ORIGINAL RESEARCH

Noise-Robust environmental sound classification method based on combination of ICA and MP features

Reona Mogi, Hiroyuki Kasai

The University of Electro-Communications,Tokyo, Japan

Correspondence: Hiroyuki Kasai. Address: 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan. Telephone: 81-42-443-5670. Email: kasai@is.uec.ac.jp.

Received: June 30, 2012 **Accepted:** October 26, 2012 **Online Published:** December 24, 2012
DOI: 10.5430/air.v2n1p107 **URL:** http://dx.doi.org/10.5430/air.v2n1p107

Abstract

This paper presents an environmental sound classification method that is noise-robust against sounds recorded by mobile devices, and presents evaluation of its performance. This method is specifically designed to recognize higher semantics of context from environmental sound. Conventionally, sound classifications have used acoustic features in the frequency domain extracted from sound data using signal processing techniques. Although the most popular feature is Mel-frequency Cepstral Coefficients (MFCC), MFCC is inappropriate for mixture sound with noise. Independent Component Analysis (ICA) can extract sound characteristics even when the source is corrupted by noise because components within the source are assumed to be independent. In recent years, Matching Pursuit (MP) has been addressed to extract time-domain features. It has been applied to various applications. The feature is effective for recognizing and classifying environmental sounds that include time-variant sound such as birdsongs, alarms, and vehicle sounds. In this way, some innovative techniques have been proposed to recognize and classify environmental sounds recorded on mobile devices. However, we have not yet obtained a decisive method to attain a higher recognition and classification rate against environmental sounds with various noises such as unintended sounds and white noise. To address this problem, we propose a noise-robust classification method using a combination of Independent Component Analysis (ICA) and MP. It is possible to reduce noise effects for feature extraction. From performance evaluations, we confirmed that the proposed method can provide about 8% better classification than that of MFCC feature extraction.

Key words

Environmental sound, Sound classification, Feature extraction, Independent component analysis, Matching pursuit

1 Introduction

Mobile devices, including smart phones, have become extremely popular. They are ubiquitous tools in our daily life, used for communication and computing. Furthermore, they have come to play an important role in the sensing and monitoring of various physical data using their equipped sensors such as Global Positioning Systems (GPS) and acceleration sensors. Actually, many papers have reported automatic monitoring systems of physical data using mobile devices. For instance, Tamminen et al. [1] recognized user actions using acceleration. Roggen et al. [2] implemented a system to recognize both surrounding and user situations. Gyorbiro et al. [3] conducted recognition of the user action situation combining body sensors and mobile devices.

Here, turning our eyes to sensory sound data as input data, we can readily perceive many noteworthy features of sound. The first feature is that, sound can provide a tremendous amount of information to understand the context and circumstances of the real world. A captured picture and video might include surrounding scenery and objects such as trees, buildings, and tables. However, it is restricted to obtaining and understanding certain higher semantics of the surrounding context because they can describe only visible information. A captured audio segment can provide invisible information related to our surroundings. They are, for example, “this place is crowded with many people,” “wind is very strong,” or “there are many insects here.” Consequently, recognizing such sounds can engender a higher level of context understanding. The second feature is that, although we often do not realize it, several sounds exist around us at any time and any place. Sound can be retrieved automatically through a microphone on devices. Therefore, it is much easier to collect surrounding sound data continuously without forcing users to conduct special actions.

From these viewpoints, we specifically aim at examining sound-based context recognition. The sounds that we describe in this paper are those that widely exist in the real world around us. Such sounds include birdsongs, chirping of insects, school chimes, traffic noises from cars or motorcycles, footsteps and announcements at a train station, sounds of children’s laughter, and others. For discussion, we define such sounds as “environmental sound.” Some earlier reports [4, 5] specifically describe examination of these sounds, and describe high classification rates in various environments. These proposals used Mel Frequency Cepstral Coefficient (MFCC) as a mode of feature extraction. MFCC is a traditional frequency-domain feature value. However, MFCC presents some drawbacks. For example, it is difficult to extract correct feature values from sound sources that include noise. This problem leads to markedly lower recognition and classification ability when attempting to recognize mixture sounds such as environmental sounds. One means to improve the recognition is to use a stereo microphone as conventional studies have done. This measure might increase the recognition capability because stereo sound has more accurate data than monaural microphones can provide. It is nevertheless unrealistic to use stereo microphones because almost all mobile devices have no stereo microphones but instead have monaural microphones.

Based on the background described above, to achieve a higher classification rate of the monaural sound, this paper presents a proposal of a noise-robust sound classification method. More specifically, we propose a new method to improve the environmental sound classification rate by combining Independent Component Analysis (ICA) [6] and Matching Pursuit (MP) [7, 8]. ICA is known as a method of blind source separation. We believe that ICA can extract sound characteristics even when the source is corrupted by noise because components within the source are assumed to be independent. Additionally, different from MFCC, where extraction is done in the frequency domain, MP is a method to extract sound characteristics in the time-domain. MP is more robust to time-variant sounds.

Consequently, this paper therefore investigates the effectiveness of the combined method using simulation experiments with actual recorded environmental sounds. We believe that the classification rate can be improved by combining the more robust ICA basis in a frequency-domain and the more robust MP feature in the time-domain. Furthermore, regarding ICA, we calculate a common ICA basis vectors using some of the collected sound data, and extract ICA feature vectors for both training data and test data based on the common ICA basis vectors. In addition to ordinal evaluations for classification rate using pre-defined classes, we analyze the relationship between classification rate and class distance. As far as authors’ knowledge, no report in the literature describes a study that attempted to classify environmental sound using MFCC that is improved by ICA and MP, and the detailed evaluation and analysis for the classification rate.

This paper is structured as follows. Section 2 presents a summary of related works. An overview of environmental sound classification methods is presented in Section 3. Section 4 explains details of the proposed method and the environmental sound classification algorithm. Sections 5 and 6 describe a sound classification experiment and discussion. Finally, we provide conclusions and describe future works in Section 7.

2 Related works

Studies of environmental sound classification are in their infancy. Goldhor [5] reported a method that recognized very short sounds such as a "barking dog sound" that exist in daily life. It used an MFCC value as an acoustic feature, which is, in general, used for speech recognition. The report showed a relation between the classification rate and the number of feature values that are extracted using MFCC, where the classification rate became higher in cases of using 12-16 values for a feature vector. Kraft et al. [9] tried to classify various kitchen sounds for a humanoid robot that understands its surroundings. Eronen et al. [10] emphasized the study of environmental sound in daily life, including sounds such as those of trains and streets. They reported various feature extraction methods and classification results, and also investigated optimal recording times for classification. Results show that sound of 1-3 min duration provided good classification results. Moreover, little difference was found in classification results obtained by humans and machines for recorded sounds. Smith et al. [4] improved the classification rate by adding differential MFCC value of velocity and acceleration.

For a sound classification implementation, Dargie [11] implemented an environmental sound classification system for restricted resource devices such as laptop computers. The sound to be classified was recorded using microphones embedded in laptop computers. The classification performance based on MFCC was high, but the specific sound result was poor. Lu et al. [12] also implemented an environmental sound classification system for mobile devices. They performed coarse classification for input signals such as speech, music, and "Ambient Sound." The "Ambient Sound" indicated unknown-origin sounds such as "driving in city", "standing in elevator" and "crowd of people". However, the system was unable to distinguish whether one is driving in the city or standing in an elevator. The coarse classification results were good, but finer classification (e.g. gender recognition in the speech) yielded lower classification rates. They implemented the system in mobile devices and performed real-time environmental sound classification. A noteworthy feature of this system is that the actual users were able to register unknown sounds with textual labels. One unknown sound was newly registered to the system once certain amounts of it had been collected in the system. Consequently, this system can expand gradually along with the collected sounds. Many efforts have been undertaken to classify environmental sounds. However, these studies did not thoroughly consider the time-directional change of the frequency.

Next, regarding more efficient feature extraction methods for classification, Chu et al. [7] proposed an environmental sound feature extraction method that emphasizes time features instead of frequency features. The method extracted features of time duration and acoustic change using matching pursuit (MP), which performs adaptive approximation of a signal in redundant functions. In fact, MP has been used in widely various applications. The combination method of MFCC and MP increased classification rates against constant sounds and short-duration sounds. Furthermore, independent component analysis (ICA) is a powerful technique that can find a linear representation of non-Gaussian data so that the components are statistically independent [13, 14]. The applications of ICA include widely various fields such as electrocardiograms (ECGs) of pregnant women. Here, ICA can be used to improve MFCC features in many studies such as those in the speech recognition field [15-17]. Huan Zhao et al. [17] proposed improved MFCC feature extraction by implementing ICA instead of Discrete Cosine Transform (DCT). It can provide robust feature extraction against noise.

As explained above, many studies and implementations have been done to date for environmental sound recognition and classification. Some innovative and powerful techniques have been proposed to recognize and classify environmental sounds recorded on mobile devices. However, we have not yet obtained a decisive method to achieve higher recognition and classification rate against environmental sounds with various noises such as unintended sounds and white noise.

3 Overview of adopted method

This section presents a description of an overview of conventional classification methods for environmental sound. Conventional feature extraction methods are first introduced. Then, an outline of classification methods is described.

3.1 Feature extraction methods

MFCC is first introduced, and ICA and MP, which are used for this study, are described in the remainder of this subsection as a feature extraction method.

3.1.1 Mel frequency cepstral coefficients (MFCC)

Known as a frequency-domain feature value, MFCC reflects perceptual characteristics of humans. Its features are widely used for speech recognition because of its effectiveness. MFCC feature extraction is depicted in Figure 1. Input sound data are pre-emphasized. Then they are separated into a short duration of time to be transformed using Fast Fourier Transform (FFT). Here, the Hamming window is applied to each separated data before FFT. Then, after Mel Filter Bank and DCT are applied, we obtain the MFCC feature.

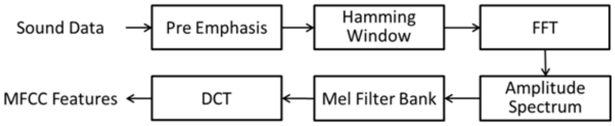

Figure 1. MFCC feature extraction

Here, considering that MFCC is applied to environmental sound classification, it is noteworthy that the environmental sound classification differs from speech recognition because environmental sound consists of multiple sounds over time. In fact, MFCC is inappropriate for robust classification against time-variant sounds.

3.1.2 Independent component analysis (ICA)

ICA, which is known as a blind source separation (BSS) method, analyzes input signals assuming independence of their components [18]. The ICA formula can be written as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where **x** is a mixed signal vector, **s** is an original signal vector, and **A** is the operator matrix (mixed matrix). BSS does not also know mixed matrix **A**. Therefore, it should estimate both the original signal s and mixed matrix **A** from mixed signal **x**. This paper performs ICA by maximizing the non-Gaussianity, a typical ICA calculation method. Non-Gaussianity, the distance from the Gaussian distribution, can be represented using differential entropy. The differential entropy H(**y**) of vector y with density *f* can be represented as written in (2).

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y} \tag{2}$$

A Gaussian distribution is known to indicate the maximum entropy in all distributions of equal variance. In other words, when entropy decreases, the distribution indicates convergence to a specific value. As the scale of non-Gaussianity, negentropy J(**y**) defines normalized differential entropy H(**y**) as presented in (3).

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{3}$$

In that equation, $\mathbf{y}_{gauss}$ is the Gaussian probability vector having the same correlation with $\mathbf{y}$. When maximizing negentropy, entropy becomes minimizing. It can therefore be estimated as matrix **W** to decompose the mixed signal. The outline of gradient method that maximizes negentropy is the following.

i. Calculate the means of data to zero (centering)

ii. Whitening data (Whitening data represents z)

iii. Set initial values both norm w and γ.

iv. Let $\Delta\mathbf{w} \propto \gamma\mathbf{z}g\left(\mathbf{w}^{T}\mathbf{z}\right)$

v. Normalize w $\left(\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|\right)$

vi. If not converged, return to step iv

In that process, g represents the function used to estimate negentropy. We repeat the gradient method for the number of signals, and obtain decomposition matrix **W**. The original signal vector **s** can be represented as shown in equation (4) using the estimated reconstruction matrix **W** ($\approx\mathbf{A}^{-1}$).

$$\mathbf{s} = \mathbf{W}\mathbf{x} \approx \mathbf{A}^{-1}\mathbf{x} \tag{4}$$

3.1.3 Matching pursuit (MP)

MP is used to extract the time-domain feature value. MP is described in detail [7, 8]. First, input signals are compared to waveforms in the dictionary data. Dictionary data D are given as D = $\{\varphi_\gamma : \gamma\in\Gamma\}$, where Γ is the parameter set and where φ is the waveform called an atom. Signal s can be represented as a residual signal and waveform φ_γ, written as shown below.

$$s = \sum_{i=1}^{m} \alpha_{\gamma_i}\phi_{\gamma_i} + R^{(m)} \tag{5}$$

Therein, m signifies the number of the waveform, α_γ stands for the coefficient of the waveform, and $R^{(m)}$ denotes the residual signal. MP approximates signal s using the atom that is some similar atoms in the dictionary. Because the frequency and time position features can be extracted from each waveform datum using an approximate signal s, it is possible to capture the time-domain feature value using MP.

3.2 Classification methods

3.2.1 k-Nearest neighbor (k-NN)

The k-nearest neighbor (k-NN) classification algorithm is used to classify a sound source from the feature value. k-NN is a simple supervised learning algorithm. k-NN classifies new data into assigned classes of k-nearest neighbors. Generally, the distance between input data and training data is calculated using the Euclidean distance, dis(x, y) as written in (7). Here, x and y respectively denote input data and reference data. i and N respectively represent a dimension index and the total number of dimensions.

$$dis(x,y) = \sqrt{\sum_{i}^{N}(x_i - y_i)^2} \tag{7}$$

3.2.2 Gaussian mixture model (GMM)

The Gaussian Mixture Model (GMM) represents data with the linear combination of multiple Gaussian distributions [19]. It is possible to represent data exactly if we can select an optimal mixture number and weighting coefficients, and adjust each distribution's mean and covariance. The linear combination of the K number of Gaussian distributions is represented in equation (8).

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N\left(\mathbf{x} \middle| \boldsymbol{\mu}_k, \Sigma_k\right) \tag{8}$$

Therein, **x** is the data, μ_k denotes the mean, and Σ_k signifies the covariance. Gaussian distribution $N(\mathbf{x}|\ \mu_k, \Sigma_k)$ is a factor of mixture, and has a parameter of each mean and covariance. π_k is a weighting parameter of the k-th Gaussian distribution. We used the Expectation-maximization Algorithm [20] to estimate parameters. A maximum likelihood estimation is used to measure the distance between the input data and each model. We calculate the classification rate by labeling test data as a class name of maximum log likelihood, $L(a)$, as shown in equation (9).

$$L(a) = \log \prod_{x \in X} f\left(a \middle| x\right) \tag{9}$$

Therein, a denotes the test data, and x is the GMM of each class, and $f(a|x)$ represents the probability density function.

4 Proposed sound classification method

We propose a new robust environmental sound classification method that combines the more robust ICA basis in a frequency domain and the more robust MP feature in the time domain. More specifically, after the MFCC feature is extracted as the frequency-domain feature, the ICA feature is extracted instead of DCT as the same as the proposal in [17]. Then, we extract the time-domain feature value using MP, which can extract event sound features even if the time duration of the event sound is short. In the remainder of this section, we describe how to improve the MFCC feature value using ICA and describe procedures of the proposed method.

4.1 ICA basis vector extraction

ICA estimates the decomposition matrix and raw signal, assuming independence of raw signals. Therefore, ICA feature values differ every time because ICA selects a basis vector to be independent. In other words, ICA features for each sound have different basis vectors. Consequently, it is impossible to compare ICA features of different sounds.

Therefore, we calculate a common ICA basis vector, and express each sound feature using the extracted basis vector. A similar strategy is used for facial recognition using ICA [21]. We first prepare each sound class datum for the ICA basis vector. Then, the ICA basis vector is derived using these sound class data, which enables comparison of sound characteristics based on the common ICA basis vector. The point is that the performance of feature extraction can be superior to MFCC because independent characteristics of sounds are selected using ICA, which removes noise effects. Figure 2 shows how to extract the ICA basis vector. Figure 2 differs from Figure 1 in terms of replacing DCT with ICA. Also, we again applied the RASTA Filter, a perceptual filter, to the transformed data. As in MFCC, the Mel Filter Bank creates feature vectors in ICA. If **x** is the resulted sound data after Mel Filter Bank and **s** is the ICA basis vector, then feature vector **A** can be shown as (10).

$$\mathbf{A} = \mathbf{X}^T \mathbf{s} \tag{10}$$

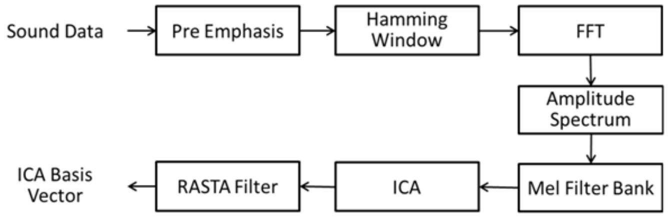

Figure 2. ICA basis vector extraction

We can calculate a transpose matrix because the quantities of ICA basis vector **s** and the dimension of sound data **x** must be equal. Therefore, all training data and test data can be represented based on the ICA basis vector. Then they can be compared. Figure 3 shows the ICA feature extraction procedure using the ICA basis vector. The block within the dotted line calculates the ICA feature vectors based on the ICA basis vector created in Figure 2.

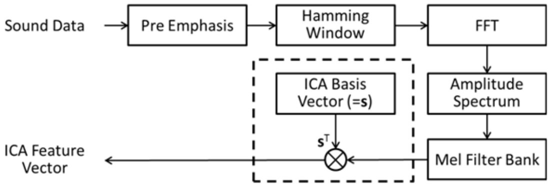

Figure 3. ICA feature extraction using ICA basis vector

As described above, this method first constructs the ICA basis vector from multiple sounds. Then it extracts feature vectors using this ICA basis vector as a supervised learning process. The feature extraction of each sound to be classified can be achieved effectively. This ICA basis vector and its combination with MP that extracts time-domain features can reduce noise and other unstable effects.

4.2 Proposed training and classification procedure

Figure 4 is a comprehensive diagram of the proposed training and classification procedures. The training procedure is depicted on the left side of Figure 4.

i. Extract feature values using MFCC and MP against the input training sound data.

ii. Apply the ICA Basis Vector for MFCC, and calculate the ICA feature.

iii. Combine the ICA feature and MP feature, and store these features in the training database.

The classification procedure from the feature extraction step to the classification step is described below as depicted on the right side of Figure 4.

i. Extract feature values using MFCC and MP against the input test sound data to be classified.

ii. Apply the ICA basis vector for MFCC, and calculate the ICA feature.

iii. Combine the ICA feature and MP feature, and compare the test data with the training data in the training database.

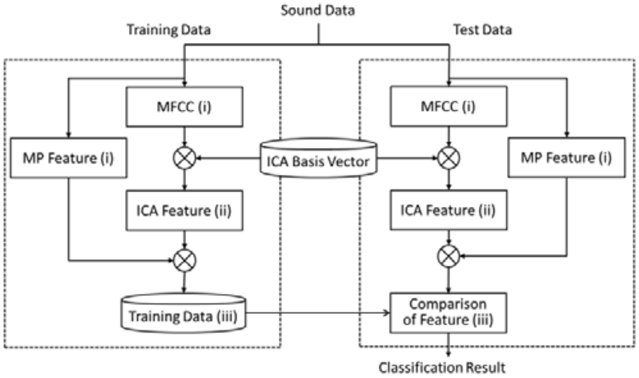

Figure 4. Diagram of the proposed training and classification procedure

5 Experiment setup

5.1 Sound data preparation

Simulation experiments were conducted using real environmental sounds. We used characteristic sounds and mutually similar sounds to confirm whether accurate feature extraction was achieved. These sounds were recorded using iPhones for different dates, times, and locations. The recording sampling rate was 44100 Hz. The recording duration for each sound was 4 s, which is the same as that described in an earlier report [7]. The sound classes to be evaluated were six classes, which were birdsong (bird), station wicket (station), park, railway crossing (crossing), traffic, and downtown, as shown in Table 1. Sounds of station wickets, parks, and downtown include voices of people and children. Sounds of birdsongs and railway crossings are very characteristic. Furthermore, those sounds include noise sounds.

Table 1. Number of each group in each class

Class	ICA basis vector data	Training data	Test data	Total
Alarm	7	56	36	99
Bird	6	89	96	191
Downtown	6	54	43	103
Park	7	180	47	234
Station	16	179	51	246
Traffic	7	174	39	220

As a simulation preparation, we classified each recorded sound manually into the corresponding class described above. Then, we divided them into three groups to calculate the ICA basis vectors, training data, and to evaluate final classification performance. The supervised data of GMM and k-NN are created using the training data grouped in Table 1. The classification is performed based on the probability model or distance. Table 1 presents the assigned number of sounds for each group.

5.2 Number of feature dimensions

We extracted 12 low-dimensional features for the MFCC feature after it extracts 20 dimensional features. This number is generally used in MFCC-based method [7]. Regarding the ICA basis vector, we created the ICA basis vector for each input datum to avoid the bad effects of small number of inappropriate sound data. Therefore, we adopted 49 dimensions, which are the same as the total number of the input data to calculate the ICA basis vector. Consequently, the ICA basis vector has

49 × 12 dimensions. This matrix is used to represent training data and test data. As described in this paper, we created ICA basis vector data of the result of FastICA [22].

Regarding MP, we used Gabor Atom as waveform of dictionary. Gabor Atom follows equation (11).

$$g_{s,u,\omega,\theta}(n)=\frac{K_{s,u,\omega,\theta}}{\sqrt{s}}\exp\left(\frac{-\pi(n-u)^2}{s^2}\right)\cos\left[2\pi\omega(n-u)+\theta\right] \tag{11}$$

Therein, function g(n) corresponds to the waveform φ_γ in section 3.1.3. Indexes of g(n) and K, a constant normalization parameter, are the parameter of Gabor atom, and (s, u, ω, θ) respectively represent scale, time, frequency, and phase. n denotes a signal position at which a selected waveform is applied. This experiment sets parameter θ as zero, and extracted MP features using Gabor dictionary so as in Table 2. Four features, which are an average and standard deviation of frequency ω and scale s of the selected atoms, are used as the MP features, as they were in reference [7].

Consequently, the dimensional quantities of each feature value used in the experiment are presented in Table 3.

Table 2. Dictionary data of Matching Pursuit

Item	Details
Dictionary type	Gabor
Frequency [Hz]	64, 128, 256, 512
Scale [samples]	64, 128, 256, 128
Window shift [samples]	16, 32, 64
Number of atom	48 (= 4×4×3) atoms

Table 3. Number of feature dimensions

Feature	Number of Features
MFCC Feature	12
ICA Feature	49
MP Feature	4
MFCC+MP	16
ICA+MP	53

6 Experimental evaluations

6.1 Preliminary experiments and overall performance evaluation

We classified the data using k-NN with k=3 by assuming that no more than three features exist of the specific interval in the feature space. We derived the distance between the test data and training data based on equation (7). Regarding GMM, before the classification rate evaluation, we conducted a preliminary experiment to ascertain the optimal mixture number of each features. Once we obtain appropriate mixtures for each method, the final comparative experiments will be conducted.

Figure 5 portrays the relation between the mixture number of Gaussian distribution and the classification rate for each feature. For this experiment, it was confirmed that the classification rate was improved using a low mixture number. Classification results of each feature value are the following. For ICA+MP and MFCC+MP, two mixture numbers were obtained for the highest classification rate. For ICA, no significant difference was found when the mixture number changed. MFCC denoted constant classification rate. MP represented the highest classification result when it selected three mixture numbers.

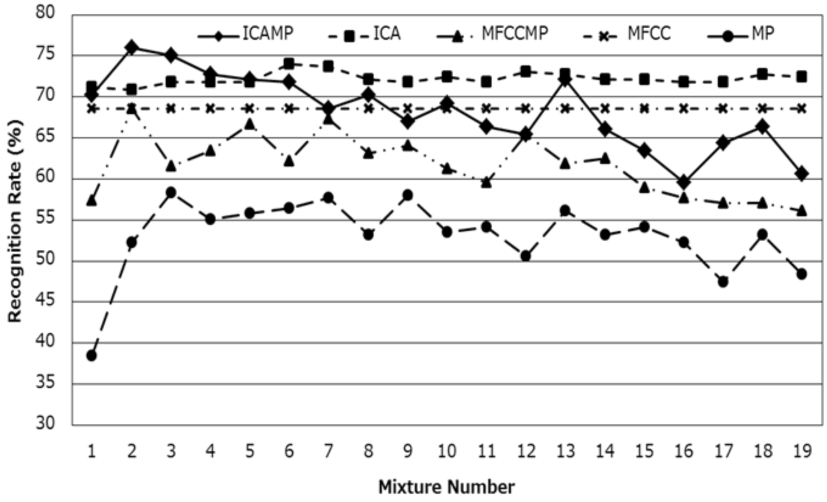

Figure 5. Classification rate by GMM Mixture Number

Based on results for the appropriate mixture number of GMM in the previous experiments, we conducted comparative performance evaluations of the classification rate. Figure 6 presents the classification result when it represented each feature value with optimal mixture numbers. The respective classification rates of ICA+MP and MFCC were 76.0% and 68.6%, respectively, which confirmed that the proposed method can improve the classification rate.

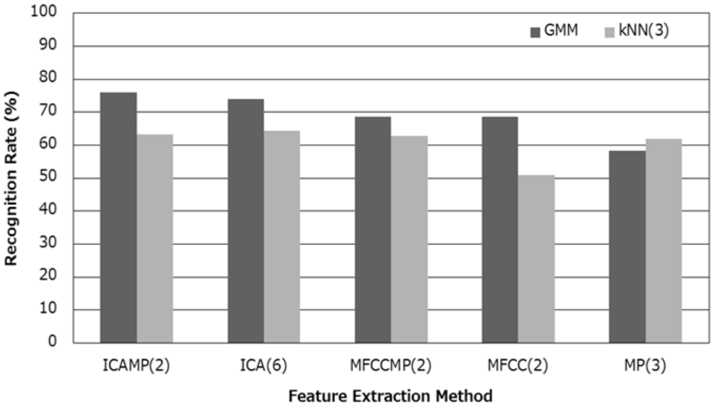

Figure 6. Overall classification rate results

6.2 Evaluation result details in each class

This subsection investigates the results of the previous section in detail. Tables 4-8 respectively present the classification results for ICA+MP，ICA，MFCC+MP, MFCC, and MP. “Target Class” is the class name of the input test data. “Classification Result” indicates the result of recognition relative to the test data.

Table 4 shows classification results obtained using the proposed method. It achieved classification accuracy of greater than 50% relative to each class. Results show that ICA+MP efficiently classifies environmental sounds: the proposed method obtained overall high classification rates. Results confirmed that the sound class of “downtown” was often misclassified for the sound class of “station”. In fact, the two sounds consisted mainly of human speech, and differ only in terms of the electric sounds of the background. Regarding the classification of the amounts of other features, they often show that only one can be confirmed.

Table 4. Classification Result of ICA+MP

ICA+MP(2)		**Classification Result**					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
Target Class	bird	79	1	15	0	5	0
	station	0	**60**	10	0	18	12
	park	0	0	87	0	11	2
	crossing	0	0	3	92	5	0
	traffic	0	5	5	0	**85**	5
	downtown	0	35	7	0	5	**53**

Table 5 shows classification results of ICA feature values. The result in the “bird” class indicates that the ICA feature values have a slightly higher rate than ICA+MP. The “traffic” class sound is often misclassified to the “park” class sound. The recorded sound source showed that the background of the “park” class sound included traffic sound. However the ICA feature values can recognize the “traffic” class better than the MFCC feature value. Apparently, the traffic class was misclassified to the “park” class because traffic sounds exist in the “park” class sound.

Table 5. Classification Result of ICA

ICA(6)		**Classification Result**					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
Target Class	bird	83	0	13	1	3	0
	station	0	**50**	4	0	2	**44**
	park	0	2	94	0	4	0
	crossing	0	0	0	97	3	0
	traffic	3	13	38	0	**31**	15
	downtown	0	**21**	0	0	0	**79**

Table 6 presents the classification result of MFCC+MP. Results show that this feature value substantially improves the classification rate of the “traffic” class. Therefore, it was noted that for the combined MP algorithm, we compensated the part which the frequency feature was unable to recognize. From Table 6, this feature value denotes that many “downtown” class data were misclassified to the “station” class. This result might signify that it is difficult to classify two class sounds.

Table 6. Classification Result of MFCC+MP

MFCC+MP(2)		**Classification Result**					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
Target Class	bird	76	0	17	2	4	1
	station	0	62	8	0	12	18
	park	4	0	83	6	2	4
	crossing	0	3	11	86	0	0
	traffic	8	5	0	3	**85**	0
	downtown	0	**72**	9	2	2	**14**

Table 7 presents classification results of the MFCC feature. It is noteworthy that MFCC features were unable to classify the “traffic” class at all. Many of the “traffic” class were misclassified to the “park” class. The reason is the same as that with ICA. Moreover, the “station” class was misclassified to the “downtown” class. However, the classification result of the “downtown” class obtained 100%. As described above, the “downtown” and “station” classes mainly include human speech. The “station” class was recognized as the “downtown” class because the MFCC feature captured the human speech feature. In connection with the recognition result of MFCC+MP, this fact denotes that MFCC features is difficult to classify sound such as “downtown” class or “station” class. On the other hand, MFCC features can obtain good classification results about the “bird”, “park”, and “crossing” class sounds.

Table 7. Classification Result of MFCC

MFCC		**Classification Result**					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
Target Class	bird	88	0	12	0	0	0
	station	0	**28**	12	0	0	**60**
	park	2	2	96	0	0	0
	crossing	14	0	8	73	5	0
	traffic	5	10	**62**	0	**0**	23
	downtown	0	0	0	0	0	**100**

Finally, Table 8 presents the classification rate attained using MP. When comparing MP with other feature values, many misclassifications occur related to the “station” and “downtown” classes. Because MP is the method of robust time feature extraction, it might be suspected that the reason for misrecognition is only slightly temporal changes in time because these experiment data consist mainly of background sounds. In fact, the “station” and “downtown” class include very similar sounds. Furthermore, these sounds had almost temporal variation.

Table 8. Classification Result of MP

MP(8)		**Classification Result**					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
Target Class	bird	63	8	11	4	14	0
	station	6	**48**	20	8	16	2
	park	4	9	60	15	6	6
	crossing	0	5	5	86	3	0
	traffic	10	5	0	5	**77**	3
	downtown	9	19	49	0	5	**19**

We next describe the consideration for classification results. Although MFCC+MP shows a higher classification rate than MFCC, the classifications of the “station” and “downtown” classes were unstable. Therefore, we confirmed that ICA can improve MFCC feature values. However, ICA was unable to distinguish between the station and downtown sounds only even if ICA is applied to MFCC. Regarding “traffic” sound, MP was higher classification rate than those of other features. Because the target sound contains temporal variation such as sound of passing car, it is believed that MP can recognize “traffic” sound well. Moreover, we believe that ICA+MP correctly classified the mixture sound because ICA+MP obtained frequency and time features of target sounds, such as a park class (park class sound consists of children voice and traffic sound). We infer that ICA+MP provides more robust classification than other features because ICA and MP compensate their mutual disadvantages. However, the proposed method indicated slightly lower performance in the station and downtown sounds. Future work is necessary to provide more robust recognition to identify similar sounds.

6.3 Relation of the classification rate and class distance

We derived the distance between class centers of gravity from training data and compared them with classification results. This investigation can indicate the positional relation of each class sound. If all interclass distances are uniformly the same without bias, its discrimination capability is higher. The interclass distance was calculated from the Euclidean distance.

Table 9 shows the class distance of ICA+MP and MFCC. ICA+MP are the right area of diagonal, and MFCC is the left area. The proposed method (ICA+MP) has some distance related to each class. Therefore, it seems that test data were identified correctly. Because the distance between classes is great, the possibility of misclassification is low. However, the “station” class exists nearby of the “downtown” class. Therefore, results show that the two class sounds might be mutually misclassified. We investigated the distance between classes by MFCC feature. Each class distance of MFCC feature is more unstable than the proposed method. For example, the “bird” and “station” classes are distant from each other. However, the “traffic” class is very close to the “park” class. Actually, Table 9 shows that the misclassified classes are mutually close.

Table 9. Class Distance of ICA+MP and MFCC

Class Distance		ICA+MP					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
MFCC	bird	-	1.55	1.22	1.69	0.96	1.62
	station	**3.72**	-	0.67	1.37	0.78	**0.17**
	park	2.04	1.67	-	1.20	0.91	0.71
	crossing	0.89	3.16	1.58	-	1.45	1.41
	traffic	2.14	1.57	**0.15**	1.63	-	0.87
	downtown	3.89	0.17	1.85	3.33	1.75	-

Table 10 shows the class distance of ICA and MP: ICA is the area right of the diagonal; MP is the left area. The distance between classes obtained using ICA is more stable than MFCC. Moreover, results show that an originally close class distance tends to be far away. A distinctive feature value was probably extracted using ICA. However, the “station” class was close to the “downtown” class. Probably, the two sounds cannot be improved using ICA. In comparing results of ICA and MFCC, ICA improved the classification rate slightly more than MFCC. The distance between classes by MP is closer to each class than that obtained using any other feature extraction method. Moreover, the recognition rate of MP is lower than that of the other method. However, it is confirmed that MP showed the usefulness of environmental sound classification in terms of the combination frequency feature extraction method.

Table 10. Class Distance of ICA and MP

Class Distance		ICA					
		bird	**station**	**park**	**crossing**	**traffic**	**down-town**
MP	bird	-	1.44	0.85	1.32	0.95	1.49
	station	0.58	-	0.60	1.29	0.51	**0.05**
	park	0.87	0.30	-	1.17	0.20	0.65
	crossing	1.05	0.47	0.25	-	1.01	1.32
	traffic	0.15	0.59	0.89	1.04	-	0.56
	downtown	0.62	0.16	0.28	0.51	0.66	-

Results show that the proposed method yielded correct feature values and provided robust sound classification. Regarding similar sounds such as stations or downtown classes, we improved the classification rate, but we were unable to provide robust classification. In future works, we shall investigate methods to identify similar sounds.

7 Conclusion

We specifically examined “environmental sound classification” to recognize higher semantics of context in the real world. We presented an actual recorded environmental sound classification method using ICA and MP. We improved the MFCC feature by ICA to classify environmental sound recorded using a single microphone on mobile devices. Experimentally obtained results show that the proposed method yielded the highest classification rate. Moreover, the proposed method identified sounds that consist mainly of the same sound but the parts of sounds mutually differ. Therefore, the proposed method provides a more robust environmental sound classification method that can provide an efficient mode of interaction between the virtual and the real world.

As described herein, we conducted classification experiments particularly addressing background sound in daily life, such as that from traffic and from crowds of people. Further study is underway to investigate the efficiency of MP using event sounds. One future challenge is to increase the number of classification classes and to eliminate the influence of the recording environment. Additionally, environmental sound classification should incorporate consideration of the change of sound configuration over time. Furthermore, we must consider the limitations of computational resources on the mobile device because the proposed method based on ICA and MP requires higher computation processes. Therefore, a cloud-based classification architecture is anticipated. We are also investigating an efficient method to separate multiple sounds that exist simultaneously in a single time.

Acknowledgement

This work is supported in part by the National Institute of Information and Communications Technology (NICT), Japan, under the new generation network R&D program for innovative network virtualization platform and its application.

References

[1] S. Tamminen, A. Oulasvirta, K. Toiskallio, and A. Kankainen, “Understanding mobile contexts,” Journal Personal and Ubiquitous Computing. 2004; 8(2): 135–143. http://dx.doi.org/10.1007/s00779-004-0263-1

[2] D. Roggen, K. Foster, A. Calatroni, T. Holleczek, Y. Fang, G. Troster, P. Lukowicz, G. Pirkl, D. Bannach, K. Kunze, A. Ferscha, C. Holzmann, A. Riener, R. Chavarriaga, and J. del R. Millan, “Opportunity: Toward opportunistic activity and context recognition systems,” IEEE International Symposium on WoWMoM. 2009: 1-6.

[3] N. Gyorbiro, A. Fabian, and G. Homanyi, “An Activity Recognition System for Mobile Phones,” Mobile Networks and Applications. 2009; 14: 82-91. http://dx.doi.org/10.1007/s11036-008-0112-y

[4] L. Ma, D.J. Smith, and B.P. Milner, “Context Awareness Using Environmental Noise Classification,” 8th European Conference on Speech Communication and Technology, Eurospeech. 2003: 2237-2240.

[5] R.S. Goldhor, “Recognition of Environmental Sounds,” IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. 1993; 1: 149-152.

[6] Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” Neural Networks. 2000; 13(4-5): 411-430.

[7] S. Chu, S. Narayanan, and C.-C.J. Kuo, “Environmental sound classification with time-frequency audio features,” IEEE Transactions on Audio, Speech, and Language Processing. 2009; 17(6): 1142-58. http://dx.doi.org/10.1109/TASL.2009.2017438

[8] S.G. Mallet, and S. Zhang, “Matching Pursuits with Time-Frequency Dictionaries,” IEEE Transactions on Signal Processing. 1993; 41(12): 3397-3415. http://dx.doi.org/10.1109/78.258082

[9] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, “Temporal ICA for Recognition of Acoustic Events in a Kitchen Environment,” Interspeech. 2005: 2689-2692.

[10] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Kalapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, “Audio-based context recognition,” IEEE Transactions on Audio, Speech, and Language Processing. 2006; 14(1): 321-329. http://dx.doi.org/10.1109/TSA.2005.854103

[11] W. Dargie, “Adaptive Audio-Based Contest Recognition,” IEEE Transactions on Systems, Man, and Cybernetics. 2009; 39: 715-725. http://dx.doi.org/10.1109/TSMCA.2009.2015676

[12] H. Lu, W. Pan, N.D. Lane, T. Choudhury, and A.T. Campbell, “SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones,” MobiSys’. 2009: 165-178. http://dx.doi.org/10.1145/1555816.1555834

[13] J.-F. Cardoso, “Multidimensional independent component analysis,” Proceedings of the ICASSP. 1998; 37.

[14] Sharma and K.K. Paliwal, “Subspace independent component analysis using vector kurtosis,” Pattern Recognition. 2006; 39: 2227-32. http://dx.doi.org/10.1016/j.patcog.2006.04.021

[15] J.-H. Lee, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, “Speech Feature Extraction using Independent Component Analysis,” IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. 2000; 3: 1631-1634.

[16] H. Park, T. Takiguchi, and Y. Ariki, “Integration of Phoneme-Subspaces using ICA for Speech Feature Extraction and Recognition,” IEEE HSCMA. 2008: 48-151.

[17] H. Zhao, L. Hu, X. Peng, and G. Wang, “An improving MFCC feature extraction Based on FastICA algorithm plus RASTA filtering,” Journal of Computers, Academy. 2011: 1477-1484.

[18] Hyvärinen, and E. Oja, “A fast fixed-point algorithm for independent component analysis,” Neural Computation. 1997; 9(9): 1483-1492.

[19] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[20] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statistical Soc., Ser. R. 1977; 39(1): 1-38.

[21] S. Marian, M. R. Javier, and S. J. Terrence, "Face Recognition by Independent Component Analysis," IEEE Transactions on Neural Networks. 2002; 13(6): 1450-1464, 2002. PMid:18244540 http://dx.doi.org/10.1109/TNN.2002.804287

[22] Hyvärinen, J. Karhunen, and E. Oja, Independent component analysis, John Wiley & Sons, 2001. http://dx.doi.org/10.1002/0471221317