IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2020, accepted June 29, 2020, date of publication July 6, 2020, date of current version July 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007199

Feature Extraction Based on the Non-Negative Matrix Factorization of Convolutional Neural Networks for Monitoring Domestic Activity With Acoustic Signals

SEOKJIN LEE[1], (Member, IEEE), AND HEE-SUK PANG[2], (Member, IEEE)
[1]School of Electronics Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Department of Electrical Engineering, Sejong University, Seoul 05006, South Korea

Corresponding author: Hee-Suk Pang (hspang@sejong.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT), Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding, under Grant 2016-0-00564.

ABSTRACT In this paper, a feature extraction method is proposed based on the non-negative matrix factorization (NMF) for classifiers for monitoring domestic activities with acoustic signals. Most of the classifiers of the acoustic signals use data-independent spectral features (e.g., log-Mel spectrum and Mel-frequency cepstral coefficients). Recently, some novel feature extraction methods have been researched, including convolution-NMF-based features combined with K-means clustering. This study proposes an enhanced NMF-based feature extraction method that is inspired by the NMF-based noise reduction algorithm. The proposed method independently estimates the frequency basis matrix for each class, and then cascades the basis matrices to form the entire frequency bases, where the acoustic signal is transformed to the proposed feature by estimating the temporal basis matrix with the trained frequency bases. In addition, this study proposes a data augmentation method for the proposed feature that is inspired by the ''mix and shuffle'' method for audio waveforms. In order to evaluate the proposed system, which consists of the proposed NMF-based feature and the convolutional-neural-network-based classifier, some evaluations were performed using the *Detection and Classification of Acoustic Scenes and Events* (DCASE) 2018 Task 5 – Monitoring of Domestic Activities Based on Multi-channel Acoustics – Database. The results showed that the proposed system has comparable performance to that of state-of-the-art algorithms and that it has enhanced the F1-score performance of 6%–12% in comparison with the conventional NMF-based feature extraction method that is based on convolutional NMF and K-means clustering.

INDEX TERMS Acoustic scene classification, convolutional neural networks, feature extraction, non-negative matrix factorization.

I. INTRODUCTION

Acoustic scene classification (ASC) is tasked to automatically recognize environments through acoustic signals. In particular, the ASC task focuses on classifying long audio segments by characterizing whole audio environments, distinguishing from sound event detection problems to detect short sound events [1]. The recognition of environments via acoustic signals is one of the main problems of computational auditory scene analysis (CASA) [2] and it has become a major area of interest in many recent machine learning techniques, including robotic navigation [3] and personal archiving [4]. As the interest in the ASC problem grows, ASC-related tasks are researched by several communities such as *the detection and classification of acoustic scenes and events* (DCASE) [5].

Both ASC and acoustic event classification (AEC) are among the main analysis problems of environments through sound signals, and the scene and event classification tasks are not clearly distinguished. Recently, ASC tasks mainly

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

tend to focus on classifying longer signals with analyzing the whole acoustic environment [1], while AEC tasks deal more with short acoustic events, such as knock sounds or laughs. Also, ASC tasks have recently expanded into the monitoring domestic activity (MDA) tasks, which classify in-door sounds into several activity classes (e.g., vacuum cleaning, cooking, or watching TV). Both the modern ASC and MDA algorithms consist of two parts: feature extraction and classification modules.

The most common choices of the audio feature for recent ASC and MDA algorithms are Mel-frequency cepstral coefficients (MFCC) [3], [6], [7] or Mel-frequency-domain spectrum [8]–[10], which are kinds of spectrum-based values processed in a psycho-acoustic frequency domain [11]. These features are motivated by their success in speech signal applications, but their performance is limited in the acoustic scene or event classification applications, as acoustic environmental signals are less structured [1]. To substitute the Mel-frequency-based features, researches of the ASC algorithm have studied with the features inspired by the computer vision [12] or the modeling of statistical distributions [13]. Recently, several psychoacoustics-based features, such as the mel-frequency discrete wavelet coefficients [14], [15], hybrid constant-Q transform [16], gammatonegram [17], and gammatone-frequency cepstral coefficients have been studied [18]. To improve the performance, a combination of multiple features, such as a DNN-based ensemble network of the log-Mel-spectrogram, gammatonegram, and constant-Q transform [19] and an ensemble of label-tree embeddings of the log-Mel-spectrogram, gammatonegram, and MFCC [20] have been studied.

MFCC and other similar features can be considered as data-independent analysis techniques, as the required analysis processes for extracting the features do not depend on the signal characteristics. Recently, data-dependent signal analysis methods, such as principal component analysis (PCA) [21] and non-negative matrix factorization (NMF) [22], have been researched. In particular, the NMF algorithm was applied to analyze the magnitude spectrogram of an acoustic signal in recent acoustic signal processing, such as music signal processing [23]–[25] and speech denoising [26]–[29], as the NMF technique can decompose the spectrogram into the frequency and temporal basis matrices.

Lee and Seung have shown that the NMF algorithm can analyze two-dimensional non-negative data by using parts-based representation [22]. For example, the NMF algorithm makes decomposed images that correspond to parts of a face, e.g. mouth and eyebrow, when the algorithm is applied to a facial image, while the vector quantization algorithm makes prototypes of the whole face and the PCA algorithm makes ''eigenfaces'' that form a distorted version of the whole face. For the analysis of a sound signal, the NMF algorithm decomposes magnitude spectrograms into frequency basis and temporal basis matrices due to the parts-based representation characteristic. Each frequency basis and temporal basis can be a frequency structure and a temporal envelope of a musical note when the NMF method is applied to a music signal [23]. In speech denoising applications, the NMF method is used to learn the frequency structures of speech and noise signals *a priori*, and the temporal basis matrix of each frequency basis is estimated from noisy speech signals [28], [29]. Most of the NMF applications take advantage of the fact that the NMF can analyze the characteristic frequency structure and the temporal activation of the same class of signals.

Recently, the NMF algorithm was applied to acoustic scene classification in both supervised [1] and unsupervised methods [1], [30], [31]. The supervised method was developed based on the task-driven dictionary learning (TDL) model with a multinomial logistic regression [32] and L-BFGS [33]. However, the model and the update algorithms were far from the recent classifiers, such as deep neural networks and gradient-based algorithms, so it was difficult to extend them using recent techniques, such as the convolutional neural network (CNN). The unsupervised methods were developed based on the NMF with time-averaged clips or convolutional NMF with K-means clustering [1], [31], but they required very a large data matrix and additional data reduction processes, such as time averaging or K-means clustering, as they have to deal with the whole un-categorized dataset. If the task is supervised, we may generate the basis matrices via simpler processes.

The task of monitoring the domestic activity [34] has a goal of classifying the audio segments to predefined classes that are composed of daily activities in home environments, e.g., cooking, dishwashing, vacuum cleaning, etc. In order to achieve this goal, in this paper, we try to develop a scene classification method based on the NMF and CNN techniques that is as simple and extensible as possible. The proposed system consists of two modules: a NMF-based feature extraction module and a CNN-based classifier module. Our main contribution is the development of simple feature extraction and augmentation methods based on NMF in a supervised manner and compatibility to common classifiers, such as the simple CNN classifier.

II. PROBLEM DESCRIPTION

A. PROBLEM DESCRIPTION

The ASC is a task used for classifying audio segments with given durations. The common ASC is defined as the recognition of the audio environments, which are defined based on

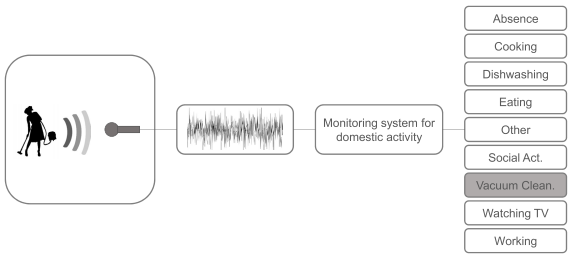

FIGURE 1. An illustrative diagram for the monitoring of domestic activity system.

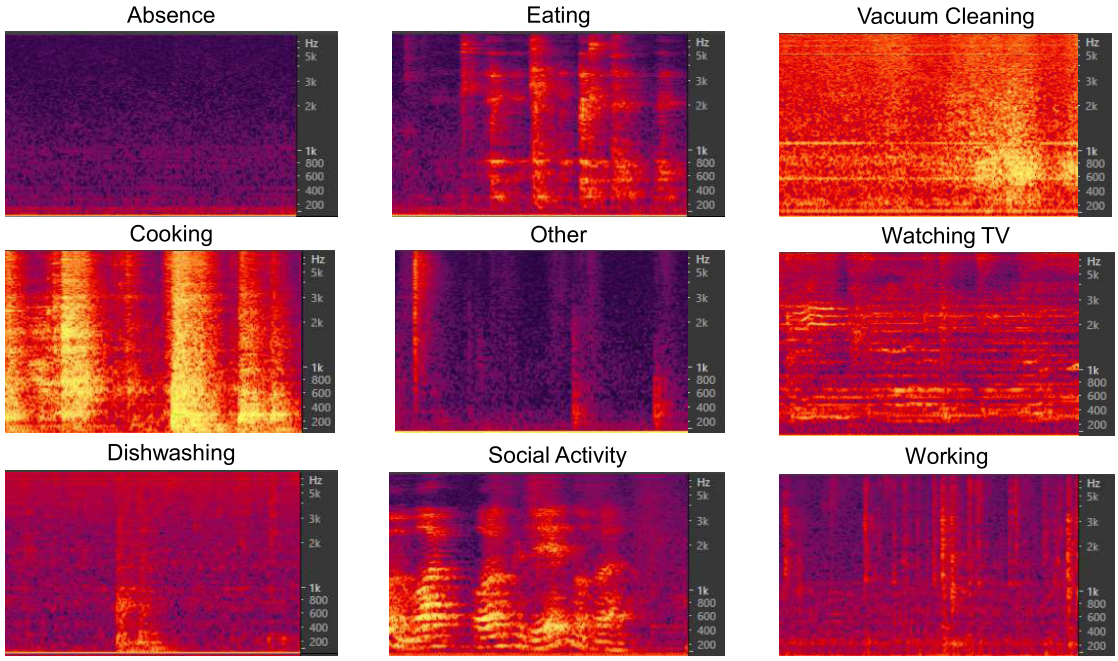

FIGURE 2. Examples of spectrograms for the activity classes.

physical or social contexts, such as parks, offices, etc [35]. However, the monitoring of domestic activity tasks has a goal, which is to classify the performed activities by people, such as cooking, dishwashing, working, etc., as shown in Fig. 1. Since the sounds of domestic activities include ensembles of multiple sound events, classifying domestic activities can be regarded as a kind of ASC tasks [34]. Moreover, the algorithm has to focus on the characteristics of sound events, such as keyboard typing and running water rather than the room environments, such as room transfer functions and background noise.

In order to classify domestic activity sounds, some features can be extracted from the sound clips and then classified into activity classes with neural-network-based classifiers just like many other recent algorithms. Although various classifiers with different network structures have been recently tried, the features adopted were still similar frequency-based features. For example, the log-Mel spectral energies were used as the input feature in 26 systems out of the 31 systems submitted to the DCASE 2018 Task 5. So, we would like to find a different feature extraction strategy that is suitable for the recent neural-network-based classifiers.

Figure 2 shows examples of the magnitude spectrograms of domestic activities. The domestic activity sounds include sets of event sounds that have distinctive characteristics in the time-frequency domain as shown in fig. 2. For example, the ''vacuum cleaning'' sound consists of broadband-noise with two tonal lines around 500 Hz and 1 kHz, and the ''watching TV'' and ''social activity'' sounds consist of various harmonic components. Both the ''eating'' and ''working'' classes may have similar structures (impulsive sounds), but their temporal characteristic (intervals between the impulsive sounds) are quite different. Therefore, we hope that the NMF method can generate distinctive features by analyzing the temporal and spectral characteristics.

B. RELATED WORKS

For the acoustic scene classification problem, V. Bisot *et al.* have developed two NMF-related methods as mentioned in the introduction. One of them is the supervised TDL model with an L-BFGS optimizer [36]. Although the TDL model demonstrated a good performance in the evaluation, it is difficult to apply it with arbitrary classifiers, as the update equations of the NMF basis and the model weight are strongly combined. The other one is an unsupervised NMF model for feature extraction [1]. The method can be applied to various classifiers, as the NMF-based feature extraction and the classifier learning process are clearly separated. Therefore, this method is consistent with the goal of this study, but it does not use the annotation data, so there is room for improvement if we use the annotation data. Recently, some networks for acoustic signals that are based on the non-negative auto encoder (NAE) [37], [38], which is a variant of the NMF, have been researched, but there is still not enough research regarding the application of NAE to the ASC tasks.

The unsupervised NMF method performs convolutive NMF to each audio clip to generate a large set of NMF bases, which are then clustered using the K-means clustering technique. Unfortunately, this process is complicated,

and it takes a long time to estimate the NMF basis matrix because of the K-means clustering technique. Therefore, in this study, we aimed to develop an NMF-based feature extraction method that is simple and easy to use. Furthermore, we also tried to enhance the classification performance by using the annotation data in the NMF basis learning step.

The NMF method has been tried in previous studies for acoustic scene classification and sound event detection tasks. However, these previous investigations utilize the NMF method as an auxiliary tool to pre-process the input signal or the activity classifier, rather than a feature extraction method. Zhou *et al.* [39] proposed the NMF-based sound event detector, but the NMF method was only used to perform noise reduction of the evaluation data. Mesaros *et al.* [40] also proposed the coupled-NMF-based sound event detector that consisted of the data analysis and classification step based on the NMF. The data analysis step has similar purpose to the feature extraction of the proposed method, but the dictionary matrix was generated in an unsupervised manner and coupled with the classifier, while the proposed method generates the dictionary matrix in a supervised manner and independently from the classifier. Chan's NMF-CNN structure [41] was developed for the weakly-supervised sound event detection task, whose dataset consisted of the data with annotations of the class and onset-offset time (strongly labeled), data with class annotations only (weakly labeled), and data without any annotation (unlabeled). Chan's method may look similar to the proposed method in that it uses both of the NMF and CNN, but the NMF method was simply used to pre-process the weakly-labeled and unlabeled data with pseudo-labeling of onset and offset time, so the design purpose and the structure are totally different compared to the proposed method.

III. PROPOSED SYSTEM

A. FEATURE EXTRACTION

1) NON-NEGATIVE MATRIX FACTORIZATION

The NMF is a method for the estimation of non-negative matrices $\mathbf{W} \in \mathbb{R}^{+}_{K\times R}$ and $\mathbf{H} \in \mathbb{R}^{+}_{R\times N}$, where the multiplication of two matrices is the same as a known non-negative matrix $\mathbf{V} \in \mathbb{R}^{+}_{K\times N}$ as [42]

$$\mathbf{V} = \mathbf{WH} + \mathbf{E}. \tag{1}$$

where $\mathbf{E} \in \mathbb{R}_{K\times N}$ is an error matrix. The matrices of $\mathbf{W}$ and $\mathbf{H}$ are estimated by minimizing the cost function between $\mathbf{V}$ and $\mathbf{WH}$ as [42]

$$\mathbf{W} = \arg\min_{\mathbf{W}} C\,(\mathbf{V}|\mathbf{WH}) \quad \text{for fixed } \mathbf{H} \tag{2}$$

$$\mathbf{H} = \arg\min_{\mathbf{H}} C\,(\mathbf{V}|\mathbf{WH}) \quad \text{for fixed } \mathbf{W} \tag{3}$$

where $C\,(\mathbf{A}|\mathbf{B})$ is the distance measure between the two matrices $\mathbf{A}$ and $\mathbf{B}$. Also, various distance measures, e.g., Euclidean distance, Kullback-Leibler divergence, Itakura-Sairo divergence, beta-divergence, etc., can be used for the NMF. The matrices $\mathbf{W}$ and $\mathbf{H}$ can be estimated to minimize the Kullback-Liebler divergence by alternating the iterations, which consist of [43]

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{[\mathbf{V}/(\mathbf{WH})]\,\mathbf{H}^T}{1_{K\times N}\mathbf{H}^T} \tag{4}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T\,[\mathbf{V}/(\mathbf{WH})]}{\mathbf{W}^T\mathbf{1}_{K\times N}} \tag{5}$$

where $\otimes$ and the fraction denote element-wise multiplication and division, respectively, and $\mathbf{1}_{K\times N}$ means a $K \times N$ matrix whose elements are all one.

In most of the NMF applications for acoustic signals, the known matrix $\mathbf{V}$ is the magnitude spectrogram of the input signals, and R is set to a small value relative to K or N so that the magnitude can be modeled as the multiplication of the matrices $\mathbf{W}$ and $\mathbf{H}$, which represent the spectral characteristics and temporal activations of acoustical events, respectively. For example, if the NMF algorithm is applied to a magnitude spectrogram of a music signal that consists of three musical events, each column vector of the matrix $\mathbf{W}$ may correspond to a frequency structure, and the row vector of the matrix $\mathbf{H}$ may correspond to a temporal envelope of a musical event, as shown in Fig. 3 (a). By focusing on these characteristics of the NMF method in the acoustic signals, several NMF applications have been developed, e.g., the speech denoising [28], [29] and the active sonar reverberation suppression [44], as shown in Fig. 3 (b). Speech denoising methods divide the bases into two classes, speech and noise, and remove the noise bases after calculating the temporal bases of each class. The active sonar reverberation suppression technique uses a similar methodology to that of speech denoising, where it divides the basis matrix into target echo and reverberation classes instead of speech and noise classes. Both the denoising and reverberation suppression methods use the NMF method as a separation tool by pre-training and merging the class-wise frequency bases. Focusing on the music signal applications, we believe that if we consider the matrix $\mathbf{W}$ and $\mathbf{H}$ as a transform matrix and a feature matrix, respectively, the generated feature matrix by the NMF can be considered as a sparse representation of the input spectrogram, because matrix $\mathbf{H}$ is a sparse representation of the input music signal in the music signal processing systems. Also, inspired by the denoising and the reverberation suppression methods, we believe that if we construct the frequency basis matrix by concatenating the class-wise frequency basis matrices, the temporal activation pattern, which is the $\mathbf{H}$ matrix, may vary depending on the class of the input signal. This is due to the fact that $\mathbf{H_s}$ and $\mathbf{H_n}$ represent the temporal activations of the speech and noise classes in the speech denoising system. Thus, we first propose a method to construct the frequency basis matrix by concatenating the class-wise bases, which greatly varies from the conventional

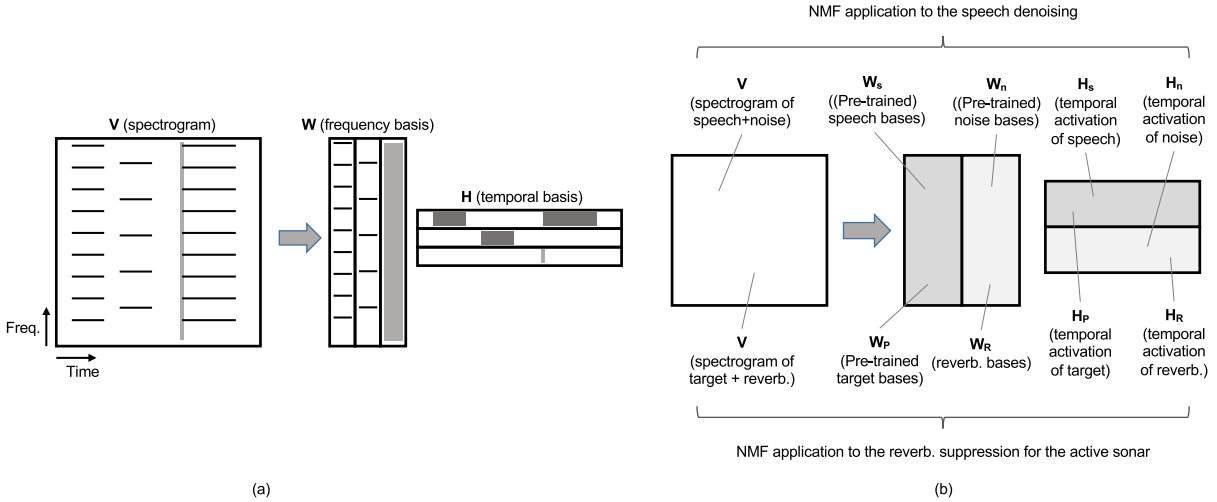

FIGURE 3. Illustrative examples of the NMF applications for (a) the music signal analysis, (b) the speech denoising (upper), and the active sonar (lower).

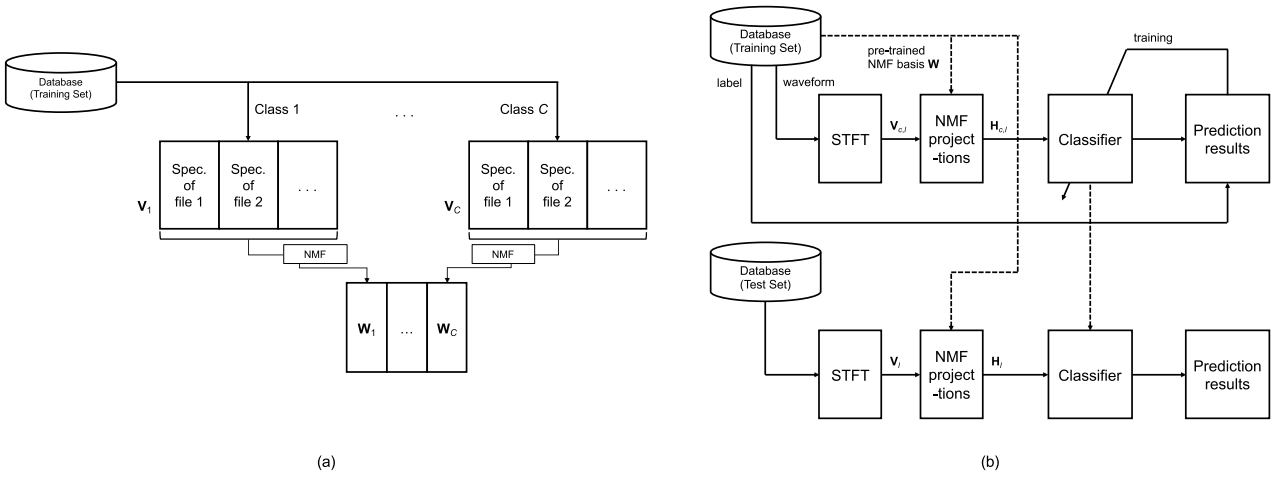

FIGURE 4. Diagrams for (a) learning procedure of the frequency basis in the proposed method and (b) classification system with the proposed feature extraction.

NMF-based feature extraction [1], as described in the next section.

2) CLASS-WISE LEARNING OF THE FREQUENCY BASIS MATRIX

As mentioned in the previous section, the NMF method decomposes the spectrogram $\mathbf{V}$ into a transform matrix $\mathbf{W}$ and a feature matrix $\mathbf{H}$. The matrix $\mathbf{W}$ may be estimated before or during the analysis process in the acoustic signal processing applications. However, we decide to learn the matrix $\mathbf{W}$ in advance because the NMF method has scale and ordering ambiguities, so it may interfere with the training and inference procedures if it is learned during the analysis.

As inspired by the NMF-based noise reduction algorithm [28], [44], we divide the basis vectors into C groups as

$$\mathbf{W} = [\mathbf{W}_1 \quad \mathbf{W}_2 \quad \cdots \quad \mathbf{W}_C] \tag{6}$$

where $\mathbf{W}_c \in \mathbb{R}^{+}_{K\times R_c}$ is a class-wise frequency basis matrix. $\mathbf{W}_c$ is estimated by iterative update equations as

$$\mathbf{W}_c \leftarrow \mathbf{W}_c \otimes \frac{[\mathbf{V}_c/(\mathbf{W}_c\mathbf{H}_c)]\,\mathbf{H}_c^T}{1_{K\times NL}\mathbf{H}_c^T} \tag{7}$$

$$\mathbf{H}_c \leftarrow \mathbf{H}_c \otimes \frac{\mathbf{W}_c^T\,[\mathbf{V}_c/(\mathbf{W}_c\mathbf{H}_c)]}{\mathbf{W}_c^T\mathbf{1}_{K\times NL}}, \tag{8}$$

where $\mathbf{V}_c \in \mathbb{R}^{+}_{K\times NL}$ and $\mathbf{H}_c \in \mathbb{R}^{+}_{R_c\times NL}$ are the class-wise spectrogram and temporal basis matrix, respectively, and R_c, K, N, and L are the number of bases per class, the number of frequency bins, the number of frames in a clip, and the number of clips in a class, respectively. The data matrix $\mathbf{V}_c$ consists of the spectrograms of the files in class c as

$$\mathbf{V}_c = \begin{bmatrix}\mathbf{V}_{c,1} & \cdots \mathbf{V}_{c,L}\end{bmatrix}, \tag{9}$$

where $\mathbf{V}_{c,l} \in \mathbb{R}^{+}_{K\times N}$ is the spectrogram of the lth file in the cth class.

The procedure for constructing the frequency basis matrix is described in Fig. 4 (a). To construct the frequency basis matrix $\mathbf{W}$, the audio clips are collected for each class, and spectrograms in each class are concatenated along the temporal axis, the NMF methods ((7) and (8)) are applied until convergence to estimate $\mathbf{W_c}$. After that, the classwise frequency matrices are concatenated by (6) to compose the frequency basis matrix $\mathbf{W}$.

3) FEATURE EXTRACTION

After the learning of the matrix $\mathbf{W}$ is completed, the feature extraction and classifier learning step can be performed. If we denote $\mathbf{V}_l$ as a magnitude spectrogram of the l-th audio clip, the feature matrix $\mathbf{H}_l$, which describes temporal activation of the basis vectors, of the clip is obtained by the iterations of

$$\mathbf{H}_l \leftarrow \mathbf{H}_l \otimes \frac{\mathbf{W}^T\left[\mathbf{V}_l/(\mathbf{W}\mathbf{H}_l)\right]}{\mathbf{W}^T\mathbf{1}_{K\times N}}. \tag{10}$$

During the estimation of $\mathbf{H}_l$, the frequency bases $\mathbf{W}$ are not changed. Thus, the feature extraction procedure requires a relatively small number of iterations. Fig. 5 shows examples of the change in the cost function with the number of iterations for training data (The details of the dataset are described in Chapter IV). The gray-colored area denotes the inter-quartile range between the 25th and 75th percentile points, and the thick solid line indicates the average values. The graphs show that the cost function may converge with about 20 iterations.

The entire structure of the classifier system with the proposed feature extraction method is shown in Fig. 4 (b). As shown in Fig. 4 (b), the proposed feature extraction method can be used by the same structure as that of the conventional features e.g. Mel-spectrogram, if the NMF frequency basis matrix $\mathbf{W}$ is pre-trained.

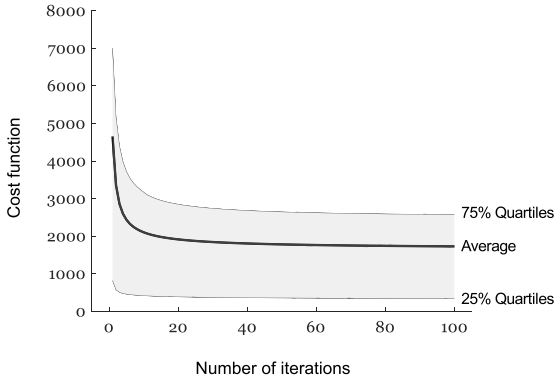

FIGURE 5. Convergence analysis of the feature extraction step.

4) DATA AUGMENTATION

Inspired by the data augmentation of the mixing and shuffling of the sound waveform [9], [45], we augment the data by mixing and shuffling the temporal basis matrix. In the waveform-based data augmentation method, the new waveform is generated by mixing two randomly chosen waveforms with a randomly shuffled order. That is, the augmented waveform $\mathbf{x}_{aug}$ is generated as

$$\mathbf{x}_{aug} = \begin{bmatrix}\underline{\mathbf{x}}^T_{b_1,c_1} & \underline{\mathbf{x}}^T_{b_2,c_2} & \cdots & \underline{\mathbf{x}}^T_{b_L,c_L}\end{bmatrix}^T, \tag{11}$$

where b_l and c_l are the block number and the clip number, respectively, l is the number of shuffle blocks in a clip, and $\underline{\mathbf{x}}_{b_l,c_l}$ is the b_l-th block of the c_l-th clip in the database of a certain class. b_l is randomly chosen in $\{l : 1 \le l \le L\}$ without duplication, and c_l is randomly chosen from two clip numbers.

If we assume that the length of each block is an integer multiple of the length of the FFT window, (11) can be presented in the time-frequency domain as

$$\mathbf{V}_{aug} = \begin{bmatrix}\underline{\mathbf{V}}_{b_1,c_1} & \underline{\mathbf{V}}_{b_2,c_2} & \cdots & \underline{\mathbf{V}}_{b_L,c_L}\end{bmatrix}, \tag{12}$$

where $\underline{\mathbf{V}}$ is the time-frequency-domain presentation, e.g., the spectrogram of $\underline{\mathbf{x}}$. According to the NMF model (1), the temporal slice of the spectrogram corresponds to the slice of the temporal basis matrix. For example, $[\mathbf{V}_1|\mathbf{V}_2] \approx \mathbf{W}\,[\mathbf{H}_1|\mathbf{H}_2]$. Therefore, (12) can be represented as

$$\begin{aligned}\mathbf{V}_{aug} &= \begin{bmatrix}\mathbf{W}\underline{\mathbf{H}}_{b_1,c_1} & \mathbf{W}\underline{\mathbf{H}}_{b_2,c_2} & \cdots & \mathbf{W}\underline{\mathbf{H}}_{b_L,c_L}\end{bmatrix}\\ &\approx \mathbf{W}\begin{bmatrix}\underline{\mathbf{H}}_{b_1,c_1} & \underline{\mathbf{H}}_{b_2,c_2} & \cdots & \underline{\mathbf{H}}_{b_L,c_L}\end{bmatrix}\\ &\triangleq \mathbf{W}\mathbf{H}_{aug}.\end{aligned} \tag{13}$$

As a result, the temporal basis matrix, which is the proposed feature, can be augmented by mixing and shuffling as

$$\mathbf{H}_{aug} = \begin{bmatrix}\underline{\mathbf{H}}_{b_1,c_1} & \underline{\mathbf{H}}_{b_2,c_2} & \cdots & \underline{\mathbf{H}}_{b_L,c_L}\end{bmatrix} \tag{14}$$

without performing an additional NMF feature extraction process. The illustrative diagram of the proposed augmentation procedure is described in Fig. 6. The conventional *mix and shuffle* augmentation method have to be applied to the waveform directly, and so the augmented data have to be processed by the NMF, which is the most time consuming part of our feature extraction procedure. However, our data augmentation method, which mix and shuffle matrix $\mathbf{H}_l$ instead of the waveform, can augment data without additional STFT or NMF calculations. While the NMF method consists of numerous multiplications, the mix and shuffle uses no multiplication, and the proposed data augmentation method can expand a large amount data with very light operations.

B. NETWORK STRUCTURE OF THE CLASSIFIER

The recent classifiers for the 2-dimensional data, e.g. Mel-Frequency spectrogram and MFCCs, are mainly based on or include the CNN structure [9], [46]. The feature matrix $\mathbf{H}_l$ is also a 2-dimensional data, so the CNN-based classifier is used in our system.

In order to compare the proposed method with the conventional feature extraction method, the classifier is similarly designed to the state-of-the-art classifier of the log-Mel energy features [9]. An example of the classifier structure of

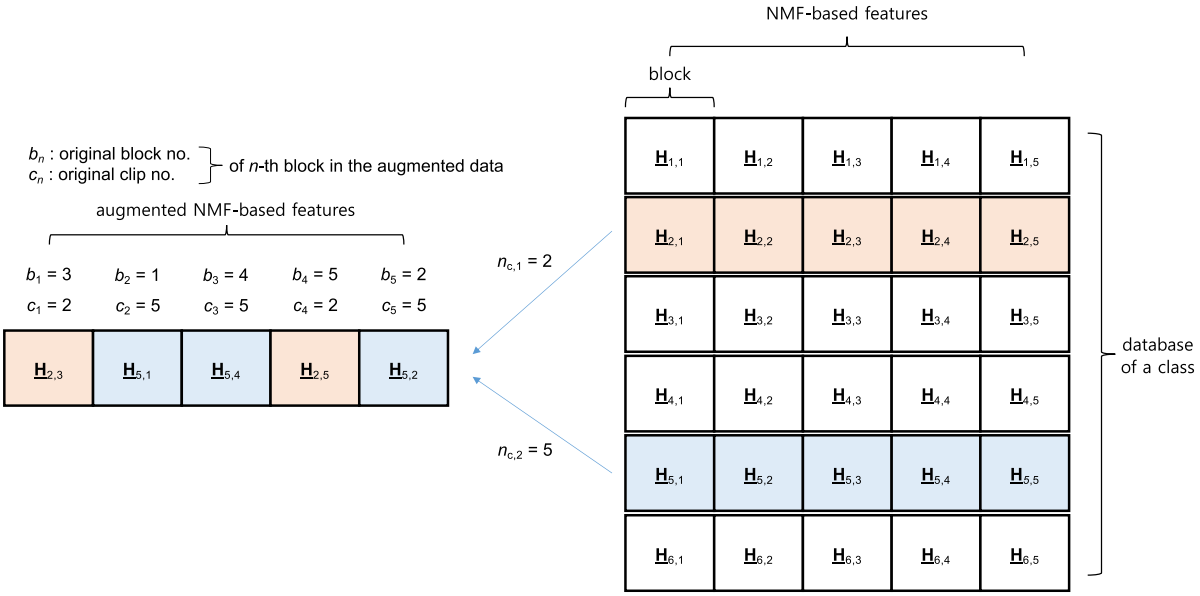

FIGURE 6. An illustrative example of the data augmentation method by mixing and shuffling the extracted NMF-based features.

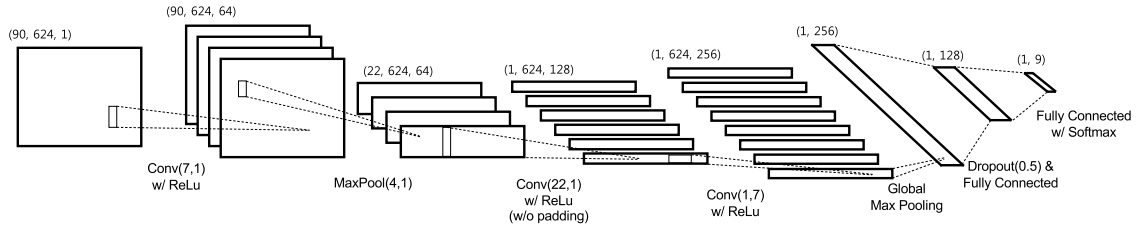

FIGURE 7. Example of the classifier structure.

the proposed NMF-based features is displayed in Fig. 7 with $R_C = 10$. Since the first-axis dimension of the input matrix (90 in Fig. 7) is defined by $N_C R_C$, where N_C is the number of classes, the filter length of the second CNN layer (Conv(22, 1) in Fig. 7) is calculated by $\left\lfloor \frac{N_C R_C}{4} \right\rfloor$, where $\lfloor \quad \rfloor$ means the floor function.

IV. EVALUATION

A. EVALUATION SETTING

In order to evaluate the proposed system for the monitoring domestic activities, some simulations were performed with the DCASE 2018 Task 5 database, which is an audio dataset for the monitoring of domestic activities that was recorded in a living room and a kitchen [34]. The audio files were recorded with 4-channel linear microphone arrays.

There were 9 activity classes: absence (nobody in the room), cooking, dishwashing, eating, social activity, vacuum cleaning, watching TV, working, and other (non-relevant activity), as shown in Table 1. Each audio file was 10-seconds long and represented one activity. The audio files were acquired with a 16-kHz sampling rate and a 12-bit quantization. The detailed recording setup, including the floorplan, can be found in [34].

The dataset consisted of development and evaluation sets. The development set approximately had 200 hours of data from 4 microphone arrays for the training and evaluation of the monitoring system. The evaluation set consisted of data from 7 microphone arrays, and the quantity of the evaluation set was similar to that of the development set. The used 4 microphone arrays to get the evaluation set were same arrays used for the development set, and the other 3 microphone arrays that were used for the evaluation set were not used for the development set.

The audio clips were short-time-Fourier-transformed by 512-samples Hamming window with 50% overlap into 512 frequency bins. The number of NMF iterations was set to 100 for learning frequency basis matrix and 30 for the feature extraction. We also tested the enhanced NMF methods by sparseness and temporal continuity [47] with various parameters, but it could not improve the performance. The classifiers were trained by Adam optimizer [48] with a learning rate of 0.0001. The batch size and number of epochs were 16 and 100, respectively. The input audio in the dataset

had a 4-channel signal, so each frequency basis matrix was independently trained and applied for each channel.

The performance was measured by the F_1-score, which is defined as

$$F_1 = \frac{2PR}{P+R} \tag{15}$$

where P and R are the precision and recall, respectively. The precision and recall are relevance measures, which are defined as

$$P = \frac{n_{TP}}{n_{TP} + n_{FP}} \tag{16}$$

$$R = \frac{n_{TP}}{n_{TP} + n_{FN}}, \tag{17}$$

where n_{TP}, n_{FP}, and n_{FN} are the numbers of true positives (relevant answers), false positives (false answers), and false negatives (missing answers), respectively. We used the *macro-averaged* score, where the class-wise scores were first calculated, and then averaged, to evaluated the performance.

The performance of the development dataset was cross-checked by 4-folds and then averaged. For example, suppose we divide the development dataset into 4-blocks, named a, b, c, and d. The first fold consists of the training data of a, b, c and the evaluation data of d, and the second fold consists of the training data of a, b, d and the evaluation data of c, and so on. The frequency basis matrix for each fold was generated by only using the training data, and the evaluation data of the development and evaluation datasets were not used. Also, the detailed cross-check configuration, including the clip list for each fold, was in accordance with the DCASE 2018 Task 5.

TABLE 1. Activity classes and their quantity in the DCASE 2018 Task 5.

Activity classes	Number of files
Absence	18860
Cooking	5124
Dishwashing	1424
Eating	2308
Social activity	4944
Vacuum cleaning	972
Watching TV	18648
Working	18644
Other	2060

B. COMPARISONS WITH THE STATE-OF-THE-ART ALGORITHM

In order to evaluate the performance, the proposed system was compared with Inoue's algorithm [9] and Liu's method [7], which have the best performances in the DCASE 2018 Task 5 competition. Inoue's algorithm consisted of log-Mel-spectrogram-based features and the CNN-based classifier with three CNN layers with batch normalization (BN) and ReLU activation and two fully-connected layers with the softmax output, which is a similar structure to that of the proposed system. In the implementation of Inoue's system, the 40-bin log-Mel spectrograms were extractced using a 64-ms window with a 20-ms overlap, and the classifiers were trained using the Adam [48] optimizer with a learning rate of 0.0001 for 100 epochs. The detailed structure of the classifier can be found in [9].

Liu's method used an ensemble structure of three sub-systems. The first sub-system used 40-bin Mel-spectrogram features per frame and a CNN-based classifier, which has a similar structure to that of the proposed and Inoue's systems. The second sub-system used 40 Mel-frequency cepstral coefficients (MFCC) per frame and a CNN-based classifier with the same structure of the first sub-system. The third sub-system used 128 extracted features by a pre-trained VGGish [49], which is a variant of the VGG [50] for audio signals, per frame and a long-short-term-memory -based (LSTM) classifier. The detailed structures can be found in [49], and the classifiers were trained by the Adam optimizer with a learning rate of 0.0001 (0.001 for the LSTM classifier) for 100 epochs.

Table 2 shows the F1-score results of the comparison and the proposed methods. NMF-CNN denotes the proposed system, and "with BN" means that each CNN layer in the classifier was combined with the BN modules. The results show that the performance of the proposed system is slightly less than that of Inoue's method and better than that of Liu's method in both the Dev and Eval2 datasets. The Eval1 performance of the proposed method is similar to that of both Inoue's and Liu's methods.

According to the results of Inoue's method with and without BN, the BN in Inoue's method can improve the performance. However, the results of the proposed method with and without BN show that the BN is not effective for the proposed system. The performance of the proposed method without BN is comparable to that of Inoue's method without BN. The proposed method is slightly better in the Dev dataset, slightly worse in the Eval1 dataset, and almost the same in the Eval2 dataset. Therefore, the performance differences between the proposed method and Inoue's method may be due to the difference in the adequacy of BN for the features.

There is one more thing to note. The performances of Inoue's and Liu's methods are about 1% and 2.5 % lower, respectively, in the Eval2 data than in the Eval1 data. This phenomenon not only occurs in the those methods but also in most of the submitted methods to DCASE 2018 Task 5. However, the performance difference between the two datasets is relatively small, about 0.4 %, in the proposed method.

C. PERFORMANCE CHANGE ACCORDING TO THE NUMBER OF BASES

The number of bases is used as a major engineering parameter in many NMF-based signal processing methods. Therefore, the performance change according to the number of bases was analyzed in this paper. Table 3 and Fig. 8 show the performances of the proposed systems with $R_C = 20$, $R_C = 10$, $R_C = 5$, and $R_C = 3$. Since the number of classes is 9 in the experiment, the dimensions of the features for a certain time frame are 180, 90, 45, and 27.

TABLE 2. Performance comparison with the state-of-the-art system.

	Dev					Eval1	Eval2
	Fold 1	Fold 2	Fold 3	Fold 4	Average	(Dev. set mic.)	(Unknown mic.)
Inoue et al. [9]	90.45	88.93	88.86	92.84	90.27	89.81	88.07
Inoue et al. (without BN) [9]	88.76	86.35	87.76	90.79	88.42	88.14	87.24
Liu et al. [7]	87.58	85.77	87.83	91.26	88.11	87.86	85.24
NMF-CNN ($R_C = 20$)	88.53	88.06	87.18	91.98	88.94	87.72	87.25
NMF-CNN ($R_C = 20$) (with BN)	88.41	87.95	88.14	92.32	89.26	86.19	85.14

TABLE 3. Performance change with the number of bases.

	Dev					Eval1	Eval2
	Fold 1	Fold 2	Fold 3	Fold 4	Average	(Dev. set mic.)	(Unknown mic.)
NMF-CNN ($R_C = 20$)	88.53	88.06	87.18	91.98	88.94	87.72	87.25
NMF-CNN ($R_C = 10$)	88.23	86.60	87.54	91.74	88.53	87.35	86.95
NMF-CNN ($R_C = 5$)	87.58	86.74	87.97	91.27	88.38	87.19	87.15
NMF-CNN ($R_C = 3$)	86.95	86.15	85.22	90.65	87.24	86.29	85.66

In the $R_C = 20$, $R_C = 10$, and $R_C = 5$ cases, the performances of the systems slightly increase with the increase in the number of bases. The performance of the Eval2 dataset does not significantly change even if the number of bases changes in those cases. However, the performance of the proposed method with $R_C = 3$ is noticeably reduced with all the datasets. We think that the performance of the proposed system was not largely affected by the change in the number of bases when $R_C \geq 5$ in this experiment.

As mentioned in the previous section, the performance differences in the proposed systems between the Eval1 and Eval2 datasets are relatively small. This property is also shown in the result of $R_C = 10$ and $R_C = 5$ cases.

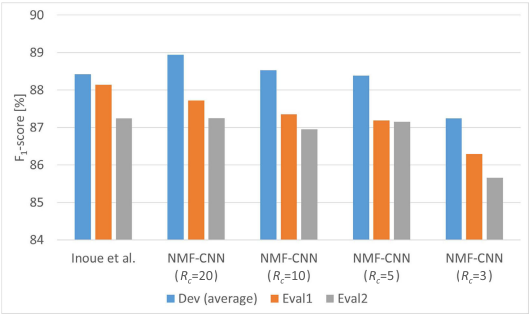

FIGURE 8. Performance comparison of the proposed systems with various numbers of bases.

D. COMPARISONS TO THE CONVENTIONAL NMF-BASED FEATURE EXTRACTION METHOD

In order to evaluate the performance of the proposed feature extraction method, we compared the proposed system with the conventional NMF-based feature extraction method [1], which consists of the convolutional NMF and K-means clustering. Just like the proposed features, the audio clips were short-time-Fourier-transformed by 512-samples Hamming window with 50% overlap into 512 frequency bins. The spectrogram of each audio clip was decomposed to 20 2D-dictionaries with 257 frequency bins and 4 time frames. Therefore, a dictionary from an audio clip was $\mathbb{R}_+^{257\times4\times20}$. The whole dictionaries were clustered into 256 and 512 centers with K-means clustering, but the K-means clustering to 512 centers failed to converge in our dataset.

Table 4 shows the comparison results between the classification performances using the conventional convolutional-NMF-based features and the proposed features. The used classifier for the conventional features are the same as those of the proposed system. The results show that the proposed NMF-based features may be more suitable for the used CNN-based architecture in the proposed system.

TABLE 4. Performance comparison to the conventional NMF-based feature.

	Dev	Eval1 (Dev. set mic.)	Eval2 (Unknown mic.)
ConvNMF + K-means ($R = 256$) [1]	83.07	79.34	75.84
NMF-CNN ($R_c = 20$)	88.94	87.72	87.25

E. COMPARISONS WITH CONVENTIONAL FEATURES

In order to compare the performace of the proposed feature with the various existing features, the performance of the proposed system was compared to the systems utilizing the conventional features, including constant-Q transforms (CQT) [51], [52], power-normalized cepstral coefficient (PNCC) [53], Mel-frequency discrete wavelet coefficients (MFDWC) [14], gammatonegram (GAM) [17], and gammatone frequency cepstral coefficient (GFCC) [18]. The length of window and overlap for the Fourier transform were set to 64 ms and 20 ms, respectively, which are the same values as in the log-Mel spectrogram case. The CNN classifiers were the same as in the proposed system, as shown in Fig. 7. Similar to the proposed system, the filter length of the second CNN layer was adjusted to $\left\lfloor \frac{N_{feature}}{4} \right\rfloor$, where $N_{feature}$ is the number of features in a frame, so that the first dimension of

the second CNN layer output was one. All training data were equally augmented by the mix & shuffle method [45] in the waveform domain.

All the compared features are frequency-based so they can be implemented with the short-time Fourier transform. The parameters of the Fourier transforms, e.g., the window/overlap length, number of FFT points, and type of the window function, were set to the same values as used by Inoue [9] and the proposed system. The number of features adopted for each system is displayed in Table 5. The lower bound frequency was set to 32.7 Hz and the number of CQT bins was 12 per octave in the CQT system, so the 96 CQT bins (= 8 octaves) could cover the whole frequency range. The number of the PNCC features was set to 40, the same as in reference [53]. The numbers of featrues of the MFDWC and GFCC were 15 and 13, respectively, as in the previous studies [14], [18], and larger numbers of coefficients (MFDWC31 and GFCC26) were also tested. The MFDWC15 system used 8, 4, 2, 1 coefficient at scale 4, 8, 16, 32, respectively, the same as in reference [14], while the MFDWC31 system used 16, 8, 4, 2, 1 coefficient at scale 4, 8, 16, 32, 64, respectively. Therefore, the number of Mel-bands was set to 64 in the MFDWC31 system, while the MFDWC15 system used 32 Mel-bands. The ensemble systems consisted of the independent networks for the log-Mel-spectrogram, CQT, and the GAM, and the prediction results of the networks were averaged.

TABLE 5. The types and numbers of conventional features.

Abbreviation	Feature type	Number of features for a frame
CQT	constant-Q transform	96
PNCC	power-normalized cepstral coefficients	40
MFDWC15	Mel-frequency discrete wavelet coefficients	15
MFDWC31	Mel-frequency discrete wavelet coefficients	31
GAM	gammatonegram	48
GFCC13	gammatone-frequency cepstral coefficients	13
GFCC26	gammatone-frequency cepstral coefficients	26
Ensemble	ensemble system of the log-Mel-spectrogram, CQT, and GAM	40 (Mel spectrogram) 96 (CQT) 48 (GAM)

In the averaged results of Dev dataset, the performance of the proposed feature is better than the results for all of the other features, except for the GAM, MFDWC31, and the ensemble system. Although the proposed feature performs slightly better than MFDWC31, the improvement is only marginal. The results from the Eval1 dataset show that the MFDWCs and the GFCC26 features perform better than the proposed algorithm, unlike the results from the Dev dataset. However, the results for the MFDWCs and the GFCC26 features exhibit significantly degraded results for the Eval2 dataset. As a result, the proposed features demonstrate a better performance than all of the compared systems, except the GAM and the ensemble system, and a very close performance to the GAM and the ensemble system.

The F_1-score of the most of the analyzed systems is about 1.5% to 5% lower for the Eval2 dataset than for the Eval1 dataset. For the MFDWC and GFCC systems, the performance assessed on the Dev and Eval1 datasets improves as the number of features increases, but the difference between the Eval1 and Eval2 datasets also increases, so the performances on the Eval2 dataset decrease. However, the proposed system shows a difference of only 0.47% between the two datasets, so the proposed system can be regarded as robust to the change of the room transfer function. The PNCC system exhibits the smallest difference (0.8%) between the datasets among the compared systems, but the performance of the PNCC system itself is inferior to the proposed system.

TABLE 6. Performance comparison to the conventional features.

	Dev	Eval1 (Dev. set mic.)	Eval2 (Unknown mic.)
NMF-CNN ($R_c = 20$)	88.94	87.72	87.25
CQT	84.06	83.50	78.33
PNCC	87.62	87.33	86.51
MFDWC15	87.80	88.47	86.27
MFDWC31	88.78	89.01	85.76
GAM	89.16	88.85	87.27
GFCC13	87.25	86.64	83.25
GFCC26	87.60	87.81	81.55
Ensemble	89.10	88.64	87.36

V. CONCLUSION

In this paper, an NMF-based feature extraction method is proposed for the monitoring domestic activity tasks by using sound signals. The proposed method was designed for supervised classifiers for domestic sounds. Inspired by the NMF-based source separation methods, the proposed method estimates class-wise frequency bases using annotated sound signals. Then, the temporal bases matrix is extracted from the input signal based on the concatenated class-wise frequency basis matrices. The temporal basis matrix was used as the feature matrix, and the features could be augmented using the proposed data augmentation method that is derived from the waveform-based mix and shuffle method without additional calculations.

In order to evaluate the proposed feature extraction method, some experiments were performed based on the DCASE 2018 Task 5 database. First, the proposed method was compared to state-of-the-art algorithms that utilize the log-Mel spectrum, Mel-spectrogram, and VGGish model output as input features. The evaluation results showed that the combined system of the proposed NMF-based feature and the CNN-based classifier has comparable performance to that of the state-of-the-art algorithms. Second, the proposed algorithm was evaluated by changing the feature dimension, and the results show that the performance of the proposed algorithm is consistent with the change in the feature dimensions, except for the extremely-small-bases case ($R_C = 3$).

Third, the proposed algorithm was compared to the conventional NMF-based feature extraction method, which consists of convolution NMF and K-means clustering, and the results showed that the proposed algorithm has better F1-score performances of 6%–12% in comparison with the conventional NMF-based features.

REFERENCES

[1] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1216–1229, Jun. 2017.

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Piscataway, NJ, USA: IEEE Press, 2006.

[3] S. Chu, S. Narayanan, C.-C. Kuo, and M. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 885–888.

[4] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proc. the 1st ACM Workshop Continuous Archival Retr. Pers. Exper.*, 2004, pp. 39–47.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, Nov. 2018, pp. 9–13.

[6] Z. Huang and D. Jiang, "Acoustic scene classification based on deep convolutional neural network with spatial-temporal attention pooling," DCASE Challenge, IEEE, New York, NY, USA, Tech. Rep., 2019.

[7] H. Liu, F. Wang, X. Liu, and D. Guo, "An ensemble system for domestic activity recognition," DCASE Challenge, IEEE, New York, NY, USA, Tech. Rep., 2018.

[8] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE Challenge, IEEE, New York, NY, USA, Tech. Rep., Jun. 2019.

[9] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, and R. Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," DCASE Challenge, IEEE, New York, NY, USA, Tech. Rep., Sep. 2018.

[10] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1547–1554.

[11] B. C. Moore, *An Introduction to the Psychology of Hearing*. Leiden, The Netherlands: Brill, 2012.

[12] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 719–723.

[13] G. Roma, W. Nogueira, P. Herrera, and R. de Boronat, "Recurrence quantification analysis features for auditory scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoust. Scenes Events*, vol. 2, 2013.

[14] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Jun. 2000, pp. 1351–1354.

[15] S. Waldekar and G. Saha, "Wavelet transform based mel-scaled features for acoustic scene classification," in *Proc. INTERSPEECH*, Sep. 2018, pp. 3323–3327.

[16] M. Wang, R. Wang, X.-L. Zhang, and S. Rahardja, "Hybrid constant-Q transform based CNN ensemble for acoustic scene classification," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1511–1516.

[17] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio scene classification with deep recurrent neural networks," 2017, *arXiv:1703.04770*. [Online]. Available: http://arxiv.org/abs/1703.04770

[18] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.

[19] L. Pham, I. McLoughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification," in *Proc. INTERSPEECH*, Sep. 2019, pp. 3634–3638.

[20] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins, "Improved audio scene classification based on label-tree embeddings and convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1278–1290, Jun. 2017.

[21] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

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

[23] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2003, pp. 177–180.

[24] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.

[25] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, Dec. 2013.

[26] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4029–4032.

[27] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement using bases update," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 450–454, Apr. 2015.

[28] H.-T. Fan, J.-W. Hung, X. Lu, S.-S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4483–4487.

[29] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 17–20.

[30] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," M.S. thesis, Dept. ATIAM, Sci. Eng., Univ. Pierre et Marie Curie, Paris, France, 2011.

[31] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6445–6449.

[32] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1, no. 10. New York, NY, USA: Springer, 2001.

[33] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, nos. 1–3, pp. 503–528, Aug. 1989.

[34] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 challenge—Task 5: Monitoring of domestic activities based on multi-channel acoustics," KU Leuven, Leuven, Belgium, Tech. Rep., 2018.

[35] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1128–1132.

[36] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoust. Scenes Events (DCASE)*, Jun. 2016, pp. 62–69.

[37] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 86–90.

[38] S. Venkataramani, E. Tzinis, and P. Smaragdis, "End-to-end non-negative autoencoders for sound source separation," 2019, *arXiv:1911.00102*. [Online]. Available: http://arxiv.org/abs/1911.00102

[39] Q. Zhou, Z. Feng, and E. Benetos, "Adaptive noise reduction for sound event detection using subband-weighted NMF," *Sensors*, vol. 19, no. 14, p. 3206, Jul. 2019.

[40] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 151–155.

[41] T. Kai Chan, C. Siong Chin, and Y. Li, "Non-negative matrix factorization-convolutional neural network (NMF-CNN) for sound event detection," 2020, *arXiv:2001.07874*. [Online]. Available: http://arxiv.org/abs/2001.07874

[42] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Hoboken, NJ, USA: Wiley, 2009.
[43] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[44] S. Lee and J.-S. Lim, "Reverberation suppression using non-negative matrix factorization to detect low-Doppler target with continuous wave active sonar," *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, p. 11, Dec. 2019.
[45] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, A. Munawar, B. J. Ko, N. Greco, and R. Tachibana, "Shuffling and mixing data augmentation for environmental sound classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, 2019, pp. 109–113.
[46] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, and K. Hamada, "Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling," DCASE Challenge, IEEE, New York, NY, USA, Tech. Rep., Sep. 2018.
[47] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[49] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[51] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.
[52] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, vol. 90, 2016, pp. 1032–1048.
[53] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.

SEOKJIN LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Seoul National University, in 2006, 2008, and 2012, respectively.

From 2012 to 2014, he was a Senior Research Engineer with LG Electronics, and from 2014 to 2018, he was an Assistant Professor with the Department of Electronics Engineering, Kyonggi University, Suwon, South Korea. Since 2018, he has been an Assistant Professor with the School of Electronics Engineering, Kyungpook National University, Daegu, South Korea. His research interests include acoustic/sound/music signal processing, array signal processing, and blind source separation.

HEE-SUK PANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Seoul National University in 1994, 1996, and 2001, respectively.

From 2001 to 2008, he worked for LG electronics as a Chief Research Engineer. Since 2008, he has been an Associate Professor with the Department of Electrical Engineering, Sejong University. His research interests include acoustics and audio signal processing.