Interspeech 2018
2-6 September 2018, Hyderabad, India

Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling

Hangting Chen[1,2], *Pengyuan Zhang*[1,2], *Haichuan Bai*[1,2], *Qingsheng Yuan*[3], *Xiuguo Bao*[3], *Yonghong Yan*[1,2]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences,Beijing 100190, China
[2]University of Chinese Academy of Sciences,Beijing 100049, China
[3]National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029 China

chenhangting@hccl.ioa.ac.cn, zhangpengyuan@hccl.ioa.ac.cn, baihaichuan@hccl.ioa.ac.cn, yqs@cert.org.cn,bxg@cert.org.cn, yanyonghong@hccl.ioa.ac.cn

Abstract

Deep learning has improved the performance of acoustic scene classification recently. However, learning is usually based on short-time Fourier transform and hand-tailored filters. Learning directly from raw signals has remained a big challenge. In this paper, we proposed an approach to learning audio scene patterns from scalogram, which is extracted from raw signal with simple wavelet transforms. The experiments were conducted on DCASE2016 dataset. We compared scalogram with classical Mel energy, which showed that multi-scale feature led to an obvious accuracy increase. The convolutional neural network integrated with maximum-average downsampled scalogram achieved an accuracy of 90.5% in the evaluation step in DCASE2016.

Index Terms: Acoustic scene classification, Scalogram, Convolutional neural network, DCASE2016

1. Introduction

Environmental sound carries a large amount of information about surroundings. Acoustic scene classification (ASC) aims to classify the sound into one of predefined classes, e.g., park, office, library[1]. Environment information enables devices and robots to be context-aware.

Acoustic feature plays an import role in ASC tasks. Raw signal is densely sampled in time, thus features are expected to character the sound without throwing away relevant information. Most features are based on the Fourier transform and sophisticated filters[2]. However, the short-time Fourier transform (STFT) is confronted with time-frequency resolution trade-off[3]. Furthermore, sound information is stored in different time scales. Pitch and timbre is at the scale of milliseconds, the rhythm of speech and music is at the scale of seconds[4]. Wavelet transform enables to sense signal at different time scales. Based on the needs of ASC task, we can independently apply wavelet filters and generate effective multiscale features. Named after spectrogram, the visual representation of the spectrum of scales varying with time is called as scalogram. Previous work[5, 6, 7] has shown its potential, but the final performance usually falls behind tradition features[8].

Deep Neural Networks (DNN) have been very successful at ASC[9], image classification[10], gesture detection[11] tasks. In computer vision, convolutional neural network (CNN) has the capability to learn appropriate filters and uncover high-level patterns. However, it remains a big challenge for CNNs to learn acoustic features from raw signal. Mel energy is usually served as CNN input[12, 13], few work has been done in wavelet features.

In this work, we explored CNNs integrated with scalogram to directly classify audio scenes. The raw signal first goes through wavelet filters in different scales, then modulus and downsampling operations to construct scalogram. The CNNs with small kernels are deployed to automatically learn high-level patterns. Compared with the published works, we have achieved the best performance for single systems with an evaluation accuracy up to 90.5% on DCASE2016 dataset.

2. Scalogram

The scalogram is locally translation invariant and stable to time-warping deformation. The properties of effective acoustic features are first reviewed, then the scalogram extraction procedure is introduced.

2.1. Background

Acoustic Features should be time-invariant and stable to time deformation[4, 14]. The former means that the audio segment belongs to the same class even if it is shifted by a constant in time, which can be written as

$$x_c(t) = x(t - c) \tag{1}$$

$$\Phi(x) = \Phi(x_c) \tag{2}$$

where $x_c(t)$ is the signal $x(t)$ shifted by a constant c and Φ transforms the origin signal to audio feature.

Stability to time warping means that small deformation in the raw signal leads to small modification in audio feature, giving

$$x_\tau(t) = x(t - \tau(t)) \tag{3}$$

$$||\Phi(x) - \Phi(x_\tau)||_2 \leq C \sup_t |\tau'| \, ||x||_2 \tag{4}$$

where function $\tau(t)$ denotes time warping satisfying $|\tau'(t)| < 1$ and there exists $C > 0$ representing a measure of stability. The modulus of STFT is translation invariant due to short window function and modulus operation, but not stable to time warping at high frequencies.

Mel scale filter bank coefficients (FBank) is the log power spectrum in Mel scales. The power of STFT ensures time invari-

10.21437/Interspeech.2018-1524

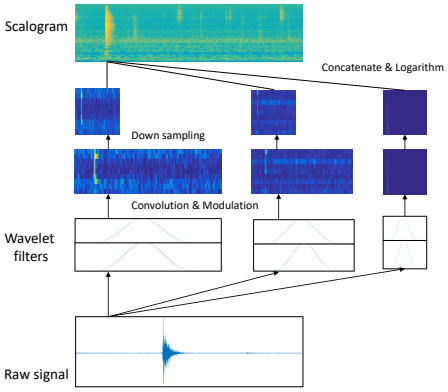

Figure 1: *Schematic diagram of scalogram.*

ant smaller than window length. The Mel filters have a constant-Q frequency bandwidth at high frequencies, which ensures its stability to time warping.

Inspired by FBank feature, we consider wavelet filters, whose support and bandwidth is logarithm spaced, similar to Mel filters. The modulus and pooling operations make feature to be time-shift invariant.

2.2. Extraction Procedure

As shown in Figure 1, the scalogram is defined as the raw signal sequentially passing through wavelet filters, modulus operation, maxpooling or average pooling, logarithm operation.

The center frequency of mother wavelet is normalized to 1. Q denotes the number of wavelets per octave. The dilated wavelets of center frequency λ is written

$$\psi_\lambda(t) = \lambda\psi(\lambda t) \tag{5}$$

$$\widehat{\psi}_\lambda(\omega) = \widehat{\psi}(\omega/\lambda) \tag{6}$$

where $\lambda = 2^{j/Q}$, $j = 0, 1, J$. The maximum scale J is calculated regarding to the maximum window width T of wavelets.

$$J = 1 + round(Qlog_2(\frac{N(T)}{4Q})) \tag{7}$$

$N(T)$ is the number of sample points in window length T. The support of $\widehat{\psi}(\omega)$ is centered at λ with a frequency bandwidth λ/Q; the energy of $\psi(t)$ is centered around 0 with a time width $2\pi Q/\lambda$. The stride of wavelets is inversely proportional to filter's bandwidth, given

$$stride = N(T)2^{-floor(\frac{j}{Q})-1} \tag{8}$$

In our experiment, we used the Morlet wavelet, which is defined as

$$\psi(t) = exp(it)\theta(t) \tag{9}$$

where $\theta(t)$ is a Gaussian filter whose bandwidth is of order Q^{-1}.

After filtered by wavelets, the modulation removes coefficients' phase and only amplitude information is preserved.

Because wavelets in different scales have different strides, a downsampling method is needed to unify coefficients into the same length. We exploit simple maximum and average pooling

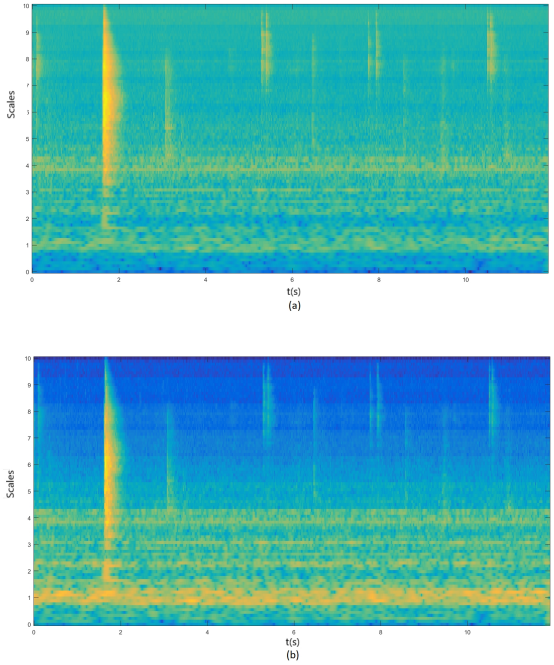

Figure 2: *Scalograms of a 12-second record which belongs to the office scene.(a)Average pooling;(b)Max pooling.*

approaches to downsample redundant coefficients. The average pooling is used to extract mean information; the maxpooling is used to capture the occurrence of strongest amplitude during a frame. For example, if the information is rhyme, average pooling removes unnecessary fluctuation. While if the information is click, maxpooling may mark the transient event. Figure 2 represents the deviation of these two downsampling approaches. The max-pooled scalogram has a higher contrast. Note that at around 1.6s, something dropped on the floor. At around 5.5s, 7.6s and 10.8s, someone clicked the mouse. We can directly seek out these events on the scalogram.

Due to its variable time-frequency resolution, FFT is applied to process the entire sequence which consumes more computation compared with windowed FFT. The extraction of scalogram becomes slow when the signal is long enough. In practice, it is recommended to split long signal with a fixed time interval, which should be much longer than wavelet's time width.

3. Methods

3.1. Dataset

All experiments were conducted on the dataset of ASC task provided by the IEEE challenge on Detection and Classification of Acoustic Scenes and Events 2016 (DCASE2016)[1]. The dataset includes development (Dev.) and evaluation (Eva.) part. The development dataset contains 15 acoustic scenes, 78 recordings for each scene, totaling 9.75 hours of WAV files (Dual Channel, Sample Rate: $44100Hz$, Resolution: 24-bit, Duration: 30 seconds). The evaluation dataset contains the same acoustic scenes as the development part, 26 recordings for each scene, totaling 3.25 hours in the same WAV format. The performance of proposed systems was first evaluated by

the mean accuracy of 4-fold cross validation on the development dataset (CV in Dev.), and then by the test accuracy on the evaluation dataset (Acc. in Eva.). Note that the testing models were trained based on the whole development dataset.

3.2. Features

We mainly created two sets of features using different signal processing methods. The FBank feature is based on windowed FFT and hand-tailored Mel filters. The scalogram is based on wavelet transform and simple downsampling methods.

3.2.1. FBank

FBank feature was extracted in contrast to scalogram. STFT was applied on the raw signal every $20ms$ over $40ms$ windows firstly. Then the coefficients were computed through 40 Mel-frequency filter banks. The delta and delta-delta coefficients to characterize variance among frames were calculated at a 9-order window both in left and right context. The dimension of FBank feature for a 30-second record is $1499 \times 2 \times 120$, where each dimension stands for frame numbers, two channels and FBank coefficients correspondingly. FBank was evaluated both on DNN and CNN.

3.2.2. Scalogram Feature

The scalogram was derived following section 2.2. The maximum window length was set as $T = 370ms$, recommend in [4]. The resolution Q was determined by cross-validation procedure using simple DNN. Two downsampling methods, average pooling and maxpooling, were experimented in parallel. The dimension of scalogram feature for a 30-second record is $162 \times 2 \times wavelet_num$, where the last dimension is determined by wavelet resolution Q.

3.3. DNN

Simple feed-forward neural networks were used to evaluate the effectiveness of features at first. The feed-forward network had an input layer, 3 hidden layers of 512 nodes respectively and a softmax output layer. Each hidden layer was composed of linear transform, batch normalization[15], ReLU[16]. This simple feed-forward network is referred as simple DNN in this paper.

3.4. CNN Architecture

Small convolutional kernels combined with maximum or average pooling enable CNN to learn high-level features. The convolution and pooling operations were conducted only on frequency/scale axis in our experiment. It was assumed that scalogram and FBank contained long-time information. For scalogram feature, the time width between frames was about $186ms$. As for FBank, the differential window of delta and delta-delta coefficients was about $380ms$.

Table 4 lists the CNN layers in order. For example, the first Conv layer represents a convolutional kernel with 2 input channels, 4 output channels and size of 3; the first Pooling layer represents a pooling kernel of size of 2. Batch normalization was applied both in convolutional and linear layers. The activation function was ReLU. The pooling methods of CNN were in accordance with the way used in the input scalogram. Every output of convolutional layers, as well as input scalogram, was concatenated into one vector, then fed to fully-connected layers, finally a softmax layer with 15 units. Scalogram were labeled and trained in a frame-wise way. To test an unknown recording, each frame's log-softmax output was summed up and then the corresponding maximum was the answer.

Table 1: *CNN model*

Input scalogram $162 \times 2 \times wavelet_num$
2×3 Conv(pad-0,stride-1)-4-BN-ReLu 2 Pooling(pad-1,stride-2)
4×3 Conv(pad-0,stride-1)-8-BN-ReLu 2 Pooling(pad-0,stride-2)
8×3 Conv(pad-0,stride-1)-16-BN-ReLu 2 Pooling(pad-0,stride-2)
16×3 Conv(pad-0,stride-1)-32-BN-ReLu 2 Pooling(pad-0,stride-2)
Concatenate input and each Conv output Flatten
Linear (512 units)-BN-ReLu
Linear (512 units)-BN-ReLu
Linear (512 units)-BN-ReLu
15-way Softmax

4. Results

4.1. FBank

We explored the classical FBank feature with simple DNN and CNN (Table 2). The CNN architecture is described in Section 3.4 . We found that CNN gave rise to the system performance.

Table 2: *Experiments on FBank*

Model	**CV in Dev.**(%)	**Acc. in Eva.**(%)
DNN	76.5 ± 2.2	86.4
CNN	77.9 ± 4.7	88.2

4.2. Scalogram

The set of wavelet filters are determined by the resolution Q. A larger Q generates a number of wavelets containing more frequency information, but the redundancy may mislead CNN. We explored different sets of wavelets with the same simple DNN. At $Q = 9$, the average-pooled scalogram gave a best CV. accuracy, while max-pooled showed little variation (Table 3). We set $Q = 9$ for its high CV. accuracy and small standard deviation.

Table 3: *DNN experiments on resolution Q*

Q	**Filter Num**	**Pooling**	**CV in Dev.**(%)	**Acc. in Eva.**(%)
8	84	Max	83.8 ± 2.3	87.2
9	92	Max	83.6 ± 1.4	88.0
10	101	Max	83.7 ± 2.0	88.5
8	84	Ave	82.9 ± 2.6	87.2
9	92	Ave	$\mathbf{84.5 \pm 2.8}$	87.4
10	101	Ave	83.6 ± 3.1	88.5

After adding convolutional layers, the CV. in Dev. and Acc. in Eva. were improved(Table 4). Here we observed the perfor-

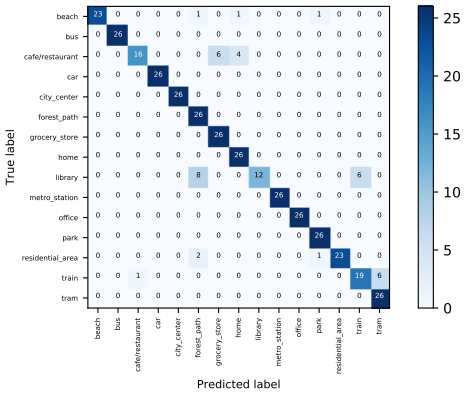

Figure 3: *Confusion matrix for best system on evaluation dataset*

mance gap between maxpooling and average pooling. Take the office scene as an example. In the cross validation procedure, the maxpooled scalogram with maxpooling CNN achieved a 98.6% accuracy in office scene; the average-pooled scalogram with average pooling CNN achieved a mean accuracy of 94.7%, which may indicate that maxpooling is good at capturing transient information and office scene may exist many short-term events.

Table 4: *DNN and CNN experiments on Scalogram with* $Q = 9$

Model	**Pooling**	**CV in Dev.(%)**	**Acc. in Eva.(%)**
DNN	Max	83.6 ± 1.4	88.0
DNN	Ave	84.5 ± 2.8	87.4
CNN	Max	85.8 ± 1.7	88.5
CNN	Ave	84.5 ± 2.3	89.7
CNN	Max & Ave	$\mathbf{85.8 \pm 2.8}$	**90.5**

Furthermore, two sequences of convolutional layers were deployed for average-pooling and maxpooling scalogram in parallel. Then the concatenated output was fed into fully-connected layers, which gave our best performance, 85.8% for CV. in Dev. and 90.5% for Acc. in Eva.(Table 4). Here we presented the confusion matrix of the best system on Eva. dataset(Figure 3).

5. Discussion

This study proposed a novel strategy using CNN combined with scalogram. It achieved an accuracy up to 90.5% for single system on the evaluation dataset in DCASE2016. As far as we know, the result has exceeded all submitted ASC systems in DCASE2016, even the fusion systems(Table 5).

Table 5: *ASC accuracy of art-of-state models on DCASE2016*

Classifier	**Feature**	**CV in Dev.(%)**	**Acc. in Eva.(%)**
CNN	Scalogram	85.8	90.5
Fusion[17]	MFCC	89.9	89.7
NMF[18]	Spectrogram	86.2	87.7
CNN[19]	FBank	79.0	86.2
SVM[20]	MFCC distribution	78.9	85.9

It is believed that wavelet-filter-based features outperform STFT-based features because wavelet can filter signal in a multi-scale way. We deployed CNN to further extract high-level information. The architecture of convolutional layers and scalogram is similar to the scattering representation in [4], but the CNN layers learn proper filters itself. Furthermore, convolutional operation can be used to avoid frequency deformation. The frequency of sound may exhibit small fluctuations related to various acoustic sources, and the convolution operation contributes to stabilize it.

The two downsampling strategies represent two different information, long-lasting and transient. We combined them with CNN to generate whole acoustic patterns. Our scalogram has relatively simple extraction procedure and few hand-tailored filters. Though the system is not strictly end-to-end, an interesting future direction is to use embedded wavelet filter in CNN with more reasonable downsampling approaches.

6. Acknowledgements

This work is partially supported by the National Key Research and Development Plan (Nos.2016YFB0801203, 2016YFB0801200), the National Natural Science Foundation of China (Nos.11590770-4, U1536117, 11504406, 11461141004), the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No.2016A03007-1), the Pre-research Project for Equipment of General Information System (No.JZX2017-0994/Y306).

7. References

[1] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference*, 2016, pp. 1128–1132.

[2] M. F. Mckinney and J. Breebaart, "Features for audio and music classification," in *Ismir 2003, International Conference on Music Information Retrieval, Baltimore, Maryland, Usa, October 27-30, 2003, Proceedings*, 2003.

[3] S. Mallat, *A Wavelet Tour of Signal Processing*. China Machine Press, 2010.

[4] J. Andn and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[5] Y. Mallet, D. Coomans, J. Kautsky, and O. De Vel, "Classification using adaptive wavelets for feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 10, pp. 1058–1066, 1997.

[6] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *The Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[7] S. Amiriparian, N. Cummins, M. Freitag, A. Qian, R. Zhao, V. Pandit, and B. Schuller, "The combined augsburg / passau / tum / icl system for DCASE 2017," DCASE2017 Challenge, Tech. Rep., September 2017.

[8] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," DCASE2017 Challenge, Tech. Rep., September 2017.

[9] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," pp. 410–414, 2017.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[11] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, *Multi-scale Deep Learning for Gesture Detection and Localization*. Springer International Publishing, 2016.

[12] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8614–8618.

[13] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *International Joint Conference on Neural Networks*, 2017.

[14] J. Bruna and S. Mallat, *Invariant Scattering Convolution Networks*. IEEE Computer Society, 2013.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," pp. 448–456, 2015.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 315–323.

[17] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[18] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[19] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[20] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," DCASE2016 Challenge, Tech. Rep., September 2016.