Real-world Acoustic Event Detection using All-Pole Group Delay and Subband Spectral Features

Chandrasekhar Paseddula

Electronics Communication Engineering
G. Narayanamma Institute of Technology and Science, Hyderabad, Telangana, India,

Abstract. In this paper, we introduce various features as acoustic event representations in audio, including log-Mel band energies (log-Mel), sub band spectral flux coefficients (SSFCs), spectral centroid magnitude coefficients (SCMCs), linear prediction cepstral coefficients (LPCCs), and all-pole group delay (APGD) features. The process of acoustic event detection (AED) in real life audio is modeled using the Gaussian Mixture Model (GMM). Various aspects for auditory event detection are briefly experimentally reviewed in this study.
We performed fourfold cross validation experiments on the TUT sound events 2016 development dataset. The phase information from the phase spectrum was left out in favor of using common spectral features as an alternative source of audio magnitude information because it is more difficult to comprehend than the magnitude information. According to the APGD method, the features deduced phase information. In comparison to the baseline mel frequency cepstral coefficients (MFCCs) and the other proposed features, the performance (F-score) of the APGD features is improved in the experimental investigations. Comparing the error rate (ER) of LPCC features to the baseline MFCC features and the other suggested features, LPCC features showed a lower ER.

Keywords: log mel band energies (log-Mel), subband spectral flux coefficients (SSFCs), spectral centroid magnitude coefficients(SCMCs), all-pole group delay (APGD), Gaussian mixture model (GMM.)

1 Introduction to Acoustic Event Detection

Acoustic event detection (AED) is the process of identifying certain audio occurrences. For instance, the sound events in an acoustic scenario include “a car driving by”, “a bird singing”, “keyboard typing”, “a coughing”, “a door knocking”, etc. It can be used for a variety of things, including context-based indexing and multimedia databases for information retrieval, sound monitoring in healthcare, security, and audio surveillance. Systems that use acoustic event detection can be divided into two categories: monophonic and polyphonic. The most prominent sound occurrences in the audio signal are trained to be detected by monophonic systems. In addition to identifying the most prominent sound

event in a segment, polyphonic systems also distinguish all the overlapping sound events [1].

The nature of sound occurrences can be described both spectrally and temporally. The spectrum energy in a specific frame are captured using the common Mel frequency cepstral coefficients (MFCC)[5]. In order to distinguish between harmonic and non-harmonic sound occurrences, sound events were classified spectrally [8]. It has been demonstrated that non-negative matrix factorization (NMF) is helpful for simulating the basic spectral features of sources hidden in an auditory scene and may also support overlapping event [16].In 2016, mel frequency cepstral coefficients were utilized as representations of acoustic events in the detection and classification of acoustic scenes and events (DCASE) challenge task3 (sound event detection in real-life audio) [15, 18].

The robust sound classification method was presented using spectrogram image features and Kullback-Leibler kernel support vector machines (SVM) [9].For AED, hidden Markov models (HMM) and spectro-temporal Gabor filter bank features were suggested [25]. For the AED challenge, deep neural networks (DNN) and frame-wise spectral-domain characteristics were proposed [6]. Log mel-band energies and a convolutional recurrent neural network (CRNN) model were used to implement the AED problem in [26]. The amplitude information of the sound unit is always the emphasis of spectral features, although phase information is also thought to be important for accurate sound unit identification. Direct processing of the phase spectra is challenging due to the phase wrapping and its dependence on the position of the window. Utilizing the group delay function from all-pole models can solve this issue. We can determine the group delay function from all-pole models using linear prediction. This technique has been applied to speaker recognition [22], formant extraction [29], etc,.

In this paper, all-pole group delay (APGD) characteristics and subband spectral features be used to characterize acoustic events. compared to traditional log mel band energies, which capture complementing information. The additional pertinent data from the phase spectrum is also taken into consideration in light of this. Acoustic events might occasionally overlap in real life rather than always happening in isolation. At the appropriate audio segment subbands, the rich information is provided. Using this as our inspiration, we developed the spectral centroid magnitude coefficients (SCMCs) and subband spectral flux coefficients (SCFCs) for audio representations of acoustic events. Acoustic events development dataset from DCASE 2016 is used to assess performance.

The inclusion of phase into the detection of acoustic events is a novel aspect of our investigation. In addition, spectral characteristics based on subbands are introduced for acoustic event recognition in practical audio tasks. The implementation of the AED task is then shown in the Figure 1.

The training and testing of the GMM model is carried out using the retrieved acoustic features and pre-emphasized audio output from the figure. To forecast the events in an acoustic scene using a trained GMM model, thresholding techniques are used in the post-processing step.

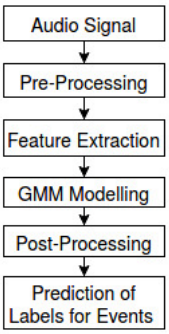

Fig. 1. Block diagram for the illustration of AED task implementation.

The remainder of the paper is laid out as follows: The baseline system is described in Section 2 together with the dataset for AED. The different recommended features extraction is described in Section 3. Section 4 of this paper provides a description of the GMM model and its parameters. The experimental setup is provided in Section 5. Results and analysis are presented in Section 6. The summary of the studies is provided in Section 7.

2 Experimental Dataset and Baseline System Description

In this section, 2016 development database and the baseline system description is presented as follows:

– TUT Sound events 2016 development dataset
This dataset contains recordings that were made in a variety of settings, including residences and streets. A 3-5 minute long audio recording was taken at each place for the recording, and it was then divided into 30 second audio tracks. The recording tools include a Soundman OKM II Klassik in-ear microphone and a Roland Edirol R-09 wave recorder. 44.1 kHz sample rate and 24 bit resolution were used to record audio tracks [18]. The selected sound event classes are:
1) Home: Under this class, the following events were considered and those are; Rustling, Cupboard, Snapping, Cutlery, Drawer, Dishes, Glass jingling, People walking, Object impact, Water tap running, and Washing dishes.
2) Residential area: Under this class, the following events were considered and those are; Banging, Car passing by, Bird singing, Children shouting, People walking, People speaking, and Wind blowing.
The dataset comprises of 22 audio recordings from two acoustic situations, and they are as follows: 10 recordings totaling 36 minutes and 16 seconds were used for the class on homes (interior environments), and 12 recordings totaling 42 minutes were used for the class on residential areas (outdoor environments).

– 2016 AED task baseline system description
The base system uses the GMM model as the classifier and the MFCC features as the representation. The MFCC static coefficients, delta coefficients, and acceleration coefficients are acoustic features that capture the spectral and dynamic properties of the sound in time. For each event class and scene, a binary classifier was created. A negative model is trained using the remaining audio, and the scene GMM model is trained using the audio chunks tagged as relating to the modeled event class. The choice is based on the likelihood ratio of the positive and negative models for each specific scene, with a sliding window of one second in length. This baseline system was used on the TUT Sound events 2016 development dataset with 4-fold cross-validation, where MFCC features are extracted once every frame with a frame duration of 40 ms and 20 ms overlap. A frame has 20 MFCC static coefficients, 20 MFCC delta coefficients, and 20 acceleration coefficients. Per frame, a feature vector with a total of 60 values was taken into consideration. For both the positive and negative sound event models, there are 16 Gaussians. The event label is forecast using the maximum likelihood method. [18].

3 Proposed Features for AED task

In this section, we first presented the numerous cepstral features extraction and their importance for AED representations. Second, we suggested using phase information to recognize sound events. The purpose of all pole group delay (APGD) calculation is specific to the execution of AED task. The group delay function's calculation and the extraction of APGD characteristics are next described in detail.

3.1 log-Mel band energies

The audio signal is represented by the log-Mel band energies since the human ear perception system models these qualities. The audio envelope in a certain segment with respect to time was captured using these features. As follows is the extraction of the log-Mel band energies: By employing windowing and frame blocking, the audio signal is divided into short duration frames. The power spectrum of each frame should be estimated using a periodogram. The energy in each filter should then be added after applying the Mel filter bank on the power spectrum. Add up all the filter bank energies, then take the logarithm. In order to properly credit Narayanan [10], [20], these factors are known as log-Mel band energies. In the Figure, the extraction of the log-Mel band energy features is shown 2.

In this study, we created a 120 dimensional feature vector per audio frame utilizing 40 mel filter banks to extract 40 static log-Mel band energies, 40 delta, and 40 double delta features. By taking into account 40 ms frame duration with 20 ms frame overlap over a 30 sec audio stream, we ultimately obtained $120X9077$ feature matrix per audio track.

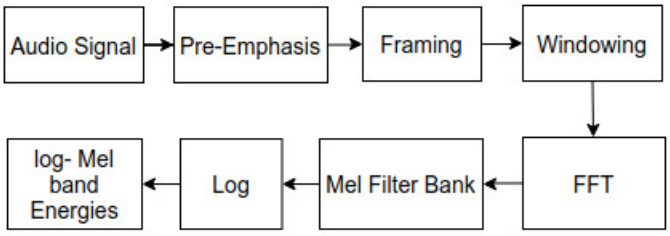

Fig. 2. Block diagram for log-Mel band energies extraction.

3.2 SSFCs

As spectral flux can be used to determine the timbre of an audio signal, which describes the characteristics of sound and allows the ear to distinguish sounds that have the same pitch and loudness, the subband spectral flux coefficients (SSFCs) are used to represent the audio signal.

By comparing the power spectrum for one frame with the power spectrum from the previous frame, the spectral flux, a measurement of how quickly a signal's power spectrum is changing, is estimated. [24, 27].
Spectral flux can be used to determine how the power spectrum changes from frame to frame. The Euclidean distance between the normalized power spectra of successive frames is used to calculate the spectrum flux. The SSFC features are extracted as follows: Computation of the subband spectral flux (SSF) of the j-th sub-band of the t-th audio frame is given by:

$$SSF_t^j = \sum_{k=1}^{N/2+1} \left\| \tilde{X_t}(k) - \tilde{X_{t-1}}(k) \right\|^2 w_j(k). \quad (1)$$

Where $\tilde{X_t}(k)$ is the magnitude of k-th frequency component of normalized power spectrum of t-th frame.

$w_j(k)$ is the spectral window function to obtain the frequency response of the j-th subband, and N is the number of bins in discrete Fourier transform (DFT). SSFCs are then obtained by performing logarithm and DCT on SSFCs [19, 23]. These coefficients are termed as SSFCs. Then the SSFC features extraction is depicted in the Figure. 3.

In this work, 40 static SSFCs, 40 delta, and 40 double delta features were extracted from each audio frame using 40 linear filter banks, resulting in a 120-dimensional feature vector overall. We obtained a $120X9077$ feature matrix representation for a specific audio track by taking into account a frame duration of 40 ms and a frame overlap of 20 ms for a 30-second audio input.

3.3 SCMCs

In order to capture the brightness of the audio segment, the spectral centroid of a subband, which contains the spectral centroid magnitude of the audio segment,

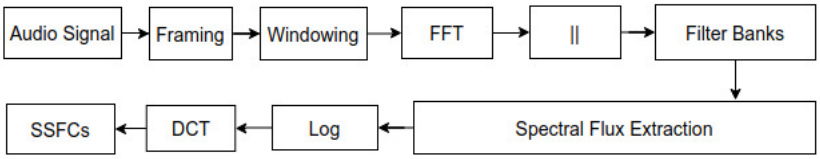

Fig. 3. Block diagram for SSFC features extraction.

is utilized to calculate the spectral centroid magnitude coefficients (SCMCs), which are used to describe the audio signal.

The SCMC features extraction as follows: For the j-th subband of the t-th audio frame, the spectral centroid magnitude is computed by using equation:

$$SCM_t^j = \frac{\sum_{k=1}^{N/2+1} g(k) X_t(k) w_j(k)}{\sum_{k=1}^{N/2+1} g(k) w_j(k)}. \quad (2)$$

Where $X_t(k)$ and $g(k)$ represent the power spectrum magnitude of t-th frame and normalized frequency ($0 \leq g(k) \leq 1$)corresponding to k-th frequency component [13, 23] and and N is the number of bins in discrete Fourier transform (DFT). Then the log and DCT operations are performed on SCM to get SCM coefficients (SCMCs). These SCMC features are encode average energy well in a subband. Due to that, subband information of audio well captured by SCMCs. Then the SCMC features extraction is depicted in the Figure. 4.

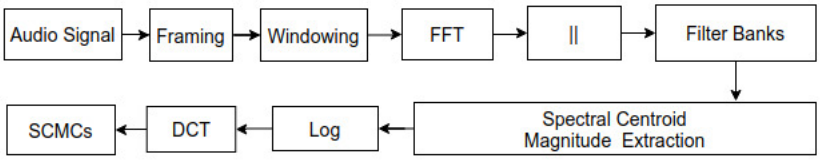

Fig. 4. Block diagram for SCMC features extraction.

In this study, we extracted 40 static SCMCs, 40 delta, 40 double delta features are considered per audio frame using 40 Bark filter bank and a total of 120 dimensional feature vector formed per frame. For each audio track, we got $120X9077$ feature matrix representation by considering 40 ms frame duration with 20 ms frame overlap over a 30 sec duration audio signal.

3.4 LPCCs

Since these features are employed to extract the environment production system characteristics from the audio signal, the linear prediction cepstral coefficients

(LPCCs) are used to represent the audio signal. The following features of the LPCC are extracted: Frame blocking and windowing are used to divide the audio signal into brief frames. Calculate each frame's auto-correlation. Utilize the all pole filter model with an order of 40 and perform the linear prediction analysis from the correlation coefficients. Use the Levinson and Durbins algorithm to solve the auto-correlation equations. These LPC parameters are transformed into cepstral domain. Extract the LPCCs for a given frame [12, 14]. Then the LPCC features extraction is illustrated in the Figure. 5. In this study, we extracted

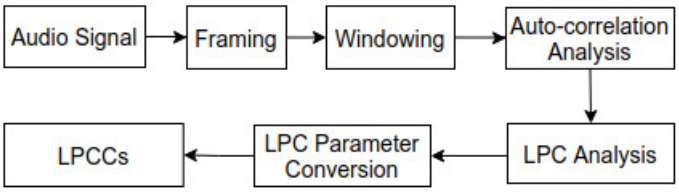

Fig. 5. Block diagram for LPCC features extraction.

40 static LPCCs, 40 delta, 40 double delta features are considered per audio frame and a total of 120 dimensional feature vector formed per frame. Each audio track has got $120X9077$ feature matrix representation by considering 40 ms frame duration with 20 ms frame overlap over a 30 sec duration audio signal.

3.5 All Pole Group Delay (APGD) Features

Phase is neglected in many audio processing methods because of the difficulties associated with the unwrapping of the phase spectrum, which is one of the reasons why APGD characteristics were chosen for acoustic event representations. Nevertheless, because of its great resolution and capacity to highlight peaks in the magnitude spectrum envelope, it can be instructive. In the context of ambient audio analysis, the variety of sound event classes and their innate polyphony renders and such argumentation are not always appropriate. Nevertheless, given the significance of phase in human perception and the harmonic structure of particular acoustic events, it is important to research the role of phase information in sound event recognition. [3, 4, 28].
The APGD features are extracted as follows: The group delay function of a signal $s[n]$ can be obtained as:

$$\tau_g(\omega) = \frac{S_R(\omega)Z_R(\omega) + S_I(\omega)Z_I(\omega)}{|S(\omega)|^2}. \tag{3}$$

Where $S(\omega)$and $Z(\omega)$are the Fourier transforms of $s[n]$ and $z[n]$, $z[n] = ns[n]$, where the real and imaginary components of the Fourier transform are indicated,

respectively, by R and I. The zeros of the transfer function of the modelled filter's model are not close to the unit circle in order for the group delay function to act properly. The magnitude spectrum dips at the appropriate frequency bins when the transfer function's zeros are close to the unit circle. By modelling analyzed environmental event sound with a source-filter model and assuming the filter all-pole, the spectrum of such filter with the frequency response $H(\omega)$ may be approximated with aid of linear prediction. Linear prediction is formulated as [14]:

$$H(\omega) = \frac{G}{1 - \sum_{m=1}^{p} b(m) e^{-j\omega m}}. \tag{4}$$

Here, G is the signal-dependent gain and p is the model order. The coefficients $b(m)$ are determined by the method of least squares in such a way that the power spectrum of $H(\omega)$ matches the power spectrum of the signal $|H(\omega)|^2$. The all-pole group delay function is computed from the phase response of this filter formed by $H(\omega)$. An optional discrete cosine transform (DCT) can be applied for decorrelation, and a number of coefficients are excluded. The feature is calculated in short frames under the assumption of spectral stationariness, and the fourier analysis is done with DFT. The overall extraction procedure of APGD features is illustrated in Figure. 6. The steps involved in the feature extraction is the following.

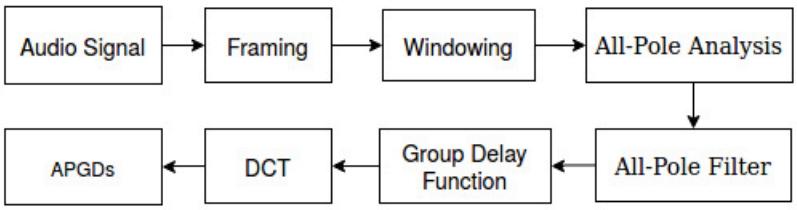

Fig. 6. Block diagram for APGD features extraction.

- Perform all-pole modelling on the frame. Obtain the filter coefficients $b(m)$.
- From the $b(m)$, form the frequency response $H(\omega)$ using above equation with G=1(for normalization purposes).
- Compute the group delay function by taking the negative derivative of the phase response of $H(\omega)$. In practice, the derivative is computed using the sample-wise difference.
- Perform DCT and keep a certain number of coefficients. Then these coefficients are known as APGD features [11].

In this study, we extracted 40 static APGDs, 40 delta, 40 double delta features are considered per audio frame using order of all pole filter is 40 and a total

of 120 dimensional feature vector formed per frame. Each audio track has got $120X9077$ feature matrix representation by considering 40 ms frame duration with 20 ms frame overlap over a 30 sec duration audio signal.

4 Proposed System Description for AED task

This section introduces the sound event recognition system, which integrates suggested characteristics into a multi-label GMM classifier. In this work, we studied the AED task in real life audio with various aspects while using the baseline system model from DCASE 2016 task 3. Mel frequency cepstral coefficient (MFCC) characteristics and a Gaussian mixture model (GMM) model-based classifier were employed as a baseline system for the 2016 AED in real-world audio problem. All audio's MFCC features were calculated using 40 ms frames, a Hamming window, and 50% overlap with 40 mel bands. The top 20 static coefficients, together with the 0th order coefficient, were taken into account. Utilizing a window length of 9 frames, the delta and acceleration coefficients were also determined. A total of 60 dimension frame-based feature vectors. The testing stage utilizes maximum likelihood decision mechanism among all acoustic scene class models [15]. Error rate (ER) and F-score in a fixed time grid was evaluated as the system performance for the AED task [17]. Using segments of one second duration, the actions of sound event classes are compared between the system output and the ground truth. If both the system output and the ground truth indicate that an event is active in a certain chunk, it is considered to have been appropriately identified in that segment. The further instances include: False positives occur when a system turnout reports an event as active when the ground truth indicates it is inert. False negative: when the system output indicates that something is dormant but the ground truth indicates that it is active. Depend on the total counts of false positives (FP), true positives (TP), and precision (P), recall (R), false negatives (FN), and F-score are computed by:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F = \frac{2PR}{P + R}. \tag{7}$$

Error rate computes the amount of errors as regards deletions (D), in-sertions (I), and substitutions (S). The system recognizes an event in a given chunk, but gives it a wrong label is known as substitution.
This is equivalent to the system output consists of one false negative and false positive in the same chunk. After counting the number of substitutions per chunk, the rest of false positives in the system output are counted as insertions, and the rest of false negatives as deletions. The error rate is then obtained

by combining segment-wise counts on top of the total number of chunks (J), with N(j) presence the number of active ground truth events in segment k is given by [21]

$$ER = \frac{\sum_{j=1}^{J} S(j) + \sum_{j=1}^{J} D(j) + \sum_{j=1}^{J} I(j)}{\sum_{j=1}^{J} N(j)}. \quad (8)$$

Event-based metrics consider false positives, true positives, and false negatives in relation to event instances. An event in the system turnout is well chosen correctly recognized if it has a temporal location overlapping with the temporal location of an event with the same label in the ground truth.
A collar of 200 ms duration was allowed for offset and the onset. Either the same 200 ms duration collar or a tolerance of 50% in respect of the ground truth event period. An event in the system output that has no match to an event with same label in the ground truth within the allowed tolerance is a false positive. An event in the ground truth that has no match to an event with same label in the system turnout within the allowed tolerance is a false negative.
Event-based substitutions are defined differently than segment-based: events with correct temporal location but erroneous class label are counted as substitutions. While insertions and deletions are the events not present for as correct or substituted in system turnout or ground truth. Recall, Precision, F-score and error rate are defined the same way, with error rate being computed with respect to the total number of events in the ground truth [18].

5 Experimental Settings

We have conducted experiments on DCASE 2016 TUT acoustic events development data with 4- fold cross validation using different features sets and GMM model. The various proposed systems are as follows:
PS1: log-Mel band energies + GMM,
PS2: SSFCs + GMM,
PS3: SCMCs + GMM,
PS4: LPCCs + GMM,
PS5: APGDs + GMM,
The same baseline system model was used for modelling the AED task but for getting optimal performance, we have varied the number of gaussians used by fixing minimum covariance is 0.001 and the number of iterations are set to 1000 for training. During testing, we have fixed the values of various parameters such as decision_threshold: 320.0, smoothing_window_length: 2.2, minimum_event_length: 0.3, and minimum_event_gap: 0.2 to predict the label for acoustic events in a acoustic scene for segment wise/event wise metric computation.

6 Results and Discussion

The performance of proposed systems (PS1-PS5) on the development dataset with 4-fold cross validation of DCASE 2016 task3 (AED in real life audio) is presented in Table 1.

Table 1. Average results of proposed systems on the 2016 DCASE task 3 development dataset by varying the number of gaussians in GMM model.

S.No.	Number of Gaussians	PS1		PS2		PS3		PS4		PS5	
		Segment-based overall metrics									
		ER	F-score	ER	F-score	ER	F-score	ER	F-score	ER	F-score
1	8	1.78	31.3%	1.07	25.7%	0.99	28.5%	0.97	32.2%	1.02	36.4%
2	16	1.48	32.3%	1.02	26.3%	1.01	28.4%	0.94	32.4%	0.97	32.6%
3	32	**1.19**	**35.7%**	**0.98**	**26.7%**	**0.99**	**29.5%**	**0.90**	**33.3%**	**1.00**	**37.1%**
4	64	1.33	34.4%	0.98	24.8%	0.95	27.3%	0.92	32.0%	1.01	32.0%

In the table, we have tabulated the average results of segment based overall parameters for AED by varying number of gaussians. From the studies, we have observed that the performance of AED system is high in-terms of average accuracy (F-score) and average error rate (ER) is low at the number of gaussians is 32.
From the table, the proposed system-PS5 had resulted highest F-score at number of gaussians 32. From the table, for the proposed system-PS4 had resulted lowest error rate at number of gaussians are 32.

The performance measured in-terms of average ER and F-score on 4-fold cross validation of development dataset for DCASE 2016 AED in real life audio task is presented in Table 2 by considering number of gaussians as 32, where at 32 gaussians the AED system got better generalization and fitting.

Table 2. Class wise average results of baseline system and proposed systems on the development dataset at the number of gaussians are 32.

S.No.	Baseline System [18]			PS1		PS2		PS3		PS4		PS5	
		Segment-based overall metrics											
1	Acoustic scene	ER	F-score	ER	F-score	ER	F-score	ER	F-score	ER	F-score	ER	F-score
2	Home	0.96	15.9 %	1.24	32.7%	1.02	21.8%	1.03	24.7%	0.94	26.9%	1.16	29.7%
3	Residential area	0.86	31.5 %	1.14	38.7%	0.93	31.5%	0.95	34.2%	0.87	39.6 %	0.84	44.5%
4	Average	**0.91**	**23.7 %**	**1.19**	**35.7%**	**0.98**	**26.7%**	**0.99**	**29.5%**	**0.90**	**33.3%**	**1.00**	**37.1%**

Here, we presented the individual class wise average results of 2016 DCASE task3 baseline system performance and proposed systems (PS1-PS5). The events of two classes i.e Home and Residential area class wise average results on development dataset with 4 fold cross validation and also average of two classes have been presented in the Table 2.

From the table, it is observed that the relative average performances (F-score) of 12.0%, 3.0%, 5.8%, 9.6%, and 1.4% got more when compared to the baseline system by proposed systems PS1-PS5, respectively. Also, we got the improved relative average performances 4.7%, 2.3%, 6.1% by the proposed systems PS1, PS4, and PS5 respectively when compared to Rank 1 state of-the-art system [2]. Where, this system was used log Mel band energies and long short-term memory (LSTM) recurrent neural network (RNN) model for AED task implementation on development data. Also, this Rank1 state of-the-art system was resulted as the performances of F-score is 31.0% and error rate is 0.91. The average error rate significantly got decreased by proposed system-PS4. The relative error rate 0.01 got reduced when compared to the baseline system and Rank 1 state of-the-art [2].

The performance results for event based F-score values of two classes have been listed in Table 3.

Table 3. Results for Event-Based F-score Calculated Class-Wise at the number of gaussians as 32

S.No.	Baseline System [18]			PS1		PS2		PS3		PS4		PS5	
		Event based over all metrics											
1	Acoustic Scene	ER	F-score	ER	F-score	ER	F-score	ER	F-score	ER	F-score	ER	F-score
2	Home	1.33	2.5%	1.72	3.4%	1.27	1.3%	1.27	2.5%	1.24	3.1%	1.54	2.7%
3	Residential Area	1.99	1.6%	2.20	1.9%	1.62	1.6 %	1.64	1.6%	1.74	3.0%	1.74	3.3%
4	Average	**1.66**	**2.0%**	**1.96**	**2.7%**	**1.45**	**1.4%**	**1.45**	**2.1%**	**1.49**	**3.0%**	**1.64**	**3.0%**

From the table, it is observed that the proposed features F-score was better than baseline features except SSFC features. The less error rate is achieved by SSFCs and SCMC features.

The segment-based class wise results at the number of gaussians is 32 have been presented in the Table 4.

Table 4. Results for Segment-Based F-score Calculated Class-Wise at the number of gaussians as 32

Residential area							Home						
	Baseline System[18]	PS1	PS2	PS3	PS4	PS5		Baseline	PS1	PS2	PS3	PS4	PS5
event class	F (%)	F (%)	F (%)	F (%)	F (%)	F (%)	event class	F (%)	F (%)	F (%)	F (%)	F (%)	F (%)
(object) banging	0.0	0.0	0.0	0.0	0.0	0.0	(object) rustling	8.3	18.7	12.6	19.2	23.8	17.4
bird singing	33.8	42.5	32.7	38.8	48.9	46.9	(object) snapping	0.0	0.0	0.0	0.0	0.0	0.0
car passing by	59.9	59.5	50.2	55.4	51.3	57.7	cupboard	0.0	0.0	0.0	4.0	0.0	0.0
children shouting	0.0	0.0	0.0	1.5	0.0	0.0	cutlery	0.0	0.0	0.0	2.3	0.0	0.0
people speaking	30.6	18.8	3.0	6.3	10.5	12.7	dishes	4.3	27.9	6.4	9.5	7.3	27.1
people walking	2.8	4.5	3.5	2.9	8.0	9.6	drawer	8.1	0.0	0.0	0.0	0.0	0.0
wind blowing	14.2	16.4	21.3	3.4	2.1	0.0	glass jingling	0.0	6.2	0.0	0.0	0.0	0.0
							object impact	22.8	31.3	11.9	21.4	20.5	22.2
							people walking	18.3	19.9	11.7	1.7	7.9	16.5
							washing dishes	24.6	46.1	20.1	17.0	27.9	30.1
							water tap running	41.2	45.7	55.3	60.1	55.6	31.8

From the table, banging event of Residential area class is not able to detect by all proposed features including baseline features.

Children shouting event of Residential area class is able to detect by proposed feature set i.e SCMCs but unable to detect using baseline features. Except banging and children shouting events of Residential area class, almost all features are well recognized. The snapping event of Home class is unable to detect all the proposed features including baseline features. The events cupboard and cutlery events of Home class can be recognized by proposed features i.e SCMCs. The Drawer event home class is able to detect SSFC features only. The event drawer of Home class is not able to detect all the proposed features. On the other hand, the proposed systems completely fails to detect some events of the classes. This is due to the simplicity of system [18].

7 Summary

We have conducted a thorough investigation using various feature extraction algorithms for the detection of auditory events. According to the results, AED can benefit from features that convey information about dynamic properties, subband information, and comprehensive spectral information. Environmental sound event recognition has been proposed using a phase-based APGD function. A diverse realistic dataset of polyphonic sound occurrences has been used to construct and test a GMM-based classifier that incorporates the feature.

The evaluation has shown that using the suggested feature sets over the MFCC features as a baseline improves the performance of the classifier. According to the results, phase is crucial for identifying various sound event types. As a recurrent neural network is capable to effectively recording potential correlations between nearby time frames of sound occurrences, we will investigate how the AED task is implemented in our upcoming work.

References

1. T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen. "Context-dependent sound event detection". *EURASIP Journal on Audio, Speech, and Music Processing*, 5:1, 01 2013.
2. S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen. "Sound event detection in multichannel audio using spatial and harmonic features". *Technical report, DCASE Challenge*, September, 2016.
3. L. D. Alsteris and K. K. Paliwal. "Short-time phase spectrum in speech processing: A review and some experimental results". *Digital Signal Processing*, 17(3):578 – 616, 2007.
4. H. Banno, Jinlin Lu, S. Nakamura, K. Shikano, and H. Kawahara. "Efficient representation of short-time phase based on group delay". In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP (Cat. No.98CH36181)*, volume 2, pages 861–864 vol.2, 1998.
5. D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley. "Acoustic scene classification". CoRR, abs/1411.3715, 2014.

6. E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen. "Polyphonic sound event detection using multi label deep neural networks". In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2015.
7. 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* , 17:1142 – 1158, 09 2009.
8. C. V. Cotton, D. P. W. Ellis, and A. C. Loui. "Soundtrack classification by transient events". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 473–476, 2011.
9. T. H. Dat, N. W. Z. Terence, J. W. Dennis, and L. Y. Ren. "Generalized gaussian distribution kullback-leibler kernel for robust sound event recognition". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5949–5953, 2014.
10. S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.
11. A. Diment, E. Cakir, T. Heittola, and T. Virtanen. "Automatic recognition of environmental sound events using all-pole group delay features". In *23rd European Signal Processing Conference (EUSIPCO)*, pages 729–733, 2015.
12. S. Furui. "Cepstral analysis technique for automatic speaker verification". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.
13. J. Kua, T. Thiruvaran, H. Nosrati, A. E, and J. Epps. "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition". *Odyssey*, 06 2010.
14. J. Makhoul. "Linear prediction: A tutorial review". *Proceedings of the IEEE*, 63(4):561–580, 1975.
15. A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley. "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, Feb 2018.
16. A. Mesaros, T. Heittola, and A. Klapuri. "Latent semantic analysis in sound event detection". In *19th European Signal Processing Conference*, pages 1307–1311, 2011.
17. A. Mesaros, T. Heittola, and T. Virtanen. "Metrics for polyphonic sound event detection". *Applied Sciences*, 6:162, 2016.
18. A. Mesaros, T. Heittola, and T. Virtanen. "TUT database for acoustic scene classification and sound event detection". In *24th European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016.
19. J. Muhammad and Akbar. "A Overview of Spoof Speech Detection for Automatic Speaker Verification". *Book*, 02 2019.
20. V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. "Computational auditory scene recognition". In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1941–II–1944, 2002.
21. G. Poliner and D. Ellis. "A discriminative model for polyphonic piano transcription". *EURASIP Journal on Advances in Signal Processing*, 3, 01 2007.
22. P. Rajan, T. Kinnunen, C. Hanilçi, J. Pohjalainen, and P. Alku. "Using group delay functions from all-pole models for speaker recognition". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2489–2493, 01 2013.
23. M. Sahidullah, T. Kinnunen, and C. Hanilçi. "A Comparison of Features for Synthetic Speech Detection". *Tech report, 09 2015*.

24. E. Scheirer and M. Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator". In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1331–1334 vol.2, 1997.
25. J. Schröder, S. Goetze, and J. Anemüller. "Spectro-temporal gabor filterbank features for acoustic event detection". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2198–2208, 2015.
26. R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah. "Large-scale weakly labeled semi-supervised sound event detection in domestic environments". In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 19–23, November 2018.
27. D. Sheng and G. Fazekas. "Automatic control of the dynamic range compressor using a regression model and a reference sound". In *The 20th International Conference on Digital Audio Effects (DAFx)*, 09 2017.
28. B. Yegnanarayana. "Formant extraction from linear-prediction phase spectra". *The Journal of the Acoustical Society of America*, 09, 1978.
29. B. Yegnanaryana, G. Duncan, and H. A. Murthy. Formant extraction from group delay function. In *IEEE Colloquium on Speech Processing*, pages 2/1–2/4, 1988.