Novel TEO-based Gammatone Features for Environmental Sound Classification

Dharmesh M. Agrawal, Hardik B. Sailor, Meet H. Soni, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India
{dm_agrawal, sailor_hardik, meet_soni, hemant_patil}@daiict.ac.in

Abstract*—In this paper, we propose to use modified Gammatone filterbank with Teager Energy Operator (TEO) for environmental sound classification (ESC) task. TEO can track energy as a function of both amplitude and frequency of an audio signal. TEO is better for capturing energy variations in the signal that is produced by a real physical system, such as, environmental sounds that contain amplitude and frequency modulations. In proposed feature set, we have used Gammatone filterbank since it represents characteristics of human auditory processing. Here, we have used two classifiers, namely, Gaussian Mixture Model (GMM) using cepstral features, and Convolutional Neural Network (CNN) using spectral features. We performed experiments on two datasets, namely, ESC-50, and UrbanSound8K. We compared TEO-based coefficients with Mel filter cepstral coefficients (MFCC) and Gammatone cepstral coefficients (GTCC), in which GTCC used mean square energy. Using GMM, the proposed TEO-based Gammatone Cepstral Coefficients (TEO-GTCC), and its score-level fusion with MFCC gave absolute improvement of 0.45 %, and 3.85 % in classification accuracy over MFCC on ESC-50 dataset. Similarly, on UrbanSound8K dataset the proposed TEO-GTCC, and its score-level fusion with GTCC gave absolute improvement of 1.40 %, and 2.44 % in classification accuracy over MFCC. Using CNN, the score-level fusion of Gammatone spectral coefficient (GTSC) and the proposed TEO-based Gammatone spectral coefficients (TEO-GTSC) gave absolute improvement of 14.10 %, and 14.52 % in classification accuracy over Mel filterbank energies (FBE) on ESC-50 and UrbanSond8K datasets, respectively. This shows that proposed TEO-based Gammatone features contain complementary information which is helpful in ESC task.

I. Introduction

The automatic recognition of an environmental sound is a growing research problem in the multimedia applications. The environmental sounds are very diverse group of everyday audio events that cannot be described as only speech or music [1]. There are various applications of the environmental sounds classification (ESC) task, such as, audio surveillance system [2], hearing aids [3], smart room monitoring [4] and video content highlight generation [5], etc. Previous approaches to address this problem include matrix factorization [6]–[9], dictionary learning [2], [10], and wavelet-based features [11], [12]. In [13], authors have used GMM classifier for Acoustic Scene Classification (ASC) task. GMM classifier is also used for sound classification task [14]. Recently, Deep Neural Network (DNN)-based classification are used for ESC task [15]–[17]. In particular, deep Convolutional Neural Network (CNN) has been observed to work better for this problem [16], [17]. CNN classifier is well suited to ESC task because of they are useful for capturing the energy modulations across time and frequency axis of audio spectrograms [16].

In [17], authors have used CNN in which the first convolutional layer is designed to capture the temporal variations in the time-frequency representation of the audio signal. The convolutional layer and pooling layer are designed to achieve a small frequency invariance. On the other hand, the CNN used in [16] do convolution in time and frequency-domain, which is designed to capture spectro-temporal variations in the time-frequency representation of the audio signal. In both of these works, log-scaled Mel spectrograms or commonly referred to as Mel filterbank energies (FBEs) were used as the input to the CNN.

In this paper, we aim to improve the classification accuracy of an ESC task by using the feature representation that is biologically-inspired and perceptually more significant. We employ the features extracted using the Gammatone filterbank with the energy estimation using TEO for ESC task such as TEO-GTCC and TEO-GTSC. The cepstral coefficients extracted using Gammatone filterbank have been previously used for speech recognition [18]–[20], and non-speech audio classification [21]. Unlike conventional energy (l^2-norm), Teager Energy Operator (TEO) profile represents both amplitude and frequency variations of a signal [22]. In experiments, we used cepstral features, and spectral features for two classifiers, namely, GMM, and CNN, respectively. Using two datasets ESC-50, and UrbanSound8K, the results show that proposed TEO-based Gammatone features work better as compared to other state-of-the-art features such as MFCC, and FBEs.

II. TEO-based Gammatone feature

A. Gammatone filterbank

The impulse response of Gammatone filter is a multiplication of Gamma distribution function and a sinusoidal tone centered at a particular frequency [21]. It is given by:

$$g(f,t) = t^{a-1}e^{-2\pi bt}\cos(2\pi ft), t > 0, \quad (1)$$

where a is the filter order, b is rectangular bandwidth, and f is the center frequency. Gammatone filter is inspired from the biologically motivated studies [23]. The Gammatone function is used for modeling of the human auditory filter response [24]. The magnitude response of a Gammatone filter is very similar to the representation of the human auditory filter response

ISBN 978-0-9928626-7-1 © EURASIP 2017

(called as *reox* function) in the cochlea [21]. The filter bandwidth of a Gammatone corresponds to the placement of filters in the basilar membrane (BM) in the human auditory system (HAS). It is measured as equivalent rectangular bandwidth (ERB) scale [25].

B. Teager Energy Operator (TEO)

Among many acoustic and perceptual features of an audio, temporal modulations are one of the important parametric representations of the audio signal. Temporal modulations describe the changes in an audio signal in terms of amplitude modulation (AM) and frequency modulation (FM) [22]. It is observed that AM and FM always co-occur and are inseparable features of the audio signal [22]. AM-FM responses can be obtained from the filterbank model of the cochlea. However, instead of separating an AM and FM responses after filterbank processing, we consider using an operator such as TEO that can track an energy due to both AM and FM components. The TEO represents the energy of the system that generates the signal or energy required to generate the signal [26]. Unlike the conventional energy (l^2-norm), the TEO is approximately equal to the squared product of amplitude and frequency. The discrete version of the TEO applied on the AM-FM signal of the form $x[n] = a[n]\cos(\phi[n])$ is defined as follows [27]:

$$\Psi_d\{x[n]\} = x^2[n] - x[n-1]x[n+1] \approx a^2[n]\omega^2[n], \quad (2)$$

where $a[n]$, and $\omega[n] = \frac{d}{dn}\phi[n]$ are discrete time-varying amplitude and instantaneous frequency (derivative of instantaneouse phase $\phi[n]$), respectively. This energy operator is useful for analyzing AM-FM signals with time-varying amplitude and frequency.

C. Proposed TEO-based Gammatone feature set

TEO cannot be applied directly to the audio signal because it works primarily on monocomponent or at least bandpass filtered signal [27]. However, an audio signal contains many frequency components. Therefore, before applying TEO, we need to filter the signal using a narrowband filterbank [28]. As seen in Figure 1, we bandpass filter the audio signal with Gammatone filterbank followed by the half-wave rectifier (HWR) on each subband of the audio signal. HWR represents a function of inner hair cell movements in the human ear [29], [30]. Then TEO is applied on each subband followed by the short-term averaging to obtain short-term spectral features. The logarithm is applied as a compressive nonlinearity, that is also found in auditory processing literature [30]. For CNN, we used directly short-term spectral features called as TEO-based Gammatone spectral coefficients (TEO-GTSC), and for GMM, we used DCT-based cepstral features called as TEO-GTCC.

As shown in Fig. 2(c) and Fig. 2(d), the TEO-based Gammatone spectral features (Fig. 2(d)) enhanced higher frequency regions as compared to Gammatone spectral features (Fig. 2(c)) as shown by upper box in Fig. 2(c) and Fig. 2(d). The TEO-based Gammatone filterbank has slightly lower resolution in lower frequency regions as compared to Gammatone spectrogram as shown by lower box of Fig. 2(c) and Fig. 2(d).

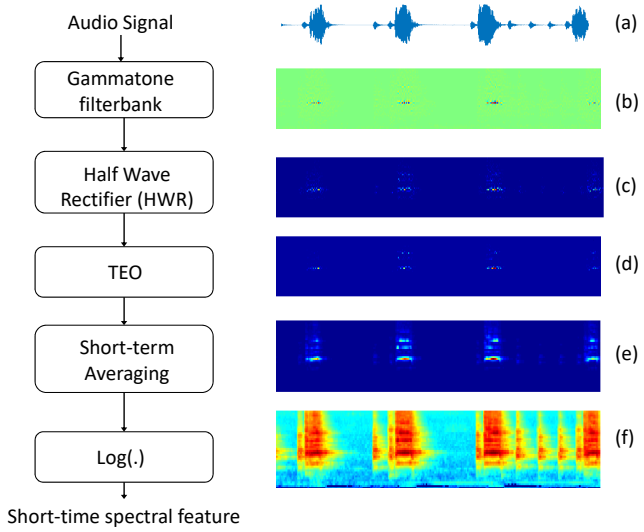

Fig. 1. Block diagram of the proposed features, (a) raw audio signal, (b) Gammatone filterbank response, (c) effect of the HWR on Gammatone filter responses, (d) TEO applid on each suband of Gammatone filterbank, (e) averaging of each subband filter response, (f) log-magnitude response of Gammatone TEO response.

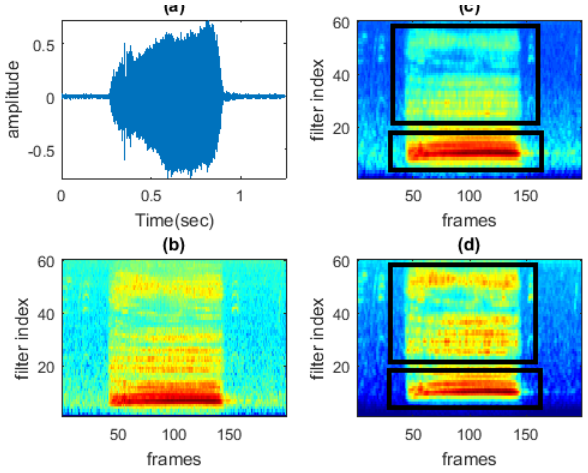

Fig. 2. Spectrographic analysis: (a) raw audio signal of cow sound, (b) Mel filterbank spectrogram, (c) Gammatone spectrogram, (d) TEO-Gammatone spectrogram. The regions indicated by black-boxes shows the differences between spectrum representation in (c) and (d).

Such representation observed improvement in classification accuracy of classes such as dog barking (DB), gun shot (GS), and street music(SM), compared to Gammatone spectral features.

III. Experimental Setup

In this paper, two publicly available standard databases, namely, ESC-50 [1], and UrbanSound8K [31] are used for the experiments. The ESC-50 dataset consists of 2000 short environmental audio recording with 44.1 kHz sampling frequency. These recordings are equally divided into 50 classes that are divided into five major categories, namely, animals,

ISBN 978-0-9928626-7-1 © EURASIP 2017

natural soundscapes and water sounds, non-speech sounds of humans, interior or domestic sounds, and exterior or urban noises. UrbanSound8K is a dataset with 8732 audio files and 10 classes. For ESC task, we have done experiments using two classifiers, namely, GMM and CNN. We use cepstral features in GMM and spectral features in CNN classifier. As compared to the cepstral features, the raw spectral features retain more information and enable the use of convolution and pooling operations that captures invariance and variability in frequency-domain [32]. In GMM, we apply cepstral features since GMM describe their statistical distribution with uncorrelated features [33].

Before, feature extraction for both classifiers, we first pre-processed the audio signal. All the audio files were downsampled to 22.05 kHz (to compare results with baseline system [17]). To extract features, the audio files were divided into frames by using 25 ms Hamming window with 50 % overlap. Then, we applied silence removal algorithm. For silence removal, we first check for more than three consecutive silence frames (approximately 50 ms duration). If silence is present in more than three frames, then we remove the silence frames else we keep frames. Simple energy thresholding algorithm was used to remove the silence regions. 60-D FBEs, GTSCs, and TEO-GTSCs were extracted from files of audio frames. For cepstral feature set, the audio spectrum envelope is converted to decibel (dB) scale, normalized with the RMS (root mean square) and finally, energy compacted with DCT.

A. GMM classifier

We have experimented with different cepstral feature sets such as MFCC [34], GTCC [21], along with TEO-GTCC in GMM classifier. We take it's Δ, and $\Delta\Delta$ components resulting in the 39-D feature vector. Class-specific GMM models with different components were trained based on the feature using the expectation-maximization (EM) algorithm. The testing stage uses maximum likelihood decision among all the class models. Classification performance is measured using an accuracy (the number of correctly classified files among the total test files) and the confusion matrix. We select 16 component GMM model since from the experiment, the highest accuracy is achieved relatively, for 16 mixtures. We have also done the score-level fusion of two different feature sets, i.e.,

$$LLk_{comb} = \alpha LLk_1 + (1-\alpha)LLk_2, \quad (3)$$

where LLk_1 is likelihood of first feature set, LLk_2 likelihood of another feature set, and LLk_{comb} is a weighted fusion of likelihoods of two feature sets. α is the weight of fusion, that varies from 0 to 1 with step size of 0.1.

B. CNN classifier

We have also used the CNN classifier with architecture as proposed in [17]. However, we have not used data augmentation technique. As studied in detail in [16], the data augmentation techniques help to improve the performance of CNN. However, for some classes, the augmentation techniques degrade the performance of CNN. Since the objective of this paper is to compare the performance of different feature sets, we have not used the augmentation to analyze that how these features perform in all the classes.

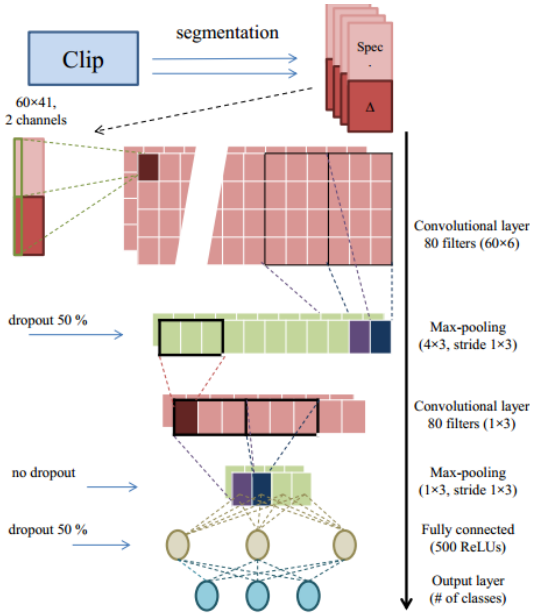

Fig. 3. CNN architecture for ESC used in this study. After [17].

Since CNN requires the input of the uniform dimensions and the length of the audio files varies across the database, the short segments of 41 frames were used as the input to the CNN. The segments were extracted with 50 % overlap from the audio files. The convolutional layers used in CNN were similar to as the ones used in [17]. Fig. 3 shows the details of each layer in the CNN architecture that we have used in ESC task. The network was implemented using Keras [35] with theano back-end on NVIDIA Titan-X GPU. A mini-batch implementation with 200 batch size was used to train the network using the stochastic gradient descent. Network parameters were similar as used in [17]. The Nesterov momentum of 0.9, learning rate of 0.002, L^2 regularization with the coefficient 0.001 and network was trained for 300 epochs on ESC-50, and UrbanSound8k databases to monitor the performance. At the testing time, the class of the test audio files was decided using the probability prediction scheme [17].

C. Experimental Results

To evaluate the performance of various feature sets, 5-fold, and 10-fold cross-validation was performed on ESC- 50, and UrbanSound8K databases, respectively. Table I shows that the results on both databases using all the feature sets for 16 component GMM. Moreover, to check the possibility of any complementary information captured by different feature sets, we have done their score-level fusion. It can be observed from Table I that TEO-GTCC gave better results than other features on both databases. The average classification accuracy of TEO-GTCC improved by 0.45 % and 1.40 % over MFCC on

ISBN 978-0-9928626-7-1 © EURASIP 2017

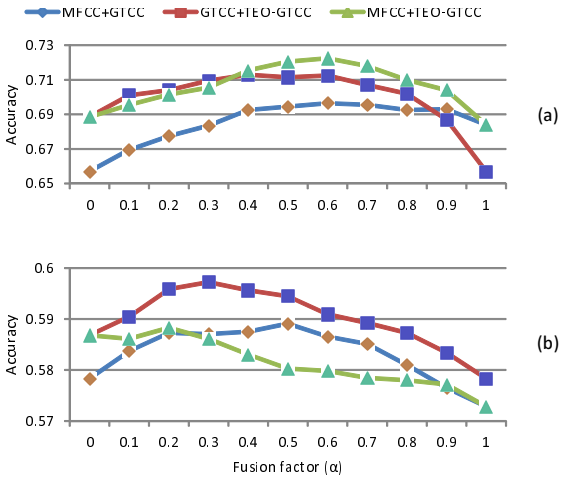

Fig. 4. Plot of an accuracy *vs*. α for different dataset on GMM for (a) ESC-50, (b) UrbanSound8K.

ESC-50 and UrbanSound8K datasets, respectively. In addition, TEO-GTCC provided an absolute improvement of 3.15 % and 0.85 % over GTCC on ESC-50 and UrbanSound8K datasets, respectively. Moreover, it can be observed that the score-level fusion of different features sets with proposed TEO-GTCC gave the improvement in an accuracy. The value of α varies with datasets and around 3.40 % and 1.04 % improvement in classification accuracy is achieved when TEO-GTCC is fused with MFCC and GTCC for ESC-50, and UrbanSound8k datasets, respectively. Fig. 4(a) and Fig. 4(b) shows the plot of an accuracy *vs*. α for ESC-50, and UrbanSound8K datasets, respectively, indicating that score-level fusion gave better performance in majority of the cases more for proposed TEO-GTCC feature set.

TABLE I
CLASSIFICATION ACCURACY (%) OF DIFFERENT FEATURE SETS ON DIFFERENT DATABASE WITH 16 COMPONENT GMM. THE $\oplus$ SIGN INDICATED SCORE-LEVEL FUSION (AS PER EQ. 3).

Feature Set	ESC-50		UrbanSound8K	
	α	Accuracy	α	Accuracy
MFCC	-	68.40	-	57.28
GTCC	-	65.70	-	57.83
TEO-GTCC	-	**68.85**	-	**58.68**
MFCC $\oplus$ GTCC	0.6	69.65	0.5	58.91
GTCC $\oplus$ TEO-GTCC	0.4	71.30	0.3	**59.72**
MFCC $\oplus$ TEO-GTCC	0.6	**72.25**	0.2	58.83

In CNN classifier, the accuracy is found to be more as compared to the GMM classifier. Table II shows the classification accuracy of ESC-50 and UrbanSound8K datasets using the CNN classifier with different feature sets. The performance of TEO-GTSC is not better as compared to GTSC in ESC-50, and UrbanSound8K. However, on ESC-50, and UrbanSound8K datasets, score-level fusion of GTSC and TEO-GTSC gave an absolute improvement of 14.10 %, and 14.52 %, respectively, compared to FBE. Since we have observed that score-level fusion of GTSC and TEO-GTSC performed better in all the experiments, we have experimented with the feature-level fusion of both feature sets. On ESC-50, and UrbanSound8K datasets,

TABLE II
CLASSIFICATION ACCURACY (%) OF DIFFERENT FEATURE SET ON DIFFERENT DATABASE WITH CNN CLASSIFIER. THE $\oplus$ SIGN INDICATED SCORE-LEVEL FUSION (AS PER EQ. 3), AND $\odot$ SIGN INDICATED FEATURE-LEVEL FUSION.

Feature Set	ESC-50		UrbanSound8K	
	α	Accuracy	α	Accuracy
FBE	-	67.85	-	73.50
GTSC	-	**79.10**	-	**85.34**
TEO-GTSC	-	74.85	-	79.65
GTSC $\odot$ TEO-GTSC	-	80.75	-	85.85
GTSC $\oplus$ FBE	0.5	79.65	0.9	86.02
TEO-GTCC $\oplus$ FBE	0.6	75.00	0.8	83.70
GTSC $\oplus$ TEO-GTSC	0.5	**81.95**	0.5	**88.02**

the feature-level fusion of GTSC and TEO-GTSC gave an absolute improvement of 12.90 %, and 12.35 %, respectively, compared to FBE. The score-level fusion and feature-level fusion of both feature sets gave better results as compared to individual one indicating that they capture complementary information. However, score-level relatively fusion performs better compared to feature-level fusion. Fig. 5 shows confusion matrix of UrbanSound8K dataset using GTSCs and TEO-GTSCs. As it can be observed from Fig.5(a),(b), TEO-GTSCs improve classification accuracy in dog barking (DB), gun shot (GS), and street music(SM). From that we can observe, TEO-GTSC captures repetitive, impulsive, and harmonic-like patterns from the audio signal much better then Mel spectral and Gammatone spectral based feature sets.

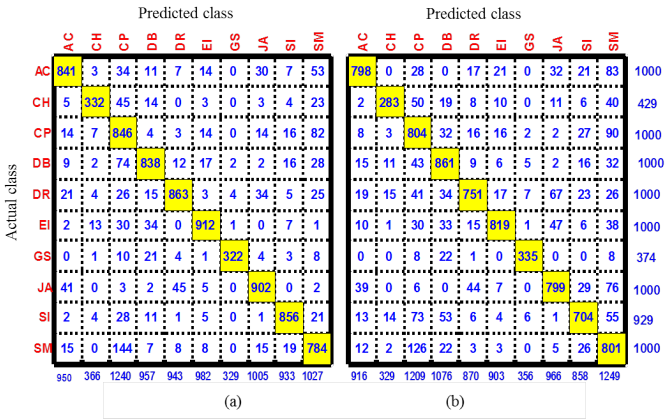

Fig. 5. Confusion matrix for Urbansound8k database using (a) GTSC and (b) TEO-GTSC. Classes are air conditioner (AC), car horn (CH), children playing (CP), dog barking (DB), drilling (DR), engine idling (EI), gun shot (GS), jackhammer (JA), siren (SI) and street music (SM).

IV. SUMMARY AND CONCLUSIONS

In this study, we proposed to use TEO-GTCC and TEO-GTSC feature sets for ESC task in GMM and CNN classifiers, respectively. Performance on ESC system was compared with different feature sets such as MFCC and FBE, GTCC and GTSC on two publicly available databases. Proposed TEO-GTCC feature set gave better results for this application with low feature dimension and GMM classifier. Moreover, the results suggested that using system combination of basic feature set and proposed feature set gave better accuracy than the individual feature sets. TEO-GTSC does not give

ISBN 978-0-9928626-7-1 © EURASIP 2017

better results by CNN classifier. However, using the score-level and the feature-level fusion of TEO-GTSC and GTSC gave better results than GTSC alone. According to the obtained class-based averaged accuracy, we conclude that TEO-based Gammatone features perform better compared to the conventional feature sets. Our feature work includes used of energy separation algorithm (ESA) to exploit AM and FM components of an audio signal for ESC task.

V. Acknowledgments

Authors would like to thank NVIDIA for providing hardware grant of Titan-X GPU for research purposes.

References

[1] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. of the* 23^{rd} *Int. Conf. on Multimedia*, Brisbane, Australia, 2015, pp. 1015–1018.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: a system for detecting anomalous sounds," *IEEE Trans. on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016.

[3] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2249–2256, 2007.

[4] M. Vacher, J.-F. Serignat, and S. Chaillol, "Sound classification in a smart room environment: an approach using GMM and HMM methods," in *The* 4^{th} *IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD), Publishing House of the Romanian Academy (Bucharest)*, vol. 1, 2007, pp. 135–146.

[5] L. Ballan, A. Bazzica, M. Bertini, A. Del Bimbo, and G. Serra, "Deep networks for audio event classification in soccer videos," in *Int. Conf. on Multimedia and Expo (ICME)*. New York, USA: IEEE, 2009, pp. 474–477.

[6] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 6450–6454.

[7] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 6445–6449.

[8] B. Ghoraani and S. Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[9] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 151–155.

[10] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 171–175.

[11] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," in *European Signal Processing Conf. (EUSIPCO)*, Nice, France, 2015, pp. 714–718.

[12] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[13] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[14] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America (JASA)*, vol. 122, no. 2, pp. 881–891, 2007.

[15] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1–7.

[16] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *CoRR*, vol. abs/1608.04363, 2016. [Online]. Available: http://arxiv.org/abs/1608.04363 {Last accessed on 26^{th} February, 2017 }

[17] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 25^{th} *Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, USA, 2015, pp. 1–6.

[18] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition." in *INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3013–3016.

[19] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, 2007, pp. 645–649.

[20] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 4625–4628.

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

[22] H. Luo, Y. Wang, D. Poeppel, and J. Z. Simon, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence," *Journal of Neurophysiology*, vol. 96, no. 5, pp. 2712–2723, 2006.

[23] D. D. Greenwood, "A cochlear frequency-position function for several species−29 years later," *The Journal of the Acoustical Society of America (JASA)*, vol. 87, no. 6, pp. 2592–2605, 1990.

[24] L. H. Carney and T. Yin, "Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model," *Journal of Neurophysiology*, vol. 60, no. 5, pp. 1653–1677, 1988.

[25] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[26] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, Apr 1990, pp. 381–384 vol.1.

[27] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[28] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, W. Hardcastle and A. Marchal, Eds. Springer, 1990, pp. 241–261.

[29] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America (JASA)*, vol. 124, no. 1, pp. 422–438, 2008.

[30] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[31] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. of the* 22^{nd} *Int. Conf. on Multimedia*, Orlando, Florida, 2014, pp. 1041–1044.

[32] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8604–8608.

[33] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[34] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[35] F. Chollet, "Keras," https://github.com/fchollet/keras { Last accessed on 26^{th} February, 2017}.