ENVIRONMENTAL SOUND CLASSIFICATION BASED ON FEATURE COLLABORATION

Byeong-jun Han and Eenjun Hwang

School of Electrical Engineering, Korea University, SEOUL, KOREA
{hbj1147, ehwang04}@korea.ac.kr

ABSTRACT

To date, common acoustic features such as MPEG-7 and Fourier/wavelet transform-based features have been frequently used for environmental sound classification. However, these transforms have difficulty dealing with specific properties of environmental sounds, due to their limited scopes. In this paper, we investigate three types of transforms as yet untried for this purpose, and show that they are more effective than traditional features. This result is mainly due to the fact that they have functionalities that were not easily treatable with traditional transforms. Experimental results show that the combination of these features with traditional features can achieve 86.09% of the maximum accuracy in environmental sound classification, compared to 74.35% of the maximum accuracy when confined to traditional features.

Index Terms—Environmental sound recognition, discrete chirplet transform, discrete curvelet transform, discrete Hilbert transform, feature extraction.

1. INTRODUCTION

Due to the limitations of acoustic signal analysis when confined to the time domain, we often transform a signal into a different space. The Fourier transform is a typical methodology used to analyze signals in the frequency domain; however, frequency representation has its own limitations. Frequency analysis usually requires that a signal within a frame is periodic and stable. However, environmental sounds usually show rapid changes and lack stability, and frequency domain-based approaches require extra steps to improve the accuracy of analysis.

In this paper, we first investigate three types of transforms used to overcome the limitations of the time and frequency domain-based approaches. The discrete chirplet transform can deal with incrementally increasing or decreasing pitches while avoiding extra processing. The discrete curvelet transform enables multi-scale analysis and an acoustic textual description of a signal. Finally, the discrete Hilbert transform recognizes various features of signal amplitude such as variations while avoiding loss of information, and hence enables detection of the time and nature of changes.

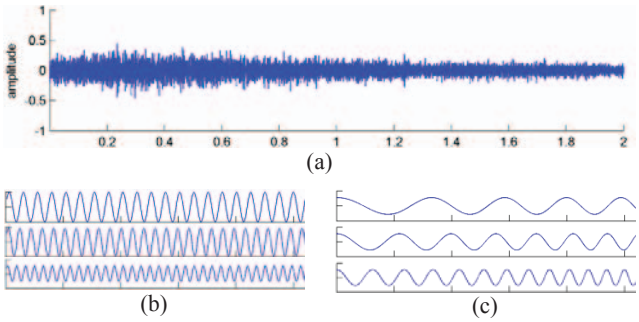

Fig. 1. (a) A sample environmental sound; (b) Wave decompositions with major components at 2552, 2027, and 3244Hz; and (c) Chirp decomposition at 101, 162, and 298 Hz.

2. RELATED WORKS

Eronen *et al.* [1] evaluated acoustic features, classifiers, and feature transformations in the context of sound classification. They extracted various acoustic features, employed k-nearest neighbor (k-NN) and the one-state hidden Markov model (HMM) as classifiers, and applied principal component analysis (PCA) and independent component analysis (ICA) for feature transformation. By organizing the context classes hierarchically, they achieved about 88% of the maximum accuracy for classification. Wang *et al.* [2] applied signal enhancement prior to recognition, and divided the recognition procedure into environmental sound classification and speech recognition. For signal enhancement, they used the perceptual wavelet analysis filterbank and the Karhunen-Loeve transform (KLT). These approaches achieved satisfactory results, when combined with traditional features and classification methodologies.

In this paper, to further enhance the accuracy for environmental sound classification, we investigated the aforementioned three types of transforms and their features. To the best of our knowledge, they are as yet untried, and this is the first attempt to employ them for this purpose.

Chirplet, curvelet, and Hilbert transforms have all received attention, due to their characteristics and specificities. The chirplet transform [3] analyzes signals in time-frequency-chirp rate (TFC) space, and effectively represents increasing parameters such as pitches. Mann and

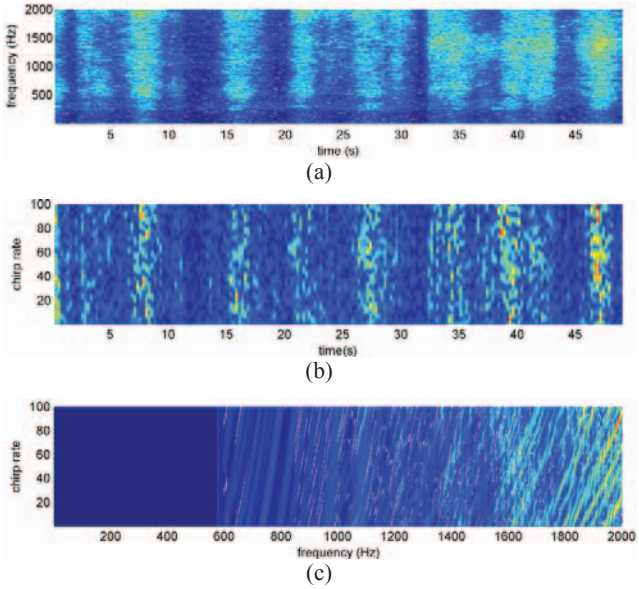

Fig. 2. Representations in (a) Time-frequency (TF), (b) Time-chirp rate (TC) at f=1,000 Hz, and (c) Frequency-chirp (FC) rate planes at t=7.5 s.

Picard first tried to apply the chirplet transform to featureless parameter estimation in image and video processing [4]. The curvelet transform [5] is famous for its excellent features in image processing. It provides analysis on diverse transfigurations and has been employed for diverse purposes [5][6]. The Hilbert transform [7] is another popular method used in various fields such as signal processing, audio source separation [8], and image denoising [9].

3. TRANSFORMS

3.1. Discrete Chirplet Transform

In many cases, an acoustic signal can be handled more effectively by considering it as a superposition of multiple waves, as shown in Fig. 1(a). This allows the analysis of a signal in terms of frequency. However, this assumption has its own limitations: frequency analysis only cannot capture all the properties of a rapidly changing signal. Thus, this approach usually requires additional heuristic steps. In contrast, the chirplet transform can effectively handle rapid signal changes and instabilities. It decomposes the signal into chirps, which consist of an initial frequency and a chirp rate. Each chirp can represent an increasing or decreasing pitch of the signal.

A discrete chirplet transform [10] can be computed as:

$$X_{Ch}[k,\hat{\alpha}] = \sum_{n=0}^{N-1} x[n] e^{-j\pi\hat{\alpha}(n-(N-1)/2)^2} e^{-j2\pi\frac{kn}{N}} \quad (1)$$

with the following condition:

$$\hat{\alpha} = \frac{\alpha}{f_s^{\,2}}; k = 0, \cdots, N-1 \quad (2)$$

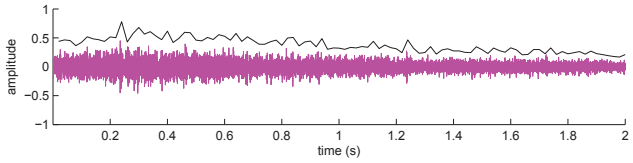

Fig. 3. Functionality of analytic signal.

In Eqs. (1) and (2), x and $X_{Ch}[\cdot]$ are the original signal and its transformation, respectively; α and k are the discrete-time chirp rate and frequency index, respectively; f_s is the sampling frequency; and N is the signal length.

Using Eq. (1), we can obtain both the time-chirp (TC) rate plane and frequency-chirp (FC) plane. Figure 2 shows a comparison between the time-frequency (TF), TC, and FC plane of a sample environmental signal.

3.2. Discrete Curvelet Transform

Usually, the main features of a signal can be represented by its texture. The discrete wavelet transform (DWT) is often employed to describe the texture property of features of sounds and images. However, in many studies the wavelet transform has shown various deficiencies in analyzing the discontinuities of signals.

The curvelet transform [5] is known to have the capability for multi-scale analysis while avoiding discontinuities. Also, it provides various functionalities for analyzing shifted, dilated, and rotated data.

3.3. Discrete Hilbert Transform

The discrete Hilbert transform (DHT) [7] can detect changes within a frame. Basically, it involves a 90 degree phase-shift of the original signal while avoiding spectral analysis, and there is lower information loss than in DFT.

The Hilbert transform is used for constructing a Gabor analytic signal $z[t]$, given by:

$$z[t] = x[t] + jX_H[t] \quad (3)$$

The l^2 norm of $z[t]$ represents the envelope of the original signal x. It cancels the magnitudes of negative frequencies and doubles the magnitudes of positive frequencies. Since it describes both the amplitudes and the differences within the original signal, we used its approximation as a classifier input, as shown in Fig. 3.

4. ENVIRONMENTAL SOUND CLASSIFICATION

This section deals with our proposed environmental sound classification scheme. Figure 4 depicts the overall steps.

4.1. Preprocessing

Since we assumed that the acoustic signal was continuous, our scheme first frames the incoming signal in a predefined unit of time and performs local framing for each frame. The

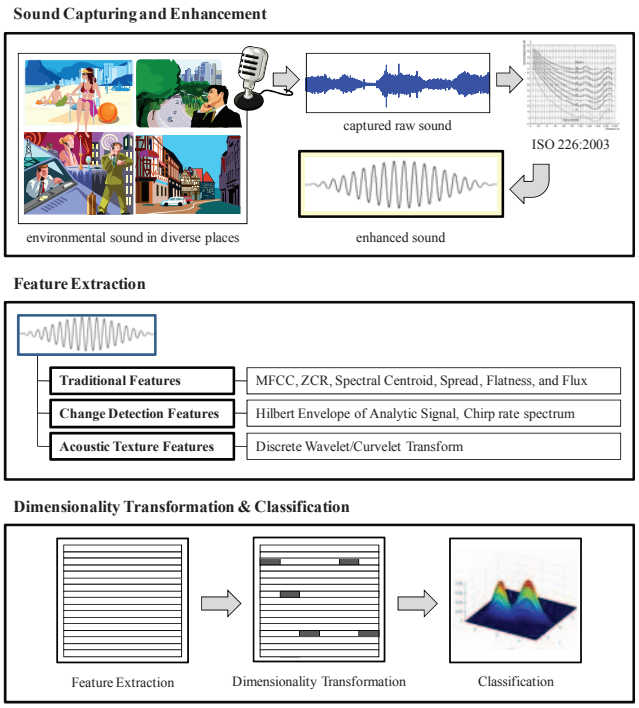

Fig. 4. Environmental sound classification scheme.

local framing divides a frame into several local frames. We then applied equal-loudness level contours [11] to each frame, to ensure that the signal more accurately represented human sound perception, and we eliminated the silence signal from the start and end points of each frame.

4.2. Feature Extraction

For classification purposes, we considered three types of features: traditional features (TFs), change detection features (CDFs), and acoustic texture features (ATFs).

For traditional features, we collected mel-frequency cepstal coefficients (MFCC), zero-crossing rate (ZCR), spectral centroid (SC), spectral spread (SS), spectral flatness (SF), and spectral flux (SFX). More details on these feature extractions can be found in [12].

Change detection features (CDFs), which are applied to each local frame in a frame, consist of a chirp rate spectrum, a Hilbert envelope of the analytic signal, and the local energy. The chirp rate spectrum is obtained by a discrete chirplet transform of the signal. This shows specific frequency changes according to its chirp rate. If the absolute value of the chirp rate at a specific frequency is high, then a chirp at the given rate and frequency has a high impact on the analyzed signal. Moreover, the Hilbert envelope of the analytic signal is computed from the absolute value of $z[t]$ in Eq. (3). Finally, the local energy of each local frame is computed by summing the squared samples of the local frame.

Acoustic texture features (ATFs) represent the DWT and discrete curvelet transform of a frame. Since the curvelet transform is basically designed for a 2-dimensional signal such as an image, we applied it to the magnitude spectrogram of the short-time Fourier transform (STFT) on the frame.

Table 1. Performance evaluation measures.

Measure	Definition	Acronyms
Sensitivity	TP / TP + FN	T = true / F = false
Specificity	TN / TN + FP	P = positive
FPR	1-Specificity	N = negative
Accuracy	(TP + TN) / ALL	ALL = all samples

Table 2. Environmental sound dataset.

Category	#	Global frames	Local frames	Total duration	Avg. duration
C1: On the street	18	167	819	239.21	13.29
C2: On the road	15	141	696	201.18	13.41
C3: Talking	22	199	879	287.44	13.07
C4: Raining	15	129	573	189.23	12.62
C5: Pub/Bar	18	158	712	230.19	12.79
C6: In car	21	181	782	265.67	12.65

4.3. Dimensionality Transformation and Classification

The high dimension of the feature data often causes performance problems, due to the execution time, memory consumption, and the influence of outliers in the features. To mitigate this problem, we used nonnegative matrix factorization (NMF) [13]. We also employed a support vector machine (SVM) [14] as a classifier. Since a SVM provides binary classification, it is difficult to classify diverse categories using only one SVM.. It is well-known that multiclass SVMs such as one-against-all SVMs, pairwise SVMs, and all-at-once SVMs can support multiclass classification functionality [14]. However, since combinations of categories can be used to describe a specific situation, we assigned a separate SVM to each category.

5. RESULTS

This section describes our experimental results. To evaluate the performance of each method, we considered the specificity, sensitivity, false positive rate (FPR), and accuracy rates, as defined in Table 1.

5.1. Setup and Dataset

We used the following parameters for local framing: 5s for framing, 1s for hopping, and 250ms for local framing.

In the experiment, we set the target dimension to 200, in order to reduce the complexity and memory consumption while avoiding significant loss of information.

We considered six different categories of sound data and collected 128 sound files from various sources such as movie clips, audio data on TV variety shows, radio, and YouTube.com. The six categories were: *on the street* (C1), *on the road* (C2), *talking* (C3), *raining* (C4), *pub/bar* (C5),

and *in car* (C6). Table 2 shows the details of our dataset. All the sound files were recorded at 22,050 Hz and 8 bit sampling rates. The total duration of the recorded file was 22 min 92.92 s, and the average duration was 12.96 s. The amplitudes of all data elements were normalized prior to extracting the features. Finally, for each category, we trained the assigned SVM, which was verified using data in other categories as the false data of the SVM.

5.2. Performance Evaluation

For the performance evaluation, we considered four different feature sets and measured their accuracy for the six categories of sound data. Table 3 shows the overall results. The table shows that each feature set was associated with distinct classification accuracy for each category.

The following findings are based on the aforementioned properties of each feature set and the experimental results: The traditional features showed the best and second best performances for C6 and C1, respectively, and had the smallest standard deviation. This implies that TFs can give decent performance in the classification of most environmental sounds. On the other hand, CDFs showed good performance for C5, C1, and C4. This is due to the fact that the discrete chirplet transform and discrete Hilbert transform can handle continuously changing pitches of sounds in these categories. Finally, ATFs showed good performance for C5 and C2. The sounds in these categories generally have various repeating segments such as car sounds, chattering sounds, and cheering. Thus, we can infer that the discrete curvelet transform and DWT successfully extracted the repeating pattern from the sounds.

Figure 5 shows that our dataset is valid for the purposes of environmental sound classification. Figure 5 depicts the FPR-sensitivity plane representation and consists of FPR-sensitivity pairs and the borderline. If the point is below the borderline, then it is not statistically applicable. As shown, our FPR-sensitivity pairs are all above the borderline.

Table 3. Performance of feature sets and overall results.

Category	Accuracy			
	TF	CDF	ATF	Overall
C1	73.13%	74.61%	72.35%	84.26%
C2	71.91%	73.39%	73.04%	83.57%
C3	72.70%	73.65%	72.43%	82.78%
C4	71.39%	74.52%	71.39%	81.39%
C5	72.52%	**75.74%**	**74.26%**	**86.09%**
C6	**74.35%**	72.87%	69.91%	83.48%
Average	72.67%	74.13%	72.23%	83.60%
S.D.	0.01027	0.01033	0.01478	0.01563

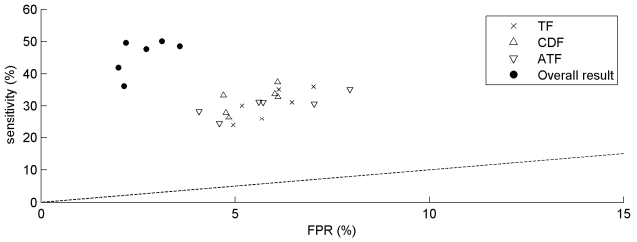

Fig. 5. FPR sensitivity plane representation of overall results.

6. CONCLUSION

In this paper, we investigated three types of transforms and compared their performance with TFs for environmental sound classification. The employed transforms are known to provide various functionalities such as sound pitch changes, texture, and variation analysis. In the experiment, we evaluated these transforms and TFs together for six different categories of environmental sounds, and showed that CDFs and ATFs are more effective than TFs for classification. Furthermore, when combined with TFs, they achieved the maximum accuracy.

REFERENCES

[1] 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," IEEE Transactions on Audio, Speech, and Language, Processing, vol.14, no.1, pp.321-329, Jan. 2006.

[2] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," IEEE Transactions on Automation Science and Engineering, vol.5, no.1, pp.25-31, Jan. 2008.

[3] S. Mann, S. Haykin, "The chirplet transform: physical considerations," IEEE Transactions on Signal Processing, vol.43, no.11, pp.2745-2761, Nov. 1995.

[4] S. Mann and R.W. Picard, "Video orbits of the projective group: a simple approach to featureless estimation of parameters," IEEE Transactions on Image Processing, vol.6, no.9, pp.1281-1295, Sept. 1997.

[5] J.L. Starck, E.J. Candes, and D.L. Donoho, "The curvelet transform for image denoising," IEEE Transactions on Image Processing, vol.11, pp.670-684, 2000.

[6] B. Zhang, *et al.*, "Wavelets, ridgelets, and curvelets for Poisson noise removal," IEEE Transactions on Image Processing, vol.17, no.7, Jul. 2008.

[7] S.L. Hahn, "Hilbert transforms in signal processing," Boston, MA: Artech House, 1996.

[8] F. Gianfelici, *et al.*, "Multicomponent AM-FM representations: an asymptotically exact approach," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.3, pp.823-837, Mar. 2007.

[9] S. C. Olhede, "Hyperanalytic denoising," IEEE Transactions on Image Processing, vol.16, no.6, pp.1522-1537, Jun. 2007.

[10] L. Weruaga and M. Képesi, "EM-driven stereo-like Gaussian chirplet mixture estimation," Proceedings of IEEE ICASSP 2005, vol.IV, pp.473-476, 2005.

[11] "Acoustics—Normal equal-loudness level contours", ISO 226:2003.

[12] Anssi Klapuri and Manuel Davy, "Signal processing methods for music transcription," Springer, 2006.

[13] D.D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401 no.6755, pp.788-791, 1999.

[14] Abe Shigeo, "Support vector machines for pattern classification," Springer, 2005.