This allows for separation of the dataset into separate training and testing sets where there is no crossover in recording locations between the two sets, avoiding the situation that gave rise to artificially inflated results in [6] where training and testing sets included segments from the same longer original recordings.

A set of 1st-order Ambisonic recordings was made as part of the DCASE 2013 challenge [1], however these were recordings of staged office environments only, and so did not provide the variety of recording environments needed for this research. The DEMAND dataset [7] contains sets of three multichannel recordings each of six different acoustic scene classes, however this is still too small a corpus for this project and their use of a nonstandard microphone grid layout could potentially make calculation of spatial features more difficult. The TUT database [8] used in DCASE challenges since 2016 features an appropriately broad range of examples of multiple acoustic scene classes, but features two-channel recordings only. A new set of recordings, the EigenScape dataset, was therefore created for this project.

The EigenScape dataset was recorded using the mh Acoustics EigenMike [9], a 32-channel spherical microphone array capable of making recordings in 4th-order Ambisonic format. Eight 24-bit/48 kHz ten-minute recordings each of eight different classes of location - Beach, Busy Street, Park, Pedestrian Zone, Quiet Street, Shopping Centre, Train Station and Woodland - were made at locations across the north of England, giving a total of 64 recordings. The location classes were inspired by the selections in the TUT dataset, but with small indoor locations discarded, reflecting the focus of this work on the acoustic scenes of public places. Only the 1st-order channels from the 4th-order recordings were used in the present work.

Detailed information on the recording process for this dataset will be published in a future paper and the data will be made publicly available in due course. It is hoped that this data will prove useful to research in both acoustic scene and event detection.

2.2. Feature Extraction

The librosa library [10] was used to extract MFCC values from the omni channel (W) of the recordings. The audio was first resampled to half the original sampling rate before 20 MFCC values were extracted. These therefore covered the frequency range up to 12 kHz. The librosa standard frame length of 2048 samples with 25% overlap was retained.

To extract spatial features, the audio was resampled as before and filtered using a bank of FIR filters into 20 mel-spaced frequency bands in order to maintain parity in terms of frequency bands with the MFCC values. This enabled use of combined MFCC and spatial audio features for each band. Directional Audio Coding (DirAC) analysis [11, 12] was used in order to gain Direction of Arrival (DOA) estimates $\mathbf{D}$ for each frequency band as follows:

$$\mathbf{D} = -\mathbf{PU} \tag{1}$$

where $\mathbf{P}$ is a matrix containing the 20 mel-filtered versions of the W-channel of each audio file and $\mathbf{U}$ is a three-dimensional matrix containing the 20 filtered versions of the X, Y and Z-channels. The resultant matrix $\mathbf{D}$ was split into time-frames corresponding to the frames used in the MFCC calculations, and mean values of $\mathbf{D}$ were calculated for each frame. Angular values for azimuth and elevation in degrees for each frame were calculated based on this and used as features.

Secondly, a figure for diffuseness ψ in each frequency band was calculated as follows [11]:

$$\psi = 1 - \frac{||-\mathbf{D}||}{c\{\mathbf{E}\}} \tag{2}$$

where c is the speed of sound, $\{.\}$ represents the mean-per-frame values as described previously, and:

$$\mathbf{E} = \frac{1}{2}\rho_0 \left(\frac{\mathbf{P}^2}{Z_0^2} + ||\mathbf{U}||^2 \right) \tag{3}$$

where ρ_0 is the mean density of air and Z_0 is the characteristic acoustic impedance of air. Combining all of these features results in a 60-dimensional feature vector output from the DirAC analysis.

2.3. Classification

For classification, each ten-minute recording was split into 30-second segments. In order to facilitate cross-validation, the data was split into four folds, whereby for each fold, six examples of each location would be used for training, with the remaining two examples used for testing. In this way, segments from the same recording location could not feature in both training and testing sets.

A Gaussian Mixture Model (GMM) system was used as classifier. Each scene class was assigned a ten-component GMM, which was trained using features extracted from the training fold audio using the expectation-maximisation algorithm [13]. GMMs with more components were tested but found to not substantially outperform the ten-component versions. Models were trained using all of the extracted data - an 80-dimensional concatenation of MFCC and DirAC features - and subsets thereof, including MFCCs alone (20-dimensional feature vector), individual DirAC features (20-dimension), and a combination of Elevation and Diffuseness (40-dimension).

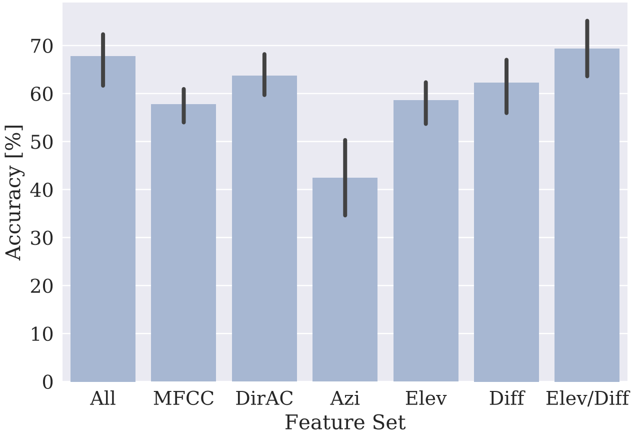

Figure 1: Mean and standard deviation classification accuracy using various feature subsets across all folds.

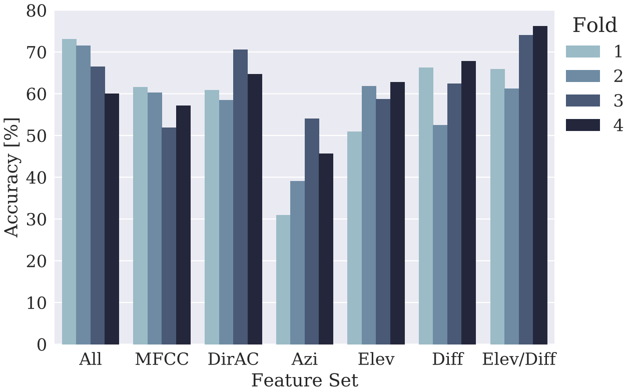

Figure 2: Accuracy of classifiers using various feature subsets across each data fold.

To classify the testing data, features from test fold frames were given probability scores by each GMM and these scores were totalled across each 30-second segment. The segments were classified based on the GMM that had given its features the highest total probability score across all frames. This is essentially identical to the simple-minded audio classifier (smacpy) [14] system used as the baseline in the DCASE 2013 challenge [1].

3. RESULTS

3.1. Overall Accuracies

Figure 1 shows the mean performance accuracy of the classifiers using MFCC features, DirAC features, a combination of all features and subsets of the DirAC features. Using the MFCC features, the classifier has an average accuracy of 58%. This is consistent with the performance of this type of classifier as reported in the literature, with Lagrange *et al* reporting 48% accuracy [15], and the DCASE 2013 baseline system giving 55% accuracy [1]. The DCASE 2016 baseline performs markedly better, though this system uses additional delta and acceleration MFCC features [8].

The DirAC spatial features outperform the MFCC features on average, at 64% accuracy as opposed to 58%. Figure 2 shows the classification accuracy in individual folds of the data. The MFCCs perform marginally better than DirAC in folds 1 and 2, but DirAC outperforms MFCCs by a larger margin in folds 3 and 4. Combining both sets of features leads to markedly improved performance relative to either alone in folds 1 and 2, with accuracies greater than 70% where each feature set alone gives accuracies closer to 60%. In the 3rd and 4th folds, however, adding the MFCCs to the DirAC features causes a decrease in accuracy of around 5% relative to using DirAC features alone.

Looking at the three sets of DirAC features individually, the elevation values alone give similar performance to the MFCCs, whereas diffuseness alone performs somewhat better than the MFCCs, except in fold 2. Using the Azimuth values alone gives the worst performance of any of the feature sets used here, averaging just 42% accuracy across all folds, and performing as badly as 31% accuracy in fold 1, though in fold 3 the accuracy is comparable to the MFCC performance. This is probably due to the fact that, whilst there should be some consistency of azimuth DOA values in similar acoustic scenes (indeed this is borne out by the fact that this classifier still performs better than chance), these azimuth values will be much more sensitive to the specific orientation of the microphone array when the recordings of the sound scenes were made. If one street scene, for instance, was recorded with the front of the microphone array facing the road, whereas another was recorded with the front parallel to the road, this will result in inconsistent azimuth values between the two scenes.

By contrast, elevation and diffuseness values should theoretically be independent of microphone rotation. This could account for the relatively high accuracy results when using these features. Because of the low accuracy results from the azimuth data, a new DirAC classifier was trained excluding this data. The results from this classifier are plotted in Figures 1 and 2 as ‘Elev/Diff’. This combination of elevation and diffuseness data was the best performing feature set on average, at 69% accuracy. With fold 4, these features give an accuracy of 76%, which was the maximum accuracy achieved in this test. There is, however, once again a marked difference in accuracy - around 10% - between the performance of the classifier in folds 1 and 2 (66%, 61%) relative to folds 3 and 4 (74%, 76%).

3.2. Classifier Confusion

Figures 3 and 4 show confusion matrices describing the classifications made by the MFCC and Elevation/Diffuseness (E/D) classifiers across all folds. Rows represent the correct classification, where columns are the labels returned by the classifiers. The values shown represent the percentages of 30-second segments classified.

From these plots, it can be seen that the E/D classifier exceeds the accuracy of the MFCC classifier for all acoustic scene classes except BusyStreet and Beach. The differences in accuracy between the two classifiers for the BusyStreet class is small. For Beach, however, the MFCC classifier classifies 36% of the samples correctly, whilst the E/D classifier classifies only 8% correctly. In fact, the Beach class is the source of the majority of the incorrect classifications from the E/D classifier. If the Beach class is excluded, the overall classifier accuracy increases from 69% to 78%.

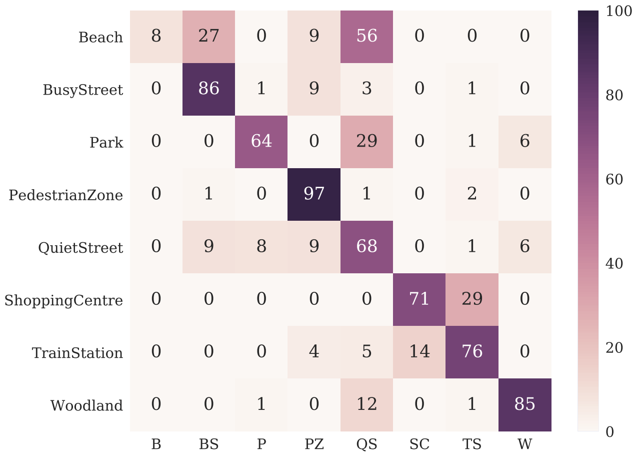

Figure 3: Confusion matrix of classifier trained using Elevation and Diffuseness features. Figures indicate classification percentages across all folds.

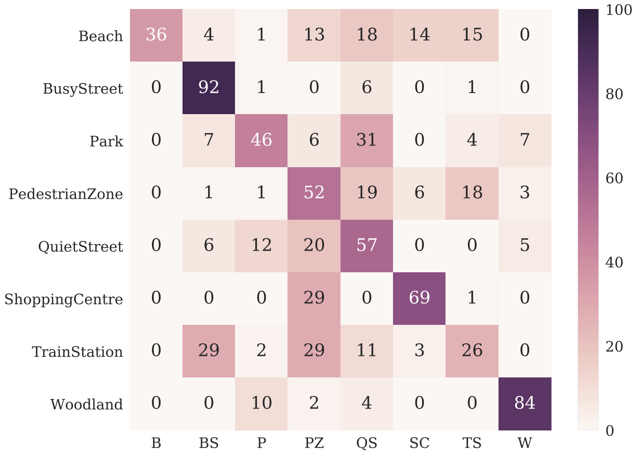

Figure 4: Confusion matrix of classifier trained using MFCC features. Figures indicate classification percentages across all folds.

This poor performance is perhaps due to the fact that in a seafront beach acoustic environment the dominant source of sound is wave motion from the sea. This will appear to DirAC analysis as a large spread of broadband noise. Looking at Figure 3, it can be seen that the E/D classifier mislabels most of the Beach clips as either BusyStreet or QuietStreet. Since one of the dominant sounds of street scenes is the broadband noise from passing cars, it is conceivable that the spatial features extracted from street environment recordings could mirror those from a beach.

Further interesting observations may be made by comparing the specific accuracies and differences in scene confusion between the two classifiers. PedestrianZone, for instance, is classified with only 52% accuracy by the MFCC classifier, with many samples classified as either TrainStation or QuietStreet, whilst the E/D classifier is 97% accurate for this class. This suggests that the spatial information present in the acoustic scene of a pedestrian zone is more unique to that scene than the spectral information, which evidently can be quite similar to that of a train station or quiet street.

It is also interesting to compare the two classifiers where there is significant confusion in both for a certain class, as often the confusion does not correspond. With ShoppingCentre, for instance, the majority of the confusion in results in the MFCC classifier is with PedestrianZone, perhaps owing to the prominent human sound (speech and footsteps) common to both locations. The E/D classifier, on the other hand, does not confuse ShoppingCentre with PedestrianZone at all, instead confusing it with TrainStation. This could be due to the nature of the recorded train stations and shopping centres as large reverberant indoor spaces, which could potentially influence the calculated values for elevation and particularly diffuseness.

4. DISCUSSION

The results of this experiment show that it is possible to classify acoustic scenes with reasonable accuracy using information on the spatial properties of the scenes as features with a basic GMM classifier. This is an important initial result as it confirms that spatial information could be a valuable feature to utilise in future acoustic scene analysis systems and is worthy of further study.

The results shown in Figure 2 indicate that the addition of MFCC features to DirAC features improves classification accuracy when the individual performance of the two feature sets is similar, but is a hinderance when the individual DirAC feature performance is better than the MFCCs alone. There is some indication of an inverse relationship between the MFCC performance and the performance of the DirAC and E/D classifiers, with the spatial classifiers performing much better when the MFCC performance is worse, though there is not enough data here to establish a trend.

Looking at the classification confusion of the MFCC classifier against the E/D classifier, it seems that in most cases spatial features more uniquely characterise acoustic scenes than spectral features. The differences in specific scene confusions between the two classifiers indicates that *spatial* similarity and *spectral* similarity between scenes are not necessarily the same.

5. CONCLUSIONS

This paper has presented a new system for the classification of acoustic scenes using spatial features extracted with Directional Audio Coding techniques. An extensive new dataset of Ambisonic acoustic scene recordings - the EigenScape dataset - was created for this research and introduced here. DirAC features extracted from EigenScape were used to train GMM classifiers and the accuracy of these classifiers was tested against classifiers trained using standard MFCC features. The DirAC-trained classifiers were shown to have comparable classification accuracy to the MFCC-trained classifiers and a subset of the DirAC features excluding azimuth estimates was shown to substantially outperform the MFCCs by over 10% on average.

Comparison of confusion matrices for the outputs of MFCC and DirAC-trained classifiers reveal many differences in specific scene confusions between the two. This indicates that acoustic scenes that have similar spatial features might not necessarily also have similar spectral features.

6. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.

[2] D. Wang and G. J. Brown, *Computation Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley, 2006.

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, May 2015.

[4] European Environmental Agency, "Good practice guide on noise exposure and potential health effects," European Union, Tech. Rep., 2010.

[5] A. L. Brown, "Soundscapes and environmental noise management," *Noise Control Engineering Journal*, vol. 58, no. 5, pp. 493 – 500, 2010.

[6] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, 2007.

[7] N. I. Joachim Thiemann and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, June 2013.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*, August 2016.

[9] mh Acoustics, *em32 Eigenmike® microphone array release notes*, mh acoustics, 25 Summit Ave, Summit, NJ 07901, April 2013. [Online]. Available: https://mhacoustics.com/sites/default/files/EigenmikeReleaseNotesV15.pdf

[10] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. of the 14th Python in Science Conference (SciPy 2015)*, 2015.

[11] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, June 2007.

[12] V. Pulkki, M.-V. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, "Directional audio coding - perception-based reproduction of spatial sound," in *International Workshop on the Principle and Applications of Spatial Hearing*, 2009.

[13] S. Marsland, *Machine Learning: An Algorithmic Perspective*. CRC Press, 2015.

[14] D. Stowell. (2012) simple-minded audio classifier in python (using mfcc and gmm). [Online]. Available: https://github.com/danstowell/smacpy

[15] M. Lagrange and G. Lafay, "The bag-of-frames approach: A not so sufficient model for urban soundscapes," *Journal of the Acoustical Society of America*, vol. 128, no. 5, November 2015.