ENVIRONMENTAL SOUND RECOGNITION USING MP-BASED FEATURES

Selina Chu, Shrikanth Narayanan and C.-C. Jay Kuo

Signal and Image Processing Institute and Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089-2564
{selinach, shri, cckuo}@sipi.usc.edu

ABSTRACT

Defining suitable features for environmental sounds is an important problem in an automatic acoustic scene recognition system. As with most pattern recognition problems, extracting the right feature set is the key to effective performance. A variety of features have been proposed for audio recognition, but the vast majority of the past work utilizes features that are well-known for structured data, such as speech and music, and assumes this association will transfer naturally well to unstructured sounds. In this paper, we propose a novel method based on matching pursuit (MP) to analyze environment sounds for their feature extraction. The proposed MP-based method utilizes a dictionary from which to select features, resulting in a representation that is flexible, yet intuitive and physically interpretable. We will show that these features are less sensitive to noise and are capable of effectively representing sounds that originate from different sources and different frequency ranges. The MP-based feature can be used to supplement another well-known audio feature, *i.e.* MFCC, to yield higher recognition accuracy for environmental sounds.

Index Terms— Environmental sounds, feature extraction, audio classification, auditory scene recognition, matching pursuit

1. INTRODUCTION

Recognizing the environment from sounds is a basic problem in audio signal processing and has important applications in navigation and assistive robotics and other mobile device-based services. The audio scene denotes a location with different acoustic characteristics such as a coffee shop, park, or quiet hallway. A stream of audio data contains a significant wealth of information, enabling the system to capture a semantically richer environment, on top of what visual information can provide. Moreover, to capture a more complete description of a scene, the fusion of audio and visual information can be advantageous, such as for disambiguation of environment and object types. Audio data could be obtained at any moment when the system is functioning, in spite of challenging external conditions such as poor lighting or visual obstruction, and is relatively cheap to store and compute than visual signals. In order to use any of these capabilities, we first have to determine the current ambient context.

Several features have been used to describe audio signals. The features of choice for most audio recognition systems typically rely on the use of mel-frequency cepstral coefficient (MFCC). The filterbanks for MFCC are based on the human auditory system and have been shown to work particularly well for structured sounds, like speech and music, but their performance degrades in the presence of noise. MFCC features are modeled based on the shape of the overall spectrum, making it more favorable for modeling single sound sources. Environmental sounds, on the other hand, typically contains a large varieties of sounds, including conditions that are characterized by narrow spectral peaks, such as chirpings of insects, which MFCCs are unable to encode effectively. Other commonly-used features for audio signals include LPC (Linear Prediction Coefficients), band energy ratio, frequency roll-off, spectral centroid, spectral bandwidth, spectral asymmetry, spectral flatness, zero-crossing, and energy [1]. Many previous efforts utilize a combination of some, or even all, of these features together, with intentions of representing all aspects of the audio signals. More importantly, the problem of using a large number of features is that there are many potentially irrelevant features that could negatively impact the quality of classification. As the feature dimension increases, data points become more sparse and some features are essentially noise.

Research on unstructured audio scene recognition has received less attention when compared to structured audio analysis such as speech or music. To date, only a few systems have been proposed that investigate modeling using raw environment audio, without pre-extracting specific events or sounds, especially those from produced movies or television tracks. Sound-based situation analysis was investigated in [2] and in the domains of wearables and context-aware applications [3, 4]. Because of randomness, high variance and other difficulties in working with environmental sounds, the recognition rates have been limited especially as the number of targeted classes increases: around 92% for 5 classes [5], 77% for 11 classes [6], and approximately 60% for 13 or more classes [2]. These works utilize MFCCs and other commonly-used features as their feature extraction method.

This paper addresses the recognition of environmental sounds, focusing particularly on feature extraction using the matching pursuit (MP) technique. MP provides a way to extract features that can describe sounds where other audio features (*e.g.*, MFCCs) fail. They are more robust with respect to background noise. The contribution of this paper is the novel use of MP for feature extraction and its application to unstructured audio processing. We investigate a variety of audio features and provide an empirical evaluation on fourteen different types of environmental sounds. We will show that the most commonly-used features do not always work well with environmental sounds. It will be shown that the MP-based feature can be used to supplement another well-known audio feature, *i.e.* MFCC, to yield higher recognition accuracy for environmental sounds.

2. FEATURE EXTRACTION WITH MATCHING PURSUIT

2.1. Matching Pursuit

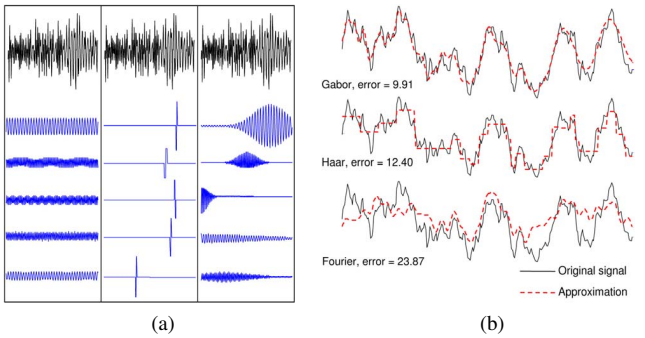
Our goal is to obtain the minimum number of bases to represent a signal, resulting in a sparse and efficient representation. This is an

NP-complete problem. Various adaptive approximation techniques have been proposed in literature, such as the method of frames, basis pursuit, matching pursuit, and orthogonal matching pursuits. All these methods utilize the notion of a dictionary that allows the decomposition of a signal by selecting basis from a given dictionary to find the best basis set. Among these, MP is a more efficient, but greedy, approach. By using a dictionary that consists of a wide variety of basis, MP provides an efficient way of selecting a small basis set that would produce meaningful features as well as a flexible representation. MP is sub-optimal in the sense that it may not achieve the sparsest solution depending on the given dictionary. However, as long as the dictionary is complete, the expansion is guaranteed to converge to a solution where the residual signal has zero energy. Elements in the dictionary are selected based on maximizing the energy removed from the residual signal at each step. Even with a few steps, the algorithm is capable of yielding reasonable approximation using only a few atoms. For further details, we refer to [7].

The MP result relies on the choice of the dictionary. A dictionary is a set of basis (or simply parameterized waveforms) for obtaining a linear combination to produce an approximated representation of the signal. Several dictionaries have been proposed for MP, including frequency dictionaries (*e.g.*, Fourier), time-scale dictionaries (*e.g.*, Haar), time-frequency dictionaries (*e.g.*, Gabor). Most dictionaries are complete or overcomplete. It is important for atoms in the dictionary to be discriminative among themselves; otherwise, similar atoms will compete with each other in the MP process, resulting in low weight value distributed among their coefficients. We will go over the Gabor function in more detail, as it will become more relevant to the details of our feature extraction method.

2.2. Time-Frequency Dictionaries

A combination of both time and frequency functions can be demonstrated in the Gabor dictionary. Gabor functions are sine-modulated Gaussian functions that have been scaled and translated, providing joint time-frequency localization. From [7], the Gabor function is defined as $g_{s,u,\omega,\theta}(t) = K_{s,u,\omega,\theta} g(\frac{t-u}{s}) cos[2\pi\omega(t-u)+\theta]$ with $g(n) = \frac{1}{\sqrt{s}} e^{-\pi t^2}$ and $K_{s,u,\omega,\theta}$ is such that $||g_{s,u,\omega,\theta}||^2 = 1$, where $\gamma = (s, u, \omega, \theta)$ denotes the parameters to the Gabor function, with s, u, ω, and θ corresponding to an atom's position in scale, time, frequency, and phase, respectively. The Gabor dictionary in [7] was implemented with the parameters of atoms chosen from dyadic sequences of integers. N is the size of the atom for which the dictionary is constructed. Scale s, which corresponds to atom's width in time, is derived from the dyadic sequence $s = 2^p$, where $1 \leq p \leq m$ and atom size $N = 2^m$.



(a) (b)

Fig. 1: (a) Decomposition of signals using MP (the first five bases) with dictionaries of: (a) Fourier (left), Haar (middle), and Gabor (right), and (b) approximation (reconstruction) using the first ten coefficients from MP with dictionaries of Gabor(top), Haar (middle) and Fourier (bottom).

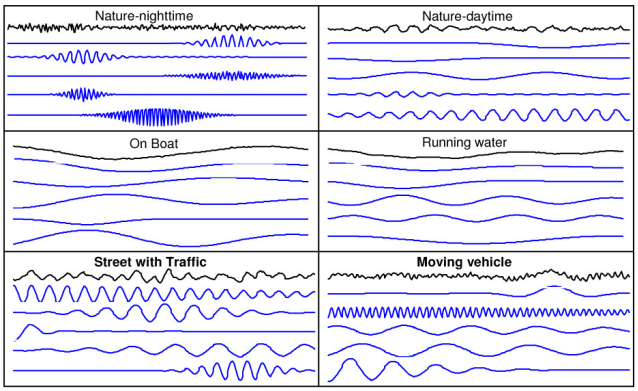


Fig. 2: Decomposition of a signal item from 6 different classes as listed, where the top-most signal is the original, followed by the first five bases.

Examples of MP decomposition using the aforementioned dictionaries are given in Fig. 1(a). Because of the nature of sine and cosine functions, it makes the Fourier dictionary more suitable for high frequency type of data, while the Haar wavelet dictionary is better for more stable, lower frequency-type of signals. The Gabor representation has advantages of these two dictionaries, characterizing the signal in both the time and frequency domain, permitting for a more general representation. Fig. 1(b) demonstrates the effectiveness of reconstructing a signal using only a small subset of coefficients. Gabor atoms result in the lowest reconstruction error, as compared with the Haar or Fourier transforms using the same number of coefficients. Due to the non-homogeneous nature of environmental sounds, using features with these Gabor properties we hypothesize would benefit a classification system. It would provide the ability to be flexible and to capture the time and frequency localization of unstructured sounds, yielding a more general representation.

In the following, we will focus on using the Gabor function. We chose $N = 256$, $m = 8$, $\omega = i^{2.6}$, where $1 \leq i \leq 35$ and distributed over [0, 0.5], $u = \{0, 64, 128, 192\}$ and $\theta = 0$. In other words, a dictionary of 1120 Gabor atoms of length 256 were generated by using atom scales of powers of two from 2 to 256 and translation of a quarter of the atom. We use a logarithmic frequency scale, where the frequency was chosen to be a parabolic function that distributes frequencies in a manner to allow for a higher resolution of histogram bins in lower frequencies and lower resolution in higher frequencies. The reason for a more subtle granularity in the lower frequencies is because more object types occur in these ranges, and we wish to capture the finer differences between them. Since we use discrete atoms, the choice of indices resolution will affect the discriminative power of atoms. The phase was kept constant. We try to keep the dictionary size small, since a large dictionary demands higher complexity. Fig. 2 demonstrates a decomposition of a signal using Gabor atoms. We observe differences in the bases between different types of environmental sounds.

2.3. MP Features

Desirable types of features should be robust, stable, physically interpretable, and sparse in the representation. We will show that using MP will make achieving these requirements possible. One of the key advantages of this representation is the ability to be potentially invariant to background noise and could capture characteristics where MFCCs tend to fail. We use MP as a tool for feature extraction. It provides an approximate representation and reduces the residual energy with as few atoms as possible. We utilize the Gabor function due to its time-frequency localization property, and because it

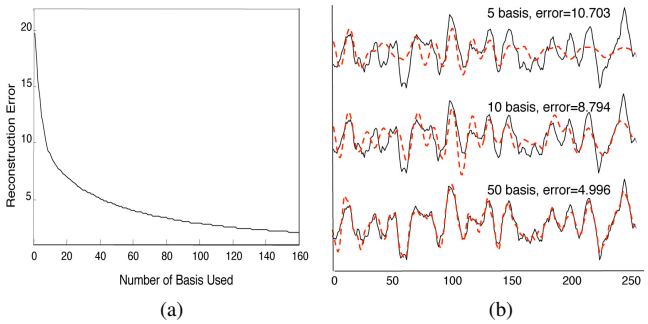


Fig. 3: Examples of reconstruction using MP with the Gabor dicionary by varying the number of atoms (bases).

is an overcomplete dictionary. Since MP selects atoms in the order of eliminating the largest residual energy, it lends itself in providing the most useful basis, even just after a few iterations. We could view each of these bases as a contribution toward the approximation, or the overall reconstruction process. Atoms with the highest residual are assumed to represent the most important and stable signal structures and, therefore, the highest contribution to the decomposition (reconstruction). We demonstrate the effectiveness of using MP with the Gabor function in Fig. 3. As shown in Fig. 3(a), the biggest drop in the residual error happens in the first few terms. We also observe from Fig. 3(b) that using only 10 atoms will provide a reasonable signal; while using the first 50 atoms produces an approximation very similar to the original one.

The feature extraction process is given as follows. For each sampling window, we decompose each segment using MP, stopping after obtaining n atoms. We then decode each atom with its original parameters, obtaining the frequency, scale, and translation positions for each atom. We accumulate all the atoms within a sampling window and take the mean and standard deviation corresponding to the each parameter separately.

Fig. 4 illustrates the classification performance as a function of the number of atoms used in the feature extraction process. It shows first a rise with increasing number of features due to the increased discriminatory power. Then at some point, around 4 or 5 atoms, the performance levels off. With a larger number of features, it increases the complexity, framing it to be more specific to each data item; thereby instances within a class appear more different from one another. With smaller number of features, it allows the data to be represented in a more general way. Therefore, we chose $n = 5$ atoms for our experiments. We use the same process to extract features for both training and test data. The advantage of using just the most prominent atoms makes the features more invariant to background noise variation. The idea is similar to that of choosing the few largest peaks in an STFT frame, which would effectively be the

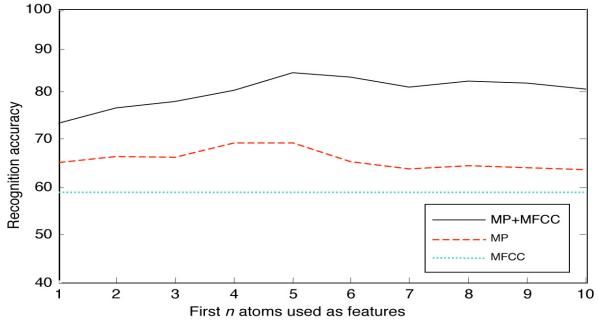


Fig. 4: Comparison of classification rates using the first 10 atoms as features while MFCC is kept constant.

Gabor features with a fixed scale axis and without shift information. The most important information in describing a signal could be found in a few bases with the highest energies, and the process in which MP selects these bases are exactly in the order in which it eliminates the largest residual energy. This means that even the first few atoms found by MP will naturally contain the most information, making them to be significant features.

3. EXPERIMENTAL EVALUATION

We will investigate the performance of a variety of audio features and provide an empirical evaluation on fourteen different types of environmental sounds. In the experiments, we use two different classification methods: K-Nearest Neighbors (KNN) and Gaussian Mixture Models (GMM) [8].

3.1. Experimental Setup

We use recordings of natural (unsynthesized) sound clips obtained from [9, 10]. Our auditory environment types are chosen so that they are made up of non-speech and non-music sounds. It is essentially background noise of a particular environment, composed of many sound events. We do not consider each sound event individually, but as the many properties of each environment. Naturally, there could be infinitely many possible combinations. To simplify the problem, we restrict the number of environment types we examined and enforce each type of sound to be distinctively different from one another and to minimize overlaps, as much as possible. The content of each type should be homogenous enough so that they provide typical representations for each environment. For example, we chose only inside of a restaurant (instead outdoor sidewalk cafes) so that we can only hear the sounds clinking of utensils and plates with incomprehensible people talking in the background, without any traffic or birds sounds. These sound clips were of varying lengths (1-3 minutes long), and were later preprocessed by dividing up into 4-second segments and downsampled to 22050 Hz sampling rate, mono-channel and 16 bits per sample. Features were calculated from a rectangular window of 256 points (11.6 msec) with 50% overlap. Each 4-sec segment makes up an instance for training/testing. We evenly distributed the data by randomly picking 100 4-sec segments to make up each class. All data were normalized to zero mean and unit variance. For KNN, we used the Euclidean distance as the dis-

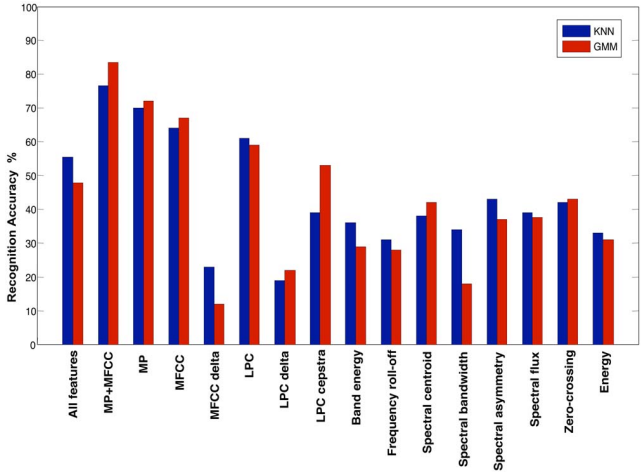


Fig. 5: Overall recognition accuracy comparing MP, MFCC, and other commonly-used features for 14 classes of sounds.

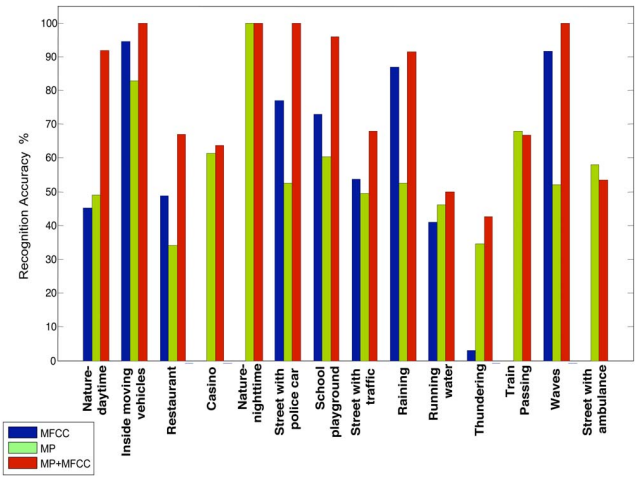


Fig. 6: Overall recognition rate comparing 14 classes using MFCC only, MP only, and MP+MFCC as features. (0% recognition for four classes using MFCC only: *Casino*, *Nature-nighttime*, *Train passing*, and *Street with ambulance*.)

tance measure and the 1-nearest neighbor queries to obtain the results. As for GMM, we calibrated the number of mixtures for each class to produce better overall results. We examine the proposed MP-features from Section 2 and a variety of commonly-used features, which includes MFCC (12), ΔMFCC (12), LPC (12), ΔLPC (12), LPCC(12), band energy ratio, frequency roll-off set at 95%, spectral centroid, spectral bandwidth, spectral asymmetry, spectral flatness, zero-crossing, and energy. The fourteen environment types considered include: *Inside restaurants*, *Playground*, *Street with traffic and pedestrians*, *Train passing*, *Inside moving vehicles*, *Inside casinos*, *Street with police car siren*, *Street with ambulance siren*, *Nature-daytime*, *Nature-nighttime*, *Ocean waves*, *Running water/stream/river*, *Raining/shower*, and *Thundering*.

In the experiments, we utilized separate source files for training and test sets. We kept the 4-sec segments that originated from the same source file separate from one another. Each source file for each environment was obtained at different locations. For instance, the *Street with traffic* class contains four source files which were labeled as taken from different cities. Because of the limited data, we require that each environment contains at least four separate source recordings. Segments from the same source file are considered a set. Therefore, we used three sets for training and one set for testing. The maximum amount of cross-validations we could perform on the data was then limited to the minimum number of source files we have. Therefore, we use a 4-fold cross-validation on the data for the experiments. To keep the source files separated, we did not use leave-one-out cross-validation because it makes the data instances for training and testing to originate from the same source file.

3.2. Experimental Results and Discussion

For the experiment, we performed a 4-fold cross validation for the MP-features and all the commonly-used features individually for feature comparison. In this setup, none of the training and the test items originated from the same source. Since the recordings were taken from a wide variety of locations, the ambient sound might have a very high variance. Results were averaged over 100 trials. Fig. 5 summarizes the findings. MP-features provided an overall classification result of 72.5%, which is slightly higher than MFCC, 70.9%. We ran the same experiments using the combination of all features together, resulting in approximately 55.2% accuracy. The performance produces poorer results than using the 12 MFCCs alone. However, when we combine the MP-features with MFCC by concatenating the two feature vectors together, we were able to achieve an accuracy rate of 83.9% for discriminating fourteen classes, even with separate source files for test and training sets. The reason for this effect could be observed in Fig. 6, where we illustrate, in detail, the recognition accuracy of each class. The figure shows that in cases where MFCC did not perform well, MP features can compliment the other, and vice versa. The clearest example is in the case of the *Nature-night-time* class, which contains many insect sounds of high frequencies. Unlike MFCCs which recognized 0% of this category, MP features were able to capture the characteristics 100%. A possible reason is its ability to capture narrow spectral peaks in high frequency signals. In general, MFCCs tend to operates on the extremes. MFCCs performed better than MP features alone in six classes, but at the same time produced extremely poor results in three other classes, with a recognition rate of 0% for four classes, *Casino*, *Nature-nighttime*, *Train passing*, and *Street with ambulance* and less than 10 percent for *Thundering*. On the other hand, MP features perform better overall, with the lowest being 35% in two classes (*Restaurant* and *Thundering*). Using the combination of both MFCC and MP features, seven classes achieved classification rate of above 90%. MFCC and MP-features provide a complimentary effect for one another, thereby correctly classifying the classes when the features are combined.

Both MFCC and LPC are excellent representations when the source properties are well behaved and consistent (such as in speech and music). Therefore, they might not be well suited for generalization. In a more realistic situation, what we may encounter will be different than what the system is being trained on, such as the focus of this paper. The goal of this work is to find the underlying structure that will allow us make generalization, even when faced with a completely new location, but similar type of environment.

4. CONCLUSION

We proposed a novel feature extraction method that utilizes MP to select a small basis set that would produce meaningful features. More importantly, it is potentially invariant to background noise and could capture characteristics in the signal where MFCC fails. To the best of or knowledge, this was the first paper to propose using MP for feature extraction and has shown to be promising in classifying fourteen different unstructured audio environments, outperforming the state of the art results in comparable experiments. Our work provides competitive performance for the multi-audio category environment recognition by using a comprehensive feature processing approach.

5. REFERENCES

[1] Tong Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, May 2001.

[2] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. on Audio, Speech and Language Processing*, 2006.

[3] A. Waibel, H. Steusloff, R. Stiefelhagen, and the CHIL Project Consortium, "Chil - computers in the human interaction loop," in *Proc. of WIAMIS*, 2004.

[4] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proc. of CARPE*, 2004.

[5] S. Chu, S. Narayanan, C.-C. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. of ICME*, 2006.

[6] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Proc. of ICASSP*, 2005.

[7] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, Dec 1993.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2nd ed. edition, 2001.

[9] "The sound effects library - original series," http://www.sound-ideas.com/bbc.html.

[10] "The freesound project," http://freesound.iua.upf.edu/index.php.