Sound-Event Classification Using Pseudo-Color CENTRIST Feature and Classifier Selection

Jianfeng Ren, Xudong Jiang and Junsong Yuan

School of EEE, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798.

ABSTRACT

Sound-event classification often extracts features from an image-like spectrogram. Recent approaches such as spectrogram image feature and subband-power-distribution image feature extract local statistics such as mean and variance from the spectrogram. We argue that such simple image statistics cannot well capture complex texture details of the spectrogram. Thus, we propose to extract pseudo-color CENTRIST features from the logarithm of Gammatone-like spectrogram. To well classify the sound event under the unknown noise condition, we propose a classifier-selection scheme, which automatically selects the most suitable classifier. The proposed approach is compared with the state of the art on the RWCP database, and demonstrates a superior performance.

Keywords: Sound-Event Classification, Pseudo-Color CENTRIST Feature, Classifier Selection

1. INTRODUCTION

Sound-event classification has recently gained the interest of the research community, which classifies the non-speech environmental sounds into one of the known classes.[1,2] It has many applications, such as acoustic surveillance,[3] environmental sound classification[4] and machine hearing.[5] In this paper, we address the challenge of sound-event classification in a noisy environment.

Time-frequency analysis such as spectrogram[6,7] well captures the power distribution of sound events, and hence is often used in sound-event classification. Many recent approaches[1,2] treat spectrogram as an image and apply image-processing techniques. However, the spectrogram is not a natural image, but a synthetic image. The differences between spectrogram and natural image are not fully explored and hence existing approaches cannot well capture the texture information of spectrogram. Local binary pattern (LBP)[8] was often used to capture image texture information. Many LBP variants have been developed,[9–20] among which CENTRIST[10] is one of the most popular features. It is often extracted from a gray-level image. To better capture the texture information, we propose to extract pseudo-color CENTRIST from spectrogram. More specifically, the gray-level spectrogram is transformed into RGB channels by pseudo-color mapping, and CENTRIST feature is extracted from each channel. CENTRIST features from three channels are concatenated as the final feature vector.

In many applications, the sound events occur in the presence of a wide variety of challenging noise conditions. The noise significantly distorts the spectrogram. In a recent approach,[2] a noise mask is estimated and used to discard the distorted regions of the spectrogram. However, useful information may be discarded as well. To address this issue, we propose a classifier-selection scheme, which automatically selects the most suitable classifier to classify the testing sample under the unknown noise condition. This technique significantly boosts the classification performance for sound-event classification.

2. PROPOSED PSEUDO-COLOR CENTRIST FEATURE FOR SPECTROGRAM

Time-frequency analysis has often been applied on the audio signal for sound-event classification, among which the logarithm of Gammatone-like spectrogram provides a rich texture representation. More specifically, we use a bank of 50 filters for Gammatone-like spectrogram, with center frequencies equally spaced between 100 Hz and $\frac{1}{2}f_s$ on the equivalent rectangular bandwidth scale, where f_s is the sampling frequency, e.g. $f_s = 48$ kHz in this paper. The Gammatone-like spectrogram is denoted as $S(f,t)$, where f is central frequency and t is the time index. To enhance the low-power elements and obtain more texture information, we take its logarithm:

Send correspondence to Jianfeng Ren: E-mail: jfren@ntu.edu.sg, Telephone: +65-67905018.

First International Workshop on Pattern Recognition, edited by Xudong Jiang, Guojian Chen, Genci Capi, Chiharu Ishii, Proc. of SPIE Vol. 10011, 100111C
 · doi: 10.1117/12.2242357

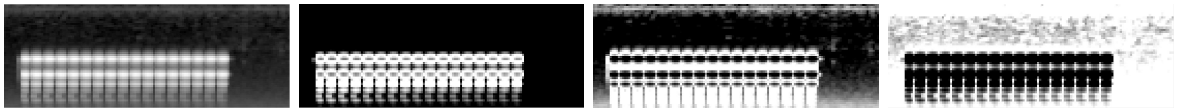
$G(f,t) = \max\{\log S(f,t), G_{min}\}$, where G_{min} is a small constant to avoid very small value of $G(f,t)$. $G(f,t)$ is then normalized to $[0,1]$ as $I(f,t) = \frac{G(f,t)-G_{min}}{\max_{f,t} G(f,t)-G_{min}}$.

There are many differences between the spectrogram and the natural image, e.g. image micro-structures such as edges, spots and corners commonly appear in a natural image, but may not appear in the spectrogram. HOG feature[21] that mainly captures edge information may not be suitable for the spectrogram. In contrast, LBP[8] can capture not only the common image micro-structures, but also other micro-structures by encoding the signs of the relative intensity of a pixel to its neighbors. CENTRIST feature[10] is a LBP variant with improved noise-robustness, and has been widely used in scene classification. It is often extracted from the gray-level image. To better capture the texture information, we propose to extract pseudo-color CENTRIST from the spectrogram.

To enhance the texture presentation, we apply pseudo-color quantization on the spectrogram. We divide the dynamic range of the spectrgram into three parts, and each part is encoded as a color channel. We use standard pseudo-color mapping function "Jet", which maps the intensity value $I(f,t)$ into one of RGB channels as:

$$Q(I(f,t)) = \begin{cases} \frac{I(f,t)-l_1}{l_2-l_1} & \text{if } l_1 < I(f,t) < l_2, \\ 1 & \text{if } l_2 \le I(f,t) \le u_1. \\ \frac{u_2-I(f,t)}{u_2-u_1} & \text{if } u_1 < I(f,t) < u_2, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\{l_1, l_2, u_1, u_2\}$ are quantization parameters, e.g. $\{\frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{9}{8}\}$, $\{\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}\}$ and $\{-\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}\}$ for RGB channels, respectively. Fig. 1 shows an example of spectrogram and its pseudo-colormapped images. As each channel emphasizes one part of the dynamic range of $I(f,t)$, more texture details are visible than the original image.



(a) Original spectrogram (b) Red channel (c) Green channel (d) Blue channel

Figure 1. Original spectrogram and its RGB channels.

The proposed pseudo-color CENTRIST feature for spectrogram is summarized in Fig. 2. We first perform time-frequency analysis on the audio signal and derive the logarithm of Gammatone-like spectrogram. To enrich texture details of the spectrogram, the image is pseudo-colormapped into RGB channels. CENTRIST feature is then extracted from each channel. In CENTRIST,[10] a spatial pyramid is used to divide image into patches at different scales. The spectrogram does not have the scale variations for an object as a natural image does, but have large variations in time. We thus divide the spectrogram into patches in frequency axis at its original scale so that the extracted feature is less sensitive to time variations. CENTRIST features of all channels are then concatenated as the final feature vector.

3. NOISE HANDLING BY CLASSIFIER SELECTION

One of the challenges of sound-event classification is robust to background audio noise. Jonathan et al.[2] assumed that for the clean samples of the RWCP database,[22] the first 60 milliseconds mainly contain silence (known as silence assumption). Thus, for a noisy sample, the first 60 milliseconds of the signal can be viewed as noise. A noise mask is then estimated and used to mask off the unreliable image regions of the subband-power-distribution image. However, the useful information residing in the distorted image regions is discarded at the same time.

To tackle this problem, we propose a classifier-selection scheme. We utilize the silence assumption to estimate the signal-to-noise ratio (SNR) instead of the noise mask, and then use the estimated SNR to determine the most suitable classifier to classify the testing sample. Fig. 3 illustrates the proposed classifier-selection scheme. During training, one classifier is trained for each noise condition using the respective training samples. During testing, we estimate the noise level of the testing sample, select the most appropriate classifier based on the estimated SNR. The core idea is to classify the testing sample using the most suitable classifier, i.e. the one trained using the samples under the similar noise condition. It is an effective approach if the estimated SNR is accurate.

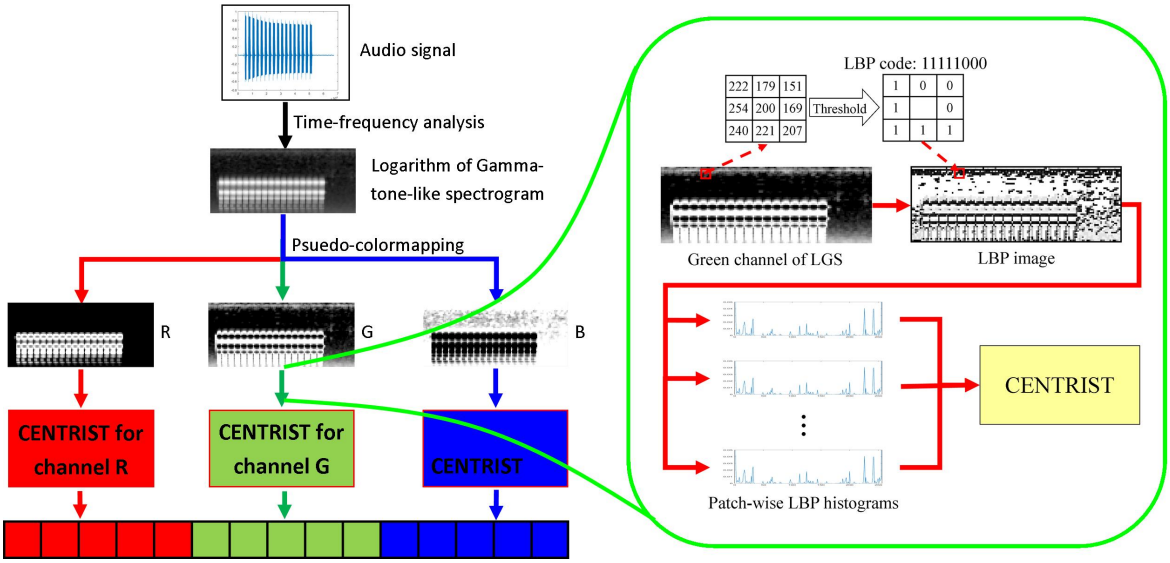


Figure 2. Block diagram of proposed feature extraction.

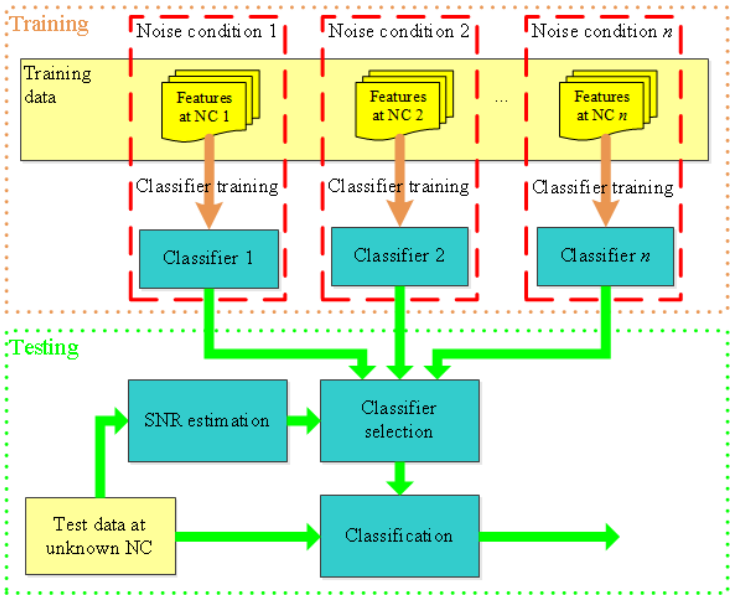


Figure 3. Illustration of the proposed classifier-selection scheme.

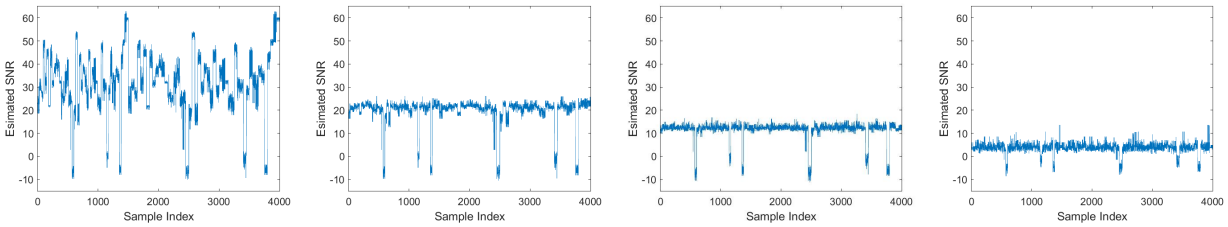
4. EXPERIMENTAL RESULTS

The proposed approach is compared with the following state-of-the-art approaches: Gabor-HMM,[23] MFCC-HMM,[2,24] SIF[1] and SPD-IF[2] on the RWCP database.[22] We conduct experiments under three noise levels besides clean condition, i.e. 20, 10 and 0 dB SNR in four noise environments: "Speech Babble", "Destroyer Control Room", "Factory Floor 1" and "Jet Cockpit 1", obtained from the NOISEX92 database.[25] The noise audio is randomly cropped and injected into the signal. As the number of frequency bands is 50, we divide the spectrogram into 5 patches in frequency axis. The dimension of CENTRIST feature of one image patch is 40. We utilize linear SVM as the classifier and the cost parameter is set as $C = 40$.

The RWCP database[22] consists of in total 9722 sound-event samples of 107 classes. The sound files have a high SNR, and each contains an isolated sound event, with some silence before and after the sound. We use the same experimental setting as in.[1,2] A total of 50 sound-event classes are selected from the RWCP database. For each trial, 50 samples are randomly selected for training and 30 for testing from each class. Overall, there are in

total 2500 and 1500 samples for training and testing, respectively. Then, those samples are injected with noise of SNR 20, 10 and 0 dB in four noise environments. We repeat the experiments 5 times.

The silence assumption is needed for SPD-IF to estimate the noise mask,[2] and also needed for the proposed classifier-selection scheme to estimate the SNR. We thus inspect this assumption on the the RWCP database. We estimate the SNRs for 4000 samples at clean, 20 dB, 10 dB and 0 dB SNR and plot them in Fig. 4, where speech noise is used. For the clean samples, most estimated SNRs are larger than 20 dB, which indicates that the silence assumption is hold for the RWCP database. This is further evidenced by the fact that the estimated SNRs are fairly close to the SNRs of the noise condition, as shown in Fig. 4 (b), (c) and (d).



(a) Clean (b) 20 dB (c) 10 dB (d) 0 dB

Figure 4. Estimated SNRs for 4000 samples of the RWCP database under clean, 20 dB, 10 dB and 0 dB SNR conditions.

The comparisons with the state of the art in terms of the average recognition rate over 4 noise environments and over 5 trials are summarized in Table 1. The recognition rates of Gabor-HMM,[23] MFCC-HMM,[24] SIF[1] and SPD-IF[2] are given by.[2] We first compare the performance under the clean condition, where the performance difference is mainly caused by different features. The proposed approach achieves a recognition rate of 99.80%, which performs better than other approaches. Gabor-HMM[23] achieves a second best recognition rate of 99.39%. However, its performance deteriorates fast with an increasing noise level. We then validate the robustness of the proposed approach to background audio noise. The best published result was achieved by SPD-IF,[2] an average recognition rate of 95.95%. The proposed approach boosts the recognition rate to 97.50%, which shows the effectiveness of proposed classifier-selection scheme.

Table 1. Comparison with the state of the art under different noise levels on the RWCP database.

Method	Clean	20 dB	10 dB	0dB	Average
Gabor-HMM[23]	99.39	91.33	92.51	56.48	84.93
MFCC-HMM[24]	97.53	95.43	91.94	67.17	88.02
SIF[1]	91.13	91.10	90.71	80.95	88.55
SPD-IF[2]	98.81	**98.00**	96.63	90.35	95.95
Proposed approach	**99.80**	97.37	**98.17**	**94.68**	**97.50**

5. CONCLUSION

In this paper, we address the challenge of sound-event classification at a noisy environment. Time-frequency analysis has been widely used in sound-event classification. We conduct texture analysis on various spectrograms and find that the logarithm of Gammatone-like spectrogram is most suitable for sound texture analysis. We analyze the difference between the spectrogram and the natural image, and propose to extract pseudo-color CENTRIST feature that can well capture the texture information of the spectrogram. To improve the robustness to audio noise, we propose a classifier-selection scheme, which can well classify a testing samples under the unknown noise condition. The proposed approach is validated on the RWCP database, which shows significant performance improvement compared with the state-of-the-art approaches.

REFERENCES

[1] Dennis, J., Tran, H., and Li, H., "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters* **18**(2), 130–133 (2011).

[2] Dennis, J., Tran, H., and Chng, E., "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. on Audio, Speech, and Language Processing* **21**(2), 367–377 (2013).
[3] Gerosa, L., Valenzise, G., Tagliasacchi, M., Antonacci, F., and Sarti, A., "Scream and gunshot detection in noisy environments," in [*15th European Signal Processing Conf.*], (2007).
[4] Ghoraani, B. and Krishnan, S., "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE Trans. on Audio, Speech, and Language Processing* **19**(7), 2197–2209 (2011).
[5] Lyon, R., "Machine hearing: An emerging field [exploratory dsp]," *IEEE Signal Processing Magazine* **27**(5), 131–139 (2010).
[6] Logan, B., "Mel frequency cepstral coefficients for music modeling," in [*Int'l Symposium on Music Information Retrieval*], (2000).
[7] Qian, S. and Chen, D., "Joint time-frequency analysis," *IEEE Signal Processing Magazine* **16**(2), 52–67 (1999).
[8] Ojala, T., Pietikainen, M., and Maenpaa, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**(7), 971–987 (2002).
[9] Ren, J., Jiang, X., and Yuan, J., "Dynamic texture recognition using enhanced LBP features," in [*IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*], 2400–2404 (2013).
[10] Wu, J. and Rehg, J., "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **33**(8), 1489–1501 (2011).
[11] Ren, J., Jiang, X., and Yuan, J., "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. on Image Processing* **22**(10), 4049–4060 (2013).
[12] Ren, J., Jiang, X., and Yuan, J., "Relaxed local ternary pattern for face recognition," in [*IEEE Int'l Conf. on Image Processing*], 3680–3684 (2013).
[13] Xiao, Y., Wu, J., and Yuan, J., "mCENTRIST: A multi-channel feature generation mechanism for scene categorization," *IEEE Trans. on Image Processing* **23**(2), 823–836 (2014).
[14] Ren, J., Jiang, X., and Yuan, J., "Learning binarized pixel-difference pattern for scene recognition," in [*IEEE Int'l Conf. on Image Processing*], 2494–2498 (2013).
[15] Satpathy, A., Jiang, X., and Eng, H., "LBP based edge-texture features for object recognition," *IEEE Trans. on Image Processing* **23**(5), 1953–1964 (2014).
[16] Ren, J., Jiang, X., Yuan, J., and Wang, G., "Optimizing LBP structure for visual recognition using binary quadratic programming," *IEEE Signal Processing letters* **21**(11), 1346 – 1350 (2014).
[17] Ren, J., Jiang, X., and Yuan, J., "Learning LBP structure by maximizing the conditional mutual information," *Pattern Recognition* **48**(10), 3180–3190 (2015).
[18] Ren, J., Jiang, X., and Yuan, J., "A Chi-squared-transformed subspace of LBP histogram for visual recognition," *IEEE Trans. on Image Processing* **24**(6), 1893–1904 (2015).
[19] Ren, J., Jiang, X., and Yuan, J., "LBP encoding schemes jointly utilizing the information of current bit and other lbp bits," *IEEE Signal Processing Letters* **22**, 2373–2377 (Dec 2015).
[20] Ren, J., Jiang, X., and Yuan, J., "Quantized fuzzy LBP for face recognition," in [*IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*], 1503–1507, IEEE (2015).
[21] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*], **1**, 886–893, IEEE (2005).
[22] Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., and Yamada, T., "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition.," in [*LREC*], 965–968 (2000).
[23] Kleinschmidt, M., "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica* **88**(3), 416–422 (2002).
[24] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing* **28**(4), 357–366 (1980).
[25] Varga, A. and Steeneken, H., "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication* **12**(3), 247–251 (1993).