Analysis of Spectrogram Image Methods for Sound Event Classification

Jonathan Dennis[1], *Huy Dat Tran*[1], *Eng Siong Chng*[2]

[1]Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632
[2]School of Computer Engineering, Nanyang Technological University, Singapore
{jonathan-dennis,hdtran}@i2r.a-star.edu.sg, ASESChng@ntu.edu.sg

10.21437/Interspeech.2014-543

Abstract

The time-frequency spectrogram representation of an audio signal can be visually analysed by a trained researcher to recognise any underlying sound events in a process called "spectrogram reading". However, this has not become a popular approach for automatic classification, as the field is driven by Automatic Speech Recognition (ASR) where frame-based features are popular. As opposed to speech, sound events typically have a more distinctive time-frequency representation, with the energy concentrated in a small number of spectral components. This makes them more suitable for classification based on their visual signature, and enables inspiration to be found in techniques from the related field of image processing. Recently, there have been a range of techniques that extract image processing-inspired features from the spectrogram for sound event classification. In this paper, we introduce the idea and structure behind six recent spectrogram image methods and analyse their performance on a large database containing 50 different environmental sounds to give a standardised comparison that is not often available in sound event classification tasks.
Index Terms: Sound event classification, spectrogram, image processing, environmental sounds

1. Introduction

The task of recognising sounds in noisy and unstructured environments is a major challenge faced by the audio processing field [1]. Recently there has been renewed interest on the topic of "machine hearing" [2], where the aim is to be able to achieve human-like recognition performance across a wide range of sounds and signal-to-noise ratios (SNRs). This can enable a range of novel applications, including acoustic surveillance [3,4], bio-acoustic monitoring [5], and meeting room transcription [6, 7]. However, the scope of the problem is different to that of the traditional ASR field, hence novel solutions are required for this challenging task.

Most conventional approaches for ASR use "frame-based" features, which model the acoustic signal as a series of fixed dimension features extracted from each time frame of the continuous audio. However, there are two significant drawbacks of such frame-based approaches for sound event classification. Firstly, sounds display a wide variety of spectral characteristics, with many examples containing a relatively sparse frequency spectrum, with most of the energy concentrated in a few frequency bands. When noise is present in the signal, the noise will mask some of the spectral information of the sound. Even for the case when the target and the noise do not overlap in the frequency domain, the frame-based feature will contain a mixture of sound and noise information and become very different from the clean training data. Secondly, frame-based features typically require Hidden Markov Models (HMMs) to capture the temporal information of the underlying frames, which uses a first order model based on the transition probability from the previous frame. However, sounds have a much more diverse temporal dependency, such that a more complete modelling of the temporal information should improve the performance.

The approach explored in this paper is to extract information from the time-frequency spectrogram, such that it captures the sound event information in a way that is robust to noise. In particular, one approach that has been attracting attention recently is the use of image processing techniques to extract two-dimensional features from the spectrogram. In this paper, we introduce and explore this idea further, and in addition we analyse the design of a number of recent spectrogram image works [8–14] that have utilised this approach to understand their structure is inspired by knowledge from the field of image processing. We then compare their performance on a large standardised dataset containing 50 different environmental sounds, which provides a fair comparison between the different techniques that is often not available amongst the wide range of sound event classification applications.

The remainder of this work is organised as follows. Section 2 first provides an overview of spectrogram image methods for sound event classification. Section 3 then describes a number of recent approaches that have utilised the technique. Section 4 details the experiments used to compare the performance of these approaches against several conventional baseline systems. Section 5 then concludes the work.

2. Utilising the Spectrogram Image for Sound Event Classification

Through visual inspection of the spectrograms of typical sound events, it is clear that a large amount of information is contained in the joint time-frequency representation. With careful analysis, it is possible to recognise similar sound events based on this visual image information. An example of this is given in Fig. 1, which shows the spectrogram images of a bottle being tapped in both clean and 0dB noise conditions. Here, the spectral information belonging to the sound event is easily distinguished from that of the background noise, as demonstrated by the highlighted areas in the figure. This is due to the consistent appearance of the bottle sound, which contains characteristic peaks and lines that are connected through a common onset corresponding to the moment of impact as the bottle was tapped. On the other hand, the noise forms the background of the images and can easily be ignored.

In the early days of speech research, visual information in the spectrogram was often studied by speech researchers, for example to analyse the phoneme structure of speech [15]. However, "spectrogram reading" has not become an automatic classification method in speech technology, due to the complicated

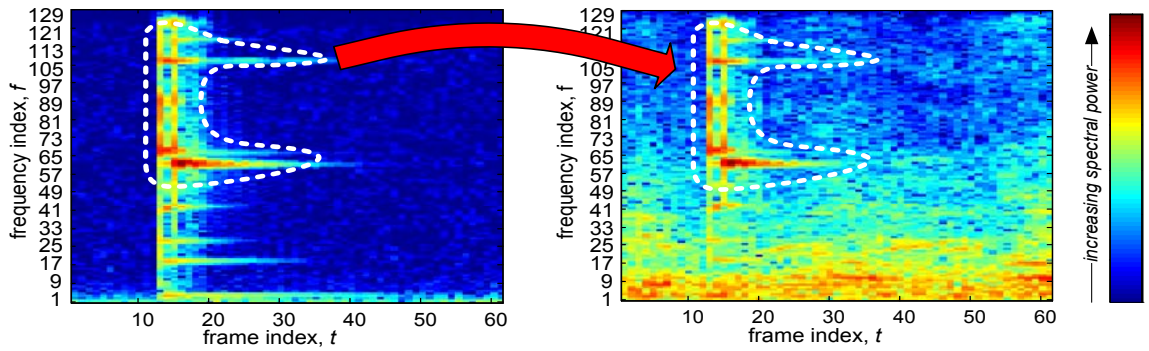

Figure 1: Examples spectrograms of a bottle sound in clean (left) and 0dB babble noise (right). The highlighted area demonstrates how the sound information is still represented clearly in severe noise conditions.

lexicon structures of speech. Unlike speech, with its connected phoneme structures, sound events often have shorter durations but with more distinctive information contained in the time-frequency image representation. It should therefore be possible to extract this information to provide a discriminative feature for classification of the sound event. Such an approach would also represent a significant departure from conventional audio processing. Here, frame-based features such as MFCCs have historically been dominant, but capture only the frequency information within a short time window. Therefore, recently there has been growing interest in capturing joint time-frequency information from the audio signal [16, 17]. Such works have demonstrated the potential advantages of operating in the spectrogram image domain, where the two-dimensional sound event information is naturally represented.

A related field that is concerned with the extraction of two-dimensional information is image processing. Here, two of the most important problems are to detect objects in an image, or classify an image into a predefined category. These problems share many similarities with the those faced in sound event classification, particularly when considering classification based on the spectrogram image. For example, the whole spectrogram image could be classified by extracting low-level pixel information, or alternatively a "sound object" could be detected by finding local correspondences in the spectrogram between training and testing. This therefore opens up the wide range of techniques that have been developed in image processing, which can provide both the inspiration, and a solid basis for developing novel approaches for sound event classification. This has inspired a number of recent works, which are introduced in the next section.

3. Recent Approaches

This section introduces six recent approaches that are inspired by the idea of using visual techniques from image processing. The techniques are grouped according to the scope of the temporal information captured within the feature, as shown diagrammatically in Fig. 2. Also, it should be noted that while several of the approaches have been applied to the task of sound event classification, others are included that have been applied to speech and music tasks. This enables a more complete comparison with a wider range of methods, which are now introduced below.

3.1. Frame-Based Features

These are designed to be similar to conventional features, such that they can combined with regular recognition systems. However, they often utilise the information over a longer segment of the spectrogram, to incorporate more temporal information, in the same way as delta and acceleration MFCC coefficients.

3.1.1. Histogram of Oriented Gradients (HOG)

This approach in [10] extracts HOG features from each frame of the spectrogram at fixed frequency locations. Such features have previously been successful for image processing tasks such as human body detection, since the image pixel gradient implicitly captures information about the underlying shapes present in the image through the direction of the intensity gradients. For the spectrogram, the idea is that the gradient information captures local spectro-temporal shapes in a way that is robust under a range of signal conditions. In particular, it should provide a way of capturing important local temporal information that improves upon the delta features commonly used with MFCCs.

3.2. Global Features

These extract a single feature to represent the whole spectrogram image. This has the advantage that it naturally captures the two-dimensional spectro-temporal information in the audio signal. The disadvantage is the features cannot easily be combined with traditional frame-based recognition systems.

3.2.1. Spectrogram Image Feature (SIF)

The idea of the SIF from [11] is to capture the stochastic nature of the spectrogram image content, using ideas from previous work in content-based image retrieval (CBIR). To achieve this, the method first quantises the dynamic range of the spectrogram into regions, corresponding to the low, medium and high power spectral information in the sound event. This process is analogous to pseudo-colouration in image processing, which is commonly used to enhance the characteristic image information for human perception. However, for machine classification it is found to improve the discrimination of the extracted feature, by allowing higher weights to be assigned to the most reliable spectral regions. The SIF is then formed by capturing the layout and distribution statistics of the quantised spectrogram image, by partitioning each image into blocks, then extracting the pixel distribution statistics from each block to form the feature for classification with SVM.

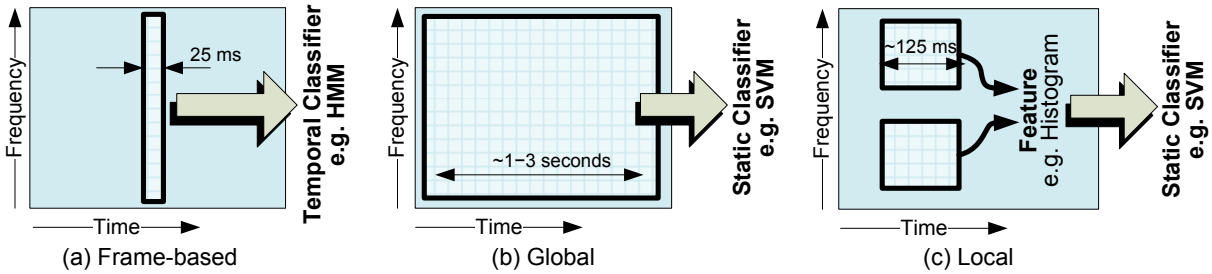

Figure 2: Diagram showing the different types of temporal processing that are commonly applied on the spectrogram.

3.2.2. Subband Power Distribution Image Feature (SPD-IF)

The SPD-IF in [14] builds upon the previous SIF framework, by first transforming the sound event spectrogram into the Subband Power Distribution (SPD) image representation before extracting the image feature. The SPD characterises the spectral power distribution over time in each frequency subband, forming a two-dimensional representation of frequency against normalised spectral power. The key advantage of the SPD representation over the spectrogram is that the signal and noise information are much more easily separated in the SPD representation compared to the spectrogram. This is because the high-powered elements of the sound event are transformed to a localised region of the SPD, enabling a missing feature mask to be generated that can better separate the signal and noise compared to the spectrogram. For classification, the unreliable SPD-IF dimensions are marginalised, and the *k*-nearest neighbours technique is used with the Hellinger distance measure to evaluate the similarity between the pixel distribution information captured in the SPD-IF extraction process.

3.2.3. Spectrographic Seam Patterns

This novel classification approach presenting in [13] is based on transforming the spectrogram into a binary image, by extracting high-energy seams in the image. These form a trace of the characteristic spectral information in the underlying signal, which should be robust in mismatched conditions. The binary seam patterns are then characterised by performing the straight-line Hough Transform on the image. This accumulates information on the shape of the lines in the image, such that it can characterise the underlying spectral peaks. The feature is simply the concatenation of the Hough accumulator bins, with classification performed using SVM. The best performance was achieved using 25 seam patterns, as originally reported in [13].

3.3. Local Features

Here each feature represents a localised time-frequency region of the spectrogram, and are subsequently combined to produce a feature for the segment, for example using a histogram. This has the advantage, as in image processing, that local features may be less susceptible to noise and occlusion than the global feature methods.

3.3.1. Ordered Spectro-Temporal BOVW

This approach combines the techniques presented in [8] and [9] to give a classification approach based on Bag-of-Visual-Words (BOVW) features. The idea in [9] is that local time-frequency patches are extracted at random during training to capture local spectro-temporal information in the sound event spectrograms. During testing, the normalised minimum mean square error (MSE) over the spectrogram is calculated, and the MSE values are used as a feature for classification using SVM. However, we found it advantageous to incorporate the idea in [8] of using capturing temporal information in the ordering of the BOVW patches. This uses a local pooling operator to find the MSE of each patch in the region surrounding its time-frequency location, which provides robustness during matching.

3.3.2. SIFT BOVW

This approach from [12] applies the SIFT feature descriptor from [18] to the tasks of music genre and sound event classification. The technique defines a local 128-dimension HOG feature, which is then clustered using *k*-means to form a BOVW codebook for matching during testing. In the approach here, a SIFT descriptor is extracted from each local 16×16 region on a regular 8×8 grid. The codeword uncertainty, measured against the 4096 visual words in the codebook, is finally used as a feature for classification using SVM.

4. Experiments

In this section, experiments are conducted to analyse the performance of the spectrogram image methods on a standardised database containing a large number of environmental sounds. This is important, because such a comparison is not common in the literature due to the wide variety of applications areas that are covered by sound event classification.

Sound Database: A total of 50 sound classes are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [19], giving a selection of collision, action and characteristics sounds. The isolated sound event samples have a high signal-to-noise ratio (SNR), and are balanced to give some silence either side of the sound. The selected categories cover a wide range of sound events, including wooden, metal and china impacts, friction sounds, and others such as bells, phones, and whistles. For each event, 50 files are randomly selected for training and another 30 for testing. The total number of samples are therefore 2500 and 1500 respectively, with each experiment repeated in 5 runs.

Noise Conditions: For each experiment the classification accuracy is investigated in mismatched conditions, using only clean samples for training. The average performance for each method is then reported in clean and at 20, 10 and 0 dB SNR for the following four noise environments: “Speech Babble”, “Destroyer Control Room”, “Factory Floor 1” and “Jet Cockpit 1”, obtained from the NOISEX’92 database [20]. The average

Group	Method	Clean	20dB	10dB	0dB	Avg.
Baseline MFCC-HMM	ETSI-AFE	99.1 ± 0.2	89.4 ± 3.2	71.7 ± 6.1	35.4 ± 7.7	73.9
	Multi-Conditional (MCT)	97.5 ± 0.1	95.4 ± 1.3	91.9 ± 2.7	67.2 ± 7.3	88.0
Spectrogram Image Processing	Ordered BOVW [8, 9]	94.8 ± 0.6	63.3 ± 9.2	32.7 ± 7.7	12.8 ± 4.0	50.9
	HOG-HMM [10]	$\mathbf{99.2 \pm 0.1}$	68.9 ± 4.6	33.2 ± 7.2	9.3 ± 5.4	52.7
	SIF [11]	91.1 ± 1.0	91.1 ± 0.9	90.7 ± 1.0	80.9 ± 1.8	88.5
	SIFT BOVW [12]	89.0 ± 0.5	45.4 ± 5.2	25.8 ± 4.5	14.6 ± 4.4	43.7
	Spectrographic Seams [13]	87.0 ± 1.6	43.2 ± 7.8	27.7 ± 5.2	15.5 ± 4.0	43.4
	SPD-IF [14]	98.8 ± 0.3	$\mathbf{98.0 \pm 0.3}$	$\mathbf{96.6 \pm 0.4}$	$\mathbf{90.3 \pm 2.0}$	**95.9**

Table 1: Experimental results comparing the classification accuracy of the baseline and spectrogram image methods.

performance across all four noise environments is reported at each SNR.

Baseline Systems: To provide a comparison against the conventional state-of-the-art, two baseline frame-based MFCC-HMM systems are implemented. Both systems use 39-dimension frame-based MFCC features, which consist of the first 13 coefficients with their deltas and accelerations appended. The first baseline pre-processes the raw MFCC features using the ESTI Advanced Front End (AFE), which performs noise reduction at the feature level. The second baseline performs Multi-Conditional Training (MCT) using the raw MFCC features vectors extracted under both clean and 10dB SNR noise for 3 out of the 4 noise environments. These two systems are selected as they are both well established techniques that provide a strong performance across a range of sound event classification tasks.

Experimental Methods: Each of the methods in Section 3 is implemented in Matlab, with the parameters taken from the original papers. For example, the Ordered BOVW implementation uses 1000 patches as in [9], while the HOG-HMM method uses eight 32-dimension HOG descriptors per frame and PCA to reduce the dimension to 50 as described in [10]. Further to this, preliminary experiment were carried out to examine their performance against the previously published results. It was found that each method gave a comparable performance to that reported by their authors on a subset of sound classes from the RWCP database that matched the size of the original dataset. For methods using HMM, the same configuration is used, with 5 states and 6 Gaussian mixtures and both training and testing carried out using the HMM Toolkit (HTK) [21].

Results and Discussion: The experimental results are shown in Table 1. In the first segment of the table, amongst the baseline MFCC-HMM systems it can be seen that the MCT system performs best on average, achieving an average classification accuracy of 88.0% compared to 73.9% for the ETSI-AFE system. The MCT system is also significantly more robust at the lowest SNRs, where the MCT method improves by over 30%.

Comparing the baseline performance against the spectrogram image methods, it can be seen that the frame-based HOG-HMM has the best performance in clean conditions, with an accuracy of 99.2%. However, over all four noise conditions the SPD-IF performs the best, with an average accuracy of 95.6%. This is a significant, as the SPD-IF achieves an absolute improvement of nearly 8% over the MCT baseline. In addition, the SPD-IF can achieve over 90% in the 0dB condition compared to 67.2% for the baseline. This result is even more significant, since the SPD-IF only requires clean data for training, hence should achieve a similar performance across a wide range of noise conditions. On the other hand, the MCT baseline method requires a large amount of data for training, and may not perform well when the noise conditions during testing are different to those observed during the training.

Amongst the spectrogram image methods, there is quite a wide variety in performance, and some of the image-based methods are less robust to mismatched noise. It can also be seen that the SIF, SIFT BOVW and spectrographic seams methods appear to be less discriminative than the other methods, achieving a classification accuracy of only 91.1%, 89.0% and 87.0% respectively in clean conditions. This is much less than the HOG-HMM method, which performs well in matched conditions, but is not robust to noise and also achieves the lowest performance in 0dB SNR conditions. This should be expected, since the local spectral gradients that are extracted to form the feature will be severely corrupted by the noise surrounding the sparse signal information. On the other hand, the SIF is designed to be strongly robust to mismatched noise, which enables it to achieve the second best overall result, with an average accuracy of 88.5%.

Overall, it can be seen that incorporating spectro-temporal information into the feature is beneficial for sound event classification. For example, the performance of the Ordered BOVW approach, which uses a local pooling operator on the time-frequency location of the patches, is significantly better than the unordered SIFT BOVW technique. The best overall performance came from the SPD-IF, which captures the temporal information through the subband distribution of the spectrogram. While the SPD does not capture the precise ordering of the temporal information in the sound, it can produce a robust and discriminative basis for feature extraction that can outperform even the state-of-the-art MCT baseline.

5. Conclusion

In this paper, we analysed the idea of using methods based upon the spectrogram image for sound event classification. The idea is to overcome the drawbacks of the state-of-the-art techniques by naturally capturing the two-dimensional spectro-temporal information in the spectrogram through techniques inspired by image processing. A range of spectrogram image methods were reviewed and implemented for comparison against traditional baseline systems on a standard data set of environmental sounds. The results showed that the approaches could outperform the baseline, with the best overall performance displayed by the SPD-IF technique which was significantly more robust than the other competing approaches.

6. References

[1] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, Oct. 2008.

[2] R. F. Lyon, "Machine Hearing: An Emerging Field," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131–139, Sep. 2010.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2005, pp. 1306–1309.

[4] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2007.

[5] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. H. Tauchert, and K.-H. H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, Sep. 2010.

[6] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems," *IV Jornadas en Tecnologia del Habla*, 2006.

[7] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, Sep. 2010.

[8] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proceedings of the Workshop on Statistical And Perceptual Audition (SAPA)*, 2008.

[9] G. Yu and J.-J. J. Slotine, "Audio classification from time-frequency texture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Apr. 2009, pp. 1677–1680.

[10] T. Muroi, T. Takiguchi, and Y. Ariki, "Gradient-based acoustic features for speech recognition," in *Proceedings of the 2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, no. Ispacs. Ieee, Dec. 2009, pp. 445–448.

[11] J. Dennis, H. D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130–133, Feb. 2011.

[12] K. Behún, "Image features in music style recognition," in *Proceedings of the Central European Seminar on Computer Graphics (CESCG)*, 2012.

[13] S. Barnwal, K. Sahni, R. Singh, and B. Raj, "Spectrographic seam patterns for discriminative word spotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Mar. 2012, pp. 4725–4728.

[14] J. Dennis, H. D. Tran, and E. S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 367–377, Feb. 2013.

[15] V. Zue, "Notes on spectrogram reading," *Mass. Inst. Tech. Course*, vol. 6, 1985.

[16] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[17] B. Ghoraani and S. Krishnan, "Time-Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.

[19] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of the International Conference on Language Resources and Evaluation*, vol. 2, 2000, pp. 965–968.

[20] A. Varga and H. J. M. Steeneken, "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*, vol. 12, no. 3, pp. 247–251, 1993.

[21] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book, version 3.4*, 2006, vol. 3.