INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Deep Neural Network Bottleneck Features for Acoustic Event Recognition

Seongkyu Mun [1], *Suwon Shon* [2], *Wooil Kim* [3], *Hanseok Ko* [1,2]

[1] Dept. of Visual Information Processing, Korea University, Seoul, Korea
[2] School of Electrical Engineering, Korea University, Seoul, Korea
[3] Dept. of Computer Science and Engineering, Incheon National University, Incheon, Korea
{skmoon,swshon}@ispl.korea.ac.kr, wikim@inu.ac.kr, hsko@korea.ac.kr

Abstract

Bottleneck features have been shown to be effective in improving the accuracy of speaker recognition, language identification and automatic speech recognition. However, few works have focused on bottleneck features for acoustic event recognition. This paper proposes a novel acoustic event recognition framework using bottleneck features derived from a Deep Neural Network (DNN). In addition to conventional features (MFCC, Mel-spectrum, etc.), this paper employs rhythm, timbre, and spectrum-statistics features for effectively extracting acoustic characteristics from audio signals. The effectiveness of the proposed method is demonstrated on a database of real life recordings via experiments, and its robust performance is verified by comparing to conventional methods.

Index Terms: acoustic event recognition, deep neural network, bottleneck feature, feature extraction, deep belief network

1. Introduction

Acoustic Event Recognition (AER) is a field of autonomously recognizing different events based on sound. It has recently attracted considerable attention due to a variety of new potential [1-7]. Among recent AER researches, a lot of efforts are focused on feature extraction for acoustic event [4-7].

Recently, DNN-based approaches have been successful in many signal processing fields [8-10]. DNN approaches aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of raw data [9-10]. Latest researches of audio signal processing used Mel-scaled spectrogram and raw FFT bin values for speech recognition [11], instead of MFCC which is widely used for modeling human auditory organ.

The aforementioned research efforts showed that deep learning algorithms have considerable potential for carrying out powerful modeling and extracting discriminative features with using only raw data input.

Despite of its potentiality, the DNN-based discriminative feature extraction with raw data input leads to challenges associated with optimizing structures (e.g. # of layers, nodes) and weight learning parameters [8-9]. It requires a significant amount of efforts to establish appropriate system structure for best possible performance.

To address the issue, this paper uses human designed high-level features in addition to raw data features. This approach helps the DNN system find optimal weight parameters more efficiently compared to using raw data feature alone. Moreover, recent deep learning system, which defeated 'Go game' human champion, also used both raw data features (e.g. stone color, turns, etc.) and human designed tactical high-level features (e.g. ladder capture, ladder escape, etc.) [12].

Based on the premise and advantages shown by aforementioned research, this paper proposes to employ both raw data features (e.g. Mel-spectrogram) and high-level features (e.g. MFCC, selected timbre/rhythm feature and spectrum statistics) for improving acoustic event recognition performance.

Additionally, in order to extract discriminative acoustic characteristics from these various features, this paper also proposes to use BottleNeck (BN) features. Bottleneck features are widely used for low information loss nonlinear feature transformation and dimensionality reduction method in speech recognition [13], speaker recognition [14] and language identification [15]. However, few works have focused on bottleneck features for acoustic event recognition. Hence, this paper proposes an acoustic event recognition system using a bottleneck feature framework for extracting discriminative features from various features including raw data and selected high-level features.

2. Proposed feature extraction

The process of acoustic event recognition system based on bottleneck features is depicted in Figure 1. Before extracting bottleneck features, additional features, which will be combined with MFCC and Mel-spectrogram, are selected by chi-square statistics and information gain measurement (Figure 1-A). After selecting additional feature set, the bottleneck features are extracted using DBN (Deep Belief Network)-DNN with input features consists of selected features, Mel spectrogram and MFCC (Figure 1-B). Figure 1-C shows a final acoustic event classifier. The bottleneck features extracted from DNN are then input alone or concatenated with the MFCC/Mel spectrogram features to train the final acoustic event classifier.

2.1. Feature Selection

In order to analyze various types of features, this paper extracts rhythm/timbre features which are widely used in music signal processing and frequency spectrum statistics. The discriminative power of each feature is evaluated by computing the value of the chi-square statistics and information gain with respect to the class [16]. From the database of real life recording (listed on Section 3), total 20 musical and statistics features were extracted and then 8 features were selected based on chi-square statistics and information gain. The feature selection was done using the WEKA toolkit [17]. Descriptions of selected features are summarized in Table 1.

http://dx.doi.org/10.21437/Interspeech.2016-1112

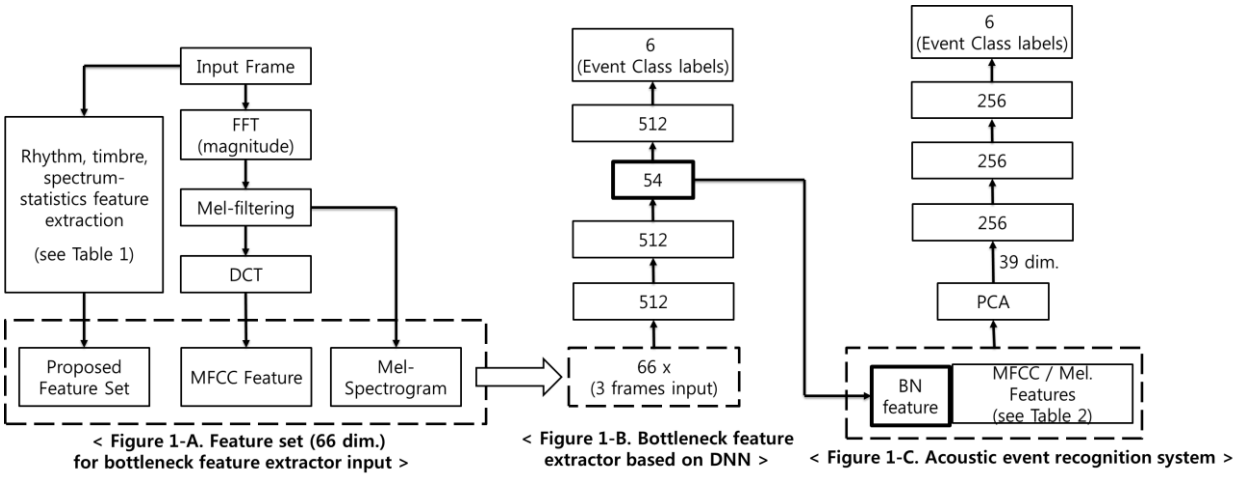

Figure 1: *Proposed acoustic event recognition framework based on bottleneck features.*

Table 1. *Descriptions of selected feature* [18]

Feature	Domain	Description
Attack slope (average in a frame)	Time	A ratio between the signal magnitude difference at the beginning and the ending of the attack period, and the corresponding time difference
Zero crossing rate	Time	Number of times the signal crosses the X-axis (or, in other words, changes sign)
Roll off	Frequency	A feature to find the frequency such that a certain fraction (e.g. 85%) of the total energy is contained below that frequency
Brightness	Frequency	Measuring the amount of energy above cut-off frequency (e.g. 1500Hz)
Roughness [19]	Frequency	An estimation of the sensory dissonance related to the beating phenomenon whenever pair of sinusoids are closed in frequency
Centroid	Frequency	Centroid of frequency spectrum
Spread	Frequency	Standard deviation of frequency spectrum
Flatness	Frequency	A ratio between the geometric mean and the arithmetic mean of frequency spectrum

2.2. Bottleneck features using DBN-DNN

Bottleneck features were extracted from a DNN [13, 18] in which one of the internal layers has a small number of hidden units relative to the size of the other layers. The DNN to extract the bottleneck features is shown in Figure 1-B. The DNN configuration 512-512-54(BN)-512-6 was used, and the selected features were normalized based on each feature's property. Mel-spectrogram (40 dim.), MFCC (18 dim.) and selected features (8 dim.) were extracted from a frame (total 66 dimensional-vector) and stacked with 3 adjacent frames. Therefore, 198-dimensional (66 x 3) vector was used as input layer of bottleneck feature extractor. In this paper, the number of hidden layers (including the bottleneck layer) is set to 4. The number of hidden units in the innermost layer is smaller than those in the other layers. This layer is called the bottleneck layer.

In the pre-training step, we trained each layer of the RBM (Restricted Boltzmann Machine) (Gaussian-Bernoulli RBM for first layer) to construct a DBN using the common DBN training [8]. With the pre-training step, the DBN achieved better initial values of the neural network. This structured bottleneck layer could be treated as a nonlinear mapping of input features.

After the pre-training step, this paper used the acoustic event labels as the target signal. DNN's can be trained by back propagating derivatives of a cost function that measures the cross entropy between the target outputs and the actual outputs produced for each training case. After supervised training, last two layers and activation function of bottleneck layer were removed. Finally, the bottleneck features extracted from the bottleneck layer were used to train the acoustic event classifier.

3. Experimental settings and results

3.1. Acoustic event recognition system description

In order to evaluate effectiveness of the proposed method, 5 different input feature sets were used. MFCC/Mel spectrogram features from a frame were extracted and stacked with 3 adjacent frames. The feature set configurations are listed on Table 2.

PCA was applied to compute the final 39-dimensional vector for feature set de-correlation [19, 20]. DNN was considered as a classifier and was trained by event DB (Table 3) added by white noise at SNR in 5, 10, 15 and 20. As shown in Figure 1-C, DNN classifier with three hidden layers of 256 units each was trained to achieve high classification accuracy for the output layer of the 6 acoustic events.

Table 2. *Feature set configuration*

Input Feature Set	Number of dimension
Mel. (Mel-scaled spectrogram)	Features from a frame : 40 stacked with 3 adjacent frames : 40 x 3 = **120**
MFCC + Δ	MFCC(18)+ Δ(18) = 36 36 x 3 = **108**
Proposed method	
BN (Bottleneck) features	BN features from 3 adjacent frames : **54**
BN + Mel.	BN features from 3 adjacent frames : 54 Mel. features stacked with 3 adjacent frames : 40 x 3 = 120 54 + 120 = **174**
BN + MFCC	BN features from 3 adjacent frames : 54 MFCC stacked with 3 adjacent frames : 18 x 3 = 54 54 + 54 = **108**
All feature sets are compressed to **39-dimensional vector** by PCA	

Table 3. *The sound event database and its size*

Event label	**# of data (3 seconds files)**
Dog bark	312
Male/female scream	302
Breaking glass	292
Siren	257
Car skidding sound	289
Whistle	299

3.2. Database description

The database consists of 6 events collected in various locations by a portable recorder with a high performance microphone and a wind screen (listed on Table 3). The average length of each data is 3 seconds. The database for model training and test were down-sampled from 44.1 kHz to 16 kHz sampling rate with 16-bit quantization in a mono-channel.

For comparing noise robustness of the feature sets, three noise types (cafeteria, business office, and crossroad) were chosen from ETSI background noise database [21] and added to the event database at SNR in 5, 10, 15 and 20 dB by using the ADDNOISE library implemented in [22].

All experiments were conducted under mismatched condition and their results were tabulated in terms of average recognition rate for 6 sound events through 5-folds cross validation test.

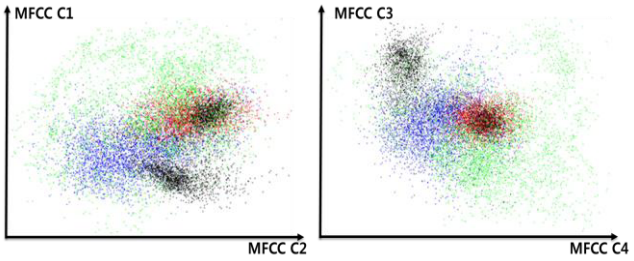

Figure 2: *Examples of MFCC feature's distributions (Red: breaking glass, Blue: dog bark, Black: whistle, Green: siren)*

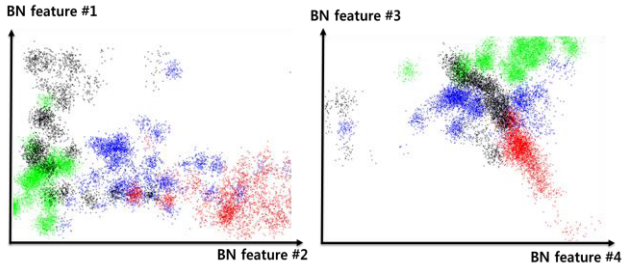

Figure 3: *Example distributions of the proposed bottleneck features*

3.3. Experimental results and discussion

The distribution comparison between the MFCC features and the bottleneck features is shown in Figures 2 and 3. The interclass distances between the proposed bottleneck features are clearly farther than conventional approach.

Experimental results for the comparison between the conventional and the proposed methods are shown in Table 4-6. The proposed methods outperformed conventional methods (14.1%, 12.6% and 11.7% improvement at 5 dB SNR compared to MFCC under cafeteria, office and crossroad noise, respectively) when the noise condition is worse, whereas the performance under relatively high SNR condition (20dB) are similar. Bottleneck features alone showed noise robust recognition under low SNR condition but slightly degraded under high SNR condition, compared to conventional MFCC/ Mel-spectrogram features alone. The combinations of bottleneck and conventional features, designed to improve recognition performance, consistently resulted in higher accuracy compared to bottleneck features alone and other conventional features alone under any SNR condition.

Table 4. *Average recognition rate [%] in 'Cafeteria' background noise according to various features*

Input feature set	SNR [dB] (Cafeteria background noise)				
	5	10	15	20	Avg.
Mel.	71.4	78.8	90.3	96.2	84.2
MFCC+ Δ	70.1	80.1	91.5	96.1	84.4
BN	81.3	87.3	91.6	95.8	89.0
BN + Mel.	83.1	89.6	**93.1**	**97.4**	**90.8**
BN+MFCC	**84.2**	**90.2**	91.5	94.6	90.1

Table 5. *Average recognition rate [%] in 'Office' background noise according to various features*

Input feature set	SNR [dB] (Office background noise)				
	5	10	15	20	Avg.
Mel.	72.3	82.2	90.2	**97.8**	85.6
MFCC+ Δ	72.6	82.6	91.6	97.2	86.0
BN	84.5	83.8	91.4	95.2	88.7
BN + Mel.	84.6	**90.6**	92.5	**97.8**	91.4
BN+MFCC	**85.2**	90.4	**93.8**	96.4	**91.5**

Table 6. *Average recognition rate [%] in 'Crossroad' background noise according to various features*

Input feature set	SNR [dB] (Crossroad background noise)				
	5	10	15	20	Avg.
Mel.	69.2	76.8	89.6	95.6	82.8
MFCC+ Δ	70.8	77.6	88.5	95.4	83.1
BN	81.9	84.6	90.3	95.2	88.0
BN + Mel.	81.3	86.6	**91.3**	**95.8**	**88.8**
BN+MFCC	**82.5**	**87.2**	90.4	95.0	**88.8**

4. Conclusions and future work

This paper proposed a novel framework of acoustic event recognition that uses bottleneck features derived from DNN. In addition to conventional features (MFCC/ Mel-spectrogram), the novel feature set was selected by analyzing chi-square statistics and information gain. To the best of our knowledge, this is the first use of bottleneck features in acoustic event recognition. Based on the additional features and bottleneck frame work, the proposed method showed overall improved acoustic event recognition performance.

Additional work will investigate improved methods for finding effective DNN bottleneck structure and optimizing learning algorithm. These are key issues for applying the proposed framework in large scale acoustic event recognition applications.

5. References

[1] W. Choi, J. Rho, D. K. Han, and H. Ko, "Selective background adaptation based abnormal acoustic event recognition for audio surveillance", *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 118-123, 2012.

[2] K. Yamano and K. Itou, "Browsing audio life-log data using acoustic and location information," *3rd IEEE Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 96-101, 2009.

[3] M. Xu, C. Xu, L. Duan, J.S. Jin and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans on Multimedia Computing and Communications and Applications*, vol. 4, no. 2, pp.1-23, 2008.

[4] W. Choi, S. Park, D. K. Han and H. Ko, "Acoustic event recognition using dominant spectral basis vectors", *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, pp. 2002–2006, 2015.

[5] C. Ludena and A. Gallardo "Acoustic event classification using spectral band selection and non-negative matrix factorization-based features" *Expert Systems with Applications*, vol. 46, pp. 77-86, 2016.

[6] J. Dennis, et al., "Temporal coding of local spectrogram features for robust sound recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 803-807, 2013.

[7] J. Ye, et al., "Robust acoustic feature extraction for sound classification based on noise reduction." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5944-5948, 2014.

[8] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups" *IEEE Signal Processing Magazine*, vol. 29, no. 6 (2012): pp. 82-97, 2012.

[9] Y. LeCun, Y. Bengio and G. Hinton., "Deep learning" *Nature* vol. 521, no.7553, pp.436-444, 2015.

[10] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, vol. 313, no. 5786, pp.504-507, 2006.

[11] L. Deng, et al., "Recent advances in deep learning for speech research at Microsoft", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8604-8608, 2013.

[12] D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search" *Nature*, vol. 529, no. 7587, pp.484-489, 2016.

[13] D. Yu and L. Michael, "Improved bottleneck features using pretrained deep neural networks", *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, Proceedings*, pp.240-243, 2011.

[14] S. Yaman, J. Pelecanos and R. Sarikaya, "Bottleneck features for speaker recognition", *IEEE Odyssey'12*, pp. 105–108, June 25-28, Singapore, 2012.

[15] P. Matejka, et al., "Neural network bottleneck features for language identification", *IEEE Odyssey'14,* pp. 299-304, Joensuu, Finland, 2014

[16] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization ", *Proceedings of the 14th international conference on machine learning ICML '97, San Francisco,* CA, USA, pp. 412–420. 1997.

[17] M. Hall, et al., "The WEKA data mining software: an update" *ACM SIGKDD explorations newsletter,* vol. 11 no. 1, pp. 10-18, 2009.

[18] O. Lartillot, T. Petri and E. Tuomas, "A matlab toolbox for music information retrieval", *Data analysis, machine learning and applications. Springer Berlin Heidelberg*, pp.261-268, 2008.

[19] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth" *The journal of the Acoustical Society of America,* vol. 38, no. 4 pp. 548-560, 1965.

[20] Z. Zhang, et al., "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification" *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015 no.1 pp.1-13, 2015.

[21] European Telecommunications Standards Institute, "ETSI: EG 202 396-1 v1.2.2," 2008.

[22] ITU, "Objective measurement of active speech level," ITU-T Recommendation P.56, 1993.