Environmental Sound Classification Using Audio Texture Features

Gregory Reardon and Rahul Shankar

Dr. Juan Bello
MPATE-GE 2623: Music Information Retrieval

Abstract— **Research into the auditory perception of sound textures has revealed the importance of time-frequency statistics in the recognition of sounds. These summary statistics play a key role in mid-level auditory processing. This work investigates whether audio texture features can be useful for classifying the larger set of environmental sounds. A baseline feature space composed of statistics of mel-frequency cepstral coefficients and their temporal derivatives is compared with an audio texture feature space. A five-way cross-validation is performed with different random splits of a dataset comprised of five different classes of environmental sounds. Statistical analyses revealed that the texture features performed significantly better than the MFCC-based feature space for environmental sound classification.**

I. INTRODUCTION

Environmental sound recognition (ESR) is concerned with the analysis, identification, and classification of everyday sounds. Improving methods for ESR has the potential to revolutionize the way humans interact with their environment. Augmenting environments and spaces with engineering systems that can listen, analyze, and make decisions based on real-time audio input has many different applications. From remote surveillance, robot navigation, and auto-tagging of audio files, it is clear that machine listening can have positive real-world impact.

In order to perform these tasks, machine listening techniques for the analysis and classification of everyday sounds must be significantly advanced. Currently, human auditory perception is superior to machine perception in many respects [10]. Bridging the gap between artificial and biological computing requires drawing from human computational auditory perception.

Audio textures compose a specific subset of environmental sounds. They are the superposition of many similar acoustic events and describe many types of different environmental sounds such as urban noise, insect sounds, and rain. These acoustic events are rich in information and, as such, the auditory system encodes the information in a compact form of time-averaged statistics [8]. These sounds have ”local structure and randomness, but the characteristics of the fine structure remain constant on the large-scale” [11]. Audio textures are thus stationary processes over some temporal window. Most environmental sounds do not fit this description. They are typically non-stationary. But, given that the auditory system uses time-averaged statistics to summarize temporal details of signals, audio texture features might prove to be a useful representation for classifying environmental sounds.

This work tests this hypothesis using a feature space of cochlear envelope marginal moments, cochlear envelope subband correlations, and modulation spectra. This feature space is compared against a baseline of statistics of mel-frequency cepstral coefficients (MFCCs) and their delta and delta-delta coefficients. Section 2 provides background on environmental sound classification and sound texture statistics. Section 3 presents the methodology. Results are presented in section 4 and discussed in section 5. Section 6 concludes and discusses future work.

II. LITERATURE REVIEW

Audio signals can be thought of as comprising three broad categories: speech, music, and environmental sound. The characteristics of each of these types of signals are distinct, requiring different techniques for retrieving information from them. Speech and music signals have been studied much more extensively than environmental sounds. More recently environmental sound recognition and more generally, acoustic scene classification, has gained attention in the larger audio information retrieval community, specifically because of its potential for real-world impact [2].

The sheer variety of environmental sounds makes their recognition a difficult task. Some events are short, such as a gun-shot, while others are much longer, such as rain. Framing-based processes that seek to classify each frame of an audio signal are less useful in environmental sound recognition because the temporal window used for analysis would have to be known *a priori*. Characterizing sounds over multiple frames of analysis is more appropriate. These procedures are known as sub-framing-based processes. In these procedures, sub-frame features are either concatenated to form a larger feature vector or averaged across time to represent a single frame. This allows for capturing some dynamic time-varying information of the signal [3].

A number of different categories of audio features have been employed for ESR. These include temporal features, such as zero crossing rate, spectral features, such as statistical moments calculated from the frequency spectrum, auditory filter bank features, and cepstral features [1] [12]. Of particular interest in this work are auditory filter bank and cepstral features.

The most common cepstral features are MFCCs. As mentioned above, in ESR, classifying static or instantaneous representations of the audio signal at frame-level (framing-based processes) is less useful. Classifying at the clip-level requires collapsing the set of MFCCs into a single feature vector, such as the mean and covariance of the MFCCs. This summarization loses information about the time-varying characteristics of the signal. Such an aggregation procedure has no respect for the sequencing of MFCCs. Often included with MFCCs are their first and second derivatives, also known as delta coefficients and delta-delta coefficients, respectively (ΔMFCCs and $\Delta\Delta$MFCCs). These features attempt to capture some of the temporal dynamics between local frames of MFCCs [7]. Collapsing these coefficients into a single feature vector does encode some time-varying characteristics of the signal, but this information is limited in scope. A number of other cepstral features have been explored for ESR and can be found in summary in [3].

A set of auditory filter bank features were investigated by McDermott and Simoncelli [9] in a subjective evaluation. The authors analyzed recorded audio textures and shaped Gaussian noise to conform to the time-averaged statistics of the recorded sounds. The authors used an auditory filter bank to deconstruct the signal and measured the first four statistical moments of each subband envelope and the cross-band correlation. Each subband was then Fourier transformed and the modulation power and modulation correlations calculated. In this study, these features are known as *audio texture features*. These measurements were used to shape the Gaussian noise signal and the synthesized stimuli were used in a perceptual texture recognition task. The authors concluded that these summary statistics play a significant in auditory texture perception and that the brain does encode auditory events in this compact form. It was also noted that if an event is sufficiently incongruous with a texture, it might be heard and remembered as a distinct event [8].

Most environmental sounds cannot be characterized as audio textures. But, given that many acoustic scenes and some environmental sounds are characterized as audio textures, have texture-like properties, or are composed of both textures and transient events, audio textures features should be investigated for use in ESR. Such a technique was investigated by Ellis, Zeng, and McDermott [4]. The authors tested audio texture features against a baseline feature space of MFCCs, ΔMFCCs,and $\Delta\Delta$MFCCs. Nine different categories of environmental sound were used (outdoor-rural, outdoor-urban, indoor-quiet, indoor-noisy, dubbed audio, intelligible speech, music, cheering, and clapping).

III. METHODOLOGY

A. Dataset and Procedure

In this work, the ESC-50 dataset for environmental sound classification was used [10]. The dataset was constructed from recordings publicly available through the Freesound project [5]. It is comprised of 2,000 labeled 5-second audio clips of various common sound events. There are 50 classes in the dataset, grouped at large into 5 classes:

- animal sounds
- natural soundscapes and water sounds
- human(non-speech) sounds
- interior/domestic sounds
- exterior/urban noise.

The dataset is balanced, with 40 clips per class and 10 classes for each of the larger 5 classes. The high dimensionality of the feature vectors used by the authors necessitated that the larger 5 classes be used in this work. Thus there were 400 5-second clips for each class.

These files were initially in .ogg format. They were first brought into MATLAB, normalized to between -1 and 1, and converted and rewritten as uncompressed .wav files. These files were then brought into Python where the feature extraction and classification occurred. The procedure, to be detailed below, closely follows that of Ellis, Zang, and McDermott [4]. Distinct from those authors, this work uses a much more diverse set of environmental sounds, many of which are not characterized as audio textures nor have textural noise in the background. It is the goal of this work to better understand the performance of audio texture features in ESR on varied types of sounds.

B. MFCC Features

The baseline feature space was composed of statistics of MFCCs, ΔMFCCs,and $\Delta\Delta$MFCCs. MFCCs were calculated over 24 ms windows with 50% overlap between adjacent frames. 14-dimensional cepstra were used to represent the signal calculated from a 40-band mel spectrum. ΔMFCCs were calculated as regression coefficients:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1}$$

where $\mathrm{d_t}$ is the delta coefficient at time t, N is the number of frames over which the derivative is estimated and $\mathrm{c_{t+n}}$ to $\mathrm{c_{t\text{-}n}}$ are static MFCCs at their given time. $\Delta\Delta$MFCCs were calculated using the same equation substituting the MFCCs with their first derivative. Thus in this baseline system each frame was represented as 42 coefficients. In this implementation, derivatives were calculated over 9 frames of data to capture some of the temporal dynamics of the signal. The entire clip is then described by the mean and covariance of these feature vectors. To transform the covariance matrix into a usable feature, the leading diagonal and the next five diagonals (42 + 41 + 40 + 39 + 38 + 37 = 237 dimensions) were grabbed and placed in a single vector of length 237. Only the first few diagonals were grabbed because the authors were concerned with relations between adjacent and neighboring subbands, not distant subbands, and it reduced the dimensionality of the feature vector. Thus the entire clip was characterized by a 279-dimensional feature vector (42 means + 237 variances and covariances) created from the statistics of MFCCs and their first- and second-order temporal derivatives.

C. Audio Texture Features

The audio texture features were calculated from the output of an 18-band auditory filter bank, created on the Mel scale. This simulates the cochlear response of the auditory system. The architecture of the procedure can be seen in

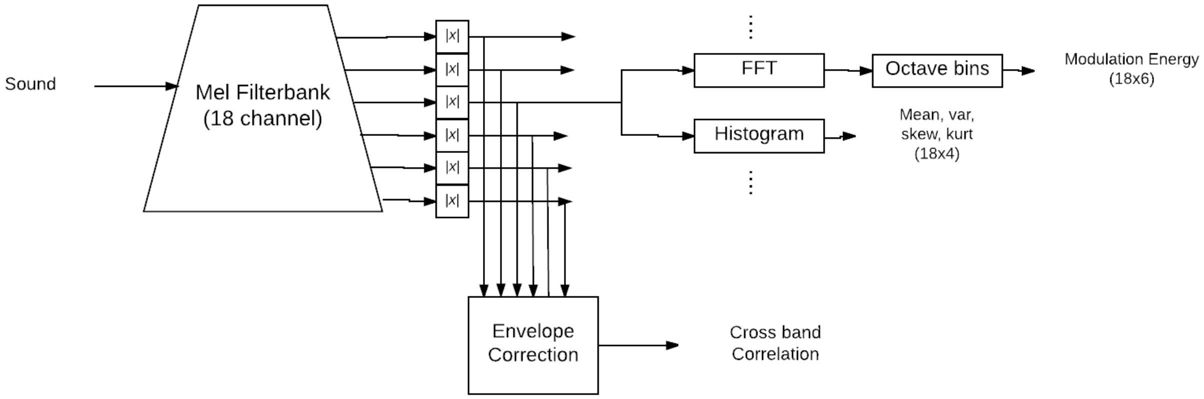

Fig. 1. Diagram of audio texture feature calculation

Fig. 1 and is as follows. The signal was first broken up via short-time Fourier Transform (STFT) with a 5.8 ms hamming window with 50% overlap between frames. The log-magnitude spectrum was filtered with the 18-band auditory filter bank. The first four statistical moments (mean, variance, skew, and kurtosis) of each subband (18 x 4 dimensions) and the normalized correlation between subbands were then calculated (18 x 18 matrix). The first five diagonals of the correlation matrix were used to represent the relation between neighboring subbands. The leading diagonal was excluded as it is exactly 1. This resulted in single vector of length 87 (17 + 16 + 15 + 14 + 13 = 75 dimensions). The subband signals were then Fourier transformed again to obtain a modulation spectrum. Each subband modulation spectrum was normalized by the variance of its subband as per [9]. The resulting magnitudes were collected into six octave-wide modulation bands (0.5 - 1 Hz, 1 - 2 Hz, 2 - 4 Hz, 4 - 8 Hz, 8 - 16 Hz, and 16 - 32 Hz). This resulted in another feature block of 18 x 6 dimensions. The feature blocks were concatenated to return a 255 dimensional feature vector for each clip ((18 x 4) + 75 + (18 x 6) = 255 dimensions).

D. Classifier

A five-way cross-validation was performed using a random forest classifier. Random forests are easier to train on a smaller dataset. The decision to use random forest classifiers was made considering the dimensionality of our feature spaces and the high performance of random classifiers on the ESC-50 database [10]. In the five-way cross-validation, the 2000 recordings were split at random into five equal-sized sets of 400 recordings. Four of the five splits were used for training the random forest classifier, while the remaining split was used for testing (80% train and 20% test). This results in 5 different 5-class accuracies for each of the feature spaces in each run of the data. The parameters used in the feature extraction procedure detailed above and the classifier were approximately optimized through manual tuning by running the extracted features through the classifier and cross-validation procedure. A third feature space that was the composite of the MFFC feature space and the audio textures was also tested.

IV. Results

The cross-validation procedure was repeated for 5 different random splits of the data. In order to understand per-class accuracy and the overall performance of each of the feature spaces, a 3 x 5 x 5 univariate repeated-measures analysis of variance (ANOVA) was conducted to analyze the results. The test indicated significant effects due to feature space (F = 57.209, p = *0.001*) and class (F = 16.727, p = *0.001*). A significant interaction was found between feature-class (F = 9.253), $p < 0.001$). All other variables and interaction terms were not significant.

The main effect of feature space on classification accuracy is presented in *Fig. 2*. The texture features performed best, when compared to both the MFCC features and the combination of the two feature spaces. A post hoc multiple comparisons test was used to analyze the relation between the feature spaces and revealed a significant difference between Textures and MFCCS and between the combined features space and MFCCs. No significant differences were found between the Texture features and the combined feature space.

The main effect of class on percentage accuracy is presented in *Fig. 3*. A post hoc multiple comparisons test was also run and revealed some significant differences between classes. These results are not reported as the authors are more concerned with the general trend in the different classes and how each feature performed on each class. As seen in *Fig. 3*, class performance is extremely variable. The average

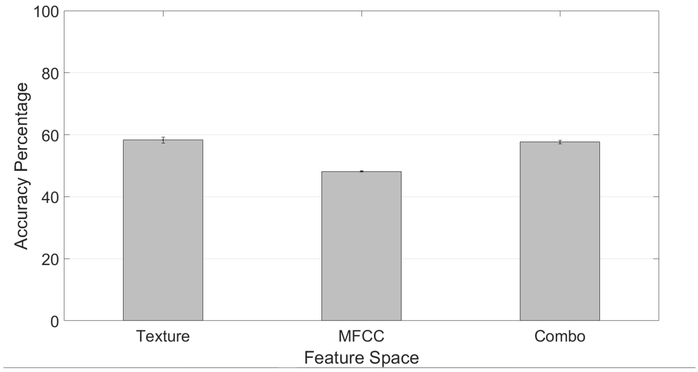

Fig. 2. Main effect for Feature Space (mean and standard error)

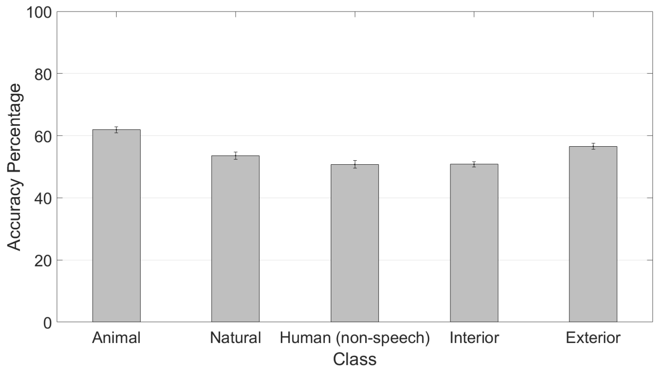

Fig. 3. Main effect for Class (mean and standard error)

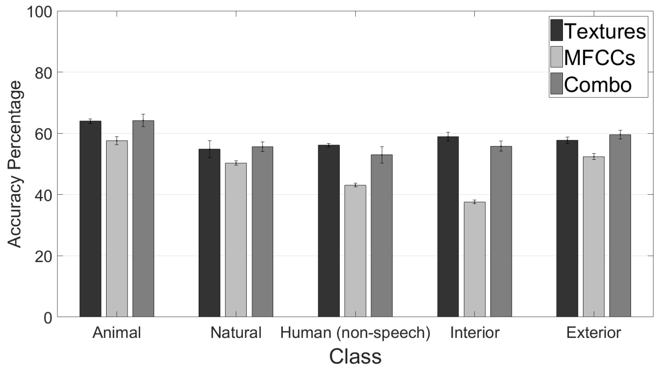

Fig. 4. Main effect for Feature-Class interaction (mean and standard error)

classification accuracy for animal sounds was over 10% greater than that of human (non-speech).

The result of most interest is the feature-class interaction. The main effect for feature-class is presented in *Fig. 4*. The variation in the performance of each of the feature spaces for the different types of classes is evident. In general, most of the variation in class accuracy comes from the poor performance of MFCCs with respect to interior and non-human speech sounds.

V. DISCUSSION

Significant differences between the baseline MFCC feature space and audio texture features point to texture features as useful representations of the audio signal for ESR. But, combining the feature spaces did not result in improved performance, except in one class of sounds - exterior. This is in contrast to those results reported by Ellis, Zeng, and McDermott [4]. The authors found that their baseline MFCC system performed marginally better than the texture features when classifying a wide range of audio soundtracks (environmental sound, speech, and music). And, the authors found that combining the feature spaces improved performance. It was concluded that texture features captured important information about the signal and could be used to augment another feature vector to improve classification.

In this work, texture features significantly outperformed the baseline MFCC system. This is likely due to the nature of the dataset. Regardless, there are improvements that can be made in the Δ and $\Delta\Delta$ coefficient calculations. When including Δ and $\Delta\Delta$ coefficients, a significant boost in accuracy of around is typically reported [6]

In this work, including these coefficients only improved classification accuracy by about 5 - 8 %. More optimization with respect to the temporal windows of derivative estimates can be performed. With respect to the texture features, the overall performance of 58.312% leaves much room for improvement. Specifically, the authors believe that further preprocessing of signals can improve results. Ellis, Zeng, and McDermott perform a time-frequency gain control procedure to balance energy in individual subbands before analyzing frames of data [4]. Including such a procedure might improve the performance of classifiers in this work. Further, when combining the feature spaces in this work, feature vectors were simply concatenated to form a large 532 dimensional vector. The authors suspect that reducing the dimensionality of this vector before classification might have improved performance. Given that only 400 sounds existed in each class and each feature vector is 532 dimensions, it is not surprising that performance did not scale accordingly.

There is much variation in environmental sounds. This remains one of the biggest difficulties in ESR and acoustic scene classification [3]. This is clearly present in the performance of all of the feature spaces and evidenced by the significant differences between class accuracies (*Fig. 3*). The MFCC system performed poorest on interior sounds. Many of these interior sounds, such as vacuum cleaners and washing machines, are best described as textural. Its likely that the stationary processes in these textures were well-represented by the texture features, but not by the MFCC features, which only looked at the first and second order derivatives over a local neighborhood. Another interesting result was that animal sounds had the highest classification accuracy. While some of the animal sounds, such as insects and frog are textures, many of sounds in this class are marked by highly transient events which are unlikely to be captured by the texture features. These results as a whole imply that there are improvements that can be made in the feature extraction stage.

VI. CONCLUSIONS AND FUTURE WORK

Our results show that audio texture features provide a moderate improvement over statistical MFCC features for environmental sound classification when using a random forest classifier. The combination of both feature spaces resulted in a non-significant drop in classification accuracy. Audio texture features clearly capture important information of environmental sound signals. Given, that many environmental sounds and more complex acoustic scenes are textural or are comprised of some sound textures, audio texture features can be used to complement other feature vectors. Further, work on auditory perception of sound textures and the role of summary statistics in sound recognition [8] imply that the brain compactly summarizes information about many sounds that are not entirely textural. Mid-level auditory processing is a piece of the larger auditory system. And while sound textures are the best class of sounds to use to observe how the brain encodes time-averaged statistics, these features might prove to be useful representations of more diverse classes of stimuli, specifically environmental sounds. The fact that animal sounds had the best classification performance in this work further bolsters this point. In the future, the authors would like to explore scattering representations for environmental sound classification. The non-stationary nature of the scattering transform might better handle the time-varying nature and variety of environmental sounds.

REFERENCES

[1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34, 2015.

[2] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. Sound analysis in smart cities. In *Computational Analysis of Sound Scenes and Events*, pages 373–397. Springer, 2018.

[3] Sachin Chachada and C-C Jay Kuo. Environmental sound recognition: A survey. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–9. IEEE, 2013.

[4] Daniel PW Ellis, Xiaohong Zeng, and Josh H McDermott. Classifying soundtracks with audio texture features. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5880–5883. IEEE, 2011.

[5] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st acm international conference on multimedia*, pages 411–412. ACM, 2013.

[6] Brian A Hanson and Ted H Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with lombard and noisy speech. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 857–860. IEEE, 1990.

[7] M Karbasi, SM Ahadi, and M Bahmanian. Environmental sound classification using spectral dynamic features. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1–5. IEEE, 2011.

[8] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013.

[9] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

[10] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM, 2015.

[11] Nicolas Saint-Arnaud and Kris Popat. Analysis and synthesis of sound textures. In *in Readings in Computational Auditory Scene Analysis*. Citeseer, 1995.

[12] Guoshen Yu and Jean-Jacques Slotine. Audio classification from time-frequency texture. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1677–1680. IEEE, 2009.