Environmental Sound Recognition: A Survey

BoreGowda H B
Asst. Prof ECE, GSSSIETW, Mysuru, Karnataka, INDIA

***Abstract*- There search in audio recognition has traditionally focused on speech and music signals, the problem of environmental sound recognition (ESR) has received more attention in recent years. Research on ESR has significantly increased in the past decade. Recent work has focused on the appraisal of non-stationary aspects of environmental sounds, and several new features predicated on non-stationary characteristics have been proposed. These features strive to maximize their information content pertaining to signal's temporal and spectral characteristics. Furthermore, sequential learning methods have been used to capture the long-term variation of environmental sounds. In this survey, we will offer a qualitative and elucidatory survey on recent developments. It includes three parts: i) basic environmental sound processing schemes, ii) stationary ESR techniques and iii) non-stationary ESR techniques. Finally, concluding remarks and future research and development trends in the ESR field will be given.**

***Index Terms*- Environmental Sound Recognition (ESR), Mel filter, Audio Features, Mel-Frequency Cepstral Coefficients.**

I. INTRODUCTION

By environmental sounds, we refer to various quotidian sounds, both natural and artificial (i.e. sounds one encounters in daily life other than speech and music). Research on ESR has significantly increased in the last decade. Recent work has focused on the appraisal of non-stationary aspects of environmental sounds, and several new features predicated on non-stationary characteristics have been proposed. Environmental sound recognition (ESR) plays a pivotal part in recent efforts to perfect machine audition. ESR can be used in automatic tagging of audio files with descriptors for keyword-based audio retrieval [9]. Robot navigation can be improved by incorporating ESR in the system [5].ESR can be adopted in a home-monitoring environment, be it for assisting elderly people living alone in their own home [3] ESR, along with image and video analysis, find applications in surveillance [7], [24]. ESR can also be tailored for recognition of animal and bird species by their distinctive sounds [1]. Among various types of audio signals, speech and music are two categories that have been extensively studied.

ESR algorithms were a reflection of speech and music recognition. For example, the speech recognition task often exploits the phonetic structure that can be viewed as a basic building block of speech. It allows us to model complicated spoken words by breaking them down into elementary phonemes that can be modeled by the Hidden Markov Model (HMM) [22]. In contrast, general environmental sounds, such as that of a thunder or a storm, do not have any apparent substructures like phonemes. Even if we were able to identify and learn a dictionary of basic units (analogous to phonemes in speech) of these events, it would be difficult to model their variation in time with HMM as their temporal occurrences would be more random as against preordained sequence of phonemes in speech. Similarly, as compared to music signals, environmental sounds do not exhibit meaningful stationary patterns such as melody and rhythm [23]. To the best of our knowledge, there was only one survey article on the comparison of various ESR techniques done by Cowling and Sitte [6] about a decade ago. For most real life sounds, even these features exhibit non-stationarity when observed over a long period of time. To capture these long-term variations, sequential learning methods have been applied. Despite increased Interest in the field, there is no single consolidated database for ESR, which often hinders benchmarking of these new algorithms

II. ENVIRONMENTAL SOUND PROCESSING SCHEMES

The three commonly used environmental sound processing schemes are

1) Framing-based processing: Audio signals to be classified are first divided into frames, often using a

Haaning or a Hamming window. Features are extracted from each frame and this set of features is used as one instance of training or testing. A classification decision is made for each frame and, hence, consecutive frames may belong to different classes. A major drawback of this processing scheme is that there is no way of selecting an optimal framing-window length suited for all classes. Some sound events are short-lived. (E.g. gunshot) as compared to other longer events (e.g. thunder).

2) Sub-framing-based processing: Each frame is further segmented into smaller sub-frames, usually with overlap, and features are extracted from each sub-frame. In order to learn a classifier, features extracted from sub-frames are either concatenated to form a large feature vector or averaged so as to represent a single frame.

3) Sequential processing: Audio signals are still divided into smaller units (called a segment), which is typically of 20- 30 ms long with 50% overlap. The classifier makes decisions on class labels and segmentation both based on features extracted from these segments. As compared to the above two methods, this method is unique in its objective to capture the inter-segment correlation and the long-term variations of the underlying environment sound.

III. STATIONARY ESR TECHNIQUES

Features developed for speech/music based applications have been traditionally used in stationary ESR techniques. These features are often based on psychoacoustic properties of sounds such as loudness, pitch, timbre, etc. A detailed description of features used in audio processing was given in [17].

Cepstral features are widely used features. They include: Mel-Frequency Cepstral Coefficients (MFCC) and their first and second derivatives (_MFCC and __MFCC), Homomorphic Cepstral Coefficients (HCC), Bark-Frequency Cepstral Coefficients (BFCC), etc. MFCC were developed to resemble the human auditory system and have been successfully used in speech and music applications. As mentioned before, due to lack of a standard ESR database, MFCC are often used by researchers for benchmarking their work. A common practice is to concatenate MFCC features with newly developed features to enhance the performance of a system. Filter-banks are often used to extract features local to smaller bands, encapsulating spectral properties effectively. On the other hand, the auto-correlation function (ACF) represents the time-evolution and has an intimate relationship with the power spectral density (PSD) of the underlying signal. Valero and Alias [33] proposed a new set of features called the Narrow-Band Auto Correlation Function features (NB-ACF). The extraction of NB-ACF features can be explained using Fig.2. First, a signal is passed through a filter bank with N = 48 bands whose center frequencies being tuned to the Mel-scale. Then, the sample ACF of the filtered signal in the ith band is calculated, which is denoted by i( ). One can calculate four NB-ACF features based on each ACF as follows.

1) i (0): Energy at lag _ = 0. It is a measure of the perceived sound pressure at the i^{th} band.

2) i1: Delay of the first positive peak which represents the dominant frequency in the i^{th} band.

3) i1 (i1): Normalized ACF of the first positive peak. It is related to the periodicity of the signal and, hence, gives a sense of pitch of the filtered signal at the ith band.

4) ie: Effective duration of the envelope of normalized ACF. It is defined as the time taken by normalized ACF to decay 10 dB from its maximum value, and it is a measure of reverberation of the filtered signal at the ith band.

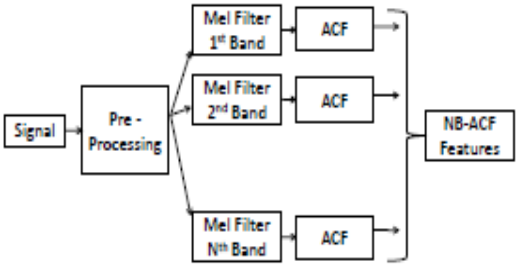

Fig 1.1: Illustration of the NB-ACF feature extraction process

IV. NON-STATIONARY ESR TECHNIQUES

The performance of commonly employed features for audio recognition, including Mel-Frequency Cepstral Coefficients (MFCC), Homomorphic cepstral Coefficient(HCC), time frequency features derived using Short-Term Fourier Transform (STFT), Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) was compared by Cowling and Sitte in [6], where the Learning Vector

Quantization (LVQ), Artificial Neural Networks, Dynamic Time Warping (DTW) and Gaussian Mixture Models (GMM) were used as classifiers. The experiments were conducted on three types of data – speech, music and environmental sounds. For the environmental sound, the data set consisted of 8 classes, and the framing-based processing scheme was adopted. It was reported that the best performance for ESR was achieved with CWT features with the DTW classifier, which was comparable to that of MFCC features with the DTW classifier. It is surprising that CWT, which is a time-frequency representation, and MFCC gave very similar results while DWT and STFT did not give good performance. It was noted in [6] that the dataset was too small to make any meaningful comparison between MFCC and CWT. Given other factors being equal, MFCC features can be more favored than CWT features because of their lower computational complexity. DTW was clearly the best classifier in the test, yet the claim should be further verified by a larger environmental sound database. Han and Hwang [13] used the Discrete Chirplet Transform (DChT) and the Discrete Curvelet Transform (DCuT) along with several other common features such as MFCC, ZCR, etc. When compared, all features gave similar performance, yet significant improvement was observed when they were used together.

V. FLOW CHART

The module is the recognition module and the system will output the speech content based on MFCC extraction, where the logic of the main flows is shown below

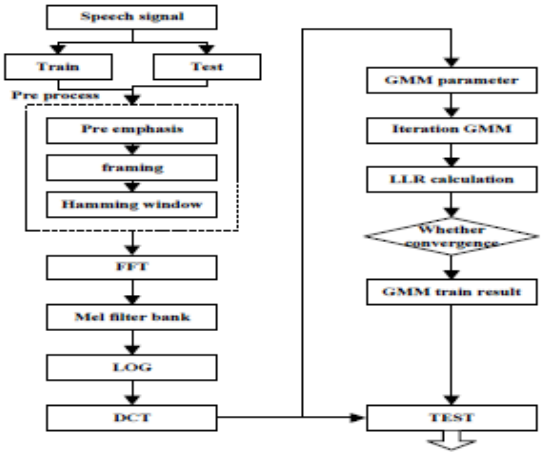

Fig1.2: Flow Chart of Sound Recognition

V. CONCLUSION AND FUTURE WORK

We did an in-depth survey on recent developments in the ESR field in this paper. Existing ESR methods can be categorized into two types: stationary and non-stationary ESR techniques. The stationary ESR techniques are dominated by spectral features. While these features are easy to compute, there are limitations in the modeling of non-stationary sounds. The non-stationary ESR techniques obtain features derived from the wavelet transform, the sparse representation and the spectrogram MFCC features are often combined with one or more features to boost classification accuracy furthermore. While the non-stationary methods give improved performance, they are often computationally expensive.

VI. RESULT

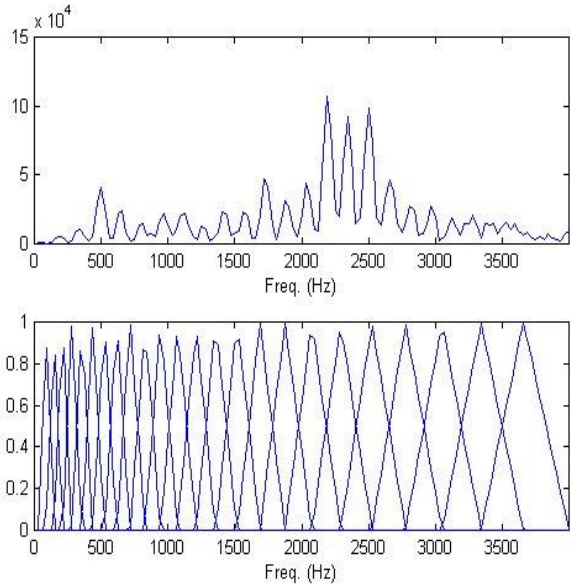

Fig 1.3: Frequency signal of Human Voice

REFERENCES

[1] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H.Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," Pattern Recognition Letters, vol. 31, no. 12, pp. 1524–1534, 2010.

[2] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor solution to the Netflix prize," KorBell Teams Report to Netflix, 2007.

[3] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on

sound," in Pervasive Computing. Springer, 2005, pp. 47–61.

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

[5] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006, pp. 885–888.

[6] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," Pattern Recognition Letters, vol. 24, no. 15, pp. 2895–2907, 2003.

[7] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," IEEE Trans. Multimedia, vol. 9, no. 2, pp. 257–267, 2007.

[8] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," IEEE Trans. Syst., Man, Cybern. B, vol. 38, no. 2, pp. 429–438, 2008.

[9] S. Duan, J. Zhang, P. Roe, and M. Towsey, "A survey of tagging techniques for music, speech and environmental sound," Artificial Intelligence Review, pp. 1–25, 2012.

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

[11] "Discriminant non-stationary signal features' clustering using hard and fuzzy cluster labeling," EURASIP Journal on Advances in Signal Processing, vol. 2012, no. 1, p. 250, 2012.

[12] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha, "Song / instrumental classification using spectrogram based contextual features," in Proceedings of the CUBE International Information Technology Conference. ACM, 2012, pp. 21–25.

[13] B.-j. Han and E. Hwang, "Environmental sound classification based on feature collaboration," in Multimedia and Expo, 2009.ICME 2009.IEEE International Conference on. IEEE, 2009, pp. 542–545.

[14] M. Karbasi, S. Ahadi, and M. Bahmanian, "Environmental sound classification using spectral dynamic features," in Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on. IEEE, 2011, pp. 1–5.

[15] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," 2013, (in press).[Online]. Available:http://www.sciencedirect.com/science/article/pii/S0020025513003113

[16] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in Proc. The Fourth Workshop on Feature Selection in Data Mining, vol. 4, 2010, pp. 4–13.

[17] D. Mitrovi´c, M. Zeppelzauer, and C. Breiteneder, "Features for contentbasedaudio retrieval," Advances in computers, vol. 78, pp. 71–150, 2010.

[18] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and Mel-Frequency Cepstral Coefficients," in Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE, 2010, pp. 11– 16.

[19] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on,vol. 2. IEEE, 2002, pp. II–1941.

[20] J. Pickens, "A survey of feature selection techniques for music information retrieval," 2001.

[21] I. Potamitis and T. Ganchev, "Generalized recognition of sound events: Approaches and Applications," in Multimedia Services in Intelligent Environments. Springer, 2008, pp. 41–79.

[22] L. R.Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[23] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music

content: a survey," IEEE Signal Process. Mag., vol. 23, no. 2, pp. 133–141, 2006.

[24] R. Sitte and L. Willets, "Non-speech environmental sound identification for surveillance using self-organizing-maps," in Proceedings of the Fourth conference on IASTED International Conference: Signal Processing, Pattern Recognition, and Applications, ser. SPPR'07. Anaheim, CA, USA: ACTA Press, 2007, pp. 281–286. [Online].Available:http://dl.acm.org/citation.cfm ?id=1331978.1332027