THE AUDIO EPITOME: A NEW REPRESENTATION FOR MODELING AND CLASSIFYING AUDITORY PHENOMENA

Ashish Kapoor[1] *and Sumit Basu*[2]

[1]Massachusetts Institute of Technology and [2]Microsoft Research

ABSTRACT

This paper presents a novel representation for auditory environments that can be used for classifying events of interest, such as speech, cars, etc., and potentially used to classify the environments themselves. We propose a novel discriminative framework that is based on the audio epitome, an audio extension of the image representation developed by Jojic et al. [3]. We also develop an informative patch sampling procedure to train the epitomes. This procedure reduces the computational complexity and increases the quality of the epitome. For classification, the training data is used to learn distributions over the epitomes to model the different classes; the distributions for new inputs are then compared to these models. On a task of distinguishing between 4 auditory classes in the context of environmental sounds (car, speech, birds, utensils), our method outperforms the conventional approaches of nearest neighbor and mixture of Gaussians on three out of the four classes.

1. INTRODUCTION

In this work, we propose a new representation and method for auditory perception that has the potential to cover a broad range of tasks, from classifying and segmenting sound objects to representing and classifying auditory environments. The core representation is an epitome, a model introduced by Jojic et al. [3] for the image domain. The basic idea was to find an optimal "palette" from which patches of various sizes could be drawn in order to reconstruct a full image. We apply this idea to the log spectrogram and log melgram with one-dimensional patches and find an optimal spectral palette from which we can take pieces to explain our input sequence. This epitome will have sound elements of a variety of timescales that it finds most appropriate to represent what it observed in the input sequence. For instance, if the input contained the relatively long sounds of cars passing by and also some impulsive sounds, like car doors opening and closing, these would both be stored as chunks of sound in the *same* epitome – without having to change the model parameters or training procedure.

Furthermore, the epitome is learned without specifying the target patterns to be classified, and attempts to learn a model of all representative sounds in the environment. To aid in this process, we have developed a new training procedure for the epitome that efficiently allows us to maximize the epitome's coverage of the different sounds. Once we have trained the epitome, we learn distributions over the epitome for each target class, which could also be applied to entire auditory environments. In other words, we treat the epitome as a continuous "alphabet" that represents the space of all possible sounds, and build models of our target classes in terms of this alphabet. We can then classify new patches and do segmentation based on these models.

2. PRIOR WORK

Thus far, there have been a variety of different approaches to recognizing audio classes and classifying auditory scenes. Most of the sound recognition work has focused on particular classes such as speech detection, and the best methods involve specialized methods and features that take advantage of the target class. For example, Zhang and Kuo [5] have described heuristics for audio data annotation. The heuristics they have chosen are highly dependent on the target classes, thus their approach cannot be extended to incorporate other more general classes. There have been discriminative approaches such as [2], where support vector machines were used for general audio segmentation and retrieval. This approach is promising but is restricted in the sense that you need to know the exact classes of sounds that you want to detect/recognize in advance at the time of training. Similarly, there are approaches based on HMMs [1],[4]. These approaches suffer from the same problem of spending all their resources in modeling the target classes (assumed to be known beforehand), thus extending these systems to a new class is not trivial. Finally, these methods were tested on databases where the sounds appeared in isolation, which is not a valid model of real-world situations.

We believe that our approach will overcome some of these limitations, since we learn a representation of all sounds in the environment at once with the epitome and then train classifiers based on this representation. In the following sections, we detail the epitomic representation, describe how we use it for classification, and show a

variety of results from our preliminary experiments on segmenting and classifying sounds.

3. OUR APPROACH

Our approach can be divided into two parts: first, learning the audio epitome itself, and second, using the epitome to build classifiers; both are described in the subsections below.

3.1. Audio Epitomes

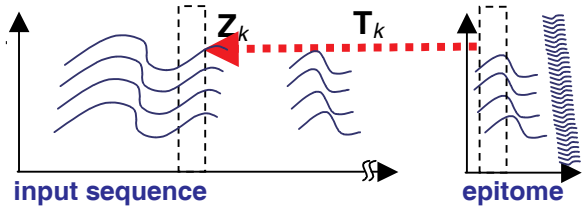

Figure 1: The audio epitome representation.

The basic principle of the audio epitome is shown in Figure 1 above: the input sequence is a log magnitude spectrogram, and the epitome is a "palette" for such spectrograms. Observed patches in the input sequence, Z_k, are explained by selecting the patch from the epitome e with the appropriate transformation (i.e., offset) T_k, i.e., where in the epitome the patch comes from. The probability of observing Z_k given this epitome and offset is a product of Gaussians over pixels as below:

$$P(Z_k \mid T_k, e) = \prod_{i \in S_k} N(z_{i,k}; \mu_{T_k(i)}, \phi_{T_k(i)})$$

where the i's are for the iteration over the individual frequency-time values or "pixels" of the spectrogram. Jojic et al. [3] describe the mechanisms by which to learn this epitome from an input sequence and to do inference, i.e., to find $P(T_k \mid Z_k, e)$ from an input patch.

The training procedure requires first selecting a fixed number of patches from random positions in the image. Each patch is then averaged in to all possible offsets T_k in the epitome, but weighted by how well it fits that point, i.e., $P(Z_k \mid T_k, e)$. The idea is that if we select enough patches then we should expect a reasonable coverage of the image. In audio, we face two problems. First, the spectrograms can be very long, thus requiring a very large number of patches before adequate coverage is achieved. Second, there is often a lot of redundancy in the data in terms of repeated sounds. We need a training procedure that takes advantage of this structure, as we describe in the following subsection.

3.1.1. Informative Patch Sampling

Rather then selecting the patches randomly, our informative patch sampling approach aims to maximize coverage of the input spectrogram/melgram with as few patches as possible. The idea is to start with a uniform probability of selecting any patch and then updating the probability in every round based on the patches selected. Essentially, the patches similar to the patches selected so far are assigned a lower probability of selection. The details are shown in figure 2.

- Initialize $P^i(k)$ to uniform probability for all positions k in the Spectrogram
- For n = 1 to Num of Patches
  - Sample a position t from P^n. The selected patch: p^n=spectrogram (: , t : t + patch_size)
  - For all positions k in the input spectrogram compute: $Err(k) = (spec(:, t : t + patch\_size) - p^n)^2$ $P^{n+1}(k) = P^n(k) * Err(k)$
  - $P^{n+1}(k) = P^{n+1}(k) / sum(P^{n+1}(k))$

Figure 2: Informative patch selection algorithm.

Once we have selected the patches representative of the input audio signal, we can train the epitome. In our implementation, all the patches used for training the epitome are of equal size (15 frames, or 0.25 seconds long). Note that in all our experiments the audio is sampled at 16 kHz; we use an FFT framesize of 512 samples with an overlap of 256 samples, and 20 mel-frequency bins for the melgram. We use the EM algorithm to train epitomes as described in [3]. One major difference is that we do epitomic analysis only in 1-D. Specifically; the patches we use are always the full height of the spectrogram/melgram, as opposed to the patches of varying width and height used in image epitomes.

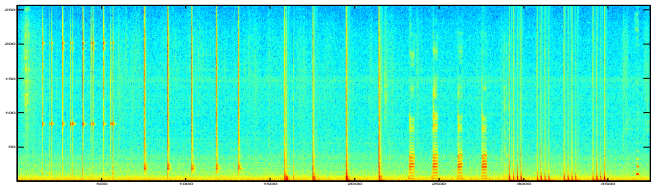

Figure 3: Spectrogram of the toy sequence showing repeated sounds.

Figure 3 shows a toy sequence which exhibits the kind of repetition we expect in natural sequences. It was collected in an office environment, and consists of repeating sounds of different objects being hit, speech, etc. From the figure

we can see not only the repetition but also a large amount of silence/background noise. If we randomly select patches, we will end up with mostly background patches and will have to select quite a few before we cover the whole spectrogram. Figure 4 shows epitomes trained from this sequence via both approaches. Figure 4 (left) is the epitome generated using random samples and Figure 4 (right) is the epitome generated using the same number of patches but now using our informative sampling scheme. Note that with our scheme, all of the individual sound elements from the input sequence have been captured, as opposed to the random sampling approach.

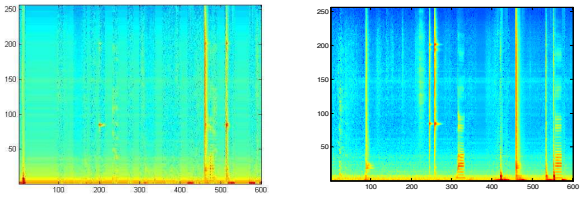

Figure 4: Epitomes learned using random (left) and informative (right) patch sampling.

3.1.2. Classification Using the Epitome

As we have shown, the learned epitome from an input sequence is a palette representing all the sound in that sequence. We now want to explore how to use this representation for classification. Since we expect different classes will be represented by patches from different parts of the epitome, our strategy is to look at the distribution of transformations T_k given a class c of interest, i.e. $P(T_k \mid c, e)$, and use this to represent the class. We can then classify a new patch by looking at how its distribution compares to those of the target classes.

In more detail, consider a series of examples from a target class that we would like to detect, e.g. a bird chirp. We first extract all possible patches of length 1-15 frames. Next we look at the most likely transformations from the epitome corresponding to each patch extracted from the given audio, i.e., $\max_k P(T_k \mid c, e)$, and then aggregate these to form the histogram for $P(T_k \mid c, e)$.

Figure 5 shows two example classes and the corresponding distributions $P(T_k \mid c, e)$. Figure 5 (left) corresponds to bird chirps and as the histogram suggests, most of the audio patches comes from only 4 positions in epitome. Note that this distribution is very different from the distribution that arises due to the acoustic event of cars passing by (Figure 5, right). Note that these distributions can be learned using very few examples for two reasons: first, we generate many patches from each example, and second, because the epitome has already compressed the input space into an optimal palette, and thus even a small number of examples should highlight the regions of the epitome that are assigned to explaining the class of interest.

Given a test audio segment to classify, we first estimate $P(T_k \mid c, e)$ using all the patches of length 1-15 from the test segment. We then seek the class $\hat{c}$ whose distribution best matches this sample distribution over all classes i in terms of the KL-divergence:

$$\hat{c} = \min_i D(P(T_k \mid c, e) \,\|\, P(T_k \mid c^i, e))$$

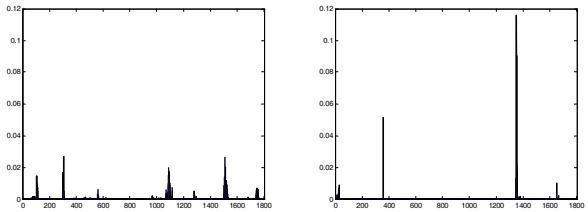

Figure 5: Distribution over transformations *T* for bird chirps (left) and cars (right)

Finally, though we have only used this framework to recognize individual sounds in our experiments, the method can also be used to model and recognize auditory environment via these distributions. In preliminary experiments, we have achieved good results on such tasks, and will report on this in a later paper.

4. EXPERIMENTS AND RESULTS

We first performed a set of experiments to compare the epitomic training using the informative patch selection with the training using random patch selection. For these experiments, we used the spectrogram shown in the figure 3. Figure 6 compares the likelihood of the input spectrogram given the epitomes trained using both the methods while varying the number of patches used for training. The higher likelihood corresponds to a better explanation of the input signal using the epitome. We averaged over 10 runs for each point in the curve. We can see that the epitome using the informative sampling always explains the input better than the epitome trained using random sampling. The difference is more prominent when the number of patches is small. Naturally, as the number of patches goes to infinity the curves will meet.

Next, we demonstrate speech detection on an outdoor sequence consisting of speech with significant background noise from nearby cars. We generated a 1 minute long epitome using 8 minutes of data. The speech class was trained as described in 3.1.2 using only 5 labeled examples of speech. Figure 7 shows the result of applying speech detection to a 10 second long audio sequence. The detector does a good job of isolating speech segments from the non-speech segments in very significant noise (around -10dB SSNR; this and other data can be heard at http://research.microsoft.com/~sumitb/ae). Note that there

is too much background noise for any intensity/frequency band based speech detector to work well.

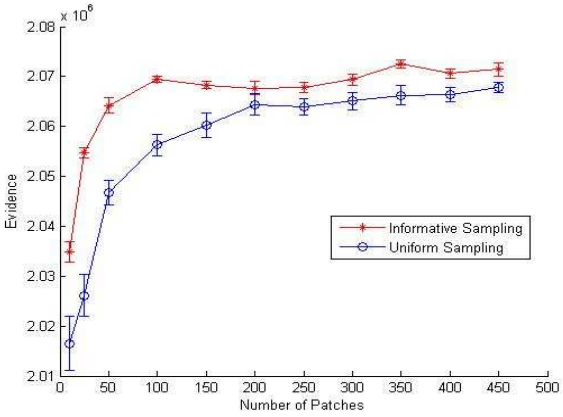

Figure 6: Evidence vs. number of training patches.

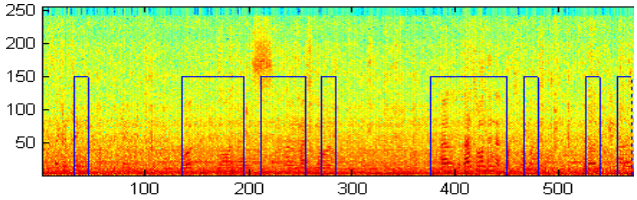

Figure 7: Speech detection example.

For the final evaluation, we collected audio data in 3 environments: a kitchen, parking lot, and a sidewalk along a busy street. On this data, we tried the task of recognizing four different acoustic classes: speech, cars passing by, kitchen utensils, and bird chirps. We segmented 22 examples of speech, 17 examples of cars, 29 examples of utensil sounds, and 24 examples of bird-chirps. Furthermore, there were 30 audio segments that contained none of the mentioned acoustic classes. We used the log mel-gram as our feature space and compared our approach with a nearest-neighbor (NN) classifier and a Gaussian Mixture Model (GMM) (both trained on individual feature frames; for the GMM the number of components were 1/10 the number of training frames, around 50 per class). For the non-epitome models, each frame was first classified using the NN or GMM, and then voting was used to decide the class-label for the segment. Note that training the epitome (which was used for all classes) took the same time as it took to train the GMM *for each class*. Table 1 compares the best performance obtained by each method using 10 samples per class for training.

Table 1: Classifier performance comparison.

	Epitome		Nearest-N		Mix of G	
	Pd	Pfa	Pd	Pfa	Pd	Pfa
Speech	**0.90**	**0.10**	0.86	0.09	0.93	0.28
Cars	**0.94**	**0.02**	**0.94**	**0.01**	1.00	0.09
Utensils	**0.94**	**0.12**	0.84	0.21	0.82	0.31
Bird Chirp	0.79	0.31	**0.94**	**0.11**	0.89	0.05

These numbers were obtained by averaging over 25 runs with a random training/testing split on every run. The proposed method outperforms both the nearest neighbor and the mixture of Gaussian in 2 out of the 4 cases. In one of the other 2 cases (cars), it is as good as the best performing method.

Finally, in Figure 8 we show the performance with increasing training data on the task of recognizing utensils. We can again see that the classification using the epitome is significantly better, especially when the amount of training data is small.

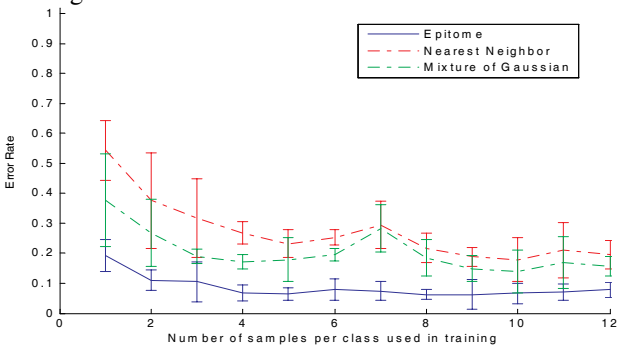

Figure 8: Error vs. number of training examples.

5. CONCLUSIONS AND FUTURE WORK

We have described a new representation for modeling audio and recognizing target classes based on the audio version of the epitome. In our future work, we plan to apply our framework to auditory environment classification and clustering.

6. ACKNOWLEDGEMENTS

Thanks to Nebojsa Jojic for helpful discussions and the image epitome code.

REFERENCES

[1] M. A. Casey, "Reduced-Rank Spectra and Minimum-Entropy Priors as Consistent and Reliable Cues for Generalized Sound Recognition," *Workshop for Consistent and Reliable Cues 2001*, Aalborg, Denmark.

[2] G. Guo and S. Z. Li, "Content-Based Audio Classification," *IEEE Transactions on Neural Networks*, Vol 14 (1), Jan. 2003.

[3] N. Jojic, B. Frey and A. Kannan, "Epitomic Analysis of Appearance and Shape," *Proceedings of International Conference on Computer Vision 2003*, Nice, France.

[4] M. J. Reyes-Gomez and D. P. W. Ellis, "Selection, Parameter Estimation and Discriminative Training of Hidden Markov Models for General Audio Modeling," *Proceedings of International Conference on Multimedia and Expo 2003*, Baltimore, USA.

[5] T. Zhang, C. C. J. Kuo, "Heuristic Approach for Audio Data Segmentation and Annotation," *Proceedings of ACM International Conference on Multimedia 1999*, Orlando, USA.