

CLASSIFYING SOUNDTRACKS WITH AUDIO TEXTURE FEATURES

Daniel P. W. Ellis, Xiaohong Zeng

LabROSA, Dept. of Electrical Engineering
Columbia University, New York
{dpwe,xiaohong}@ee.columbia.edu

Josh H. McDermott

Center for Neural Science
New York University, New York
jhm@cns.nyu.edu

ABSTRACT

Sound textures may be defined as sounds whose character depends on statistical properties as much as the specific details of each individually-perceived event. Recent work has devised a set of statistics that, when synthetically imposed, allow listeners to identify a wide range of environmental sound textures. In this work, we investigate using these statistics for automatic classification of a set of environmental sound classes defined over a set of web videos depicting “multimedia events”. We show that the texture statistics perform as well as our best conventional statistics (based on MFCC covariance). We further examine the relative contributions of the different statistics, showing the importance of modulation spectra and cross-band envelope correlations.

Index Terms— Sound textures, soundtrack classification, environmental sound.

1. INTRODUCTION

Sound textures, as produced by a river, or a crowd, or a helicopter, can readily be identified by listeners. But listeners may not be able to distinguish between different excerpts from a single texture: what has been recognized is something relating to the overall statistical behavior of the sound, rather than the precise details. The principles underlying our perception of this statistical structure could be valuable in the design and construction of automatic content classification, as textures are common in real-world audio signals. One application would be a system to classify web videos as belonging to particular categories of interest, on the basis of their soundtracks including relevant or tell-tale sound textures.

In [1], sound texture perception was investigated by measuring various statistics in a real-world sound texture, imposing the measured statistics on a noise signal, and then testing whether the result was perceived to sound like the original. A number of statistics computed from an auditory subband analysis were found to allow perceptual identification of a wide range of natural sound textures. For instance, moments (the variance, skew, and kurtosis) of the amplitudes in each subband were found to be important (for capturing sparsity), as was the correlation between subband envelopes. Subjects

scored better than 80% correct in identifying a 5 second synthesized texture drawn from a pool of 25 classes.

In this paper we investigate whether statistics of this kind can also be useful in the *automatic* recognition of environmental sound textures. Our task is to label the soundtracks of short clips extracted from web videos with a set of 9 broad labels such as “outdoor-rural”, “indoor-quiet”, “music”, etc. We compare features modeled after [1] with a conventional baseline that uses the statistics of Mel-frequency cepstral coefficient (MFCC) features and their derivatives.

Prior work on sound textures has investigated different representations and methods including wavelet trees [2] and frequency-domain linear prediction [3]. These approaches have considered perceptual and biological aspects only indirectly, unlike the direct perceptual validation at the basis of this work. They are also concerned primarily with synthesis, not classification. Environmental sound classification has been addressed by our earlier work using MFCC statistics [4] and Markov models [5], among many others [6, 7, 8], but not with a range of features specifically aimed at sound textures.

Section 2 describes the task and our data in more detail. Section 3 describes our approach, including the texture feature set, the baseline MFCC features, and the SVM classifier. Section 4 reports our results on both feature sets and their combination. We draw conclusions in section 5.

2. TASK AND DATA

The soundtrack classification system was developed as part of a system for the TRECVID 2010 Multimedia Event Detection task [9]. This evaluation is aimed at systems able to detect complex events in videos, using the specific examples of “Making a cake”, “Batting a run”, and “Assembling a shelter”. The task includes a development set of 1746 videos, including 50 positive examples of each class. These videos come from a range of public video websites, and span the styles and qualities of video typically encountered online.

As part of a wider effort to detect these events based on mid-level semantic classes, this project was to develop a set of soundtrack classifiers trained on specific labels applied to the video data. To this end, a set of 9 semantic properties

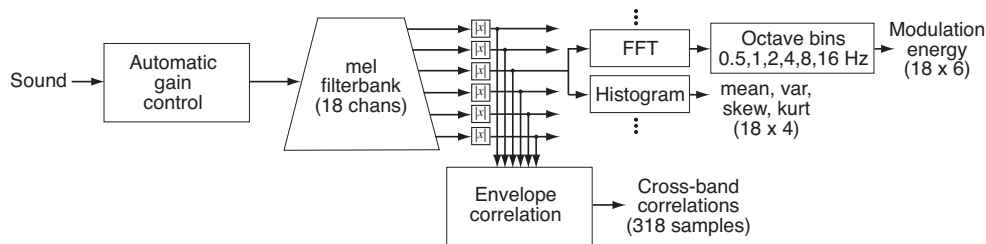


Fig. 1. Block diagram of the texture feature calculation.

was defined, as listed in table 1. To create training data for these classifiers, each of 534 development set videos was segmented into nonoverlapping 10 s clips. Each of the resulting 6630 clips was manually annotated with the nine attributes. Because clips were annotated separately, different clips from the same video could have different labels.

3. APPROACH

3.1. Sound texture features

The sound texture features are calculated as shown in figure 1. The input sound file is first put through a frequency-dependent automatic gain control to reduce the impact of different recording conditions and channels [10]. Temporal smoothing has a time constant of 0.5 s, and spectral smoothing is over a sliding 1 mel window. The signal is then broken into 18 subbands on a mel scale to simulate an auditory filterbank; in practice, this is achieved by combining the bin magnitudes of a short-time Fourier transform (STFT) operating with a 32 ms window and 16 ms hop time. Each channel’s log-magnitudes are accumulated over 8.2 s (256 of the 16 ms frames); after discarding values more than 40 dB below the peak, the histogram of magnitudes is described by its first

Table 1. The nine labels applied to each 10 s segment from the TRECVID MED development set. Note that the first four classes are mutually exclusive, i.e., each clip carries at most one of these labels. Each video is divided into multiple 10 s clips.

Concept	# videos	# clips
Outdoor - rural	278	1387
Outdoor - urban	146	570
Indoor - quiet	225	1905
Indoor - noisy	265	1735
Dubbed audio	249	2074
Intelligible speech	333	2882
Music	249	2538
Cheering	151	416
Clapping	99	261

four moments – the mean, variance, skew, and kurtosis – to give the first block of 18×4 features.

The sequence of 256 subband magnitudes is also Fourier transformed to obtain a modulation spectrum. The magnitudes are collected into six octave-wide modulation bands spanning 0.5-1 Hz, 1-2 Hz, 2-4 Hz, 4-8 Hz, 8-16 Hz, and 16 Hz to the Nyquist rate of 32 Hz. This constitutes a second block of 18×6 features. Finally, the normalized correlations between all the subband envelopes are analyzed. This 18×18 matrix is represented by its first 12 diagonals, such that correlations between spectrally distant channels are not included (we also exclude the leading diagonal, which is identically 1) to give a further $17 + 16 + \dots + 6 = 138$ dimensions. Thus, each 8.2 s chunk of sound is transformed into $18 \times 4 + 18 \times 6 + 138 = 318$ dimensions. For longer sounds, the analysis is repeated every 4.1 s, although the clips in this study were each just 10 s long, so were analyzed as a single frame.

These features follow the results of [1], who found that synthetic sounds shaped to match such statistics of an original sound were generally recognizable to listeners, with recognition improving as more statistics were matched. Using subband histogram moments was superior to simply matching the energy in each subband, and cross-band correlations and subband modulation spectra further improved the realism and recognizability of the synthesis. Subband histogram moments help distinguish between sound textures that have fairly steady power in a subband (like classic filtered noise) versus power that has a few, sparse, large values (large variance, skew and kurtosis, as in a crackling fire). The modulation spectrum helps to capture the characteristic rhythm and smoothness of these variations within each subband (e.g. fast and rough in clapping vs. slow and smooth in seawash). Cross-band correlations can identify subbands that exhibit synchronized energy maxima (e.g., crackling, speech), as distinct from independent variations in each band (many water sounds).

3.2. MFCC features

Our baseline system models the second-order statistics of common MFCC features, calculated over 32 ms windows on a 16 ms grid. We used 20-dimensional cepstra calculated

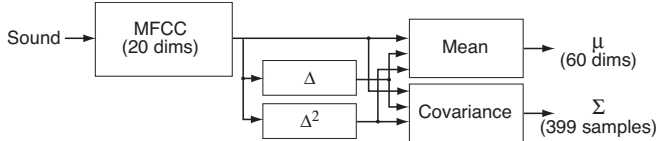


Fig. 2. Block diagram of the MFCC feature calculation.

from a 40-band mel spectrum. The feature calculation is as illustrated in figure 2; to include some information on temporal structure, we calculate delta and double-delta coefficients (over a 9 frame window) to give a total of 60 dimensions at each time frame. The entire clip is then described by the mean and covariance of these feature vectors; the 60×60 covariance matrix is represented by its leading diagonal and next 6 diagonals, giving $60 + 59 + \dots + 55 = 399$ unique values, and each of these is treated as a separate clip-level feature dimension. Thus, each 10 s clip is represented by $60 + 399 = 459$ dimensions. These numbers, as well as those of the texture features, were approximately optimized through trial and error.

3.3. SVM classifier

To build the audio classifiers, we take feature vectors from clips that have been manually labeled as reflecting a particular class from table 1 (the positive examples), a second set that do not belong to the class (negative examples), and train the parameters of a generic discriminative classifier. We use support vector machines (SVMs) with a Gaussian kernel. Such classifiers calculate the Euclidean distance between all training examples (positive and negative), scale them with a parameter γ , and optimize a decision plane in the implied infinite-dimensional space, trading misclassifications for “margin width” according to a weighting parameter C .

The tolerance of misclassification limits the success of SVMs trained with a large imbalance between positive and negative examples. To avoid this, we took the simple measure of discarding examples from the larger class (usually negative examples) until we had a number equal to the smaller class. This also meant that we could evaluate the performance of all classifiers by accuracy, where random performance would give an accuracy of 50%.

To set the parameters γ and C , we performed a coarse grid search, with the classifier trained on half the positive and negative examples, then tested on the other half, and then trained and tested again with training and test sets interchanged (2 way cross-validation). We were careful to assign all the clips cut from any single video to the same cross-validation half, to avoid a test set containing clips with properties unrealistically similar to items in the training set.

The parameters giving the best accuracy on cross-validation were retained. This cross-validation accuracy is also the performance figure we report below; although this is an over-

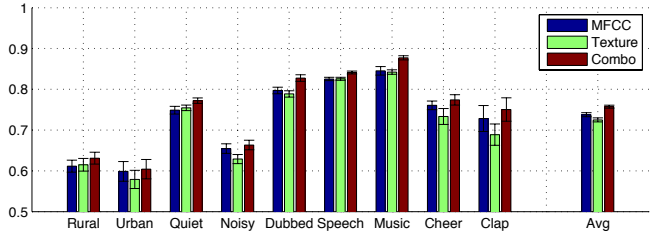


Fig. 3. Classification accuracy by class for the baseline MFCC system, the texture-based system, and a combination of both classifiers. Error bars show the standard deviation of results from 20 runs with different random train/test splits.

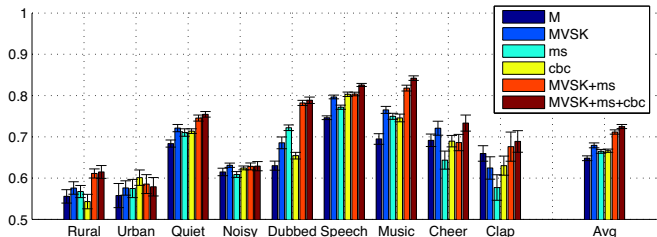


Fig. 4. Classification accuracy by class for various combinations of the texture feature subblocks. M = mean subband energies; MVSK = all four subband moments; ms = modulation spectrum; cbc = cross-band correlations.

estimate of the performance on truly unseen test data, it is sufficient for the comparison between different systems and configurations.

4. RESULTS

Figure 3 shows the overall accuracy for the nine classifiers, comparing the baseline (MFCC) system, the system based on the full set of texture features, and a combination system formed by averaging the distance-to-margin estimates of both systems prior to making the final classification. (This scheme interprets the distance-to-margin coming out of the SVM as a kind of confidence or mapped posterior). We see a wide range of performance across classes, with relatively poor performance for the classes with the fewest training examples (“Urban”, “Cheer”, “Clap”), and strong performance for the classes with clear acoustic properties (“Speech”, “Music”), as well as for classes strongly correlated with these attributes (“Quiet” frequently occurs with “Speech”, and “Dubbed” frequently occurs with “Music”). The MFCC and texture systems have very similar average performance, although the texture system appears to have the edge for “Quiet” and “Rural”, and the MFCC system is superior for “Urban”, “Noisy”, “Cheer”, and “Clap”. The simple margin combination scheme outperforms either system alone in every case, giving an overall accuracy averaged over all classes

of $75.5 \pm 0.4\%$, versus $73.8 \pm 0.5\%$ for the baseline, and $72.5 \pm 0.5\%$ for the texture system.

Figure 4 provides some additional insight into the texture features by showing the accuracies by class for systems built from different subsets of the texture feature blocks. Using higher order moments (i.e., variance, skew, and kurtosis) gives a clear advantage over subband mean alone – although most of this gain is provided by just the variance. Modulation spectra and cross-band correlations showed large differences between classes, but performed roughly the same as each other, and a little worse than the moments, when averaged across all classes. Combining them with the moments gave a significant gain for all classes (except the difficult “Urban” case), indicating complementary information. Interestingly, modulation spectra are not particularly useful for “Speech”, but cross-band correlations are.

5. CONCLUSIONS

We have shown that the perceptually important statistics in sound textures are a useful basis for general-purpose soundtrack classification. They can be used to recognize foreground sound categories like speech, music, and clapping, as well as more loosely-defined contexts such as outdoor-rural, and indoor-noisy. Classifiers based on texture statistics are able to achieve accuracies very similar to those based on conventional MFCC features, and the two approaches can be easily and profitably combined.

The similarity in performance between MFCC and texture features raises the question as to whether they are truly modeling different aspects of the sound. Apart from the DCT involved in cepstral calculation, the MFCC features resemble the mean and variance moments from the texture features. The MFCC features also measure the covariance between different feature dimensions, something represented separately by cross-band correlations in texture features. The deltas and double-deltas in the MFCCs give a limited view of the temporal behavior of each dimension, whereas the modulation spectra in the texture set describe the temporal structure at a broad range of time scales. This may be behind the benefits obtained by combining MFCC and texture-based classifiers.

The particular task we have investigated is not the ideal test for texture statistics, since the categories we sought to distinguish are not crisply distinguished by texture. A class like “Indoor-noisy” might consist of restaurant babble or machine noise without distinguishing between them, even though they would be perceived as very different textures. On a test of more precise categorization over a wider range of sounds – such as recognizing or describing the particular characteristics of a soundtrack – texture features might show a greater advantage. This will be the focus of our future work, including an approach to characterizing soundtracks by their textual similarity to a large set of reference sound ambiences obtained from a commercial sound effects library.

6. ACKNOWLEDGMENTS

Many thanks to Yu-Gang Jiang and the other members of Columbia’s Digital Video MultiMedia lab for providing access to the manual labels of the MED data. This work was supported by a grant from the National Geospatial Intelligence Agency.

7. REFERENCES

- [1] Josh H. McDermott, Andrew J. Oxenham, and Eero P. Simoncelli, “Sound texture synthesis via filter statistics,” in *Proc. IEEE WASPAA*, Mohonk, 2009, pp. 297–300.
- [2] S. Dubnov, Z. Bar-Joseph, Ran El-Yaniv, D. Lischinski, and M. Werman, “Synthesizing sound textures through wavelet tree learning,” *IEEE Computer Graphics and Applications*, vol. 22, no. 4, pp. 38–48, Jul/Aug 2002.
- [3] Marios Athineos and Daniel P. W. Ellis, “Sound texture modelling with linear prediction in both time and frequency domains,” in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, 2003, pp. V–648–651.
- [4] K. Lee and D.P.W. Ellis, “Audio-based semantic concept classification for consumer video,” *IEEE TASLP*, vol. 18, no. 6, pp. 1406–1416, Aug 2010.
- [5] Keansub Lee, Daniel P. W. Ellis, and Alexander C. Loui, “Detecting local semantic concepts in environmental sounds using markov model based clustering,” in *Proc. IEEE ICASSP*, Dallas, 2010, pp. 2278–2281.
- [6] Lie Lu, Hong-Jiang Zhang, and Stan Z. Li, “Content-based audio classification and segmentation by using support vector machines,” *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 04 2003.
- [7] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, “Audio-based context recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 321–329, Jan. 2006.
- [8] S. Chu, S. Narayanan, and C.C.J. Kuo, “Environmental Sound Recognition With Time–Frequency Audio Features,” *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 17, no. 6, pp. 1142–1158, 2009.
- [9] NIST Multimodal Information Group, “2010 TRECVID Multimedia Event Detection track,” 2010, <http://www.nist.gov/itl/iad/mig/med10.cfm>.
- [10] D. Ellis, “Time-frequency automatic gain control,” 2010, http://labrosa.ee.columbia.edu/matlab/tf_agc/.

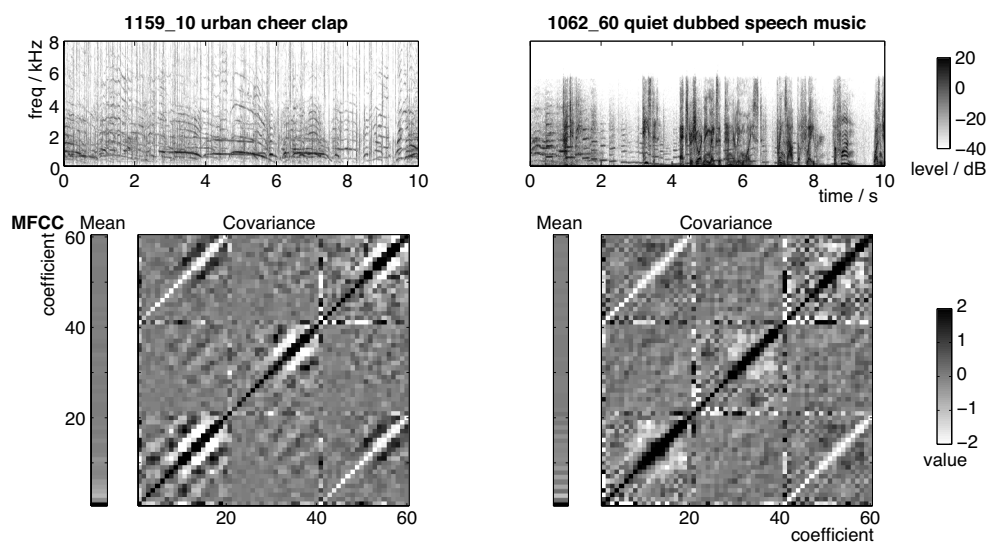


Fig. 5. Two example soundtracks along with their MFCC-based representations. The top row shows conventional spectrograms for each 10 s segment. Below are the 60 element mean vectors (means of the 20 dimensional MFCCs over the entire clip, as well as their deltas and double-deltas), and the 60×60 covariance matrix. (Only a 7-cell-wide strip of the upper-diagonal nonredundant part of this matrix is used as features, for 399 dimensions). The audio segment on the left consists of cheering and clapping at a sports event. It is annotated with the labels “outdoor-urban”, “cheering”, and “clapping”. The segment on the right comes from a 1960s TV commercial for cake mix (containing both music and voice-over) and is annotated as “indoor-quiet”, “dubbed audio”, “intelligible speech”, and “music”.

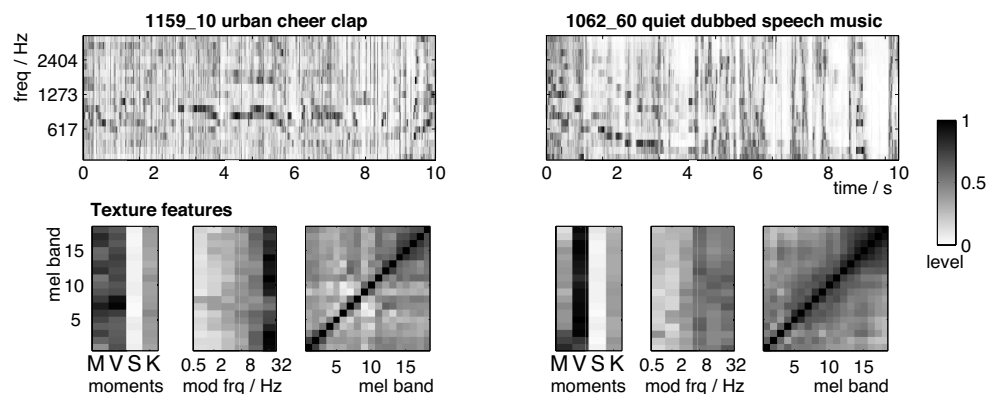


Fig. 6. The two examples of figure 5 along with their texture-feature representations. The top row shows the post-agc mel-scaled 18 band spectrograms. The bottom row shows, for each clip, the four subband histogram moments (Mean, Variance, Skew, and Kurtosis), the six modulation frequency bins (spanning 0.5 to 32 Hz in octaves), and the 18×18 cross-band correlation coefficients. In the interests of data compactness, only a 12-cell wide off-diagonal stripe of this matrix, for 138 dimensions, is recorded as features.