Empirical Mode Decomposition Based Feature Set Optimization for Categorization of Real-World Noisy Environments

Sujay D. Mainkar
Assistant Professor, E & TC Engineering Department,
Finolex Academy of Management & Technology, Ratnagiri,
Maharashtra, India
smart.extc@gmail.com

Abstract

In inter-human or human-machine verbal interactions, information misinterpretation frequently occurs due to disturbance signals, e.g., acoustic echo signals, background noises, etc. In most of the practical situations, background noise is inevitable and dominantly distracts attention of the listeners. Such noisy atmosphere can significantly reduce the intelligibility and reliability of the speech recognition system thereby making speech recognition as really challenging task. Moreover, non-stationary noise signals in day-to-day life such as office noise, train noise, factory noise, restaurant/cafeteria noise constantly vary their properties making the speech enhancement more difficult. So, one approach is to classify noise and then to train speech recognizers with that noise in the background. This paper attempts to classify environmental background noise using Huang Transform also known as Empirical Mode Decomposition (EMD) which considers inherent non-stationarity of noisy speech signal by decomposing the signal into Intrinsic Mode Functions (IMFs). These IMFs are used for feature extraction. This work suggests two types of composite feature vectors formed by- (i) Uni-Feature Multi-IMF and (ii) Multi-Feature Uni-IMF for classification and propose an optimized best suitable feature set for classification of different noisy environments. For classification, Maximum-Likelihood Gaussian Mixture Model (ML-GMM) and k-Nearest Neighbor (k-NN) classifiers are used. Utilization of this optimized best suitable feature set yields the maximum accuracy of 98.33 % in multiclass noise classification. Further, this proposed optimized feature set proved to be independent of speakers, gender of speaker and utterances to identify surrounding environment of the speaker.

Keywords: *Empirical Mode Decomposition, Intrinsic Mode Function, k-Nearest Neighbor classifier, Maximum-Likelihood Gaussian Mixture Model.*

1. Introduction

In recent years, advancements in fields of information and communication technology(ICT), multimedia technology and rapid development of digital networks has opened new research avenues in the field of audio classification. An audio signal classification system should be able to classify different audio input formats. In this task of audio classification, noise is a major problem. There are several problems which noise imposes in audio processing. To deal with practical noisy environments, it is worth to have an idea about properties of the noise itself. Environmental background noise signals are mostly categorized into two classes: stationary and non-stationary noise signals. Fan noise is one of the examples of stationary noise, where statistical characteristics remain unchanged over time. Examples of non-stationary noise include traffic noise or crowd of people speaking in the background, etc., where statistical characteristics constantly change w.r.t. time. Such noise affects audio applications beyond just speech recognition. Maithani and Tyagi (2008) note that noise affects e.g. speech compression, cell phone comfort and hearing aids [10]. They suggest noise classification as one solution to problems in these areas as well as for speech recognition. The fact that noise signals can be characterized by their evolvement over time and frequency motivates to consider

correct temporal and spectral characteristics to obtain unique signature of corresponding source.

The five audio classes: silence, speech, music, speech with music and speech with noise is classified using feature extraction matrix in [1]. The classification of traffic noise sources: motorbikes, cars and heavy trucks are made in [2] by using spectral features like spectral centroid, spectral roll-off, sub-band energy ratio and zero-crossing rate as temporal feature. Three feature sets for representing timbral texture, rhythmic content and pitch content of music signals were proposed and evaluated using statistical pattern recognition classifiers with 61% accuracy in [3] for classifying ten musical genres. Environmental sound classification is done in [4] with the help of Chirplet, curvelet, and Hilbert transforms. In [5], authors have used discrete wavelet transform (DWT) to discriminate between speech and music. In [6], authors have presented an algorithm for audio classification that is capable of segmenting and classifying an audio stream into speech male, speech female, music, noise and silence. They also have put forward best suited features for multiclass classification yielding accuracy of 96.34% in audio discrimination. Four types of background noise sources are classified using EMD with discrimination success rate of 77% to 85% in [7]. This literature review identified a need for an implementation of multiclass noisy environment classification with accuracy improvement using EMD to deal with inherent non-stationarity of audio signal.

In this paper, we propose an optimized most appropriate feature set to distinguish between different categories of environmental background noise sources irrespective of speakers, their gender and utterances to understand the surrounding environment of the speaker.

The further outline of this paper includes overview of EMD in section 2. The basic idea behind proposed methodology is presented in section 3. Section 4 describes feature extraction and feature selection. The classification algorithms used are discussed in Section 5. Section 6 presents experimental results along with database description followed by conclusion and future scope in section 7.

2. Empirical Mode Decomposition

Norden E. Huang introduced a promising tool for data analytics in the form of Huang transform also commonly known as Empirical Mode Decomposition (EMD). In contrast to almost all the conventional transform methods, EMD works in temporal space directly rather than in the corresponding frequency space; it is adaptive, with an a posteriori defined basis derived from the data. EMD is an adaptive data analysis method that is based on local characteristics of the data, and hence, it catches nonlinear, non-stationary oscillations more effectively. EMD method is able to decompose a complex signal into a series of intrinsic mode functions (IMFs) and a residue [8].

2.1. Intrinsic Mode Function

EMD decomposition has implicitly a simple assumption that, at any given time, the data may have many coexisting simple oscillatory modes of significantly different frequencies, one superimposed on the other. Each component is defined as an Intrinsic Mode Function (IMF) satisfying the following conditions:

1. In the entire data set, the number of extrema and the number of zero-crossings must either equal or differ at the most by one.
2. At any data point, the mean value of the envelope defined using the local maxima and the envelope defined by the local minima must be zero.

2.2. Sifting Process

The purpose of sifting is to subtract the large-scale features of the signal repetitively until only the fine-scale features remain. First, the original noisy audio signal, x(t),

should be enclosed by the upper and lower envelope in the time domain. Using cubic-spline interpolation, the local maxima is connected forming the upper envelope U(t) and the local minima is connected forming the lower envelope L(t). These two envelopes cover up all the data points. The local mean envelope m(t) is determined as follows:

$$m(t) = \{U(t) + L(t)\}/2 \quad (1)$$

The first component is described as,

$$h_1(t)=x(t)- m(t) \quad (2)$$

The component $h_1(t)$ is now examined to see if it satisfies the conditions to be an IMF. If $h_1(t)$ does not satisfy the conditions, $h_1(t)$ is regarded as the original data and the sifting process would repeat, obtaining the mean of the upper and lower envelopes, which is designated as m_{11}; so:

$$h_{11}(t)=h_1(t)–m_{11}(t) \quad (3)$$

We must repeat this procedure until $_{h1k}$ is an IMF,

$$h_{1k}(t)=h_{1(k-1)}(t)–m_{1k}(t) \quad (4)$$

After k siftings, we get the first IMF component as;

$$c_1 = h_{1k} \quad (5)$$

Finally, c_1 revealed the higher frequency component of IMF. To obtain enough physical definitions of IMF, the sifting stoppage criterion, known as the stop condition is of key importance found by determining standard deviation (SD), given by:

$$SD_k = \frac{\sum_{t=0}^{T} \left| h_{k-1}(t) - h_k(t) \right|^2}{\sum_{t=0}^{T} h_{k-1}^2(t)} \quad (6)$$

If SD is smaller than a predetermined threshold, the sifting process is to be stopped. The typical values of SD are 0.2 and 0.3. To obtain the second and subsequent intrinsic mode functions of noisy audio stream, the residual signal can be calculated as:

$$x(t)–c_1(t)=r_1(t) \quad (7)$$

r_1 considers the original data, and by repeating the above procedures, x(t) could be obtained by the second IMF component c_2. The procedure as described above is repeated for n times, so as to obtain n-IMFs of signal x(t) as:

$$r_{n-1}(t) – c_n(t) = r_n(t) \quad (8)$$

This sifting process can be stopped by any of the following predetermined criteria: either when the component (c_n), or the residue(r_n), becomes so small that it is less than the predetermined value of substantial consequence, or when the r_n, becomes a monotonic function from which no more IMF can be extracted. We finally obtain:

$$x(t) = \sum_{i=1}^{n} c_i + r_n \quad (9)$$

by adding all IMFs and residue.

The entire sifting process serves two purposes: to eliminate riding waves and to make the wave-profiles more symmetric.

3. Proposed Methodology

The input noisy audio signal is non-stationary in nature. The time-frequency analysis of such a signal is possible through proper pre-processing. In the proposed approach, initially, the silence period is removed from the noisy audio clip under test. The remaining noisy audio stream is then decomposed into frames with frame period of 50ms and overlap period of 25ms. This frame overlapping ensures that audio features occurring at a discontinuity are at least considered whole in the subsequent overlapped frame. These frames are then decomposed into number of Intrinsic Mode Functions (IMFs). Different temporal and spectral features are extracted from these IMFs to form composite feature vectors as described in section IV below. Finally, classification is done using Maximum

Likelihood Gaussian Mixture Model (ML-GMM) and k-Nearest Neighbor (k-NN) classifiers, followed by performance characterization of each composite feature set so as to conclude with unique optimized robust feature set which is best suitable for efficient discrimination of various noisy environments.

4. Feature Extraction and Feature Selection

4.1. Feature Extraction

The feature extraction is an essential processing step in multiclass audio classification tasks. The goal is to extract that set of features from the noisy audio stream of interest which is capable of conveying maximum information regarding desired characteristics of the original signal. Feature extraction involves the analysis of the noisy input audio stream. The feature extraction techniques can be classified as temporal analysis and spectral analysis technique. Temporal analysis uses the time-domain waveform of the audio signal itself for analysis. Spectral analysis utilizes spectral representation of the audio signal for analysis. All audio features are extracted from IMFs generated by breaking the input signal into a succession of analysis windows or frames, and computing one feature value for each of the windows.

4.2. Feature Selection

From a large set of features it is important to select particular set of features that would determine the nature and hence the class of the audio signal. These features determine the dimensionality in the feature space. It is important therefore to select optimum number of features that not only keeps accordance with the accuracy and the level of performance but also reduces the computation costs. Thus there is no point in just increasing the number of features as it would not have a drastic impact on the accuracy but would pave for more complexities in computation. Therefore a selected feature must have the following properties,

1) Invariance to irrelevancies: Any good feature should exhibit invariance to irrelevancies such as noise, bandwidth or the amplitude scaling of the signal. It is also upon the classification system to consider such variations as irrelevant to achieve better classification across a wide range of audio formats.

2) Discriminative Power: The purpose of feature selection is to achieve discrimination among different classes of audio patterns. Therefore a feature must take round about similar values within the same class but different values across different classes.

3) Uncorrelated to other features: It is very important that there are no redundancies in the feature space. Each new feature that is selected must give altogether different information about the signal as possible [9].

In this work, we have selected Short Time Autocorrelation Function (ACF), Short Time Energy (STE) and Zero Crossing Rate (ZCR) as temporal features and Spectral Centroid (SC), Spectral Roll off (SR) and Spectral Flux (SF) as spectral features [3]. Further, we also have chosen Mel Frequency Cepstral Coefficients (MFCC) as one of the feature.

4.3. Composite Feature Set Formation

In this work, we perform experimentations by taking into account two types of composite feature vectors as explained here:

1) Uni-Feature Multi-IMF (UFMI) Feature Set:

In this type, we form a feature set by considering single feature associated with all frames of multiple IMFs.

2) Multi-Feature Uni-IMF (MFUI) Feature Set:
In this type, we construct a feature set by combining multiple features for all frames of first IMF.

5. Classification

For the classification purpose, we have used k-Nearest Neighbor (k-NN) classifier which is an instance based classifier and Maximum-Likelihood Gaussian Mixture Model (ML-GMM) classifier.

5.1. k-Nearest Neighbor Classifier

The k-NN algorithm (k-Nearest Neighbor) can be classed as a nonlinear non-parametric classification method. This algorithm is based on very simple principle that similar data are close to each other in the searching or data space. In other words, for every object from test data set of k objects the k-NN finds the training data that are closest to the test object (nearest neighbors). The label assignment is usually based on the rule of majority voting, e.g. the most frequent class from the k nearest neighbors for given test object determines the class where this object should belong. A value of k dictates a number of closest objects from training data that are taking into account at the label decision. If the value is too small, then the result can be sensitive to noise points. If it is too large, then the neighborhood may include too many points from other classes [13].

Example of k-value impact to classification result is shown in Figure 1, where, k-NN classifier classifies two dimensional data into two classes. First circle represents a region with three neighbors are involved into making decision where orange point is belonging. In this case k value is set to three (k=3) and classified data/point belongs to "red" class. Second circle represents six neighbors (k = 6) considered in classification task. In the second case, the classification result is an opposite and unknown data/point belongs to "blue" class [11].

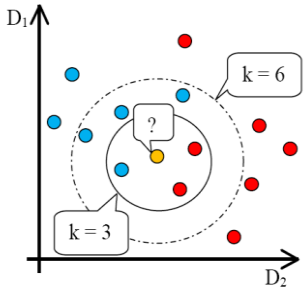

Figure 1. Illustrative Example of k-NN Classification [11]

Besides a k value, the distance metric is important to the k-NN algorithm. As can be clearly seen, the distance metric represents the measure of data similarity. The choice of particular distance metric usually depends on the given classification problem. Regardless simplicity of k-NN, this method is well suitable for multi-modal classes, very flexible and belongs to top 10 data mining algorithms (IEEE Conference on data mining 2007 [12]).

5.2. Gaussian Mixture Model Classifier

In GMM classification Gaussian mixture model is used for statistical representation of noisy audio patterns. The distribution of feature vectors extracted from noisy audio stream is modeled by a mixture of Gaussian density functions (Figure 2). Complete GMM is defined by mean vector, covariance matrix and mixture weights. Every recognized noisy environment type has its own model which

is then used as its characteristic representation instead of speakers and utterances to understand the surrounding environment of the speaker [13].

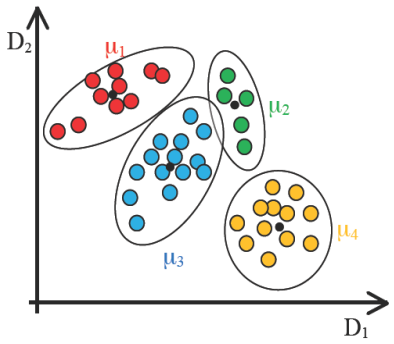

Figure 2. Illustrative Example of modeling 2-dimensional data using 4-Gaussian mixtures [11]

The identification assignment is maximum likelihood classifier. The main task of the system is to make a decision if input noisy audio belongs to one of the set of noisy environments, which are represented by its models. This decision is based on computation of maximum posterior probability for input feature vector [11].

6. Experimental Results

For our experimentation purpose, we have used NOIZEUS database. NOIZEUS is a noisy speech corpus recorded in laboratory to facilitate comparison of speech enhancement algorithms among research groups. The noisy database, corrupted by eight different real-world noises at different SNR, contains 30 IEEE sentences produced by three different male and three female speakers. Thirty sentences are taken from the IEEE sentence database as these sentences are phonetically balanced with relatively low word-context predictability and recorded in a sound proof chamber using Tucker Davis Technologies (TDT) recording equipments. The sentences were originally sampled at 25 KHz and down sampled to 8 kHz. The noise signals were added to the clean speech signal at SNR of 5 dB, 10dB and 15dB. Out of these 30 samples, first 10 noisy samples are associated with male speakers followed by next 10 noisy samples include female speakers and in samples 21 to 30, first five represent male speakers whereas last five represent female speaker [13].

The system is used for multiclass classification of 4 representative noisy environment types namely-babble, car, exhibition hall and train noise (30 samples each) for 0 dB SNR level. We have used first 15 samples (10 male speakers and 5 female speakers) for training and next 15 samples (5 male speakers and 10 female speakers) for testing. Results of different experimentations performed under varying conditions are presented here:

6.1. Classification using UFMI feature set

In this, we have formed a composite feature vector comprising of ZCR (single feature) of first 3 IMFs (multiple IMFs) for all the frames of all 30 samples associated with 4 noisy environments and the results are presented in Figure 3.

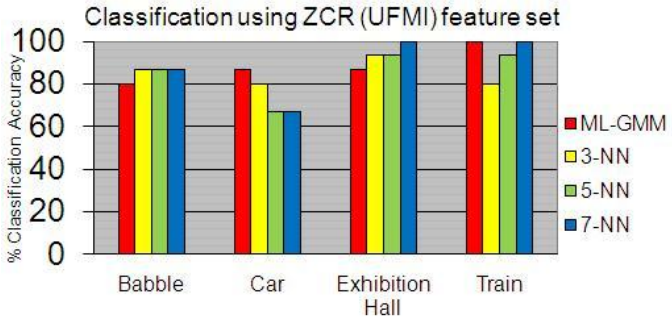

Figure 3. Classification using ZCR for first 3 IMFs

Experimental results in figure 3 show that, ML-GMM and 7-NN classifiers perform equally well giving overall classification accuracy of 88.33% when ZCR is used. Similar experimentation is carried out for SF and the results are shown in figure 4. It can be observed from figure 4 that, SF is best suited for classifying babble and car noises giving 100% accuracy and 5-NN classifier yields best results with an average accuracy of 91.67% when SF is used.

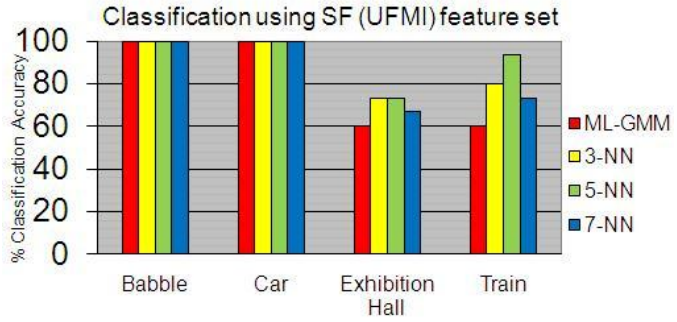

Figure 4. Classification using SF for first 3 IMFs

Similar procedure is repeated for STE and MFCC and corresponding results are shown respectively in figure 5 and figure 6.

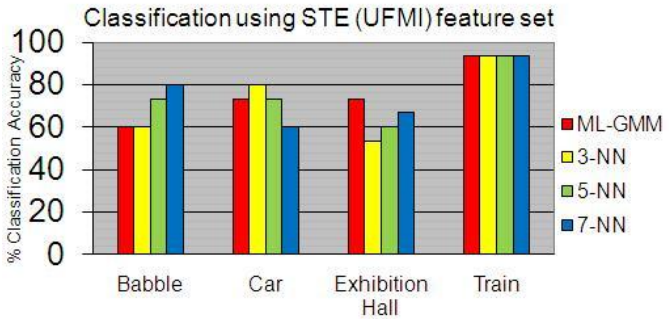

Figure 5. Classification using STE for first 3 IMFs

Figure 5 illustrates that although, in case of STE based classification all 4 classifiers give moderate accuracy of around 70-75% still it is most appropriate to classify train noise because irrespective of variation in classifier constant accuracy of 93.33 % is achieved.

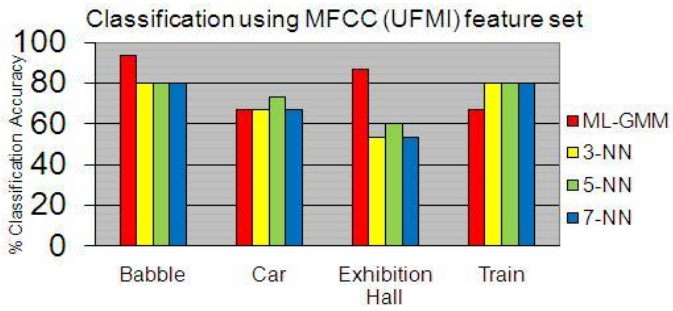

Figure 6. Classification using MFCC (max.) for first 3 IMFs

Mel-frequency cepstral coefficients (MFCC) are a set of perceptually motivated features that have been popularly used in audio recognition [9]. Figure 6 indicates that, classification using MFCC represents moderate behavior with maximum accuracy of 78.33% for ML-GMM classifier, when it is considered as single feature. So, for improvement in success rate of classification we put forth the concept of Multi-Feature Uni-IMF (MFUI) Feature Set.

6.2. Classification using MFUI feature set

During this work, we have formed a composite feature vector consisting of multiple features corresponding to first IMF (uni-IMF), for all the frames of all 30 samples associated with 4 noisy environments and the results are presented here:

Figure 7 highlights the performance of 4 classifiers when 3 features namely- ZCR, SF and MFCC combined together with their respective first IMF, so as to form a feature vector. Further, figure 8 and figure 9, respectively, put a glance on classification accuracy achieved by combining first IMF of SF-MFCC and ZCR-MFCC to form a feature set.

From figure 7, we can observe that ZCR-SF-MFCC feature set yields considerable variation in accuracy from as low as 73.33% with 7-NN classifier to as high as 98.33% for ML-GMM classifier. Moreover, typically this MFUI set performs poor while classifying exhibition hall noise, achieving classification accuracy as low as 26.67% using 7-NN classifier. At the same time, this MFUI set gives 100% accuracy for both car and train noises with all classifiers.

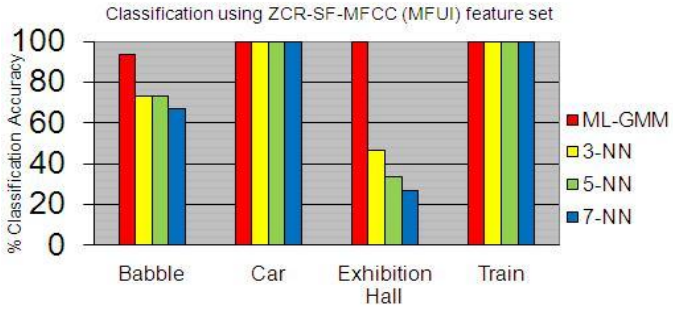

Figure 7. Classification using ZCR-SF-MFCC for first IMF

Out of 3 features used in figure 7, when ZCR is omitted in MFUI set formation(Figure 8), overall classification performance decreases in the range of 60% to 70% with drastic reduction in classification accuracy of ML-GMM classifier from 98.33% to 61.67%. But, at the same time it must also to be noted that, SF-MFCC feature pool provides 100% classification accuracy for car noise, irrespective of variation in type of classifier used.

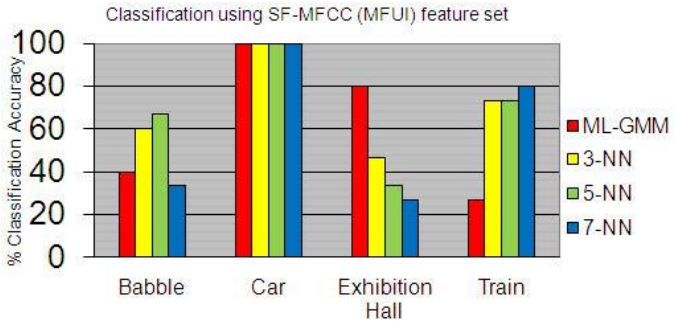

Figure 8. Classification using SF-MFCC for first IMF

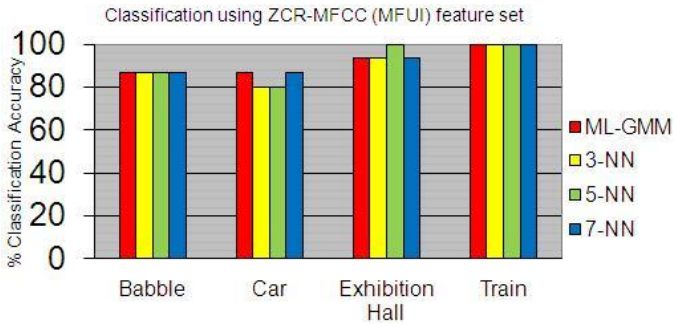

Figure 9. Classification using ZCR-MFCC for first IMF

On the other hand, as shown in figure 9, out of 3 features represented in figure 7, when SF is dropped forming ZCR-MFCC feature set then average accuracy is improved to around 90% and above, with all classifiers and irrespective of change in classifier type, this MFUI set shows robustness for babble noise and train noise with categorization accuracy of 86.67% and 100%, respectively.

Figure 10 represents the performance of 4 classifiers when 3 features namely- ZCR-STE-MFCC are used together with their respective first IMF to form a MFUI feature pool.

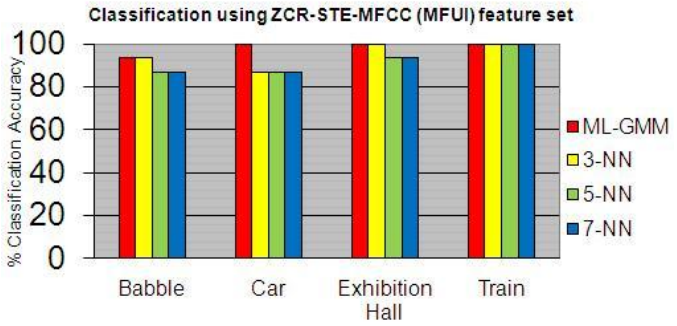

Figure 10. Classification using ZCR-STE-MFCC for first IMF

As can be seen from figure 10, ZCR-STE-MFCC (MFUI) feature set shows best performance for discrimination of all 4 noisy environments using all 4 classifiers, with average classification accuracy of 94.17% and maximum accuracy of 98.33% for ML-GMM classifier. Further, in order to test for robustness, we have evaluated the performance of this MFUI set for 2 different distance metrics of k-NN (k=3,5,7) classifier, namely- Euclidean distance (second order norm between two points) and Manhattan (city block) distance (sum of absolute differences) and found that accuracy remains almost unchanged. Moreover, by using this best feature set, we have experimented for training with only male speaker noisy samples and testing for only female noisy samples. The performance results so obtained are presented in figure 11.

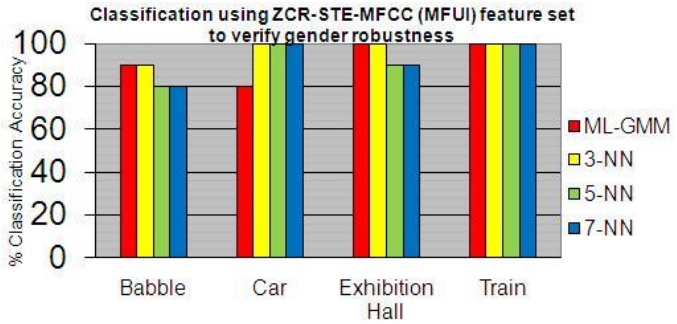

Figure 11. Classification using ZCR-STE-MFCC for first IMF when only male speaker noisy samples used for training and only female noisy samples used for testing

From figure 11, we can see that, classification ability of ZCR-STE-MFCC (MFUI) feature set still remains well above 90%. So, we can consider that, proposed optimized feature set is best suited for multiclass noisy environment classification irrespective of speakers, gender of speaker and utterances to identify surrounding environment of the speaker.

7. Conclusion

In this paper, we explained and evaluated ML-GMM and k-NN (k=3,5,7) classifiers used for multiclass noisy environment classification task. Classification accuracy for the 4 noises- babble, car, exhibition hall and train from NOIZEUS speech corpus was computed for UFMI feature set as well as MFUI feature set, formulated on the basis of EMD. The best classification accuracy of 98.33% was attained by ML-GMM classifier using ZCR-STE-MFCC as MFUI feature set; irrespective of speakers, their gender and utterances to know surroundings of the speaker. Further, this feature set has also provided accuracy of more than 90% for k-NN (k=3,5,7) classifier; with both Euclidean and Manhattan distance metric. Thus, robustness of ZCR-STE-MFCC as MFUI feature set along with its great multiclass discrimination accuracy has been proven by our experiments.

An obvious direction for future research is expanding the number of noisy classes for multiclass classification. Noisy audio streams with higher SNR levels such as 5 dB and 10 dB from NOIZEUS database representing less noisy environment will be investigated for multiclass categorization and for testing robustness of proposed MFUI feature set. Alternative classifiers such as support vector machine (SVM) could also be used for performance evaluation.

References

[1] N. Nitanda, M. Haseyama, and H. Kitajima, "Accurate audio segment classification using feature extraction matrix," *Proc. ICASSP, 2005*.

[2] M. A. Sobreira- Seoane, A. R. Moleras and J. L. A. Castro, "Automatic classification of traffic noise", Proc. Acoustics'08, Paris, June 29 – July 4, 2008.

[3] George Tzanetakis, Perry Cook, "Musical Genre Classification of Audio Signals", *IEEE Transactions on Speech And Audio Processing, Vol.10, No. 5, July 2002.*

[4] B. Han and E. Hwang, "Environmental sound classification based on feature collaboration", *Proc. ICME, 2009.*

[5] Thiruvengatanadhan Ramalingam and P. Dhanalakshmi, "Speech/Music classification using wavelet based feature extraction techniques", *Journal of Computer Science 10(1): 34-44, 2014.*

[6] S.P.Mahajan, Jyotsana Sahu, M.S.Sutaone, V.K.Kokate, "Improving Performance of Multiclass Audio Classification using SVM", *CIIT International Journal of Data mining and Knowledge Engineering, Volume 2,No 5,pp.95-103,ISSN 0974-9683, May 2010.*

[7] Deepak Jhanwar, Kamlesh K. Sharma and S. G. Modani, "Classification of Environmental Background Noise Sources Using Hilbert-Huang Transform", *International Journal of Signal Processing Systems Vol. 1, No. 1 June 2013.*

[8] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung and H. H. Liu, "The Empirical Mode Decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, vol. 454, pp. 903-995, 1998.*

[9] J. J. Burred and A. Lerch, Hierarchical Automatic Audio Signal Classification, *Journal of Audio Engg. Soc., Vol. 52, pp. 724-739, July/August 2004.*

[10] Sunita Maithani and Richa Tyagi. Noise characterization and classification for background estimation, *IEEE-International Conference on Signal processing, Communications and Networking, pp. 208–213, Chennai, India, 2008.*

[11] Hric, M.; Chmulik, M.; Jarina, R.; "Comparision of Selected Classification Methods in Automatic Speaker dentification", *COMMUNICATIONS Scientific Letters of the University of Zilina, vol. 13, 2011.*

[12] X. Wu, Vipin K., The Top 10 Algorithms in Data Mining, Chapman & Hall/CRC, 2009.

[13] http://ecs.utdallas.edu/loizou/speech/noizeus/.