Enhanced Local Feature Approach for Overlapping Sound Event Recognition

Jonathan Dennis and Huy Dat Tran
Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632
{jonathan-dennis,hdtran}@i2r.a-star.edu.sg

***Abstract*—In this paper, we propose a feature-based approach to address the challenging task of recognising overlapping sound events from single channel audio. Our approach is based on our previous work on Local Spectrogram Features (LSFs), where we combined a local spectral representation of the spectrogram with the Generalised Hough Transform (GHT) voting system for recognition. Here we propose to take the output from the GHT and use it as a feature for classification, and demonstrate that such an approach can improve upon the previous knowledge-based scoring system. Experiments are carried out on a challenging set of five overlapping sound events, with the addition of non-stationary background noise and volume change. The results show that the proposed system can achieve a detection rate of 99% and 91% in clean and 0dB noise conditions respectively, which is a strong improvement over our previous work.**

I. Introduction

The topic of Sound Event Recognition (SER) covers the detection and classification of sound events in unstructured environments, which may contain multiple overlapping sound sources and non-stationary background noise. Many sounds contribute to the understanding and context of the surrounding environment, and therefore should not be regarded simply as noise, as is common in automatic speech recognition (ASR). Instead, such sounds are useful in many applications, such as security surveillance [1], bioacoustic monitoring [2], meeting room transcription [3], and ultimately "machine hearing" [4].

In this paper, we address the problem of recognising overlapping sound events by utilising the visual information in the time-frequency spectrogram representation of the audio signal. The spectrogram has historically been used to analyse the phonetic structure of speech using a technique known as "spectrogram reading" [5], where a person is able to pick out the important spectral structures and use these to recognise the underlying speech. Despite this, visual-based techniques for automatic classification of speech have not been heavily researched, in part due to the complicated lexical structure. Sound events, on the other hand are typically more sparse and distinct, making the visual information more tractable for automatic classification.

Previous work on recognition of overlapping sounds can be separated into several distinct methodologies. The first are multi-microphone techniques, which use one or more microphone arrays combined with detection and beamforming to isolate specific sounds from the overlapping mixture [6]. However here we focus specifically on the task of recognising overlapping sounds in single channel audio. The second is blind source separation, where factorisation is commonly used to decompose the input signal into its constituent sources. For example, [7] use unsupervised non-negative matrix factorisation (NMF) to process the input audio into four component streams, where different sound events may be separated into different streams for recognition. The recogniser is then applied to all four streams to find occurrences of the 61 trained sound classes. The final group of methods are based on feature-based classification. One approach developed for ASR is Factorial HMMs (FHMMs) [8], based on the MixMax model of source interaction [9], where the best combination of hidden states is found among the trained models to explain the observed feature. However, the combinatorial nature of the problem results in extremely high computational complexity, which limits the number of simultaneous sources that can be recognised. A more recent approach uses hierarchical SVM [3], where the first SVM classifies the input as either isolated events or a combined "overlapped" class, which is then expanded in a second SVM to identify the overlapped combination. This requires sufficient training samples of the overlapped sounds in advance, covering each possible degree of overlap, which may not be available in advance.

Our feature-based system extends our previous work [10] combining local spectrogram feature extraction with the Generalised Hough Transform (GHT) [11] for recognition [12]. The idea is to first extract local features surrounding interest-points in the spectrogram. These features, and their time-frequency locations, are then used to create a model which is used for recognising each sound class. The GHT is used as a scoring mechanism, which maps consistent interpretations of each sound model into local maxima in the Hough space, enabling us to detect multiple overlapping sound events in a single clip. The key in this work is to use the output of the GHT as a feature that can be used to train a classifier to map the output of the GHT to a probability that the segment contains the target sound class. In particular, we use the random forest classifier for this purpose, which is able to avoid overfitting the sparse training data.

The paper is organised as follows. Section II first introduces the idea behind our GHT approach. Section III then details the approach used to map the GHT output into a recognition probability. Section IV describes our experiments and the results obtained. Finally, Section V concludes this work.

978-616-361-823-8

II. Generalised Hough Transform System for Overlapping Sound Event Recognition

The approach presented in this paper takes inspiration from works in the field of object detection from image processing [12], where finding objects in a cluttered real-world scene can be seen as having many parallels with that of overlapping SER. The central idea is to characterise a spectrogram by a set of independent local features, where each feature represents a glimpse [13] of the local spectral information. The GHT is then a summation of the local evidence provided by each local feature, based on a model that learns the distribution of the observed features in the spectrogram for each class during training. Since the GHT voting is additive, a sound can still be recognised even when a proportion of features is missing or corrupted due to noise or overlapping sounds. The representation of each sound in the Hough accumulator space is also sparse and separable, such that overlapping sounds occurring at the same time will produce distinct spikes in the separate accumulators for each class so that both can be detected.

Figure 1 details the steps required for training and testing. It also highlights the contribution in this paper in the shaded boxes, whereby the output of the GHT is treated as a feature and used for training a random forest classifier to map the GHT output into a recognition probability. The three main steps are now summarised in this section, while the details of the proposed scoring system are given in Section III.

A. Local Spectrogram Feature Extraction

We first detect "keypoints" in the spectrogram to locate characteristic spectral peaks and ridges. Keypoints are then detected at locations that are local maxima across either frequency or time, subject to a local signal-to-noise ratio (SNR) criterion to remove the majority of those occurring on the background noise. For each keypoint, we extract an LSF and local missing feature mask to represent the local spectral region. We use a plus-shaped LSF, composed of the local horizontal and vertical spectral shapes in the spectrogram, as this was found to be more suitable than including the full two-dimensional region which may become dominated by non-stationary noise or overlapping sounds.

B. Training

The extracted LSFs are first clustered to generate a codebook, where each entry in the codebook represents characteristic local patterns found in the training samples. After codebook matching, the LSF detected in the spectrogram can considered to be replaced by the matching codebook entry. Each sound event is then modelled through the occurrence distribution of the matched codebook clusters in the training spectrograms, over time and frequency. Note that we do not model the distribution over spectral power as in [10] as the proposed scoring method was found to be sufficiently discriminative. The sound onset is used as a reference point for the temporal information to overcome the problem of time-shifting, while the frequency dimension is considered to be

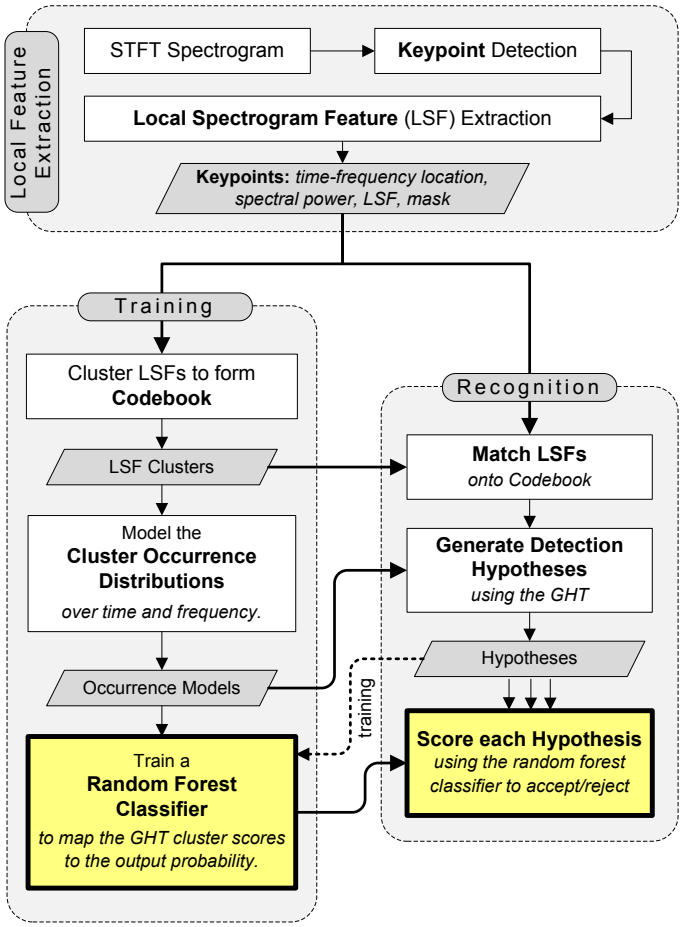

Fig. 1. Overview of the proposed LSF enhanced scoring system, which uses a random forest classifier to map the GHT score to a probability.

fixed. Previously, we would then extract scoring parameters for verification during testing, however here we propose an alternative approach by training a random forest classifier to perform the mapping between the GHT output and the probability, as introduced in the next section.

C. Recognition

The LSFs are first matched onto the codebook, with the best matching codebook selected. We then generate sound onset hypotheses using the Generalised Hough Transform (GHT) [11], which is a voting system that sums the matching keypoint-cluster distribution information in a separate Hough accumulator space for each class. The idea is that all keypoints belonging to the same sound event in the spectrogram will share a common onset, and their time-frequency distribution relative to the onset should match that modelled in the training. Therefore the occurrence distribution model, learned during training, is used as the voting function for the GHT. Here we propose to use the output of the GHT as a feature, and the score from the random forest classifier is used as a metric for accepting the hypothesis.

III. Proposed Probability Mapping and Detection System

In this section, we introduce our proposed method for mapping the output of the GHT into a metric that is suitable for accepting or rejecting each individual sound event hypothesis. We first discuss the drawbacks of the previous knowledge-based approach, before introducing our proposed classification-based approach.

A. Previous Knowledge-based Scoring Approach

The previous approach in [10] was to extract scoring parameters, which were then used as a threshold for hypothesis verification during testing. The first scoring parameter was the voting count of the cluster, to represent the average log-spectral power assigned to each cluster. This should capture how important a particular cluster is for classification of a given class, such that if the cluster is missing the hypothesis will be assigned a lower weight. The second scoring parameter was the cluster score, to represent the relative weight that the cluster contributes to the sound class. Since the cluster score was normalised to sum to unity for each sound class, a threshold could be set for accepting a hypothesis to provides a trade-off between false rejection and acceptance.

The drawback of this approach is that several manually defined thresholds and factors were required to determine whether the clusters could be considered matched in the spectrogram. The optimum thresholds may vary significantly between datasets, hence a more automated approach should improve the generalisation capability of the system.

B. Proposed Random Forest Mapping

Here we propose to take the output of the GHT as a feature for classification, and use the random forest classifier to provide the mapping between GHT output and hypothesis probability score. We can denote the output of the GHT as follows:

$$ght\_output = H^X(k,t) \tag{1}$$

where H represents the Hough accumulator, $k = 1 \dots K$ is the cluster index, t is the current time frame, and X is the given sound class. For each time frame, this is a vector of length K, containing the summation of the cluster distribution voting functions in the GHT. Note that in all of our experiments, we use $K = 200$.

A separate random forest classifier is trained for each class X, using the GHT output for the given class $H^X(k,t)$ as the input feature. The classification is performed frame-wise, and the training labels are generated by finding the frame with the maximum GHT voting score over the training clip. This frame represents the onset of the sound and is assigned a positive class label, along with the 4 adjacent frames. As the training data is sparse, additional training examples are generated by randomly setting 50% of the feature to half of their original value. This is repeated five times, and was found to improve the generalisation ability of the classifier to unseen examples. The output is a probability, and we were able to set a threshold at $\Omega = 0.1$, as the random forest produced few false positives.

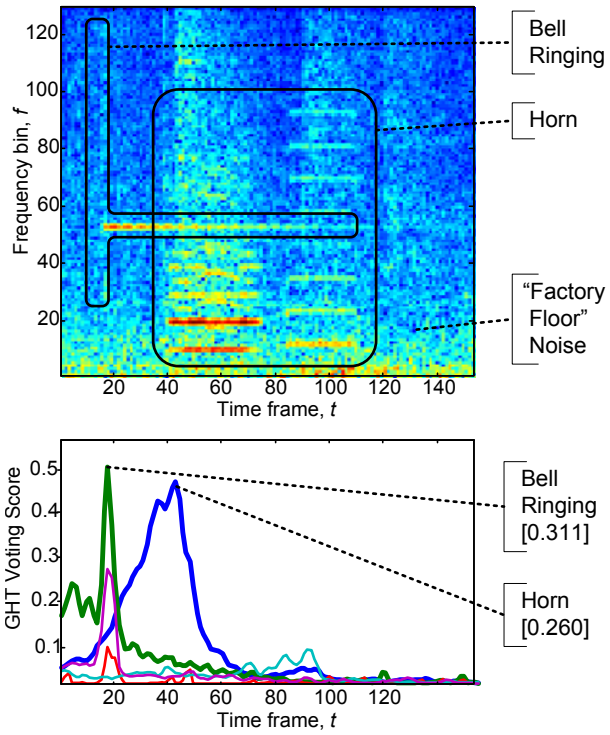

Fig. 2. Example of two sounds overlapping in 0dB Factory Floor noise *(above)* and the output GHT voting score *(below)*. The detection hypotheses for the two sounds can clearly be seen, and the GHT cluster scores surrounding these points are input to the random forest classifier for verification. The verification score for each detection is shown in brackets.

IV. Experiments

Database: For our experiments, we generate a database of overlapping sound events from the Real Word Computing Partnership (RWCP) Sound Scene Database [14]. We select the following five classes: horn, bells5, bottle1, phone4 and whistle1. Amongst the sounds, the bottle1 class contains the most variation, with five different bottles being struck by two different objects, although there is some variation across all classes. From this, we generate all 15 overlapping combinations, each consisting of one or two sound events, using randomly chosen onset times ensuring between $50 - 100\%$ temporal overlap.

Baseline: The first baseline we call MixMax-GMM, which can be seen as a simplification of the FHMM approach [8] using one state, as in [10]. We model the PDF using a 6-component GMM, and take the maximum log-likelihood summed across all frames in the clip as the classification result. The second baseline we call Overlap-SVM, which is based on the approach proposed by [3]. Note that Overlap-SVM also requires 20 samples for each of the 10 overlapping combinations, which we generate from the isolated samples selected for training. Here, the mean and variance of the 60-dimension frame-based features is taken over the clip, giving a final feature with 120 dimensions. Finally, the performance of our previous LSF-based approach from [10] is compared.

Experimental Methods: For training, we randomly select

TABLE I

EXPERIMENTAL RESULTS ACROSS THE VARIOUS TESTING CONDITIONS. THE VALUES FOR TP/FA ($\% \pm$ STD) ARE REPORTED OVER 5 RUNS OF THE EXPERIMENTS. RESULTS HIGHLIGHTED IN BOLD INDICATE THE BEST PERFORMANCE FOR EACH CONDITION.

Experiment Setup		Proposed LSF-RF		LSF		Overlap-SVM		MixMax-GMM	
		TP	FA	TP	FA	TP	FA	TP	FA
Isolated[a]	Clean	99.9 ± 2.7	$\mathbf{0.4} \pm 2.4$	99.3 ± 2.7	0.4 ± 2.4	$\mathbf{100} \pm 0.0$	1.5 ± 3.4	99.6 ± 1.4	1.3 ± 5.8
Overlapping[b]	Clean	$\mathbf{98.9} \pm 3.5$	$\mathbf{0.3} \pm 1.1$	98.0 ± 3.4	0.8 ± 3.6	96.5 ± 7.3	1.3 ± 2.8	84.0 ± 29.3	5.2 ± 17.0
Overlapping[b] + added noise	20dB	$\mathbf{98.5} \pm 5.0$	$\mathbf{0.5} \pm 2.2$	97.2 ± 5.0	0.7 ± 3.2	76.9 ± 39.0	18.6 ± 35.1	52.8 ± 44.9	27.8 ± 42.6
	10dB	$\mathbf{96.9} \pm 7.9$	$\mathbf{0.7} \pm 2.6$	95.5 ± 9.1	0.9 ± 3.5	74.7 ± 40.9	20.9 ± 36.8	37.8 ± 42.9	25.1 ± 41.2
	0dB	$\mathbf{91.0} \pm 14.5$	$\mathbf{0.8} \pm 2.8$	90.2 ± 17.6	2.5 ± 8.2	65.7 ± 41.9	25.8 ± 36.1	22.9 ± 38.8	20.9 ± 35.7
Overlapping[b] + volume change	$\times 0.5$	$\mathbf{98.9} \pm 3.7$	$\mathbf{0.1} \pm 0.6$	98.1 ± 3.0	0.7 ± 3.3	84.0 ± 24.8	1.5 ± 5.0	56.0 ± 43.4	12.4 ± 27.8
	$\times 0.75$	$\mathbf{99.2} \pm 3.1$	$\mathbf{0.1} \pm 0.6$	98.4 ± 2.9	0.5 ± 1.8	92.8 ± 13.1	1.1 ± 2.9	80.6 ± 30.0	4.4 ± 13.7
	$\times 1.5$	$\mathbf{99.3} \pm 3.0$	$\mathbf{0.2} \pm 0.7$	98.4 ± 2.7	0.6 ± 2.1	95.9 ± 9.9	4.0 ± 11.4	82.0 ± 29.8	8.3 ± 21.6
	$\times 2$	$\mathbf{99.4} \pm 3.0$	$\mathbf{0.2} \pm 0.9$	98.0 ± 3.3	0.7 ± 2.0	94.1 ± 14.7	7.0 ± 18.3	68.7 ± 40.7	23.7 ± 39.3
Average		**98.0%**	**0.4%**	97.0%	0.9%	86.7%	9.1%	64.9%	14.3%

[a] Results are averaged over the 5 isolated sound classes
[b] Results are averaged over the 15 overlap combinations

20 clean isolated samples of each sound event from the database. For testing, we generate 50 overlapping samples for each of the 15 overlapping combinations, using samples excluded from the training set. We then investigate the performance under mismatched noise and volume conditions. In particular, we add "Factory Floor 1" noise, from the NOISEX'92 database [15], to the testing samples at 20, 10 and 0 dB SNR. This noise is chosen for its challenging, non-stationary nature. We also change the volume by pre-multiplying the waveform by the factors $\{0.5, 0.75, 1, 1.5, 2\}$ prior to taking the STFT of the signal, to simulate a channel transfer function. As evaluation measure, we calculate the recognition accuracy (TP) and false alarm (FA) over each of the sound classes, over 5 runs of the experiment. TP is calculated as the ratio of correct detections to the number of clips containing occurrences of that class. Analogously, FA is the ratio of incorrect detections to the number of clips not containing that class.

Results: The results show that the proposed LSF-RF hypothesis verification approach performs significantly better than the original knowledge-based approach. The system can now achieve close to 99% TP in clean conditions across the 15 overlapping conditions, and 91% in 0dB noise. The FA is also reduced and is maintained below 1% for all experiments.

An example in Figure 2 shows the challenging conditions present for the overlapping recognition system in 0dB conditions. The GHT voting is able to identify the onset of two sound events in the mixture, while the random forest scoring system is able to convert the GHT output to a probability.

V. CONCLUSION

In this paper, we propose a method for the simultaneous recognition of overlapping audio events. Our approach is based on local spectrogram features and the Generalised Hough Transform (GHT), but treats the output of the GHT as a feature for classification using a random forest classifier, which maps the GHT output into a probability for recognition. The experiments show an improvement over the previous approach, and demonstrate the simplicity of the random forest mapping compared to a knowledge-based approach.

REFERENCES

[1] 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.

[2] 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.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.

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

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

[6] R. Chakraborty and C. Nadeu, "Real-time multi-microphone recognition of simultaneous sounds in a room environment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8672–8676.

[7] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound Event Detection in Multisource Environments Using Source Separation," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, 2011, pp. 36–40.

[8] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of Eurospeech*, vol. 7, Geneva, 2003, pp. 1009–1012.

[9] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[10] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, Jul. 2013.

[11] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[12] A. Lehmann, B. Leibe, and L. Van Gool, "Fast prism: Branch and bound hough transform for object class detection," *International journal of computer vision*, vol. 94, no. 2, pp. 175–197, 2011.

[13] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[14] 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.

[15] 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.