



Effects of modelling within- and between-frame temporal variations in power spectra on non-verbal sound recognition

Nobuhide Yamakawa¹, Tetsuro Kitahara², Toru Takahashi¹,
Kazunori Komatani¹, Tetsuya Ogata¹, Hiroshi G. Okuno¹

¹Graduate School of Informatics, Kyoto University,
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

² College of Humanities and Sciences, Nihon University,
3-25-40, Sakurajousui, Setagaya-ku, Tokyo 156-8550, Japan

nyamakaw[at]kuis.kyoto-u.ac.jp, kitahara[at]cssa.chs.nihon-u.ac.jp

Abstract

Research on environmental sound recognition has not shown great development in comparison with that on speech and musical signals. One of the reasons is that the sound category of environmental sounds covers a broad range of acoustical natures. We classified them in order to explore suitable recognition techniques for each characteristic. We focus on impulsive sounds and their non-stationary feature within and between analytic frames. We used matching-pursuit as a framework to use wavelet analysis for extracting temporal variation of audio features inside a frame. We also investigated the validity of modeling decaying patterns of sounds using Hidden Markov models. Experimental results indicate that sounds with multiple impulsive signals are recognized better by using time-frequency analyzing bases than by frequency domain analysis. Classification of sound classes with a long and clear decaying pattern improves when HMMs with multiple number of hidden states are applied.

Index Terms: audio signal classification, non-speech sound recognition, environmental sound recognition, time-frequency analysis, Matching-Pursuit

1. Introduction

The sounds we perceive in our daily life convey an enormous amount of information for communication with other people and understanding our environment. Speech signals are obviously indispensable for our daily communication, and music is an essential tool for our cultural activities. Environmental sounds are especially useful for obtaining non-verbal information which convey changes in surroundings. For example, we can predict someone's approach by the sound of his/her footsteps as well as his/her entrance into the room by the banging of a door.

Human beings are capable of discriminating speech, music and environmental sounds. Several pieces of research on neuroscience revealed that each sound category provokes excitation of different combinations of brain regions [1], [2], and [3]. Research on computational analysis of sound signal has also been carried out with the same category criteria as those for human auditory perception. The methods of computational environmental sound recognition, however, they have not extensively been studied. Research on the other sound categories has frequently attracted a large number of researchers.

Since the nature of environmental sounds is diverse and handling them in general is difficult, most of the studies for environmental sound recognition were involved limiting the applications of the system [4], [5], and [6]. To deal with environmental sounds more generally, some applied speech/music recognition techniques onto non-speech or non-musical sounds [7] [8], and the other tried to develop environmental audio "scene" recognition system [9] [10], e.g., a recognition scheme to make mobile robots aware of surrounding information that used audio signals.

Recent research by Cowling [11] and Chu [12] mainly focused on the fact that environmental sounds have a non-stationary nature which cannot be effectively detected by frequency-based feature such as Mel-frequency cepstrum coefficient (MFCC) and Linear predictive coding (LPC) used in speech and music sound recognition. Cowling presented the effectiveness of the use of non-stationary continuous wavelet transform (CWT) together with dynamic time warping (DTW) against environment sounds. Chu also used a wavelet-based technique, Matching-Pursuit (MP), to extract audio features as well as a Gaussian mixture model (GMM) as a classifier. Both concluded that the use of time-frequency audio features improves the recognition of non-stationary environmental sound recognition. However, they only carried out recognition tests with the Gabor (or complex Morlet) wavelet and the effects of using other wavelet bases on non-stationary audio features are still unknown. Additionally, in Cowling's experiment although a GMM was compared with other classification methods such as an artificial neural network (ANN) and K-nearest neighborhood (k-NN), Hidden Markov models (HMMs) were not taken into account because environmental sounds are non-speech and lack the phonetic structure. Since environmental sounds contain a large amount of temporal interdependence among sequential analytic frames, expressing it using HMMs with multiple hidden states can be an effective method.

We focus on the recognition of impulsive environmental sounds, e.g., sounds generated when two hard materials are collided. First, we discuss the recognition performance of the Gabor wavelet (time-frequency domain) together with the Fourier (frequency domain) and the Haar (time domain) using MP so that we can examine the effectiveness of time-frequency analysis on environment sounds. Moreover, the effect of the number of hidden states in HMMs on classification performance for impulsive sounds is also discussed.

2. Characteristics of Environmental Sounds

The characteristics of environmental sounds are fairly diverse and differ from other sound categories such as voice and music. The characteristics of each sound category are listed in Table 1.

Unlike MFCC and LPC, for some classes of environmental sounds, stationarity and clear harmonic structure inside an analysis frame cannot be assumed. For example, impulsive environmental sounds generally have unclear harmonic structure and fast temporal variation inside (intra) or between (inter) analysis windows.

Table 1: Characteristics of sound categories

Acoustical Characteristics	Voice	Music	Environmental Sounds
No. of Classes	No. of Phonemes	No. of tones	Undefined
Length of Window	Short (fixed)	Long (fixed)	Undefined
Length of Shift	Short (fixed)	Long (fixed)	Undefined
Bandwidth	Narrow	Relatively Narrow	Broad Narrow
Harmonics	Clear	Clear	Clear Unclear
Stationarity	Stationary	Stationary (except percussions)	Non-stationary Stationary
Repetitive Structure	Weak	Weak	Strong Weak
.	.	.	.

2.1. Temporal Variation of Sound Signal: Intra-Window

A portion of a clink of coins has been analyzed using short-time fourier transform (STFT) with different lengths and shifts rate of analysis windows, and their spectrograms are plotted in Figures 2 and 1.

As shown in Figure 1, the signal is analyzed by a narrower window (8 msec) and shorter shift rate (4 msec), thus the spectrogram has high resolution in the time domain. The signal spectra noticeably varies with time. On the other hand, as shown in Figure 2, where the window width and shift rate are 25 and 10 msec, the temporal variance of the spectrum observed in Figure 1 appears to be smoothed over windows. If the non-stationary components characterize a clink sound, rougher time analysis is not suitable for a sound source recognition task. The window parameters in Figure 2 are equal to the ones typically used in MFCC. This is the reason why speech/music recognition techniques are not always applicable to environmental sounds. Moreover, time analysis with a shorter window length and shift rate is applicable for detecting temporally variant audio signals; however, due to the uncertainty principle of Fourier transform, a short analysis window width coarsens frequency resolution, and this leads to degraded information in the frequency domain.

To overcome the time and frequency resolution trade-off, we used wavelet analysis to extract a non-stationary audio feature since wavelet bases detect the signal localized on a narrow area both in the time and frequency domains. We extracted feature by using wavelet bases with MP and explain the process in Section 3.

2.2. Temporal Variation of Sound Signal: Inter-Window

Non-stationarity of impulsive sounds exists not only inside analysis windows but also between sequential windows. In

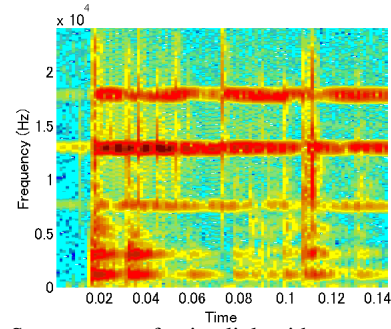


Figure 1: Spectrogram of coin clink with narrower window (8 msec) and shorter shift rate (4 msec)

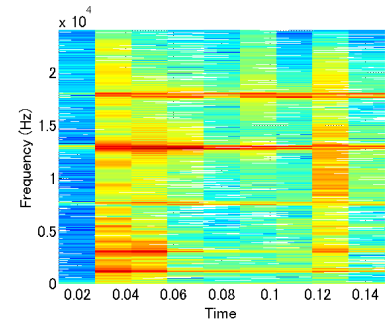


Figure 2: Spectrogram of coin clink with larger STFT rate (length = 25 msec, shift = 10 msec)

other words, extracted feature vectors may significantly vary with the transition of windows.

As well as time variance of signal spectra, Figure 2 illustrates time variance of feature vectors, hence hidden states of each HMM. The transition of states due to decay of the signal is not stationary and it can be a global pattern of the signal. To make HMMs to classify the unique decay patterns, the classifier needs to have multiple number of hidden states.

The suitable number of hidden states for the impulsive sound classification task will be investigated in the experiment.

3. AudioFeature Extraction With Matching-Pursuit

Matching-pursuit is a technique for sparse signal decomposition using an over-complete base dictionary. The main advantage of MP is that it is used to analyze a signal using an arbitrary dictionary consisting of a wide variety of bases. For example, a dictionary can contain multiple types of wavelet bases, and base widths, shift rates in the signal, and many other analytic parameters that can be specified by users. This feature of MP means that the analysis results strongly depend on the design of the dictionary. In addition, if the Fourier base is selected as an analyzing base, the MP algorithm essentially behaves as a Fourier series expansion; thus, the Fourier Transform can also be modeled with this algorithm.

Here, MP is an algorithm for approximating a given signal s as a linear sum of m bases $\phi_{\gamma_1} \dots \phi_{\gamma_m}$ where m is an arbitrary number:

$$s = \sum_{i=1}^m \alpha_{\gamma_i} \phi_{\gamma_i} + R^{(m)}. \quad (1)$$

Note that R represents a residual signal and m bases are selected from the dictionary $D = \{\phi_{\gamma_1} \dots \phi_{\gamma_{m'}}\}$ which contains $m' (\geq m)$ bases.

The base selection procedure is as follows:

1. Calculate correlation of s and all the bases in D
2. Extract the base with the maximum correlation ϕ_{γ_1} from s together with the correlation coefficient α_{γ_1}
3. Apply the same procedure as 1. to $R^{(1)} = s - \alpha_{\gamma_1} \phi_{\gamma_1}$, and $\alpha_{\gamma_2} \phi_{\gamma_2}$ is obtained
4. Iterate the procedure above until the required number, m , of bases are decomposed.

Each base may have parameters such as base width, frequency, amplitude and so on, depending on the type of base. Accordingly, these parameters can be used as feature vectors. With the MP algorithm, m bases from the highest energy can be extracted in order from s in theory and computational complexity increases linearly with respect to m . Therefore, by effectively designing the dictionary and if the meaningful features can be gained with small a m , computational complexity does not noticeably increase.

On the basis of these advantages, we conclude that MP is a suitable framework to implement audio feature extraction using wavelet bases. For more theoretical detail, the reader should refer to Mallat et al.'s work [13].

4. Experimental Evaluation

We investigated the validity of applying time-frequency analysis using the Gabor wavelet and decaying pattern classification using HMMs to impulsive environment sounds by evaluating its sound-source recognition performance for 12 environmental sound classes that were both stationary and non-stationary. To justify the effectiveness of time-frequency analysis against time and frequency only ones on this task, the Fourier and the Haar wavelet bases were used for the comparison of recognition performance. The effect of HMMs with multiple hidden states on classification accuracy was investigated by observing the change in recognition rate with the number of the states.

4.1. Experimental Setup

The representation of three wavelet bases, the Haar, Fourier, and Gabor are illustrated in Table 2.



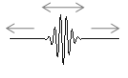
As Haar mother wavelet base can only detect amplitude of a signal sample, we used it as a model of time-domain analysis. The Fourier base, which covers the whole analysis window, was used to simulate the Fourier transform of the whole window, so that we can conduct frequency-domain analysis of each analysis frame. The Gabor wavelet base was used to extract non-stationarity by scaling and shifting inside the window. The base is defined as sine-modulated Gaussian functions with the capability of being scaled and shifted; thus, they can detect time-frequency localization of non-stationary signals. For the Gabor base a feature vector contains center frequency and base width, whereas only the width or the frequency values are required for vectors of Haar and Fourier base respectively. Accordingly, for the Fourier and Haar base, the size of a feature vector is equal to the number of bases extracted using MP and, for the Gabor base, the number needs to be doubled.

In this experiment, Gabor bases were designed to have base width varying from 2 to 1024 samples and the width of the Fourier bases were fixed to the length of a window. Features were extracted for each unit window, and three window-lengths,

25, 50, and 100 msec with a 40% shift rate, were prepared so that we could investigate the effect of the amount of non-stationarity on identification performance for all bases. These processes were implemented using the Matchin-Pursuit Toolkit (MPTK) [14].

We used HMMs with 1, 3, 6, 9, and 12 hidden states in the models of signal spectral evolution. All HMMs were mixtures of 16 Gaussian distributions and had a left-to-right transition rule and a general covariance matrix. The Hidden Markov-Model Toolkit (HTK) [15] was used to define HMMs and perform classification.

Table 2: Haar, Fourier and Gabor wavelet bases

	Time	Frequency	Time-Frequency
MP Analysis Base	 Haar	 Fourier	 Gabor

Sound source identification was carried out using the audio files obtained from Real World Computing Partnership's (RWCP's) sound scene database in a real acoustical environment [16]. The audio files used in this experiment had the length of time no longer than 1 sec and consisted of 12 classes including collision sounds of a piece of wood, a book, a metal plate, a glass cup, coins, hands clapping, dices, a drum, a lock of a door and, as sustaining sounds, particle droppings, a spray, and a phone beeping. They were monaurally recorded under anechoic conditions and sampled at 16 kHz with 16 bit depth. Each sound class had 100 files and consisted of sounds from the same sound source but with slightly different recordings. The first nine classes were categorized as relatively non-stationary and fast-decaying sounds. The rest of the classes were non-decaying signals which were expected to have a large amount of stationarity. The identification rates for each sound class were evaluated using a 10-fold cross validation with 1080 train data and 120 test data.

4.2. Results and Discussion

4.2.1. Effects of MP/Gabor and HMM for non-stationary signal

The overall recognition results using 64 MP Gabor and the Fourier bases with respect to increase of window width and shift rate is shown in Figure 3. The result of Haar wavelet base is removed because its recognition rate was 23% over the all window conditions. The number of hidden states of HMMs is chosen to one so that the effect of modeling by multiple hidden states is eliminated. The recognition rate of the Gabor wavelet bases do not decrease when window width becomes wider, but the performance of the Fourier bases degrade when the window widens, i.e., non-stationarity becomes greater inside each window. The performance difference between the two is 8.7% at 25 msec and becomes 19.7% at 200 msec. The results indicate that the Gabor wavelet base can extract meaningful non-stationary features while the Fourier base cannot.

The relationship between overall recognition performance of each MP analysis base and number of hidden states used for classification is shown in Figure 4. In this plot, the window length is fixed to 50 msec. Again, the result of Haar wavelet base is removed for the reason above. When the number of

states is one, i.e., the classifier is a GMM, recognition performance is the worst on both bases. As the number of states increases, the performance improves until it reaches around six for both bases.

Improvement in recognition rate is more significant on the Fourier base and the performance becomes almost equal to that of the Gabor at 12 states. The results show that using multiple number of hidden states and modeling decay patterns of impulsive sounds enhances the recognition performance on both bases, in particular, on the Fourier base.

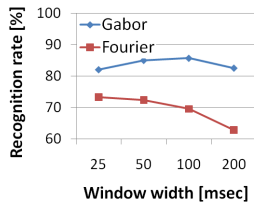


Figure 3: Recognition rate of 64 Gabor and Fourier bases with increasing width of window.

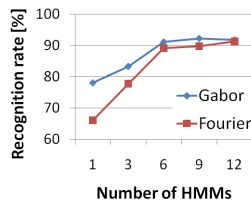


Figure 4: Recognition rate of 16 Gabor and Fourier bases with increasing # of HMMs.

4.2.2. Identification rate for each class

The recognition rate of each sound class of each MP analysis base is illustrated in Figure 5. The number of the extracted MP base is 16, the window width is 50 msec, and the number of hidden states of HMMs is six. The most significant tendency is that the Gabor wavelet bases outperform the Fourier bases when the sound class contains multiple sound events in short time (more than window length), i.e., a sound signal contains multiple sound occurrence, such as dices, a cup, and coins. The Fourier base shows the best performance on the sound classes with single sound event such as wood, metal, and claps. Both perform well on the sustaining sound classes, particles, spray, and phone. For the Fourier on the spray class, the performance improves significantly when the number of MP base extraction becomes large.

The results indicate that the Gabor can detect the non-stationary features that the Fourier base smooths out over an analysis window.

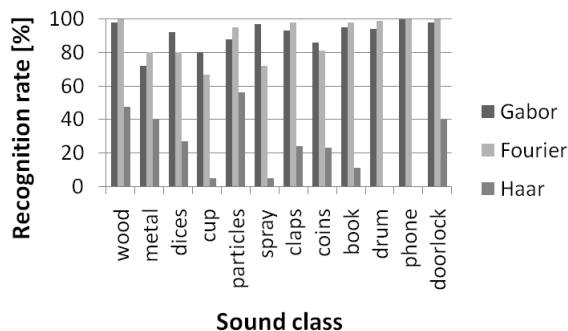


Figure 5: Recognition rate of each sound class where No. of MP base is 16, window width is 50 msec, and # of hidden states is 6.

5. Conclusions

We first explained how environmental sounds differ from speech and music sound and the existence of non-stationarity inside and between (an) analysis window(s) in impulsive environmental sounds.

A non-stationary signal, which is difficult to analyze using Fourier transform-based techniques, was detected using MP with the Gabor wavelet bases. Moreover, non-stationarity between analyzing windows, e.g., signal decay, can be modeled using HMMs and thus improve sound-source identification.

6. References

- [1] Taniwaki, T., Tagawa, K., Sato, F. and Iino, K., "Auditory agnosia restricted to environmental sounds following cortical deafness and generalized auditory agnosia," *Clinical Neurology and Neurosurgery*, 102(3):156-162, Sep 2000.
- [2] Patel, D.A., "Language, music, syntax and the brain," *Nature Neuroscience*, 6, 674-681, 2003. Elsevier, 1991.
- [3] Lewis, J.W., Wightman, F.L., Breczynski, J.A., Phinney, R.E., Binder, J.R. and DeYoe, E.A., "Human Brain Regions Involved in Recognizing Environmental Sounds," *Cerebral Cortex*, 14(9):1008-1021, Sep 2004.
- [4] Jahns, G., Kowalczyk, W. and Walter, K., "Sound analysis to recognize individuals and animal conditions," *Proceedings of VIII CIGR Congress on Agricultural Engineering:1-8*, 1998.
- [5] Ashiya, T., Nakagawa, M., "A Proposal of a Recognition System for the Species of Birds Receiving Birdcalls—An Application of Recognition Systems for Environmental Sound—," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, E76-A(10):1858-1860, Oct 1993.
- [6] Exadaktylosa, V., Silvab, M., Aertsb, J.M., Taylor, C.J., Berckmansb, D., "Real-time recognition of sick pig cough sounds", *Computers and Electronics in Agriculture*, 63:207-214, 2008.
- [7] Goldhor, R.S., "Recognition of environmental sounds," *Proceedings of ICASSP, NY, USA*, 1:149-152, 1993.
- [8] Cowlng, M., Sitte, R., "Recognition of environmental sounds using speech recognition techniques," *Advanced Signal Processing for Communication Systems*, 703:31-46, 2002.
- [9] Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J., "Audio-based context recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, 14(1):321-329, Jan. 2006.
- [10] Peltonen, V. and Tuomi, J. and Klapuri, A. and Huopaniemi, J. and Sorsa, T., "Computational auditory scene recognition," *Proceedings of ICASSP, 2*, 2002.
- [11] Cowlng, M. and Sitte, R., "Comparison of techniques for environmental sound recognition," *Pattern Recog Letters*, 24(15):2895-2907, 2003.
- [12] Chu, S. and Narayanan, S. and Kuo, C.C.J., "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Trans. Audio, Speech, Lang Process.*, 17(6):1142-1158, Aug 2009.
- [13] Mallat, S., Zhang, Z., "Matching pursuits with time-frequency dictionaries", *IEEE Trans. on Signal Processing*, 41(12), 1993.
- [14] Krstulovic, S., Gribonval, R., "MPTK: Matching Pursuit made Tractable", *Proceedings of ICASSP, Vol. 3*, Toulouse, France, 2006.
- [15] Young, S., Young, S., "The HTK hidden Markov model toolkit: Design and philosophy", *Entropic Cambridge Research Laboratory, Ltd*, 2, 1994, 2.44.
- [16] Real World Computing Partnership, "RWCP Sound Scene Database", <http://tosa.mri.co.jp/sounddb/index.htm>