INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Locality-constrained Linear Coding based Fused Visual Features for Robust Acoustic Event Classification

Manjunath Mulimani, Shashidhar G. Koolagudi

Dept. of CSE, National Institute of Technology Karnataka, Suarthkal, India, 575 025
manjunath.gec@gmail.com, koolagudi@nitk.edu.in

Abstract

In this paper, a novel Fused Visual Features (FVFs) are proposed for Acoustic Event Classification (AEC) in the meeting room and office environments. The codes of Visual Features (VFs) are evaluated from row vectors and Scale Invariant Feature Transform (SIFT) vectors of the grayscale Gammatonegram of an acoustic event separately using Locality-constrained Linear Coding (LLC). Further, VFs from row vectors and SIFT vectors of the grayscale Gammatonegram are fused to get FVFs. Performance of the proposed FVFs is evaluated on acoustic events of publicly available UPC-TALP and DCASE datasets in clean and noisy conditions. Results show that proposed FVFs are robust to noise and achieve overall recognition accuracy of 96.40% and 90.45% on UPC-TALP and DCASE datasets, respectively.

Index Terms: Acoustic Event Classification (AEC), Fused Visual Features (FVFs), Gammatonegram, Scale Invariant Feature Transform (SIFT), Locality-constrained Linear Coding (LLC)

1. Introduction

Acoustic Event Classification (AEC) is the task of recognizing a specific sound in an environment. It has many emerging applications such as; robotics [1], wildlife monitoring [2], audio surveillance [3] [4], machine hearing [5] and so on.

Study on AEC is still in its infancy as compared to the speech/speaker recognition tasks. However, different feature encoding methods from computer vision are used recently, to encode the frame-based features into higher level feature representation for AEC with improved recognition accuracy [6–9]. The traditional frame-based features such as Mel-frequency cepstral coefficients (MFCCs) are extracted from a continuous acoustic event signal frame-by-frame. Further, a sequence of these frame-based feature vectors of an acoustic event is encoded into a fixed dimensional higher level feature representation (code) for AEC using two popular feature encoding methods: Vector Quantization (VQ) [10] and sparse coding [11].

VQ is widely used in Bag-of-Audio-Words (BoAW) approach [7][12], which even outperforms the emerging Deep Neural Networks (DNNs) [10]. However, VQ suffers from the large quantization error [11]. Hence, sparse coding is used over VQ for AEC with improved recognition accuracy [9]. However, many times real-world acoustic events are overlapped with high-background noise. Traditional frame-based features are sensitive to background noise. Hence, codes (either from VQ or sparse coding) of frame-based features may not be suitable for AEC, especially during noisy conditions.

Speech is different from the acoustic events when one considers its phonetic structure [13][14]. Acoustic events are brief and have more discriminative Time-Frequency Representations (TFR). Hence, in our previous work [6], a grayscale spectrogram (TFR) is represented as Bag-of-Visual-Words (BoVW) using traditional VQ for robust AEC. It hints that features from spectrogram are robust to noise and suitable for AEC. However, Gammatonegram reveals more spectral information of acoustic events than spectrogram [15][16]. Hence, in this work, rows of intensity values and Scale Invariant Feature Transform (SIFT) vectors of grayscale Gammatonegram of an acoustic event are separately represented (encoded) as codes of Visual Features (VFs) using Locality-constrained Linear Coding (LLC) [17]. LLC is an extension of sparse coding evaluated based on the assumption that locality of features is more discriminative for AEC than sparsity. Locality contains sparsity but not vice-versa. Further, VFs (LLC codes) of row vectors and SIFT vectors of grayscale Gammatonegram are combined to get Fused Visual Features (FVFs). Performance of the FVFs is compared with state-of-the-art methods, and its robustness is verified in different noisy conditions.

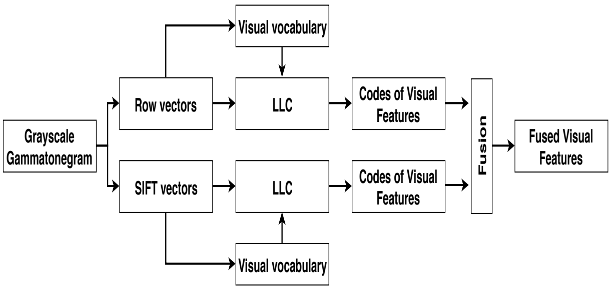

Figure 1: *Overview of the proposed approach.*

The rest of the paper is organized as follows, Section 2 contains explanation of the proposed Fused Visual Features in brief. Section 3 describes the experiments carried out. Results are discussed in section 4. The conclusions are given in section 5.

2. Fused Visual Features

Overview of the proposed approach is given in Fig. 1. The row vectors and SIFT vectors of the grayscale Gammatonegram are encoded separately into codes of Visual Features (VFs) using LLC. Further, these VFs are combined to get Fused Visual Features (FVFs). The steps involved in the generation of FVFs are explained below in brief.

2.1. Grayscale Gammatonegram generation

In this step, the acoustic events are represented as Gammatone spectrogram or Gammatonegram $C(f,t)$, where f (ranging from 1 to F) is the center frequency of the Gammatone filter and t is the time frame obtained by windowing the signal into frames using the Hamming window of length 20ms with 50% overlap. The sampling rate is 44100 Hz and 64 ($F = 64$)

http://dx.doi.org/10.21437/Interspeech.2019-1421

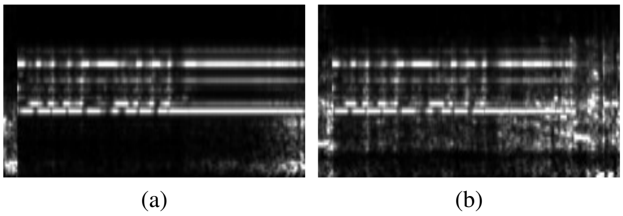

Figure 2: *Acoustic event ringing phone. (a) Grayscale Gammatonegram $GI(f,t)$ at clean condition; (b) grayscale Gammatonegram $GI(f,t)$ at 0dB SNR.*

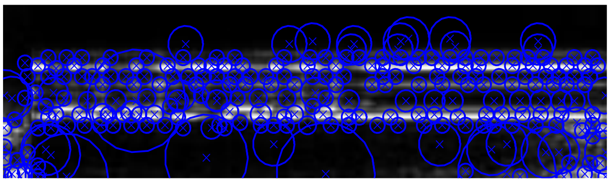

Figure 3: *SIFT keypoints of acoustic event ringing phone.*

filters are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale. The logarithmic Gammatonegram is obtained from (1).

$$C_{Log}(f,t) = log(C(f,t)) \tag{1}$$

Further, a grayscale intensity Gammatonegram image (see Fig. 2) is generated by normalizing the values of Time-Frequency matrix $C(f,t)$ between [0, 1] using (2).

$$GI(f,t) = \frac{C_{Log}(f,t) - min(C_{Log})}{max(C_{Log}) - min(C_{Log})} \tag{2}$$

Acoustic events are highly variant with respect to time, which may cause dimensional variations. Hence, grayscale Gammatonegram $GI(f,t)$ is transposed as given in (3) to get fixed 64-dimensional row vectors.

$$G(t,f) = GI(f,t)^T \tag{3}$$

In this work, intensity values of each 64-dimensional row vector of a $G(t,f)$ are considered as a feature vector per frame for VFs representations using LLC.

2.2. Extraction of SIFT feature vectors

David G. Lowe, first introduced the SIFT algorithm for object recognition in a digital image [18]. SIFT algorithm includes two main stages: keypoint detection and feature extraction. Algorithm first localizes (detects) interesting keypoints in the grayscale Gammatonegram of an acoustic event (see blue color keypoints in Fig. 3) and the second stage extracts a set of 128-dimensional feature vectors associated with detected keypoints for VFs representations using LLC.

2.3. Visual vocabulary generation

Let Y be a set of D-dimensional feature vectors (row vectors or SIFT vectors) extracted from a grayscale Gammatonegram, i.e., $Y = [y_1, y_2, ..., y_N] \in \mathbb{R}^{N \times D}$. A set of feature vectors Y of five randomly selected acoustic events per class are grouped into mutually exclusive clusters using the K-means clustering algorithm. The centroids of these clusters are referred to as visual words or codewords. All the visual words together constitute a vocabulary V or a codebook. It is worth to point out that five acoustic events per class are sufficient enough to build discriminative visual vocabulary with less computational time. There is no known best way to select the size of the vocabulary M, i.e., the number of visual words. In this work, the size of the vocabulary ranging from 64 to 512 is considered and its impact on the performance of AEC is analyzed.

2.4. Locality-constrained Linear Coding

Given a set of D-dimensional feature vectors Y (row vectors or SIFT) from a grayscale Gammatonegram and a vocabulary V with M entries, $V = [v_1, v_2, ..., v_M] \in \mathbb{R}^{M \times D}$, Locality-constrained Linear Coding (LLC) converts each feature vector into a M-dimensional code for AEC [17]. The objective function of LLC is formulated as:

$$\begin{aligned} &\min_C \sum_{i=1}^{N} \|y_i - Vc_i\|^2 + \lambda \|d_i \odot c_i\|^2 \\ &s.t.\ 1^T c_i = 1, \forall i \end{aligned} \tag{4}$$

Where $C = [c_1, c_2, ..., c_N]$ is a set of coding coefficients (codes) for Y, $\odot$ represents the element-wise multiplication, $d_i \in \mathbb{R}^N$ is a distance vector (evaluated using Euclidean distance) between y_i and each visual word of the vocabulary V, λ is a regularization parameter, which weights the locality constraint.

Minimizing the Eq. (4) tends to encode each feature vector $y_i \in Y$ into a M-dimensional code using visual words of vocabulary V. Locality adaptor d_i penalizes the visual words far away from the feature vector. Specifically, a larger d_i suppresses the value of corresponding code c_i to zero. Hence, the resulting set of codes $C \in \mathbb{R}^{N \times M}$ is a sparse representation of Y with few non-zero elements. It can also be interpreted as each feature vector $y_i \in Y$ only responses to nearby visual words of vocabulary V.

2.5. Visual feature vector generation

In this step, a single M-dimensional code $Z = [z_1, z_2, ..., z_M] \in \mathbb{R}^{1 \times M}$ is obtained from the set of codes C (the result of Eq. (4)) using max pooling function $Z = max(C)$. Specifically, a maximum value is chosen from C in a column-wise manner as given in (5).

$$z_j = max\{c_{1j}, c_{2j}, ..., c_{Nj}\} \tag{5}$$

Where N is number of feature vectors in Y. Each column of C corresponds to the responses of all feature vectors in Y to a visual word in V.

The pooled features (Z) are also known as codes of visual features, which are normalized using ℓ_2 norm. Normalized visual features of row vectors and SIFT vectors of grayscale Gammatonegram are fused (concatenated) to get Fused Visual Features (FVFs).

3. Experiments

3.1. Acoustic event datasets

Performance of the proposed approach is evaluated using two publicly available datasets namely UPC-TALP dataset [22] and DCASE-2013 dataset [23].

Table 1: *Comparison of overall Recognition accuracy (%) of proposed FVFs with other methods at clean and different SNR conditions on UPC-TALP and DCASE datasets.*

		UPC-TALP					DCASE				
Method	Ref.	Clean	20dB	10dB	0dB	Average	Clean	20dB	10dB	0dB	Average
MFCCs	-	74.79	66.46	58.55	46.88	61.67	50.94	43.68	38.44	26.93	39.99
GTCCs	[19]	76.87	74.83	72.54	59.02	70.81	54.06	51.25	48.13	37.75	47.79
DNNs	[20]	70.42	69.38	58.55	35.63	58.49	39.69	38.12	33.44	25.00	34.07
CNNs	[21]	86.33	84.05	80.17	65.97	79.13	65.25	61.30	57.17	45.32	57.26
BoVW	[6]	93.54	92.51	88.76	79.54	88.58	68.75	66.25	59.37	46.88	60.31
SIFT vectors - VFs	-	94.06	92.96	90.21	80.71	89.48	75.01	70.63	61.57	47.19	63.06
Row vectors - VFs	-	95.01	93.34	92.30	86.26	91.72	79.07	74.34	65.63	53.76	68.02
FVFs	-	**99.17**	**99.15**	**98.12**	**89.17**	**96.40**	**96.32**	**93.19**	**90.69**	**81.63**	**90.45**

3.1.1. UPC-TALP dataset

A 12 different isolated meeting room acoustic events, namely: applause, chair moving, cough, door knock, door slam, keyboard typing, key jingle, laugh, paper wrapping, phone ring, spoon cup jingle and steps are selected for AEC. Approximately 60 acoustic events per class, recorded using 84 microphones, namely: an array of 64 Mark III microphones, 12 T-shape clusters microphones, 8 table top and omni-directional microphones are used. In this work, only the third channel of Mark III array is considered for the experiments.

3.1.2. DCASE dataset

A 16 different isolated acoustic events namely, alert, clearing throat, cough, door slam, drawer, keyboard, keys, knock, laugh, mouse, page-turn, pen drop, phone ring, printer, speech and switch are used for AEC. Approximately 27 acoustic events per class are presented in the DCASE dataset. In this work, microphone channel one from stereo recordings is considered for the performance evaluation.

Acoustic events of both the datasets are trimmed to the length of given annotations and resulting data is divided into five disjoint folds to perform five-fold cross-validation. Each fold has the equal number of acoustic events per class. To compare the robustness of the proposed approach, 'speech babble' noise from NOISEX'92 database [24] is added to the acoustic events at 20, 10 and 0dB SNR. The 'speech babble' noise is diffuse and most of its energy is distributed at lower frequencies. All acoustic events are processed at 44100 Hz sampling rate.

3.2. Evaluation methods

Performance of the proposed approach is compared with the following baseline system and the other state-of-the-art methods of AEC.

1. Baseline system : Mean and standard deviation of 13 MFCCs and their first and second-order derivatives are taken over each frame, resulting in 39 × 2 dimensional feature vector.
2. GTCCs : Mean and standard deviation of 13 Gammatone Cepstral Coefficients (GTCCs) and their first and second-order derivatives are taken over each frame [19], resulting in 39 × 2 dimensional feature vector.
3. DNNs : Mel band energies are used as features to DNNs, which has three fully connected layers followed by a softmax output. Each layer uses 500 units with ReLU activation function and 10% dropout. Categorical cross-entropy used as a loss function [20].
4. CNNs : 60-dimensional log Mel features used as input features to Convolutional Neural Networks (CNN). The network had two CNN's of 32, 64 and 128 filters. Each CNN is followed by batch-normalization and Rectified Linear Unit (ReLU). CNN's with same number filters followed by max pooling. A softmax activation function is used as the output layer [21].
5. BoVW : Grayscale spectrograms of acoustic events are represented as BoVW [6] features to train Chi-square kernel SVM for AEC.

The MFCC and GTCC features used in our experiments are extracted using 20ms hamming window with 50% overlap and normalized to zero mean and unit variance. Results of all the methods except DNNs, CNNs and BoVW are reported using linear kernel SVM, which achieves higher recognition accuracy with lower computation cost. Hence, one-versus-rest multi-class SVM is chosen as a classifier and its optimal parameters (such as C) are selected using five-fold cross-validation.

4. Results and Discussion

The experimental results given in Table 1 demonstrate that proposed FVFs-SVM combination significantly outperforms all the methods in clean and different noisy conditions with average recognition accuracy of 96.40% and 90.45% on UPC-TALP and DCASE datasets, respectively. The recognition performance of all the methods on DCASE dataset is lower compared to performance on UPC-TALP dataset. This is due to the quality of DCASE audio recordings, which are already quite noisy (there are no clean acoustic events) and real SNR might be lower than denoted. Hence, the performance of all the methods drastically reduces in clean and noisy conditions. However, the proposed FVFs discriminate acoustic events from the noise and largely outperform all other methods in clean and noisy conditions.

As we had mentioned earlier, the speech babble noise is diffuse and its maximum energy concentrated at lower frequencies. The MFCCs are sensitive to noise at lower frequencies. Hence, the performance of the MFCCs baseline system significantly drops at 0dB SNR.

The GTCCs are also obtained from Gammatone filter bank as our grayscale Gammatonegram. Hence, GTCCs are considered for performance comparison with FVFs in this work. Gammatone filterbank resolution at lower frequencies is much higher with ERB scale than Mel filter bank with Mel scale. Hence, GTCCs discriminate the spectral components at lower frequencies belong to the acoustic event and noise accurately than MFCCs. However, the performance of GTCCs still worse when compared to proposed FVFs.

Further, the performance of FVFs-SVM combination is also compared with the emerging DNNs and CNNs. The CNNs are effective and outperform DNNs in clean and noisy conditions (see Table 1). However, CNNs/DNNs require huge data for training and do not perform very well in the cases of limited training data.

Our previous BoVW approach [6] performs significantly better in clean and noisy conditions. However, BoVWs are the histograms of traditional VQ codes obtained from Grayscale spectrograms of acoustic events. In VQ, similar (nearby) feature vectors of acoustic events have different VQ codes due to large quantization errors, in contrast to that of LLC which ensures similar LLC codes for the nearby feature vectors of the acoustic events. Hence, LLC codes or VFs from row vectors and SIFT vectors of Gammatonegram significantly discriminate acoustic events in clean and noisy conditions than VQ codes. It is worth to point out that, proposed VFs achieve impressive recognition accuracy with linear SVM. The BoVW approach performs well with non-linear kernel classifiers such as Intersection and Chi-square kernels SVMs, which demand higher computational time than simple linear SVM.

The magnitude of the intensity values of an acoustic event in the grayscale Gammatonegram is much higher than that of noise. The noise is distributed over Gammatonegram compared to the acoustic event and its maximum energy is concentrated at the lower region of Gammatonegram. However, higher intensity values (strongest peaks) of the acoustic events are unaffected by noise (see Fig. 2 at clean and 0dB SNR), which are effectively discriminated by VFs (row vector VFs) in all noisy conditions.

The SIFT feature extraction algorithm blurs the grayscale Gammatonegrams using different scales before keypoint localization that highlights the strongest peaks of the acoustic events and reduces the effect of noise. Further, localizes the keypoints around the strongest peaks of an acoustic event and extracts the SIFT feature vectors from detected keypoints. Hence, VFs from SIFT feature vectors of grayscale Gammatonegrams are robust to noise and outperform the BoVW in different noisy conditions.

However, VFs from row vectors or SIFT feature vectors does not perform alone as expected. The FVFs are the combination of significant VFs from row vectors and SIFT vectors of the Gammatonegram. Hence, the FVFs contains significantly much more information about the acoustic events than VFs alone and they are observed to outperform all other approaches in both clean and noisy conditions.

4.1. Recognition accuracy versus size of the codebook

The acoustic event recognition accuracy of VFs from row vectors and SIFT vectors of Gammatonegrams with linear SVM to the number of visual words i.e., size of the vocabulary at clean condition on UPC-TALP dataset is shown in Fig. 4. One can observe that recognition accuracy improves as there is an increase in the number of visual words. The smaller vocabulary groups the dissimilar acoustic events into the same visual word. Hence, the smaller vocabulary is not much discriminative and gives poor performance. On the other hand, the larger vocabulary is more discriminative and achieves improved performance for both VFs. However, as the size of the vocabulary increases computational complexity also increases. In this work, we use 256 visual words (i.e., the size of the vocabulary) which generate 256-dimensional VFs from row and SIFT vectors for AEC with higher recognition accuracy and lower computational cost. Further, both 256-dimensional VFs are concatenated to get 512-dimensional FVFs used for AEC.

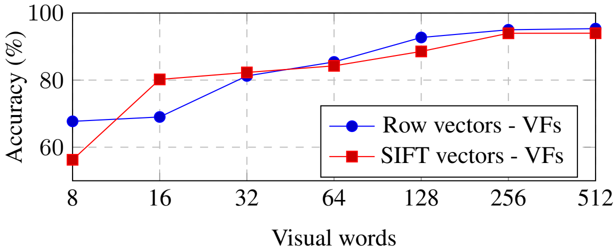

Figure 4: *Recognition accuracy versus number of visual words.*

Table 2: *Comparison of overall Recognition accuracy (%) of proposed FVFs with state-of-the-art methods at clean condition on UPC-TALP and DCASE datasets.*

Dataset	Method	Ref.	Accuracy
UPC-TALP	BoAW - SVM	[25]	96.80
UPC-TALP	Proposed FVFs - SVM	-	**99.17**
DCASE	Gabor Filterbank Features - HMM	[26]	80.00
DCASE	Global statistical Features - HMM	[27]	70.31
DCASE	Proposed FVFs -SVM	-	**96.32**

4.2. Performance of the FVFs versus state-of-the-art methods on UPC-TALP and DCASE datasets

The overall recognition accuracy of the FVFs at clean condition on UPC-TALP and DCASE datasets is also compared with the state-of-the-art methods reported in the literature (see Table 2). The traditional speech features are represented as BoAW in [25] and achieves 96.80% of acoustic event recognition accuracy, which is less than that using proposed FVFs (i.e., 99.17%) in clean condition, for UPC-TALP dataset. However, these speech features are sensitive to noise and performance of their BoAW representations is expected to reduce in noisy conditions. The effect of FVFs is clearly observed on DCASE dataset and it largely outperforms state-of-the-art HMM-based methods [26] [27], which are also based on traditional speech features, may not be suitable for AEC.

5. Conclusions

In this paper, a novel FVFs are proposed for AEC. FVFs are the combination of significant VFs, obtained from row vectors and SIFT vectors of grayscale Gammatonegram of an acoustic event. Intensity values of the acoustic event in the grayscale Gammatonegram are higher than that of noise and those are effectively discriminated by FVFs. FVFs achieve overall 96.40% recognition accuracy in clean and different noisy conditions. It indicates that proposed FVFs features are robust and have a significant contribution towards the AEC. In future, the concatenation of spectral and other image features to FVFs may further improve the performance of the proposed approach.

6. Acknowledgements

The authors would like to thank Ministry of Electronics & Information Technology (MeitY), Government of India for their support in part of the research.

7. References

[1] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 509–514.

[2] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[4] M. Mulimani and S. G. Koolagudi, "Extraction of MapReduce-based features from spectrograms for audio-based surveillance," *Digital Signal Processing*, vol. 87, pp. 1–9, 2019.

[5] R. F. Lyon, "Machine hearing: An emerging field [exploratory DSP]," *IEEE signal processing magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[6] M. Mulimani and S. G. Koolagudi, "Robust Acoustic Event Classification using Bag-of-Visual-Words," in *INTERSPEECH*, 2018, pp. 3319–3322.

[7] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012, pp. 2105–2108.

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Sparse representation based on a bag of spectral exemplars for acoustic event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6255–6259.

[9] L. Zhang, J. Han, and S. Deng, "Unsupervised temporal feature learning based on sparse coding embedded boaw for acoustic event recognition," in *INTERSPEECH*, 2018, pp. 3284–3288.

[10] R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-features methods for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1242–1252, 2017.

[11] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1794–1801.

[12] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 3325–3329.

[13] M. Mulimani and S. G. Koolagudi, "Segmentation and characterization of acoustic event spectrograms using singular value decomposition," *Expert Systems with Applications*, vol. 120, pp. 413–425, 2019.

[14] M. Mulimani, U. Jahnavi, and S. G. Koolagudi, "Acoustic event classification using graph signals," in *Region 10 Conference (TENCON)*. IEEE, 2017, pp. 1812–1816.

[15] R. V. Sharan and T. J. Moir, "Subband time-frequency image texture features for robust audio surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2605–2615, 2015.

[16] M. Mulimani and S. G. Koolagudi, "Robust acoustic event classification using fusion fisher vector features," *Apllied Acoustics*, vol. 155, pp. 130–138, 2019.

[17] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3360–3367.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[20] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for DCASE challenge 2016," *Proceedings of DCASE*, 2016.

[21] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 126–130.

[22] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 311–322.

[23] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[24] A. Varga and H. J. 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.

[25] H. Phan, M. Maass, L. Hertel, R. Mazur, I. McLoughlin, and A. Mertins, "Learning compact structural representations for audio events using regressor banks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 211–215.

[26] 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*, vol. 23, no. 12, pp. 2198–2208, 2015.

[27] S. Jayalakshmi, S. Chandrakala, and R. Nedunchelian, "Global statistical features-based approach for acoustic event detection," *Applied Acoustics*, vol. 139, pp. 113–118, 2018.