SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 19, No. 2, June 2022, 239-259

UDC: 534.87:517.984.4 **DOI: https://doi.org/10.2298/SJEE2202239K**

Hilbert Spectrum Based Features for Speech/Music Classification

Arvind Kumar[1], Sandeep Singh Solanki[1], Mahesh Chandra[2]

Abstract: Automatic Speech/Music classification uses different signal processing techniques to categorize multimedia content into different classes. The proposed work explores Hilbert Spectrum (HS) obtained from different AM-FM components of an audio signal, also called Intrinsic Mode Functions (IMFs) to classify an incoming audio signal into speech/music signal. The HS is a two-dimensional representation of instantaneous energies (IE) and instantaneous frequencies (IF) obtained using Hilbert Transform of the IMFs. This HS is further processed using Mel-filter bank and Discrete Cosine Transform (DCT) to generate novel IF and Instantaneous Amplitude (IA) based cepstral features. Validations of the results were done using three databases –Slaney Database, GTZAN and MUSAN database. To evaluate the general applicability of the proposed features, extensive experiments were conducted on different combination of audio files from S&S, GTZAN and MUSAN database and promising results are achieved. Finally, performance of the system is compared with performance of existing cepstral features and previous works in this domain.

Keywords: EMD, Hilbert Spectrum, Hilbert Huang Transform, Cepstral Features, Speech/Music Classification.

1 Introduction

In the last few years, with the rise of social networking and various digital platforms, there has been an exponential growth in the multimedia content. This has increased the need for robust multimedia processing technique to process the audio signal and extract various speech and music signal-based information. Speech/music classification is the fundamental task in multimedia processing where the incoming audio signal is categorized in relevant classes. The output of speech/music classification may be used for different speech processing task like speech recognition, speaker identification, emotion detection or different music processing task like genre detection, mood detection or note identification.

[1]Department of ECE, Birla Institute of Technology, Ranchi-835215, India; E-mails: arvind9835@gmail.com, sssolanki@bitmesra.com
[2]Department of ECE, Reva University, Bengaluru; E-mail: shrotriya69@rediffmail.com

Speech/music classification is usually the first block followed by further audio processing. Some of the applications of speech/music classification are [1, 2]:

– Discrimination of different audio scenes in broadcast news;
– Selection of radio broadcast channel as per listener preference;
– Automatic adjustment of hearing aids for speech and music environment;
– Preprocessing of audio files helps in minimization of computations for non-speech files for speech-based applications;
– Speech/music segmentation helps in content-based audio storage and
– Prior knowledge of speech and music segments will help in effective use of audio compression techniques.

Researchers have proposed various techniques over the year for designing an effective speech/music classifier. Most of these techniques are based on traditional spectral and temporal features like Zero Crossing Rate (ZCR), Spectral Roll off, MFCC, Chroma, Harmonic Ratio, Entropy, Energy, Glottal based excitation features and wavelet decomposition [1–8]. However, in this article features extracted from Hilbert Spectrum are proposed to design a speech/music classifier and its performance is tested for different sets of audio files. The paper is organized as follows: Section 1 covers the related works, motivation and methodology. Section 2 gives an overview of Emperical Mode Decomposition (EMD) and Hilbert Huang Transform. Section 3 briefs the existing cepstral features proposed across different domains. Section 4 introduces the proposed features. Study of feature importance and feature selection is covered in Section 5 and 6 respectively. Section 7 introduces the classifiers and Section 8 presents the experimental results and discussion. Section 9 concludes the work.

1.1 Related work

Earlier work for speech/music discrimination has exploited different temporal and spectral features [2–8]. Features likes zero crossing rate (ZCR), Mel Frequency Cepstral Coefficients (MFCC), spectral centroid, syllabic rate and excitation source of speech have been widely exploited for classification of speech and music giving a classification accuracy of 92–98%. However, the proposed work focuses on performance evaluation of different features formulated from IE and IF of Hilbert Spectrum. Both music and speech signal differ in the way they are produced. Music signal is composed of different notes produced while playing various instruments whereas speech signal consists of voiced and unvoiced sounds. Zhou [9] proposed speech/music discriminator using similarity based classifiers. Speech signal is band limited to 4 kHz containing most of the information in low frequencies whereas music signal is more spread in spectral domain ranging from (0–20) kHz [10]. These variations can be easily seen in the IMF extracted using EMD and is exploited for

speech/music discrimination. Early works in speech/music classification is proposed by Saunders [10]. Energy and zero crossing rate (ZCR) based statistical features were extracted from 2.4 seconds audio segment. Accuracy up to 98% was reported when probability measures on signal energy was used over skewness of ZC rate distribution [10]. Scheirer and Slaney [5] explored various temporal and spectral features for speech/music classification. An accuracy of more than 90% was reported for audio classification and 95% for audio segmentation. Pikrakis [4] proposed a three step approach for speech/music discrimination using hybrid of MFCC, energy and chroma based features and reported an accuracy of 96%. In another work, an accuracy of 96.94% discrimination rate was reported using hybrid of temporal and spectral descriptors and decision tree classifier [3]. Ruiz-Reyes N proposed a new technique for speech/music discrimination using fundamental frequency and obtained a classification rate of 97% [11]. A fast and robust speech, music discrimination approach was proposed in [7] using modified low energy ratio (MLER) and a Bayes MAP classifier reporting an accuracy of 91.4% and 90.2% for speech/music discrimination. In [2], hybrid of wavelet-based features, MFCC and ZCR were used for audio classification reporting an accuracy of 96.69%. In another work, data mining method using decision tree is proposed for speech/music discrimination reporting an accuracy of 97.9% [8]. Musical features like harmony and signal continuity were explored [6] for audio classification. Didiot E [1] proposed energy features for different wavelet families. In another work, 97% accuracy was reported for audio classification using Support Vector Machines (SVM) and Artificial Neural Network (ANN) using genetic algorithm [12]. EMD based features were explored in [13] achieving an accuracy of 90.03%. In [14], two new robust features based on Energy Variance of Filter Bank were proposed achieving a discrimination rate of 98.75%. Chroma based features like chroma high frequency and chroma difference were explored in [15] for audio classification reporting a classification accuracy of 97.1% and 93.0% for speech and music signal, respectively. Khonglah [16] proposed speech specific features for speech/music discrimination. This study exploited vocal tract information achieving an accuracy of 96.75%. Application of i-vector for speech and music classification is explored in [17] reporting an accuracy of (99.2−100)%.

Popular choice of classifiers among researchers over the year for speech/music discrimination are support vector machine (SVM) [2, 16−19] and Gaussian mixtures models (GMM) [15−17, 20]. Although, in some of the works, Artificial Neural Network (ANN) [2], k-Nearest Neighbours (k-NN) [5], Naïve Bayes [4], Decision tree [3] and Dynamic Time Warping (DTW) [9] based classifiers had also been explored. The proposed study evaluates the performance of the proposed features using two classifiers. Regarding datasets, both Scheirer and Slaney (S&S) and GTZAN databases had been extensively used by

researchers for evaluation [1, 5, 15, 16, 19, 21]. However, some of the work used their own created database [2, 7, 22].

1.2 Motivation

This work attempts to explore robust features from a non-linear and non-stationary data analysis technique by performing its Hilbert Spectral Analysis. EMD iteratively decomposes any complex signal into its Amplitude Modulated-Frequency Modulated (AM-FM) components, known as IMFs. IMFs have been explored in the past for Speech/Music Discrimination (SMD) task using statistical features [28]. Researchers have also found that IMFs contain different speech production information like formant tracking, glottal source information and vocal tract structure [24]. Earlier work evaluated different statistical and energy based features computed from IMFs of a signal for SMD task. The proposed work explores another representation of IMFs i.e. Hilbert Spectrum (HS) for extracting useful features efficient in speech/music classification. Hilbert Spectrum gives instantaneous frequencies which are functions of time resulting in instantaneous energies distribution over both time and frequency [29]. In the past, HS has been used for different speech processing task like emotion classification, formant extraction from speech signal, speaker verification and pitch estimation. Since, these are important speech characteristics, it can be assumed that HS based features can be used for SMD task too.

The main objective of the work is to explore the efficacy of HS for designing and testing SMD system for standard databases. Features are extracted from instantaneous frequencies and amplitudes of the IMFs obtained from HS. The proposed algorithms are tested on three different databases. The results obtained are compared with state-of-the-art techniques to evaluate the usefulness of HS representation for SMD task.

1.3 Methodology

This section describes the complete process of building a speech/music discrimination illustrated in Fig. 1. Audio sample from speech-music corpus is down sampled to 8 kHz. The processed files are decomposed into 10 IMFs using EMD. This is followed by extraction of different cepstral based features. The extracted features are then fed to different classifiers and their performances are tabulated.

2 EMD and Hilbert Huang Transform

EMD and Hilbert spectral analysis (HSA), together has proved to be a powerful technique for non-stationary and non-linear data analysis. EMD reduces complicated multicomponent data into IMFs which can be processed to generate both instantaneous frequencies and energies. Together EMD and HSA, utilizes the data in the most effective way in defining the longest period

component. Hilbert Transform optimally fits the local data in sine and cosine form, uniformly defining the frequency resolution of any point by finding the local derivative of the phase. This technique takes benefit of the combination of EMD and HSA in extracting low-frequency oscillation over wavelet analysis [33].

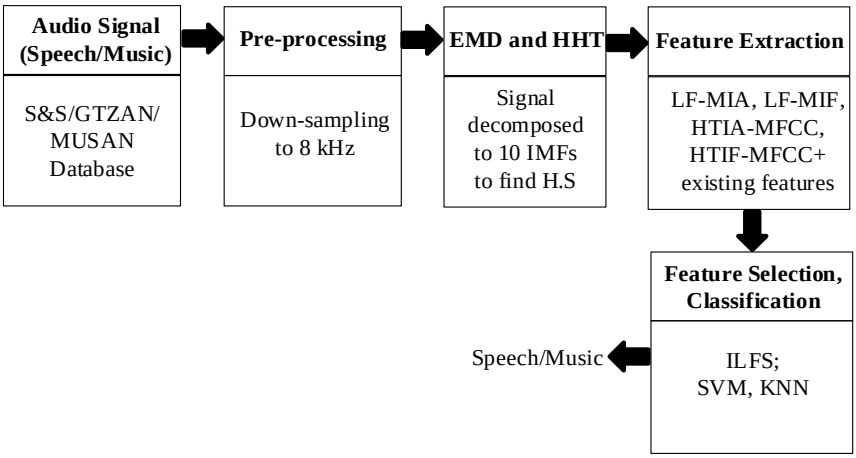


Fig. 1 – *Methodology.*

2.1 EMD

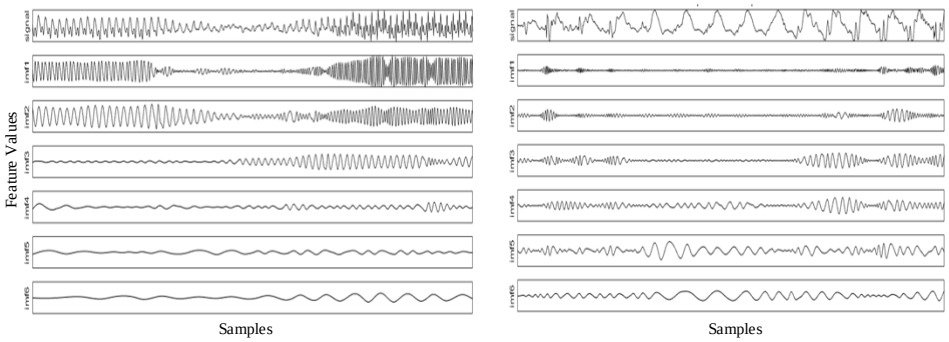
EMD is a data adaptive signal analysis technique which breaks a complex signal into its constituent AM-FM components called IMFs. Unlike wavelet transform, this technique decomposes a signal without any prior information about the nature of signal and hence is suitable for analysis of complex real world signals [33]. Both speech and music signal are time varying signals composed of multiple phones and notes played by different musical instruments respectively. EMD has found great advantage in capturing the dynamics of time varying signals. It adaptively decomposes a complex signal into various AM-FM components called Intrinsic Mode Function (IMF) using dyadic filter bank [33, 37, 38]. For an IMF, the number of local maxima and zero crossing are at most one and mean value of upper and lower envelope is equal to zero. Equation (1) represents the decomposed IMFs of an audio signal.

$$s(n) = r_k(n) + \sum_{i=1}^{k} c_i(n), \tag{1}$$

where $r_k(n)$ is residue and $c_i(n)$ is intrinsic mode function of i^{th} mode.

Fig. 2 illustrates first 6 IMFs generated from EMD for samples of both music and speech signal from S&S database. Difference in the nature of the IMFs in respect to energies and frequencies for both speech and music signal is clearly evident from the figures. This difference is further explored by extracting features

based on instantaneous frequencies and instantaneous energies of the IMFs generated by Hilbert Huang Transform (HHT).

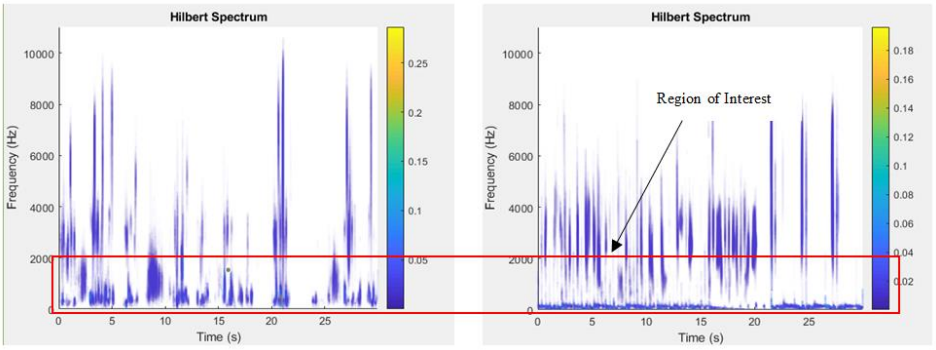


(a) (b)

Fig. 2 – *IMFs 1-6 generated from EMD for* (a) *Speech* (b) *Music.*

2.2 Hilbert Huang Transform (HHT) and Hilbert Transform (HT)

HHT is used to find the Hilbert Spectrum (HS) of a signal identified by several IMFs. HHT takes IMFs as an argument and performs Hilbert spectral analysis to generate different temporally and spectrally localized features. HHT is efficiently used to perform joint time-frequency analysis of non-stationary and non-linear data. Brief algorithm to find HHT can be seen at [33]. Hilbert Transform (HT) is another demodulation technique similar to Teager Kaiser Energy Operator (TKEO) and is also used to estimate both instantaneous amplitude and frequencies of a mono-component signal. HT suppresses the negative spectrum of the signal to zero and doubles the positive spectrum without changing the energy of the signal [29]. Fig. 3 illustrates Hilbert Spectrum of speech and music signal computed using HHT on MATLAB using *hht* function.



(a) (b)

Fig. 3 – *Hilbert Spectrum of* (a) *Speech and* (b) *Music signal obtained using HHT.*

The plot shows the variation of instantaneous frequencies and instantaneous energies for a raw sample of speech and music signal from S&S database. As evident from the Fig. 3, energy distribution for speech signal is dominant mostly in low frequency highlighted by dark regions of blue with region of silence between two spectra. Music signal on the other hand, because of its rhythmic nature has continuous energy distribution in low frequency region with no region of silence.

3 Existing Cepstral Features

Over the years, researchers have proposed different temporal and spectral domain features for building an efficient Speech/Music discrimination model. However, this work primary focuses on testing the performance of the system on different cepstral domain features proposed over the years along with new features presented in Section 4.

Table 1
Existing IMF based cepstral feature.

No.	FEATURE	DESCRIPTION
1	MFCC	MFCC are short-time features computed by taking the cosine transform of the logarithm of the power spectrum projected over mel-scale filter banks. This non-linearity is designed to replicate the response of human auditory system.
2	EMDCC [23]	Emperical Mode Decomposition Cepstral Coefficients (EMDCC) features are proposed by Tapkir et al. for replay spoof detection. In this work, authors replaced the mel-filter bank in MFCC with EMD. The dyadic nature of EMD was used to handle non-linear and non-stationary speech signals.
3	IMFCC [24]	Karan B et al. proposed IMF based Cepstral Coefficients (IMFCC) for Parkinson disease prediction. In this work too, the author replaced the Mel-filter bank in MFCC with EMD.
4	EMD-MFCC [25]	Alipoor G. et al. proposed robust speaker gender identification in noisy environment using EMD based cepstral features. In this work, the author proposed Complete Ensemble EMD (CEEMD) as a filter bank to decompose the speech signal.
5	SMFCC [39]	SMFCC features proposed by Li et al. reflect the distribution of energies in frequency domain more accurately. These features were proposed to remove the short comings of MFCC features in presence of signal trend in an audio signal.
6	ESA-IFCC [27]	Kamble M et al. proposed Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (ESA-IFCC) for spoof speech detection. This work also used TKEO to estimate the energy of the signal as product of amplitude square and frequency.
7	HT-IACC, HT-IFCC [40]	In one of the recent work, Kamble M et al. proposed two different features HT-IFCC (Hilbert Transform-IFCC) and HT-IACC (Hilbert Transform-Instantaneous Amplitude Cosine Coefficients) for replay detection. In this work, both HT and Teager Energy Operator (TEO) based decompositions were used to estimate IE and IF of the speech signal.

Cepstral features are based on human auditory system and have been exploited for different speech and music processing application. They are primarily derived from signal spectrum and carry vital information about the inherit nature of the source. **Table 1** tabulates different IMF based cepstral features proposed in literature.

4 Proposed Features for Speech/Music Classification

This section discusses novel features based on HS to design a speech/music classifier. Inspired by the characteristics of Hilbert spectrum for speech and music signal in low frequency, Section 4.1 and 4.2 introduces two novel features. Further, Section 4.3 introduces Hilbert Transform based Mel scaled features cascading EMD, HHT and Mel-filter bank and promising results are reported.

4.1 Low Frequency Mean Instantaneous Frequency/Amplitude (LF-MIF/LF-MIA)

As evident from Fig. 3, Hilbert Spectrum of speech and music signal differs mostly in low frequency region. To separate IMFs of interest, a primarily study is conducted to visualize IMFs containing low frequency signal information using FFT for both speech and music signal. After studying the spectral distribution of IMFs, it was found that IMFs 5-10 reflects low frequency signal content in the range of (0-1000) Hz. However, for music signal, IMFs 4-10 shows similar characteristics. Nevertheless, to generate feature vectors of uniform dimension for both speech and music samples, IMFs 5-10 is processed for both classes. Hence, this work explores only these 6 IMFs to derive feature to classify an audio signal into speech/music sample. Additionally, unlike music samples, Hilbert Spectrum of speech samples have region of silence between spectra because of unvoiced regions. This may lead to difference in normalized mean instantaneous amplitude (MIA) and can be utilized as discriminatory evidence between speech and music samples.

Fig. 4 illustrates the process of extracting LF-MIF/LF-MIA features. Down-sampled audio samples from standard databases are decomposed into 10 IMFs. To extract low frequency information, out of these 10 IMFs, IMFs 5-10 are further processed through Hilbert Transform to find Instantaneous Frequencies (MIF)/ Instantaneous Amplitudes (MIA) of these IMFs. Further, each of these 6 IMFs are framed using a framing window of 500 milliseconds and their mean are evaluated. Fig. 5 illustrates the variation of MIF for IMFs 5-10. Discriminatory evidence between speech and music signal is clearly visible in all the IMFs. Vector of means of each frame is further passed through Log and DCT block to extract relevant information. First ten coefficients from output of DCT are stored for each IMF to form the feature vector.

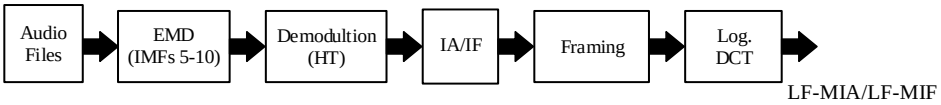


Fig. 4 – *Block diagram to find LF-MIF.*

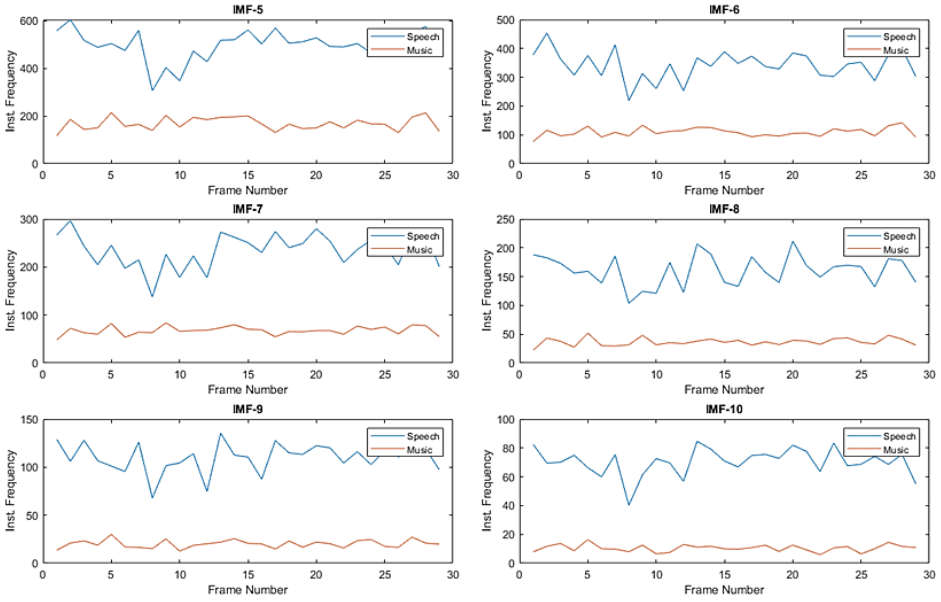


Fig. 5 – *Variation of MIF for IMFs 5-10.*

4.2 HTIA-MFCC and HTIF-MFCC

In this section we propose two new features based on Hilbert spectral analysis of IMFs i.e. Hilbert Huang Transform-Instantaneous Amplitude Mel-Frequency Cepstral Coefficients (HTIA-MFCC) and Hilbert Huang Transform-Instantaneous Frequency Mel-Frequency Cepstral Coefficients (HTIF-MFCC) to capture advantage of both EMD based HT and Mel filter bank. Music signals are more rhythmic than speech signals with smoother variation in both pitch and energy. The proposed features attempt to captures this variation. Fig.6 illustrates the process of finding the proposed coefficients.

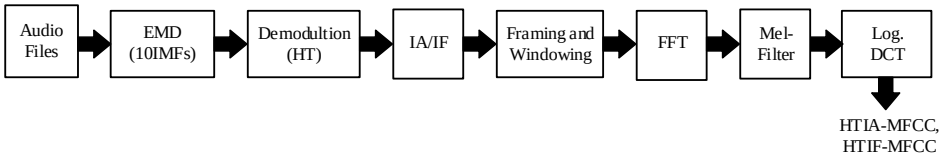


Fig. 6 – *Block diagram of HTIA-MFCC/HTIF-MFCC feature.*

Both IE and IF of decomposed IMFs are estimated through HT. IE are further processed to evaluate Instantaneous Amplitude (IA) based features. Unlike different cepstral feature (EMDCC and EMD-MFCC) discussed in Section 4, which directly process the IMFs, HTIA-MFCC/HTIF-MFCC process instantaneous amplitudes and frequencies of IMFs. This work proposes to map the feature vectors to Mel-filter bank to improve the resolution of the signal in low-frequency. Moreover, mapping the IA/IF values to triangular filter bank improves the harmonic structure. We expect to collect different information using this algorithm effective for SMD task as compared to previously reported SMD based features directly extracted from IMFs. Both IA/IF are framed and windowed using an overlapping window of 30 ms with 50% overlap. Fast Fourier Transform (FFT) of these framed samples are mapped to Mel-filter bank and passed through log and DCT block for feature reduction. 39 coefficients from each of the 10 IMFs are appended to form the feature vector. Feature importance is studied using t-SNE and ROC plot in Section 5. Performances of the proposed features are compared with various existing cepstral based feature introduced in Section 4.

5 Study of Feature Importance

Before feeding the features into classifiers and evaluating their classification accuracies, a study of feature importance is carried in this section.

5.1 Feature Visualization using t-SNE

t-distributed Stochastic Neighbour Embedding (t-SNE) is an algorithm for visualizing high dimensional data using dimensional reduction. t-SNE helps in visualization of clusters in original high dimensional data using low dimensional points. The algorithm maps higher dimensional points to lower dimensional points in respect with similarities between points and works in five main steps to embed data in lower dimension [42]. Fig. 7 illustrates the scatter plot of lower dimension features obtained using t-SNE for MUSAN database. This visualization helps in understanding the relevance of features. Features with less interclass overlap are assumed to perform better classification task. Scatter plot of the proposed features HTIA-MFCC, HTIF-MFCC along with MFCC, EMDCC, EMD-MFCC, HT-IACC and SMFCC shows clear distinction between two different classes and hence are assumed to give good classification accuracy. Scatter plot of ESA-IFCC, HT-IFCC and IMFCC shows an overlap amongst few data. The relevance of feature is further tested using ROC analysis.

5.3 Feature importance using ROC analysis

This section analyses the feature importance using Receiver operating characteristic (ROC) curve. The parameter to judge the efficiency of a classifier is AUC (Area under Curve). A perfect classifier has an AUC of 1. Larger AUC indicates better performance of the classifier. **Table 2** tabulates the AUC of ROC

for different features for all three different databases along with feature dimension. The proposed features HTIA-MFCC shows comparatively improved performance for all the three databases at a cost of larger feature dimension. Performance of baseline features (MFCC, EMD-MFCC, and SMFCC) is also found to be competitive.

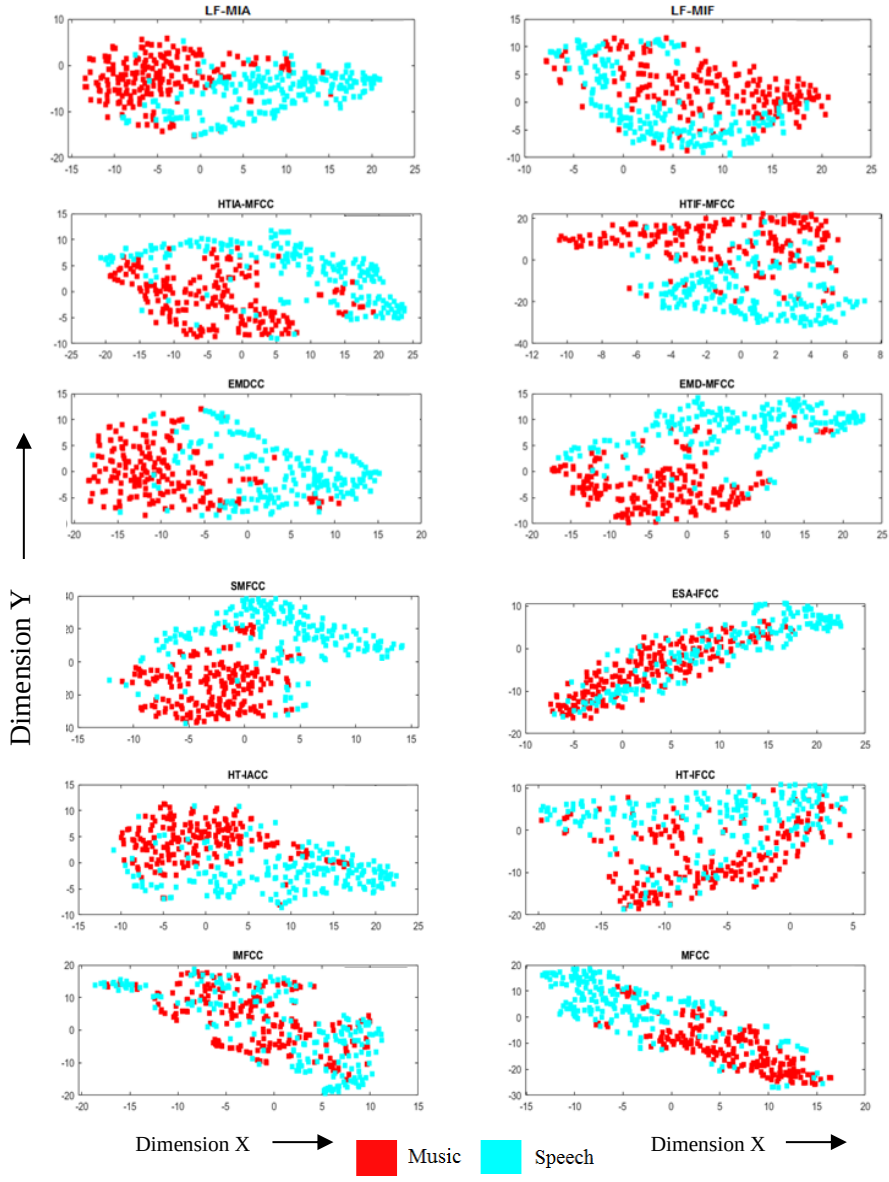


Fig. 7 – *Scatter Plot of lower dimensional features obtained using t-SNE.*

Table 2
Analysis of ROC Curve.

Feature	Area Under Curve (ROC)		
	S&S	GTZAN	MUSAN
MFCC(Baseline)	1.0000	1.0000	0.9995
EMDCC	0.9981	0.9990	0.9991
IMFCC	0.9953	0.9949	0.9884
EMD-MFCC(Baseline)	1.0000	1.0000	0.9995
SMFCC(Baseline)	1.0000	1.0000	0.9995
ESA-IFCC	0.9997	1.0000	0.9926
HT-IACC	1.0000	0.9988	0.9984
HT-IFCC	0.9975	0.9990	0.9833
LF-MIA	0.9986	0.9988	0.9984
LF-MIF	0.9997	0.9988	0.9869
HTIA-MFCC	1.0000	1.0000	0.9998
HTIF-MFCC	1.0000	1.0000	0.9990

6 Feature Selection using Infinite Latent Feature Selection (ILFS)

Role of feature selection has increased in machine learning application to improve the performance of the model. The aim is to select relevant feature vectors amongst the total feature set removing the redundant features. Feature Selection not only helps in reducing the dimension of the training matrix decreasing the computational complexity but also helps in improving classification performance. In this paper, a Probabilistic Latent Graph-Based Ranking Approach is employed for feature selection. This algorithm ranks the features as per relevance by observing all subsets of the features and studying the convergence properties of power series of matrices [44]. ILFS technique works by assigning score of importance to each feature considering all features as nodes on the graph modelling a pairwise relationship between all possible combinations of features by weighing the edges joining them. ILFS method is applied to the input features which ranks the feature as per relevance. The top ranked features are then fed into the classifier and the performance is tabulated.

7 Classifiers

This section briefly describes the classifiers used in this work. Speech/Music discrimination is a task of binary classification where the incoming audio signals will be either classified as speech or music. Conventional classifiers like SVM, GMM, KNN and ANN have been widely explored in this domain with promising results [20, 2, 16, 5]. Hence, the proposed features were evaluated for its efficiency on SVM and k-NN classifiers for a comparative analysis.

7.1 Support Vector Machines (SVM)

SVM uses a hyperplane to optimally separate two classes. The algorithm tries to find a hyperplane which maximizes the margin of separation. In 2D, this hyperplane is a line which divides the plane in two different parts where each class lies on either side. For more complex data which are not linearly separable in two dimensions, SVM uses kernels to map the data in higher dimension. A function φ is used to map data points to a feature space. This function can be linear, polynomial or Gaussian. For a set of data with training vectors x_j and their categories y_j in some dimension d where $x \in R^d$ and $y_j = \pm 1$, the equation of hyperplane is

$$f(x) = x'w + b = 0, \tag{2}$$

where, *w* is the weight vector and *b* is a bias. More information can be found in [41]. All the simulations were done with *fitcsvm* function in MATLAB. Using Sequential Minimal Optimization (SMO), *fitcsvm* optimally finds the value of BoxConstraint and KernelScale.

7.2 k-Nearest Neighbours (k-NN)

k-Nearest Neighbours (k-NN) classifier uses data spread in multidimensional feature space each having class labels for training. 'k' in k-NN is a user defined constant and indicates the number of neighbors voting is made from. A test sample is assigned a class by assigning the label of the training samples which occurs maximum number of times among the training samples closest to the test point. Use of k-NN has been explored in [5]. This algorithm finds the distance of the test samples with k-nearest neighbors. Euclidean distance is a commonly used distance metric and is used in this work too. All the simulations were done with *fitcknn* function in MATLAB with $\mathrm{k} = 2$.

8 Simulation Results and Discussion

This section evaluates the classification accuracy of proposed feature using three different databases. Initially, a brief introduction of the database is laid followed by the performance of speech/music discrimination system using SVM and k-NN classifier. Further, feature selection technique is applied, and the performance accuracy is observed for variety of data. In the end, performance of the proposed feature is compared with existing state of art techniques.

8.1 Experimental database

In this work, three different databases are used to evaluate the reliability of the proposed feature set for classification of speech and music signal. Scheirer and Slaney (S&S) Database is a music-speech corpus and consists of 240, 15-seconds audio recordings of radio broadcast in MSWAVE format recorded at a

sampling frequency of 22.05 kHz. It was collected by Eric Scheirer during his internship at Interval Research Corporation in the summer of 1996 under the supervision of Malcolm Slaney [30]. GTZAN is a music-speech corpus consisting of 120, 30-seconds 16 bit audio files in WAV format with each class having 64 wav files. It was collected by George Tzanetakis and is publicly available for evaluating speech/music discrimination algorithms. Music corpus consists of wide range of music forms like classical, jazz, guitar etc. while speech corpus has voice sample from variety of male and female speakers [31]. MUSAN is music, speech and noise corpus containing in total 660 music samples spread across several genre, 426 speech samples spread across 12 languages and close to 900 noise samples suitable for training models for SMD task [32]. For fair comparison, length of audio samples from each database is controlled to 15 seconds with sampling rate of 8 kHz. **Table 3** tabulates the description of training and testing data.

Table 3
Description of database

Database	TRAINING		TESTING	
	Speech	Music	Speech	Music
S&S	60	60	20	20
GTZAN	48	48	16	16
MUSAN	150	150	50	50

8.2 Performance accuracy for proposed feature

Each of the features was analysed with two different classifiers on three standard database and the results were tabulated in **Table 4** and **5**. To evaluate the algorithms, whole data set is divided into training and testing data. To avoid biasing of results, performance of the model is iterated ten times for different combinations for training and testing data and the average percentage accuracy is reported along with standard deviation.

Table 4 tabulates the performance accuracy of proposed raw features HTIA-MFCC and HTIF-MFCC using SVM classifier on three different databases. Best performance accuracy of 97.66%, 93.12% and 94.80% is observed for S&S, GTZAN and MUSAN database respectively for the proposed feature. An improvement of 5.01%, 10.42% and 3.83% is seen for best performing proposed features over MFCC features. A slight improvement over EMD-MFCC and SMFCC features is also observed across all the three datasets.

Table 5 tabulates the performance accuracy of proposed features HTIA-MFCC and HTIF-MFCC using k-NN classifier on three different databases. Best performance accuracy of 95.33%, 87.80% and 89.00% is observed for S&S,

GTZAN and MUSAN database respectively for the proposed feature. An improvement of 10.84%, 20.28% and 13.37% is seen for best performing proposed features over MFCC features.

Table 4

Performance accuracy (%) *of proposed feature using SVM (RBF) classifier.*

Feature	S&S			GTZAN			MUSAN		
	Speech	Music	ACC(σ)	Speech	Music	ACC(σ)	Speech	Music	ACC(σ)
MFCC (Baseline)	92.00	94.00	93.00(3.66)	84.66	84.00	84.33(5.45)	91.80	90.80	91.30(1.94)
EMDCC	85.33	73.33	79.33(6.04)	78.00	75.33	76.66(8.01)	89.00	82.20	85.60(3.33)
IMFCC	93.33	94.66	94.00(3.78)	68.12	83.75	75.93(6.59)	84.60	73.00	78.80(2.65)
EMD-MFCC (Baseline)	98.00	95.33	96.66(2.72)	96.66	85.33	91.00(3.53)	88.60	89.60	89.10(2.02)
SMFCC (Baseline)	95.33	92.00	93.66(3.31)	91.87	90.00	90.93(1.50)	93.20	94.00	93.60(3.99)
ESA-IFCC	90.00	98.00	94.00(4.09)	69.37	84.37	76.87(5.92)	83.00	82.00	82.50(3.17)
HT-IACC	85.33	67.33	76.33(6.37)	68.12	90.00	79.06(5.32)	85.40	87.20	86.30(3.80)
HT-IFCC	92.00	84.66	88.33(5.71)	90.00	92.50	91.25(2.87)	59.00	86.60	72.80(4.69)
LF-MIF	97.33	88.66	93.00(3.31)	84.37	83.12	83.75(7.90)	92.00	80.66	87.00(4.28)
LF-MIA	94.66	89.33	92.00(5.92)	86.87	88.12	87.50(5.51)	88.00	81.33	84.66(5.01)
HTIA-MFCC	97.33	96.00	96.66(2.72)	90.62	95.62	**93.12(3.84)**	90.80	91.00	90.90(3.63)
HTIF-MFCC	96.66	98.66	**97.66(3.44)**	92.62	93.25	92.93(3.10)	94.60	95.00	**94.80(1.81)**

Table 5

Performance accuracy (%) *of proposed feature using k-NN* (*k* = 2) *classifier.*

Feature	S&S			GTZAN			MUSAN		
	Speech	Music	ACC(σ)	Speech	Music	ACC(σ)	Speech	Music	ACC(σ)
MFCC (Baseline)	84.66	87.33	86.00(5.62)	78.00	68.00	73.00(5.97)	93.20	63.80	78.50(4.35)
EMDCC	88.00	79.33	83.66(5.07)	95.33	46.00	70.66(5.62)	87.00	78.60	82.80(4.36)
IMFCC	96.66	72.66	84.66(6.12)	86.87	40.00	63.43(6.91)	94.60	35.20	64.90(3.03)
EMD-MFCC (Baseline)	98.00	90.00	94.00(2.62)	94.66	84.66	89.66(4.56)	94.20	84.60	89.40(2.50)
SMFCC (Baseline)	86.00	87.33	86.66(6.28)	93.75	58.12	75.93(8.46)	97.80	71.80	84.80(2.78)
ESA-IFCC	63.33	87.33	75.33(8.19)	71.87	78.75	75.31(7.99)	57.60	86.80	72.20(4.34)
HT-IACC	92.66	44.00	68.33(6.13)	91.25	12.50	51.87(4.93)	93.80	54.40	74.10(3.54)
HT-IFCC	100.00	0.00	50.00(0.00)	93.75	0.00	46.87(0.00)	85.40	36.80	61.10(3.34)
LF-MIF	82.00	94.00	88.00(2.72)	76.87	85.62	81.24(5.75)	70.00	82.66	76.33(5.48)
LF-MIA	88.66	97.33	92.00(4.91)	74.37	81.25	77.81(6.59)	69.33	89.33	79.33(9.53)
HTIA-MFCC	98.66	92.00	**95.33(2.81)**	89.37	85.62	87.50(3.89)	93.80	78.20	86.00(2.53)
HTIF-MFCC	90.66	94.00	92.33(3.86)	88.75	86.87	**87.81(4.28)**	95.20	82.80	**89.00(2.78)**

On comparing LF-MIF/LF-MIA with baseline features, it is observed that although the performance of LF-MIA/LF-MIF in **Table 4** and **5** is satisfactory, these features couldn't outperform the baseline features. HTIA-MFCC and HTIF-MFCC had better classification accuracies against the baseline features for most of the scenarios. However, for MUSAN dataset, SMFCC performed better than

raw HTIA-MFCC for SVM classifier with an accuracy of 93.60%. For k-NN classifier, EMD-MFCC outperformed raw HTIA-MFCC/ HTIF-MFCC for both GTZAN and MUSAN dataset. To further improve the discriminatory capabilities of these features, feature selection techniques were employed on raw HTIA-MFCC and HTIF-MFCC features only with SVM classifier.

8.3 Effect of feature selection

In this paper, ILFS feature selection is employed for selecting best performing features. Raw feature matrix is processed and ranked according to their significance. **Table 6** tabulates the performance accuracy of best performing feature for all the three databases using SVM classifier with RBF kernel.

Table 6
Performance accuracy (%) *of proposed feature using feature selection.*

Feature	S&S			GTZAN			MUSAN		
	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall
HTIA-MFCC	100.00	100.00	**100.00**	100.00	100.00	**100.00**	95.00	100	**97.50**
HTIF-MFCC	100.00	100.00	**100.00**	100.00	100.00	**100.00**	95.00	100	**97.50**

For S&S and GTZAN database, best efficiency of 100% is achieved for both HTIA-MFCC and HTIF-MFCC feature. For MUSAN database, best efficiency of 97.50% is achieved for both HTIA-MFCC and HTIF-MFCC feature. **Table 7** tabulates the number of top ranked feature used to achieve best performances for the proposed feature. Although for both S&S and GTZAN database, best performance accuracy is observed for first few features, comparatively more number of feature vectors was required to achieve best performance for MUSAN database.

Table 7
Number of features used to achieve best performance.

Feature	S&S	GTZAN	MUSAN
	Number of features used	Number of features used	Number of features used
HTIA-MFCC	6	14	14
HTIF-MFCC	11	14	109

8.4 Combined database result

This section evaluates the performance accuracy for the proposed features HTIA-MFCC and HTIF-MFCC when trained and tested with audio samples from all the three-database using SVM classifier with rbf kernel. **Table 8** depicts the performance accuracy of the system for combined database with and without feature selection technique. Best efficiency of 100% and 96.90% is observed for HTIA-MFCC and HTIF-MFCC with feature selection technique whereas best efficiency of 94.13% and 92.16% is observed for the proposed features without

feature selection technique. This study also analyses the performance of the system for mismatched data using cross data training and testing. **Table 9** shows the classification accuracy of the model when it is trained with one database and tested with another database. An overall accuracy of 86.67% is observed when model is trained on GTZAN database and tested with S&S database whereas an overall accuracy of 87.50% is observed when vice-versa.

Table 8

Performance accuracy (%) *for complete database.*

Feature	S&S+GTZAN+MUSAN					
	With Feature Selection			Without Feature Selection		
	Speech	Music	Overall	Speech	Music	Overall
HTIA-MFCC	100.00	100.00	**100.00**	93.08	95.18	**94.13**
HTIF-MFCC	96.88	96.92	**96.90**	92.59	91.72	**92.16**

Table 9

Performance accuracy (%) *for cross data training and testing.*

	Models trained on					
	GTZAN Database			S&S Database		
Models tested with	Speech	Music	Overall	Speech	Music	Overall
GTZAN Database				93.75	81.25	87.50
S&S Database	80.00	93.33	86.67			

8.5 Comparison with previous work

Table 10

Comparison of different speech/music discrimination algorithms.

Algorithm	Feature Used	Raw Feature Dimension	Database	Accuracy (%)
[22]	ZCR, Energy and Periodicity	9	S&S	95.5
[19]	ZCR, Spectral roll-off, RMS energy, MFCC, Spectral flatness	72	GTZAN	95.9
[16]	Excitation source, vocal tract system and syllabic rate of speech	3	S&S	88.09
[15]	Chroma vector based features	10	GTZAN	93.5
[21]	Minimum Energy Density (MED) features	1	GTZAN	95.8
[1]	Wavelet energy features	28	GTZAN	96.6
Proposed	HTIF-MFCC	390	S&S, GTZAN, MUSAN	S&S-100 GTZAN-100 MUSAN-97.50
Proposed	HTIA-MFCC	390	S&S, GTZAN, MUSAN	S&S-100 GTZAN-100 MUSAN-97.50

This section compares the performance of the proposed feature with existing state-of-the-art technique. **Table 10** tabulates the comparison chart for different speech/music discrimination algorithms proposed in literature [43].

9 Conclusion

Features derived from Hilbert Spectrum are proposed in this work for classification of speech and music files. Each of these audio files is decomposed into 10 IMFs using EMD. These IMFs are further processed for feature extraction. Hilbert Transform is used to evaluate IA and IF of these IMFs. Inspired from the Hilbert Spectrum, initially low frequency based features are proposed by using the IMFs 5-10. Further, we also propose HTIA-MFCC/HTIF-MFCC feature vectors by processing the IA/IFs from IMFs 1-10 through Mel-filter bank. Initially, feature importance was studied using ROC analysis and t-SNE plot. Both existing and proposed features are evaluated on three standard datasets i.e. S&S, GTZAN and MUSAN dataset. Performance of the proposed feature is evaluated on SVM and k-NN classifier. Infinite Latent Feature Selection is employed to improve performance accuracy and reduce feature dimension. Variation of performance accuracy for different number of feature vectors is also studied. To evaluate the general applicability of the proposed features, extensive experiments were conducted on different combination of audio files from S&S, GTZAN and MUSAN database. Classification accuracy of 100.00%, 100.00% and 97.50% is observed for S&S, GTZAN and MUSAN database respectively for both HTIA-MFCC and HTIF-MFCC feature. Further, cross data training and testing is also studied for S&S and GTZAN dataset. The highest efficiency of 87.50% is observed for cross data training and testing. This study highlights the significance of Hilbert Spectrum representation of IMFs for designing an effective speech/music classifier.

10 Acknowledgments

We would like to thank HOD, management and supporting staffs of Dept. of ECE, Birla Institute of Technology for providing us funding and facilities for conducting our research work. We would also like to thank different authorities for giving us permission to work on standard speech-music database.

8 References

[1] E. Didiot, I. Illina, D. Fohr, O. Mella: A Wavelet-Based Parameterization for Speech/Music Discrimination, Computer Speech & Language, Vol. 24, No. 2, April 2010, pp. 341 – 357.

[2] M. K. S. Khan, W. G. Al-Khatib: Machine-Learning Based Classification of Speech and Music, Multimedia Systems, Vol. 12, No. 1, August 2006, pp. 55 – 67.

[3] Y. Lavner, D. Ruinskiy: A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation, EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2009, June 2009, pp. 239892.

[4] A. Pikrakis, T. Giannakopoulos, S. Theodoridis: A Speech/Music Discriminator of Radio Recordings based on Dynamic Programming and Bayesian Networks, IEEE Transactions on Multimedia, Vol. 10, No. 5, August 2008, pp. 846 – 857.

[5] E. Scheirer, M. Slaney: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, April 1997, pp. 1331 – 1334.

[6] J. Shirazi, S. Ghaemmaghami: Improvement to Speech-Music Discrimination Using Sinusoidal Model Based Features, Multimedia Tools and Applications, Vol. 50, No. 2, November 2010, pp. 415 – 435.

[7] W. Q. Wang, W. Gao, D. W. Ying: A Fast and Robust Speech/Music Discrimination Approach, Proceedings of the 4th International Conference on Information, Communications and Signal Processing, Singapore, Singapore, December 2003, pp. 1325 – 1329.

[8] Q. Wu, Q. Yan, H. Deng, J. Wang: A Combination of Data Mining Method with Decision Trees Building for Speech/Music Discrimination, Computer Speech & Language, Vol. 24, No. 2, April 2010, pp. 257 – 272.

[9] H. Zhou, A. Sadka, R. M. Jiang: Feature Extraction for Speech and Music Discrimination, Proceedings of the International Workshop on Content-Based Multimedia Indexing, London, UK, June 2008, pp. 170 – 173.

[10] J. Saunders: Real-Time Discrimination of Broadcast Speech/Music, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, USA, May 1996, pp. 993 – 996.

[11] N. Ruiz-Reyes, P. Vera-Candeas, J. E. Muñoz, S. García-Galán, F. J. Cañadas: New Speech/Music Discrimination Approach based on Fundamental Frequency Estimation, Multimedia Tools and Applications, Vol. 41, No. 2, January 2009, pp. 253 – 286.

[12] A. Pikrakis, T. Giannakopoulos, S. Theodoridis: A Speech/Music Discriminator of Radio Recordings based on Dynamic Programming and Bayesian Networks, IEEE Transactions on Multimedia, Vol. 10, No. 5, August 2008, pp. 846 – 857.

[13] A. Ghosal, B. C. Dhara, S. K. Saha: Speech/Music Classification Using Empirical Mode Decomposition, Proceedings of the 2nd International Conference on Emerging Applications of Information Technology, Kolkata, India, February 2011, pp. 49 – 52.

[14] M. Kos, Z. Kačič, D. Vlaj: Acoustic Classification and Segmentation Using Modified Spectral Roll-Off and Variance-Based Features, Digital Signal Processing, Vol. 23, No. 2, March 2013, pp. 659 – 674.

[15] G. Sell, P. Clark: Music Tonality Features for Speech/Music Discrimination, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014, pp. 2489 – 2493.

[16] B. K. Khonglah, S. R. Mahadeva Prasanna: Speech/Music Classification Using Speech-Specific Features, Digital Signal Processing, Vol. 48, January 2016, pp. 71 – 83.

[17] H. Zhang, X.- K. Yang, W.- Q. Zhang, W.- L. Zhang, J. Liu: Application of I-Vector in Speech and Music Classification, Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, December 2016, pp. 1 – 5.

[18] C. Lim, J. H. Chang: Efficient Implementation Techniques of an SVM-Based Speech/Music Classifier in SMV, Multimedia Tools and Applications, Vol. 74, No. 15, August 2015, pp. 5375 – 5400.

[19] N. Tsipas, L. Vrysis, C. Dimoulas, G. Papanikolaou: Efficient Audio-Driven Multimedia Indexing Through Similarity-Based Speech/Music Discrimination, Multimedia Tools and Applications, Vol. 76, No. 24, December 2017, pp. 25603 – 25621.

[20] G. Fuchs: A Robust Speech/Music Discriminator for Switched Audio Coding, Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France, August 2015, pp. 569 – 573.

[21] S. Kacprzak, M. Ziółko: Speech/Music Discrimination via Energy Density Analysis, Proceedings of the 1st International Conference on Statistical Language and Speech Processing, Tarragona, Spain, July 2013, pp. 135 – 142.

[22] A. Ghosal, S. Dutta: Speech/Music Discrimination Using Perceptual Feature, Proceedings of the International Conference on Computational Science and Engineering, Beliaghata, India, October 2016, pp. 71 – 76.

[23] P. Tapkir, H. A. Patil: Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection, Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, September 2018, pp. 721 – 725.

[24] B. Karan, S. S. Sahu, K. Mahto: Parkinson Disease Prediction Using Intrinsic Mode Function Based Features from Speech Signal, Biocybernetics and Biomedical Engineering, Vol. 40, No. 1, January 2020, pp. 249 – 264.

[25] G. Alipoor, E. Samadi: Robust Speaker Gender Identification Using EMD-Based Cepstral Features, Asia-Pacific Journal of Information Technology and Multimedia, Vol. 7, No. 1, June 2018, pp. 71 – 81.

[26] E. Samadi, G. Alipoor: Efficient Band Selection for Improving the Robustness of the EMD-Based Cepstral Features, Sādhanā, Vol. 44, No. 3, March 2019, p. 54.

[27] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. Ali Mahjoub, C. Cleder: Automatic Speech Emotion Recognition Using an Optimal Combination of Features based on EMD-TKEO, Speech Communication, Vol. 114, November 2019, pp. 22 – 35.

[28] B. K. Khonglah, R. Sharma, S. R. Mahadeva Prasanna: Speech vs Music Discrimination Using Empirical Mode Decomposition, Proceedings of the 21st National Conference on Communications (NCC), Mumbai, India, February 2015, pp. 1 – 6.

[29] R. Sharma, R. K. Bhukya, S. R. Mahadeva Prasanna: Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification, Speech Communication, Vol. 96, February 2018, pp. 207 – 224.

[30] Dan Ellis: The Music-Speech Corpus, Available at:

[31] https://labrosa.ee.columbia.edu/sounds/musp/scheislan.html

[32] MARSYAS, Available at: http://marsyas.info/downloads/datasets.html

[33] D. Snyder, G. Chen, D. Povey: MUSAN: A Music, Speech, and Noise Corpus, arXiv:1510.08484 [cs.SD], October 2015, pp. 1 – 4.

[34] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.- C. Yen, C. C. Tung, H. H. Liu: The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis, Proceedings of the Royal Society of London A, Vol. 454, March 1998, pp. 903 – 995.

[35] P. Cosi: Evidence Against Frame-Based Analysis Techniques, Proceedings of NATO Advance Institute on Computational Hearing, Il Ciocco, July 1998, pp. 163 – 168.

[36] C.- S. Jung, K. J. Han, H. Seo, S. S. Narayanan, H.- G. Kang: A Variable Frame Length and Rate Algorithm based on the Spectral Kurtosis Measure for Speaker Verification, Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Chiba, Japan, September 2010, pp. 2754 – 2757.

[37] M. S. Deshpande, R. S. Holambe: Speaker Identification based on Robust AM-FM Features, Proceedings of the 2nd International Conference on Emerging Trends in Engineering & Technology, Nagpur, India, December 2009, pp. 880 – 884.

[38] R. Sharma, S. R. Mahadeva Prasanna, R. K. Bhukya, R. Kumar Das: Analysis of the Intrinsic Mode Functions for Speaker Information, Speech Communication, Vol. 91, July 2017, pp. 1 – 16.

[39] P. Flandrin, G. Rilling, P. Goncalves: Empirical Mode Decomposition as a Filter Bank, IEEE Signal Processing Letters, Vol. 11, No. 2, February 2004, pp. 112 – 114.

[40] X. Li, X. Li: Speech Emotion Recognition Using Novel HHT-TEO Based Features, Journal of Computers, Vol. 6, No. 5, May 2011, pp. 989 – 998.

[41] M. R. Kamble, H. Tak, H. A. Patil: Effectiveness of Speech Demodulation-Based Features for Replay Detection, Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, September 2018, pp. 641 – 645.

[42] C. Cortes, V. Vapnik: Soft Margin Classifier, U.S. Patent, Patent No. 5,640,492, June 1997.

[43] L. van der Maaten, G. Hinton: Visualizing Data Using t-SNE, Journal of Machine Learning Research, Vol. 9, No. 86, November 2008, pp. 2579 – 2605.

[44] G. K. Birajdar, M. D. Patil: Speech and Music Classification Using Spectrogram Based Statistical Descriptors and Extreme Learning Machine, Multimedia Tools and Applications, Vol. 78, No. 11, June 2019, pp. 15141 – 15168.

[45] G. Roffo, S. Melzi, U. Castellani, A. Vinciarelli: Infinite Latent Feature Selection: A Probabilistic Latent Graph-Based Ranking Approach, Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 2017, pp. 1407 – 1415.