Recognition of Overlapping Sound Events

Jayalaxmi*, Hegde Abhijna Satish, Harshitha N Kotari, and Deeksha

Department of Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangaluru-575007
*Email:jaya06297@gmail.com

ABSTRACT
In this paper, we address the challenge of recognizing the isolated sound in the noisy background. Here we propose an approach where we extract the local spectrogram features for each isolated acoustic sound events. The local spectrogram features are extracted by using the keypoints which are unique for each sound event. The "keypoints" are the peak values for each sound event where the sound is maximum. These local spectrogram features are then clustered to form a codebook. The codebooks are used for training purpose. The features of the sound events which will be used for testing are extracted separately using spectrogram. The extracted feature is then mapped with the local spectrogram features in the codebook to recognize the sound event. The experimental setup has 12 isolated sounds, 12 overlapped sound events, and 11 mixed noises to determine the accuracy of our approach.

***Keywords:* Cluster, Codebook, Keypoint detection, Local spectrogram features, Time Frequency Location, visual word.**

1. INTRODUCTION

In any environment there can be many overlapping sound events which will be present along with many background noises. In many cases the background noise will be as important as the structured sound events, so they cannot be simply neglected and considered as unstructured sound events. In cases like surveillance camera, hearing machine and also automatic speech recognition the unstructured surrounding background noises are as important and useful as the structured sound events. So, the concept of Sound Event Recognition (SER) is used to detect and also to classify the sound events which are present in the unstructured environment. Detecting and classifying these sounds based on the sound events are helpful in case of security cameras, monitoring of bioacoustics, meeting room transcription and is also very helpful in case of "hearing machines".

Different Technologies have been developed regarding sound event recognition, most popular techniques among them are based on frame-based features, such as Mel-frequency cepstral coefficients (MFCCs) from ASR, or MPEG-7 descriptors (Casey,2001). All these techniques can then be modeled with Gaussian Mixture Models (GMMs) and combined with Hidden Markov Models (HMMs) in order for recognition, and it can also be used to train SVM or Support Vector Machine for the different classifications based on the features. But these methods may not perform best in the case of mismatched conditions which occur in sound events recognition tasks.

To overcome these challenges Missing Feature Recognition systems were developed. The task here is to identify how to mask the sound so that it stands out separately from the background noise. The performance of this system depends on how well the mask can separate the sound from background. This technique may not be helpful in case of overlapping sounds as there will be information about two or more sounds.

In a research of humans understanding of speech traces of frame-based feature is found, and it is also found that the human auditory system may be based on the partial feature extraction that are uncoupled and also local across the frequency of the speech. This helps humans to recognize the speech even if there is lot of disturbance and distortion across the different regions of the spectrogram of the sound event. Thus, based on Local Spectrogram Features we develop a Sound Event Recognition system, where we will making use of frame-based features.

Here we try to address a task of simultaneous recognition of the sound events which are from single channel audio. Conventional Frame-based methods cannot be used here as each time the frame will contain information mixed as the sound event will have different sounds or from multiple sources.

Another method which can be used to detect the overlapping sound is Missing Feature Recognition technique. The drawback with this technique is that the recognition of the sounds will be based on the way mask is created or in other words the recognition of sound depends on the accuracy of the mask. Here we try to make use of Local Spectrogram Features which represent the local spectral feature of each sound and this is extracted from spectrogram which is covered with keypoints. "Keypoints" represents the peaks in the spectrogram. Based on these keypoints we can form LSF clusters and their occurrences can be shown using spectrogram.

We have conducted experiments on the isolated sound without background noise, and also on isolated sounds with factory floor noise as background noise with the background noise taken in different decibels, and finally on overlapped sound events.

2. RELATED WORK

Sound event classification is used for applications like security surveillance [1], bioacoustics monitoring [2], meeting room transcription [3] and mainly in machine hearing [4]. Sound events original feature can be extracted by using visual signature which is the representation of sound's frequency. These features can be extracted using spectrogram which is gray scale normalized [5]. Another way of extracting local spectrogram features is by making use of keypoints concept. The keypoints are the peak values of the sound. The local spectrogram features can be extracted by extracting the values around keypoints. These extracted values are unique for each sound. These extracted local spectrogram features along with the label name is used to train the SVM model [1].

To separate the sounds, audio event detection is used. To classify the audio events the system uses two parallel GMM classifier. The classifier is trained initially with the audio features which are obtained using 2-step process. At different signal to noise ratio such as 0dB, 10Db, and 20dB the experiments were done. The approach is applicable to separate two sounds for the noisy background. Firstly, the features are extracted from the audio events and these features are used for analysis [6].

Spectral subtraction is used to separate the noise from the sound events. Training is used to improve the performance of automatic speech recognition. In multi condition training the system is trained in different situations where it can work. Hierarchical spectro-temporal processing is used to extract features from the noisy background [7].

Invariant features present in the sound event does not change its application under any circumstances. These invariant features of the sound should match with the objects in the surrounding. Feature should be correctly matched with features in the database of features of the known sound [8].

Hearing machines are present to detect the speech from the musical environment and background noises. Using these features the machine can recognize the speech of the sound in the noisy environment [9]. Built on the key developments in statistical modelling of natural language processing and involuntary identification schemes, there is extensive submission in works which need a humanoid mechanism interface, such as auto call processing in the telephony system and enquiry-based data system which does work such as providing upgraded portable data, stock rate extracts, climate report [10].

Based on the description of the probabilistic mixed prototype for a frame of speech the recognition of speech is performed. Every part of prototype is the naming stage of Hidden Markov model grounded identification of speech [11].

3. IMPLEMENTATION

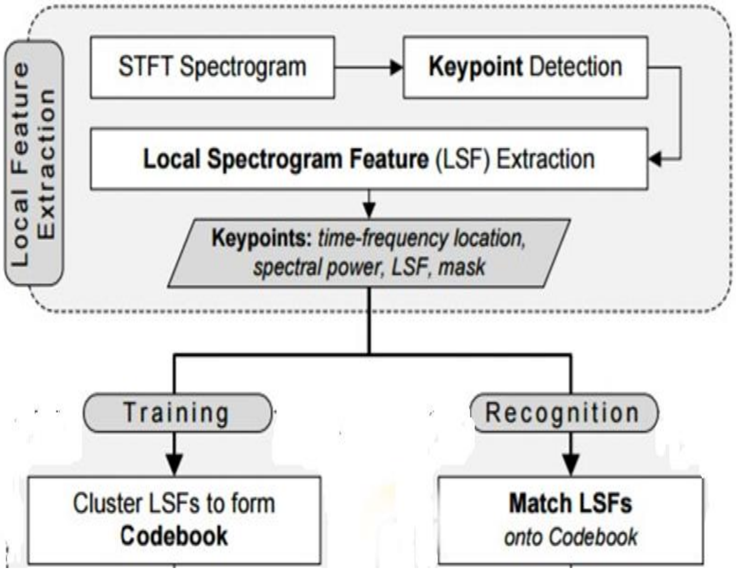
Local Feature Extraction
STFT Spectrogram
Keypoint Detection
Local Spectrogram Feature (LSF) Extraction
Keypoints: time-frequency location, spectral power, LSF, mask
Training
Recognition
Cluster LSFs to form Codebook
Match LSFs onto Codebook

Figure 1. Overview of proposed method

3.1 Keypoint detection

The feature of the sound is first extracted using MATLAB and is stored into mat file. This mat file is then converted into a csv file. This csv file is then given as an input to spark program. The spark program identifies the peak values of the given sound. The identified peak values are then used to detect the keypoints corresponding to the sound. These keypoints are used to extract local spectrogram features.

The Keypoint Detection is summarized as follows:

Algorithm: Detecting Keypoints

```
Input: csv file contenting the feature of the sound.
Output: Detected Keypoints for that sound.
1. Extract the values from .csv file and split them as comma-separated value.
2. Make frames of size -6 to 6.
3. Find max for each frame.
       max(float(x) for x in x.split())
4. Find the sum of each frame which is divided by 40.
5. Keypoints should be greater than max value and sum value.
6. Print the keypoints.
```

The frame size -6 to 6 and the sum value divided by 40 gives more accurate Keypoint values.

3.2 Local spectrogram feature extraction

Local spectrogram features for a sound are the values surrounding the keypoints. Once the keypoints are detected we can then extract the local spectrogram features using these keypoints. The extracted local spectrogram features is then grouped into clusters. Clustering is done using K-Means clustering algorithm. This makes use of Euclid's distance to form the clusters. The Euclid's distance formula is as follows:

$$\sqrt{(x_2^2 - x_1^2) + (y_2^2 - y_1^2)} \qquad \text{…. (1)}$$

Where (x_1,y_1) and (x_2,y_2) are the co-ordinates of keypoint selected for clustering. The clusters are grouped based on the distance obtained using Euclid's formula. In K-Means, K refers to number of clusters. In our approach we are making use of 500 clusters which are formed using 50 iterations

3.3. Codebook

For each of the clusters formed mean is calculated which is called visual word. The collection of all those visual word is called codebook. Then we are matching extracted features or the input sound with this codebook to predict the sound event. Fig 2 shows the generation of codebook.

The extracted local spectrogram features for the input is matched with the codebook line by line and features with least distance will be plotted in the histogram along with the corresponding labels. The label for which the histogram value is highest is predicted as the output. Fig 3 shows the plotting of histogram. In the figure, x-axis represents labels and y-axis represents values.

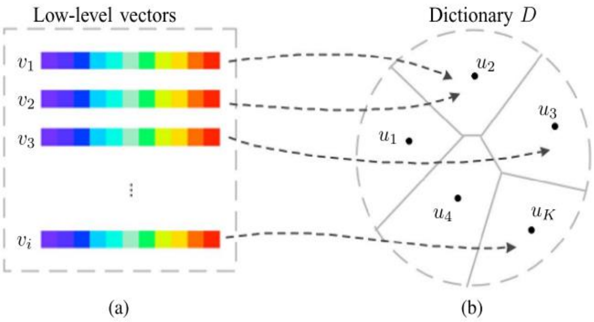

Figure 2. Codebook generation

3.4. Training and testing

From each sound class 32 sounds are taken for training and 8 sounds are taken for testing. During training the features and the labels are given to support-vector machine (SVM).

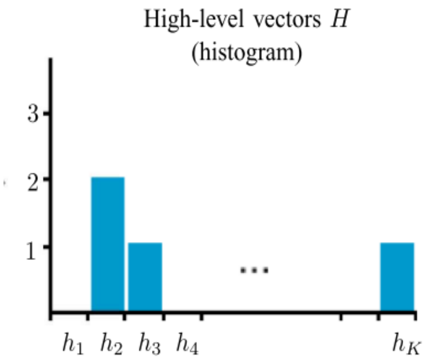

Figure 3. Histogram

During testing only, features are given and the corresponding labels are given as output from trained SVM. To classify the sound events and to predict the accuracy SVM Chi-squared Kernel is used. Chi-squared Kernel is very popular for training non-linear SVM. The Chi-squared Kernel is calculated as follows:

$$k(x, y) = \exp\left(-\gamma \sum_i \frac{(x[i] - y[i])^2}{(x[i] + y[i])}\right) \qquad \text{…. (2)}$$

Where x and y need to be non-negative and should be normalized.

Accuracy is calculated by predicting how correctly the sound is recognized. SVM classifiers are used to maintain a balance between accuracy of training and the strength of the classifier.

4. EXPERIMENTAL RESULTS

4.1. Datasets

For our experiment we are using the following 12 classes of sound: Applause, Cup Jingle, Chair Moving, Cough, Door Slam, Key Jingle, Knock, Keyboard Typing, Phone Ringing, Paper Work, Steps, and Laugh. For convenience in table 2 it is represented as Ap, CJ, CM, Co, DS, KJ, Kn, KT, PR, PW, St, La respectively. Each of this single class consists of 40 sounds. Firstly, we have considered all the above classes in its isolated form amongst which 32 sounds were used for training and 8 sounds were used for testing. Next we mixed the above isolated sounds with the factory floor noise which is from NOISEX'92 database. The sounds were mixed in 0dB, 10dB, and 20dB. Next, we have considered single sound from 1 class (i.e. Laugh) and mixed with other classes to form mixed sound events.

4.2. Results

As estimation we are measuring the accuracy of the recognized sound events. The results of our experiment can be found in Table I. As shown in the Table 1 isolated sounds have more accuracy compared to overlapped sounds. The factory floor noise of 20dB has more accuracy compared to the 10dB and 0dB.The sound events with 20dB has less noise compared to the 0dB.Mixed events have less accuracy compared to the other overlapped sound since there will be confusion to recognize the sounds. Here we are taking a single sound from single class and mixing with all the other 11 classes. Our approach finds it difficult to distinguish the two sound events present in the mixed environment. In Table 2 has a confusion matrix for mixed sound events where the confusion gives the clear picture of how well the sound has been recognized correctly and where it has been mis-predicted. In table 2 we have taken confusion matrix of mixed sound events that has given the accuracy of 71.591%. In this 8 applause sounds that are given for testing is correctly recognized as applause itself. But in cup jingle sound only 4 out of 8 sounds were recognized correctly, remaining 4 sounds are mispredicted one as chair moving, one more as keyboard typing and the remaining 2 as paper work. Because of these types of mispredictions, the accuracy has been decreased.

5. CONCLUSION

In this paper a technique to recognize the event is proposed in overlapped noisy form. Our motivation is derived from social observation, which has been recommended for human listening built on confined evidence, and also after picture entity identification that will make parallels with overlying SER. The methodology we made use of is to discover keypoints in the spectrogram, later portray the sound conjointly via the LSF and the key-point dispersal in relation with the sound inception. Further as future deed, our goal is to improve the accuracy of mixed sound events. Also the work may comprise reconstruction of the recognized acoustic events.

Table 1 Experimental Result under various test conditions

Experimental setup	**Acoustic events**				
	Isolated Sound Events	***Noisy sound***			***Mixed Sound Events***
		0 dB	***10 dB***	***20dB***	
Accuracy	88.542	74.583	79.167	82.292	71.591

Table 2 Confusion Matrix for mixed events

	Ap	CJ	CM	Co	DS	KJ	Kn	KT	PR	PW	St
Ap	8	0	0	0	0	0	0	0	0	0	0
CJ	0	4	1	0	0	0	0	1	0	2	0
CM	0	0	5	1	0	0	0	0	1	1	0
Co	0	0	0	7	0	1	0	0	0	0	0
DS	0	0	0	0	8	0	0	0	0	0	0
KJ	0	0	0	1	0	6	0	0	0	1	0
Kn	0	0	0	0	1	0	6	0	0	0	1
KT	0	1	0	0	0	1	0	4	0	2	0
PR	0	0	0	3	0	0	1	1	3	0	0
PW	0	1	0	0	0	1	0	0	0	5	1
ST	0	0	0	1	0	0	0	0	0	0	7

ACKNOWLEDGMENT

This research was supported by Sahyadri College of Engineering and Management. We are grateful to Mr. Sunil B.N and Dr. Pushpalatha K who moderated this paper and in that line improved the manuscript significantly.

REFERENCES

[1] Gerosa, L., Valenzise, G., Antonacci, F., Tagliasacchi, M., Sarti, A., 2007. Scream and gunshot detection in noisy environments, in: 15th European Signal Process. Conf. (EUSIPCO-07), Sep. 3–7, Poznan, Poland.

[2] Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K., Frommolt, K., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognition Lett. 31, 1524–1534.

[3] Temko, A., Nadeu, C., 2009. Acoustic event detection in meeting-room environments. Pattern Recognition Lett. 30, 1281–1288.

[4] Dennis, J., Tran, H., Li, H., 2011. Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process. Lett. 18, 130–133.

[5] Dennis, J., Tran, H., Chng, E., 2012. Overlapping sound event recognition using local spectrogram features with the generalised hough transform, in: Proc. Interspeech 2012.

[6] Heckmann, M., Domont, X., Joublin, F., Goerick, C., 2011. A hierarchical framework for spectro-temporal feature extraction. Speech Comm. 53, 736–752.

[7] Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vision 60, 91–110

[8] Lyon, R., 2010. Machine hearing: an emerging field. IEEE Signal Process. Mag. 27, 131–139.

[9] Nádas, A., Nahamoo, D., Picheny, M., 1989. Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoustics Speech Signal Process. 37, 1495–1503.

[10] O’Shaughnessy, D., 2008. Invited paper: automatic speech recognition: history, methods and challenges. Pattern Recognit. 41, 2965–2979.