Digital Signal Processing 48 (2016) 71–83

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Speech / music classification using speech-specific features

Banriskhem K. Khonglah *, S.R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Available online 21 September 2015

Keywords:
Speech / music classification
Excitation source
Vocal tract system
Zero frequency filtered signal
Hilbert envelope

ABSTRACT

This paper proposes the use of speech-specific features for speech / music classification. Features representing the excitation source, vocal tract system and syllabic rate of speech are explored. The normalized autocorrelation peak strength of zero frequency filtered signal, and peak-to-sidelobe ratio of the Hilbert envelope of linear prediction residual are the two source features. The log mel energy feature represents the vocal tract information. The modulation spectrum represents the slowly-varying temporal envelope corresponding to the speech syllabic rate. The novelty of the present work is in analyzing the behavior of these features for the discrimination of speech and music regions. These features are non-linearly mapped and combined to perform the classification task using a threshold based approach. Further, the performance of speech-specific features is evaluated using classifiers such as Gaussian mixture models, and support vector machines. It is observed that the performance of the speech-specific features is better compared to existing features. Additional improvement for speech / music classification is achieved when speech-specific features are combined with the existing ones, indicating different aspects of information exploited by the former.

1. Introduction

Audio data obtained from the broadcast news channels generally consists of complex scenarios. Some of these include speech recorded in studio which is of good quality, speech recorded in the field which is mostly in outdoor environments and may contain background noise, speech with background music, vocal and non-vocal music. Hence processing audio data for different multimedia applications is a challenging task. Among different issues, the fundamental one to pursue is speech / music classification, needed for separation of speech and music regions for further processing. The current work explores the task of speech versus music classification.

The speech / music classification task has been explored in several ways in literature using different features and classifiers [1–8]. This work proposes to explore the speech-specific features for speech / music classification motivated from the use of music-specific features explored in [9]. There are several reasons for looking at this task from the speech-specific point of view. The music signal cannot be generalized so easily due to the presence of different types of music sources. Hence selecting robust features relating to music is a difficult task. Speech is produced by humans and extensive work has been done to study the speech production and perception systems in terms of the excitation source, vocal tract system, and the dynamics associated with them. The gross mechanism for producing speech remains the same across the human race. In order to produce a particular sound unit, the shape of the vocal tract (lowering jaw) and glottal vibration as excitation source remain mostly the same for a particular speaker. Even though there are other factors like fundamental frequency, pronunciation and speaker's individual anatomy that influence the production of a particular sound unit across different speakers, the major factors involved in the production are the shape of the vocal tract and the nature of the excitation source. Hence, exploring the behavior of speech-specific features which exploit the characteristics of excitation source, vocal tract system, and syllabic rate of the speech signal may be a better option for speech / music classification. The behavior of these speech characteristics in music segments is expected to be different compared to the speech segments.

The quasi-periodic and impulsive nature of the glottal vibration (a major excitation source in speech production) are unique to speech production. The normalized autocorrelation peak strength (NAPS) [10,11] of the zero frequency filtered signal (ZFFS) represents the quasi-periodic nature of the excitation source information of speech. The peak-to-sidelobe ratio (PSR) [12] of the Hilbert envelope (HE) of linear prediction (LP) residual feature represents the impulsive nature of the excitation source information. The majority of energy in case of speech is in the vowel-like sounds and

* Corresponding author.
E-mail addresses: banriskhem@iitg.ernet.in (B.K. Khonglah), prasanna@iitg.ernet.in (S.R. Mahadeva Prasanna).

http://dx.doi.org/10.1016/j.dsp.2015.09.005
1051-2004/© 2015 Elsevier Inc. All rights reserved.

concentrated in the low frequency region of audio-spectrum. The log mel energy feature can be used to exploit this property, and hence to represent the vocal tract information. Due to the physical limitation of the speech production system, the number of sound units that can be generated per unit time is also limited. The rate of speech production can be measured using the modulation spectrum. This feature, which has already been exploited in [2], represents the changes in the slowly varying temporal envelope corresponding to the speech syllabic rate [13]. These speech-specific features are different and carry independent evidence for speech / music classification.

The main idea of the work is the use of speech-specific features for the speech / music classification task. The NAPS of ZFFS has been explored for the task of foreground speech segmentation in [10], where the periodicity of the ZFFS for foreground speech is higher compared to the background speech and noise. The NAPS was used to measure the periodicity. The ZFFS was extracted based on the average pitch period of speech [14]. The pitch period of speech is lower than the pitch period of music and a study related to finding the pitch period of speech and music can be found in [15], which is an autocorrelation based method. This method has no upper frequency limit search range and hence it is possible to use this algorithm to find the pitch period of music. Extracting ZFFS of music according to the method in [14] will be interesting considering the higher pitch of music and the nature of the ZFFS of music may be different compared to speech as seen in the case of background speech and noise in [10]. This difference may be exploited for classifying speech from music signals.

The PSR of HE of LP residual has been explored in the case of processing degraded speech [16], where the spurious instants of significant excitation detected from small random peaks in the Hilbert envelope are eliminated using this measure. It is also used as a quantity to compare the cross correlation function of different methods in [12]. There is impulse-like excitation for speech signals, but such an excitation is completely different for the music signal. The HE of LP residual of speech has been used to find the impulse-like excitation in speech in earlier works. However, since the nature of excitation in music may be different, it will be interesting to study the behavior of this feature in music. Hence this feature is explored for the speech / music classification task. The best way to measure the differences of the HE of LP residual in speech and music is in terms of the PSR which is found to act as an effective measure for different tasks explored in [12,16].

There is an alternating nature of vowel and non-vowel regions in speech and this kind of nature may not be present in music. In [17], the gross vocal tract information was represented in terms of the sum of ten largest peaks of the DFT spectrum for the task of vowel onset point detection. The log mel energy can be considered to be an extension of this feature. However, for the log mel energy feature, the source information is smoothed out by passing the DFT spectrum through the mel filter banks. The difference of the vowel nature in speech and music regions can be captured in terms of the log mel spectrum energy, which represents the vocal tract information for the speech / music classification task.

Except for modulation spectrum, to the best of our knowledge, the other features have not been explored in speech / music classification task. In particular, their behavior in music needs to be studied. Since these features are speech-specific, their behavior in the music regions may deviate from speech regions. This gives a kind of discrimination between the speech and music regions. Accordingly, the novelty of the work may be summarized as follows:

- The overall concept of using the speech-specific features for speech / music classification. These features may have been explored in earlier literature in the context of speech. The behavior of these features in music is analyzed in this work. Their ability to discriminate speech from music is examined, so as to achieve an effective speech / music classification system.

The significance of the proposed speech-specific features may be explained as follows: For instance, the NAPS feature is estimated (to be described later) using *a priori* knowledge of human pitch range. As a result, the NAPS will be able to localize speech regions better compared to music regions. This may result in a distinct behavior of NAPS for speech and music regions. Alternatively, existing features for speech / music classification like zero crossing rate banks on the generic characteristics of the audio signal in the time domain, and not on any specific properties of speech production and perception. Thus NAPS may be more effective compared to zero crossing rate. These reasons motivate us to explore the speech-specific features for speech / music classification.

Considering the complexity of the audio data in broadcast news, and since the task involves only a two class classification, certain regions in the audio signal are defined as to be either speech or music. This task involves obtaining the speech regions as much as possible and hence speech in all kinds of scenarios (indoor, outdoor and with music background) belongs to the speech class. The music class contains either vocal or non-vocal music, where majority of the broadcast news music segments considered consists of non-vocal or instrumental music. The speech with background music refers to news headlines and advertisements, and the vocal music category includes songs and advertisements having singing voice. Although, there are different scenarios within the speech or music class, the focus of this work is mainly on classifying audio[1] into speech and non-vocal music segments. The music segments which contain singing mixed in with musical instruments are also considered. The music segments which do not involve musical instruments but contain only singing are not considered in this work.

Initially, the classification task is performed using a threshold based approach and non-linear mapping on the proposed speech-specific features. The classifiers such as Gaussian mixture models (GMMs) and support vector machines (SVMs) are then considered with the speech-specific features given as the input. Existing features like zero crossing rate (ZCR), spectral roll-off, spectral flux, spectral centroid and percentage of low energy frames which are popularly used in the literature have also been considered for the classification. The speech-specific features are then concatenated along with the existing features and the effect of this combination on the classification task is studied. The rest of the paper is organized as follows: The description of the speech-specific features and their significance are given in Section 2. Speech / music classification using the speech-specific features is described in Section 3. Section 4 describes the results and discussions. Finally the conclusion of the work is given in Section 5.

2. Speech-specific features for speech / music classification

This work focuses on the use of speech-specific features which will be described in a detailed manner. The other existing features consisting of spectral flux, spectral centroid, spectral rolloff [6], zero crossing rate (ZCR) [5], and percentage of low energy frames [2] have been used extensively for the speech / music classification task. Therefore they will not be described in detail here due to their availability in the literature.

[1] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/speech_music.php.

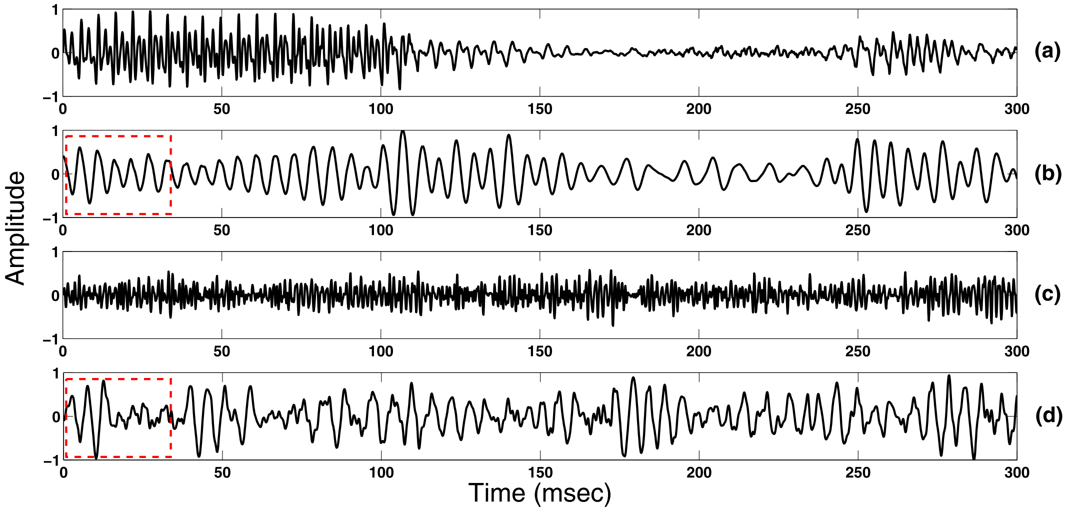

Fig. 1. (a) Speech signal, (b) ZFFS of speech, (c) Music signal, and (d) ZFFS of music.

2.1. *Speech-specific excitation source features*

The quasi-periodic and the impulsive nature of the excitation source of speech are exploited for deriving features which are described below.

2.1.1. *Normalized autocorrelation peak strength*

The ZFFS gives information about the epoch locations in the speech signal [14]. The speech signal is passed through a resonator located at the zero frequency which preserves signal energy around zero frequency and significantly attenuates all other information, mainly due to the vocal tract resonances. The trend in the output of zero frequency resonator is removed further by considering a window of length one or two pitch periods. The trend removed signal is termed as the ZFFS [14]. The positive zero crossings of ZFFS are demonstrated to give the location of epochs. The ZFFS is obtained as follows [14]:

- Difference the speech signal $s[n]$

$$x[n] = s[n] - s[n-1] \tag{1}$$

- The differenced speech signal $x[n]$ is passed through a cascade of two ideal zero frequency (digital) resonators, i.e,

$$y[n] = -\sum_{k=1}^{4} a_k y[n-k] + x[n] \tag{2}$$

where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$, $a_4 = 1$
- Remove the trend i.e.,

$$y_1[n] = y[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y[n-k] \tag{3}$$

$$\widehat{y}[n] = y_1[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y_1[n-k] \tag{4}$$

where $2N+1$ corresponds to the average pitch period over a longer segment of speech
- The trend removed signal $\widehat{y}(n)$ is termed as ZFFS.

In this work, the above method is applied to obtain the ZFFS of the audio signal. Fig. 1 displays the ZFFS plots for speech and music. It may be noted that the nature of the ZFFS for speech and music is different. The periodic nature of the ZFFS is more evident in speech compared to music and is unique for speech. The short term autocorrelation analysis is performed to exploit the differences in the periodic nature of the ZFFS of speech and music. The ZFFS is processed in blocks of 30 ms with 1 ms frame shift. The value of the first peak (after the central peak) in the autocorrelation sequence is an indication of the level of correlation in the frame. The central peak is the peak of the autocorrelation sequence at the origin and is indicated by the arrows in Fig. 2. The value of the first peak is normalized with the central peak resulting in normalized autocorrelation peak strength (NAPS) feature. The marked rectangles in Fig. 1 represent the region over which the NAPS is computed. It may be noted from Fig. 2 (a) and (b) that the NAPS of ZFFS of speech and music is 0.74 and 0.49, respectively.

The NAPS is higher for speech compared to music reflecting better the periodic nature of the ZFFS of speech compared to music. The presence of glottal activity in the speech regions gives a nearly periodic nature of the ZFFS and this type of periodic nature may not be present in music regions due to the different glottal activity like action in music compared to speech. This feature was developed as a speech-specific feature, not taking into account the other signals like music. Hence the behavior of this feature in music may be different compared to its behavior in speech as is evident in Fig. 1.

The reason for the difference is due to the pitch. The extraction of ZFFS may be viewed as a low pass filtering process (DC resonator) followed by a high pass filtering process (trend removal) to form a band pass filter. The center of the band pass filter depends on the average pitch period. The average pitch period considered is around the typical range for speech. Since the band pass filter is a function of the pitch period of speech, it will result in an extraction of a periodic signal corresponding to the pitch period of speech. In case of music, the band pass filter will emphasize a portion of the spectral energy around the pitch of speech. This spectral portion does not carry any information relating to the pitch of music, since the pitch of music is higher than that of speech in most cases. There may be cases of down-tuned instruments used in modern rock music, where the pitch of music may be lower than that of speech, but such cases are very few in our considered database. The spectral energy of music emphasized by the band pass filter has a different nature from the spectral energy of speech emphasized by that same filter. This results in a lesser periodic nature of the ZFFS for music compared to speech. Fig. 6 (b) shows the NAPS of ZFFS of the audio signal sample, Fig. 6 (a), taken from the Indian broadcast news where the first 5 s of the audio signal cor-

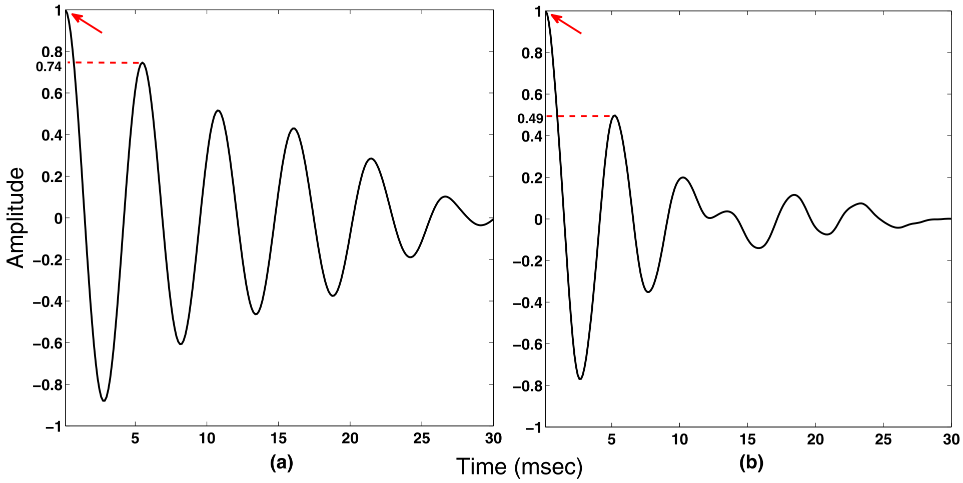

Fig. 2. Normalized autocorrelation plot for a selected portion of ZFFS of (a) speech and (b) music, respectively. The selected regions are shown in Fig. 1.

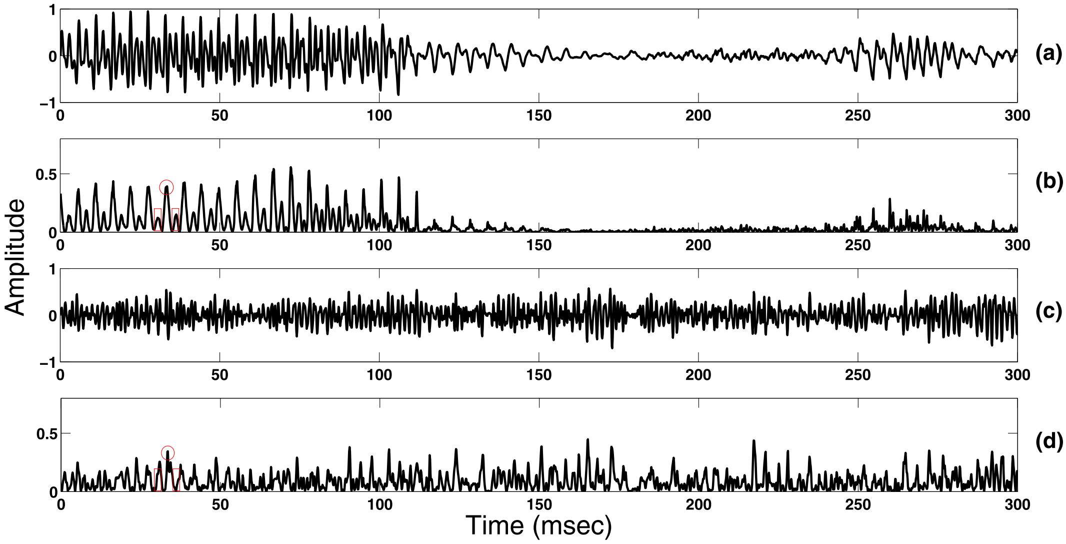

Fig. 3. (a) Speech signal, (b) HE of LP residual of speech, (c) Music signal and (d) HE of LP residual of music. The marked circles indicate the peaks while the marked rectangles indicate the region over which the side lobe variance is computed.

responds to speech while the next 5 s corresponds to music. It can be seen that the NAPS of ZFFS gives larger values in the speech regions compared to the music regions. Thus NAPS may be used as a feature to discriminate between speech and music regions.

2.1.2. Peak-to-sidelobe ratio

The LP analysis is a method of extracting the vocal tract and excitation source information in speech [18]. The effect of this analysis on the audio signal is studied in this work, where each frame size of 30 ms is processed with a frame shift of 1 ms. For each block of 30 ms, 10th order LP analysis (audio is sampled at $Fs = 8$ kHz) is performed to estimate the LP coefficients. The audio signal is passed through the inverse filter to extract the LP residual signal. In speech, the time-varying changes in the excitation source characteristics are smeared in the LP residual due to its bipolar nature [19]. These changes are further enhanced by computing the HE of LP residual [19].

The HE ($h_e[n]$) of LP residual ($e[n]$) is defined as [20]

$$h_e[n] = \sqrt{e^2[n] + e_h^2[n]} \tag{5}$$

where $e_h[n]$ is the Hilbert transform of $e[n]$, and is given by

$$e_h[n] = \mathrm{IDFT}(E_h[k]) \tag{6}$$

where

$$E_h[k] = \begin{cases} -jE[k], k = 0, 1, \ldots, (\frac{N}{2}) - 1 \\ jE[k], \quad k = \frac{N}{2}, (\frac{N}{2}) + 1, \ldots., (N-1) \end{cases} \tag{7}$$

and IDFT denotes the inverse discrete Fourier transform with $E[k]$ computed as the DFT of $e[n]$, and N is the number of points used for computing DFT.

The HE of LP residual for the audio signal is computed in this work. The HE of LP residual contains peaks corresponding to the excitation source information along with side-lobes around the peak. These peaks are higher in case of speech compared to music as shown in Fig. 3 (b) and (d). The side-lobe variation of the speech and music regions is almost the same although in most cases this variation is higher in music as compared to speech. This motivated the idea of computing the peak-to-sidelobe ratio (PSR) [12] of the HE of LP residual. The PSR is computed by first obtaining the peaks of HE of LP residual marked as circles in Fig. 3 (b) and (d). These peaks can be located by searching around the epoch locations obtained from the ZFFS [14]. A frame size of 3 ms around the epoch is considered for searching the peaks and the maximum value of the peaks in that frame is considered as the peak of the HE of LP residual. The side-lobe variance is computed over a frame size of one pitch period, which consist of half pitch period before

4 samples to the left and half pitch period after 4 samples to the right of the peak marked as rectangles approximately in Fig. 3 (b) and (d). Dividing the peak of HE of LP residual of the speech signal by the side-lobe variance gives the ratio of peak-to-sidelobe. For the marked segments in the Fig. 3 (b) and (d), the PSR for speech and music, respectively, is 53.34 and 11.76. Thus the speech regions have a higher PSR compared to the music regions as shown in Fig. 6 (c), where the PSR in the figure has been normalized over the entire duration of the 10 s audio clip.

The HE of LP residual has higher peaks in the speech regions which represent the impulse-like excitation of the glottal source. This impulse-like excitation may not be present in the music signals and is evident from the nature of the HE of LP residual wherein the peaks in the music region are much lower or even absent (Fig. 3). The main reason for this is attributed to the estimation of the LP residual signal. In LP analysis, each sample is predicted as a linear weighted sum of past p samples, where p is the order of prediction. The residual which is the difference between the predicted sample and the actual sample, gives a high value for the speech signals at the instant of significant excitation. In music, the error is nearly the same throughout the signal since the excitation source for music is different compared to speech. The side-lobe variance of the speech regions also tend to be lower than the music regions due to the noise-like nature of some of the music signals like the one in Fig. 3. The PSR of HE of LP residual hence has a higher value in the speech regions compared to the music regions and this feature can be used for discriminating speech from music regions.

2.2. Speech-specific vocal tract system features

The presence of a majority of high energy vowel-like sounds in speech is exploited to define a feature in terms of the vocal tract system which is described below.

2.2.1. Log mel spectrum energy

Speech is produced as a sequence of sound units. These sound units are produced as a result of changes in the vocal tract shape. At a gross level the sound units can be grouped into vowel and non-vowel-like sounds. Thus there will be continuous change in the vocal tract shape from the production of vowel to non-vowel-like sounds and vice versa. Distinct vocal tract shapes are associated with the production of vowels in speech. The vowel sound units are high energy regions and have most of their energy concentrated in low frequency range ($\leq$2.5 kHz). The dominance of vowel-like sound units in speech having energy concentrated in the low frequency range makes it different compared to music components. The energy in the low frequency range can be represented in terms of the mel filterbank energies. Due to the multiplication of the magnitude spectrum by the mel filterbank and summing the values obtained in each filter, most of the source information is smoothed out while computing mel filterbank energies. Hence the resulting evidence may be treated as a representation of vocal tract shape information. The vocal tract shape is manifested in the log mel spectrum energy values of the speech signal.

In this work, the audio signal is processed in blocks of 30 ms with a shift of 1 ms. For each block of 30 ms, a 512 point DFT is computed to obtain the spectrum of each block. The spectrum is then passed through 22 triangular filters (audio is sampled at Fs = 8 kHz) having central frequencies on the linear scale converted from the evenly distributed central frequencies on the mel scale to obtain the mel filter energy values. The logarithm of mel-filter energy values is then calculated. The sum of the first 18 log mel filter energy values are computed which covers about 2.5 kHz, the range of first 2 to 3 formant frequencies of the vowel sound units of speech, and this sum represents the log mel spectrum energy. Mathematically this is expressed as:

$$E[i] = \sum_{g=1}^{18} \log \left[\sum_{k=1}^{M} \left| S[k,i] f_g[k] \right|^2 \right] \tag{8}$$

where i is the frame number, g is the filter number, k is the frequency bin and M is the total number of bins.

$$S[k,i] = \sum_{n=0}^{N-1} s[n+iR] w[n] e^{\frac{-j2\pi nk}{N}} \tag{9}$$

where $s[n]$ is the speech signal, $w[n]$ is a rectangular window, R is the frame shift, and N is the total number of points for computing the Fourier transform. $f_g[k]$ is the triangular filter which has the central frequency $f_{\text{cent}}[g]$ on the linear scale converted from the mel scale as,

$$f_{\text{cent}}[g] = 700(e^{\frac{m}{1125}} - 1) \tag{10}$$

where m is the index number in the mel scale and g is the filter number.

Fig. 4 (c) and (f) represents the log mel spectrum energy values obtained from the 22 triangular filters, which are marked as triangles in the figure for speech and music, respectively. These values have been computed for a frame of speech as well as music which is indicated as rectangles in the figure. It can be seen that for speech, the marked rectangle represents a vowel-like region, which is also evident from the magnitude spectrum. The log mel energy values are high for the lower order mel filters indicating the high energy concentration in the low frequency range for speech segments containing vowel-like regions. There is a continuous change in shapes from vowel to non-vowel-like units production while moving from one frame to the next. The vowel-like regions exhibit higher energy and the non-vowel-like units exhibit lower energy. Thus there is a high variation of the log mel spectrum energy as seen in Fig. 6 (d). The log mel spectrum energy in the figure has been normalized over the entire duration of the audio clip. Alternatively, for music the energy distribution is random for a particular frame. The nature of the energy distribution which is evident in vowel-like regions in speech, may not be present in the music regions (Fig. 4 (f)). There is also a lower variation in the log mel spectrum energy of music as seen in Fig. 6 (d).

The high variation of the log mel spectrum energy in the speech regions compared to the music regions may be attributed to the fact that the speech regions contain an alternative nature of the high energy vowel-like regions and other types of non-vowel-like regions. This kind of nature may not be present in the music regions. It may be observed that the log mel spectrum energy is related to the first mel frequency cepstral coefficient (MFCC) c_0. The first MFCC, c_0 is computed by summing the log mel filter energy values of all the filters covering the entire frequency range. However, for the log mel spectrum energy in our work, the sum of the first 18 log mel filter energy values is taken and these filters cover about 2.5 kHz, the range of first 2 to 3 formant frequencies of the vowel sound units of speech. The variation of the log mel spectrum energy is higher for speech compared to music and this variation is unique in the case of speech. The variance in the log mel spectrum energy feature may therefore act as a good discriminator for the speech / music classification task.

2.3. Speech-specific modulation spectrum features

The syllable rate of speech is exploited to define a feature in terms of the modulation spectrum as described below.

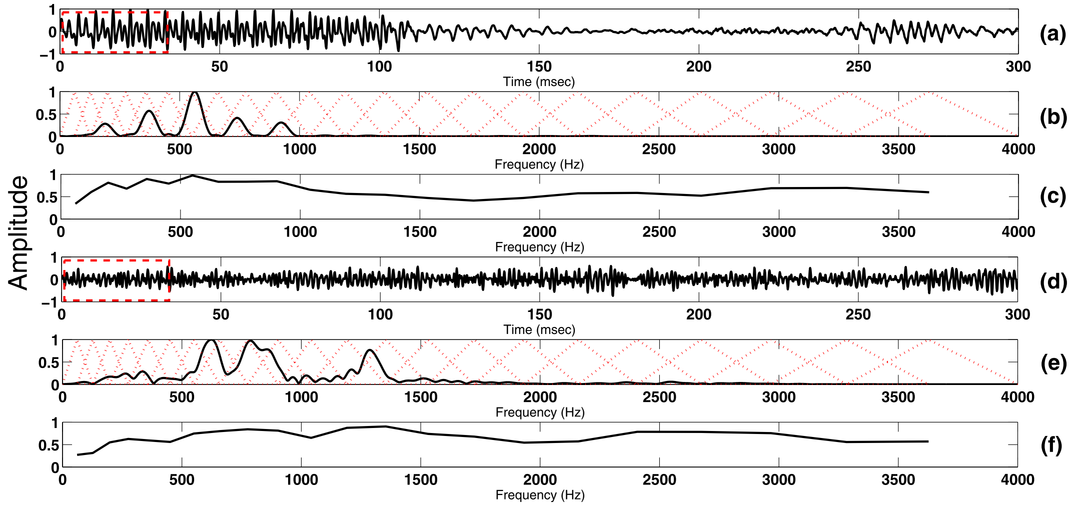

Fig. 4. (a) Speech signal, (b) Fourier transform spectrum of speech, (c) Log mel filter energy values of speech, (d) Music signal, (e) Fourier transform spectrum of music, (f) Log mel filter energy values of music. The marked rectangles indicate the regions over which the log mel filter energy values are computed. The marked triangles indicate the distribution of the mel filter banks. The first 18 filters cover the 0 to 2.5 kHz range of frequencies.

2.3.1. *Modulation spectrum energy*

The slowly varying temporal envelope components in the speech signal generally represent the modulation spectrum [21]. Low frequency components of several Hz mostly constitute the temporal envelope of speech signal. This kind of representation has compelling parallels to the speech production dynamics, where the articulators move at the rates of 2 to 12 Hz [22], and to the sensitivity of auditory cortical neurons to amplitude modulations at the rates below 20 Hz. Several studies have been explored earlier to show the importance of modulation spectrum in speech related tasks [23,24]. The use of modulation spectrum in speech / music classification has also been explored in [2] which exploits the idea that speech has a characteristic energy peak around the 4 Hz syllabic rate and music does not have this kind of nature. A detailed focus on the modulation spectrum including the development of the modulation spectrogram has been demonstrated in [21,25].

Given the audio signal, the modulation spectrum energy is computed as follows [21,25]: The audio signal is first analyzed into 18 critical band filters between 0 and 4 kHz frequency band. The filters are generally trapezoidal in shape, and the overlap between adjacent bands is minimum. Half-wave rectification and filtering with a low pass filter having cutoff frequency of 28 Hz is performed in each band to obtain an amplitude envelope signal. Down-sampling of each amplitude envelope signal to 80 samples/s is performed. Each down-sampled amplitude envelope signal is then normalized by the average envelope level in that channel, measured over the entire audio signal clip. In order to capture the dynamic properties of the signal, the modulations of the normalized envelope signals are analyzed by computing DFT over a Hamming window of length 250 ms with a window shift of 12.5 ms. Finally, the 4 Hz components are summed together, across all critical bands. Mathematically the modulation transfer function energies are expressed as

$$\mathrm{MTF}[m] = \sum_{c=1}^{18} \left[\left| \widehat{X}_c[k1, m] \right|^2 \right] \tag{11}$$

where m is the frame index, c represents the critical band number, and $k1$ represents a frequency index of 4 Hz. $\widehat{X}_c[k, m]$ is computed as

$$\widehat{X}_c[k, m] = \sum_{n=0}^{N-1} \widehat{x}_c[n + mR] w[n] e^{-\frac{j2\pi nk}{N}}; c = 1, 2, ..18. \tag{12}$$

where $\widehat{x}_c[n]$ represents the normalized envelope of cth filter output, $w[n]$ is a Hamming window, R is the frame shift and N is the number of points used for computing the DFT. The modulation energy components computed are up-sampled to 8000 samples/s.

The distribution of the 4 Hz modulation energy is shown in Fig. 5 (b) and (d) for speech and music, respectively, computed for a frame of speech and music, shown as rectangles in the figure. It can be clearly observed that there is higher energy at the 4 Hz frequency in speech compared to music. Fig. 6 (e) shows the plot of the 4 Hz modulation spectrum energy, where its values have been normalized over the entire duration of the audio clip. The 4 Hz modulation spectrum energy feature represents the slowly varying temporal envelope corresponding to the speech syllabic rate. Speech is more characterized by the 4 Hz syllable rate compared to music and hence this feature is higher in speech regions compared to the music regions. The 4 Hz modulation spectrum energy has been used for the speech / music classification task in earlier work. It has been used in the current work, since it represents the long term aspect of speech which is different from the source and the system.

So far, the behavior of the features for a single frame and for a single audio file has been described. Their behavior over a larger number of frames computed over all the audio files from the Scheirer and Slaney database [2] is seen by the histogram plot in Fig. 8, where the thick line indicates speech and the dashed line indicates music. It can be seen that there is higher separability in the speech and music distribution for the log mel energy variance feature with minimum overlap than the rest of the speech-specific features, although there is visible separation for the other speech-specific features as well. The nature of the histogram for the log mel spectrum energy variance and the PSR of HE of LP residual is similar to the existing features. The nature of the histogram for the NAPS mean and modulation spectrum mean is different compared to the other features.

3. Speech / music classification using speech-specific features

The previous section described the different speech-specific features that are considered for speech / music classification. The illustrations indicated that each of the features indeed show discrimination between speech and music regions. The evidence from each of these features need to be effectively combined for speech / music classification. The explorations for the same are presented in this section.

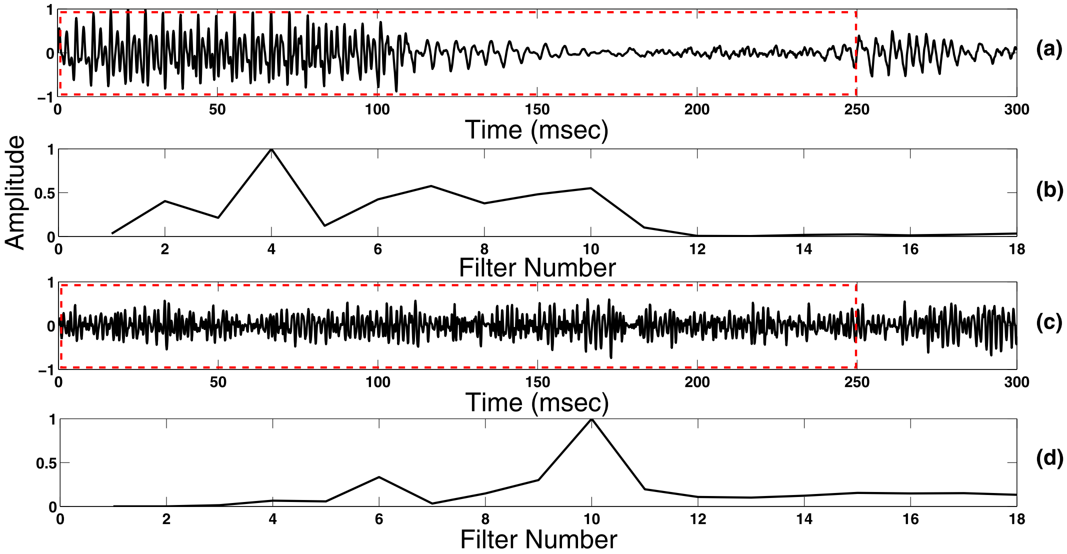

Fig. 5. (a) Speech signal, (b) 4 Hz Modulation spectrum energy from the critical band filters for speech, (c) Music signal, (d) 4 Hz Modulation spectrum energy from the critical band filters for music. The marked rectangles indicate the regions over which the 4 Hz modulation spectrum energy is computed.

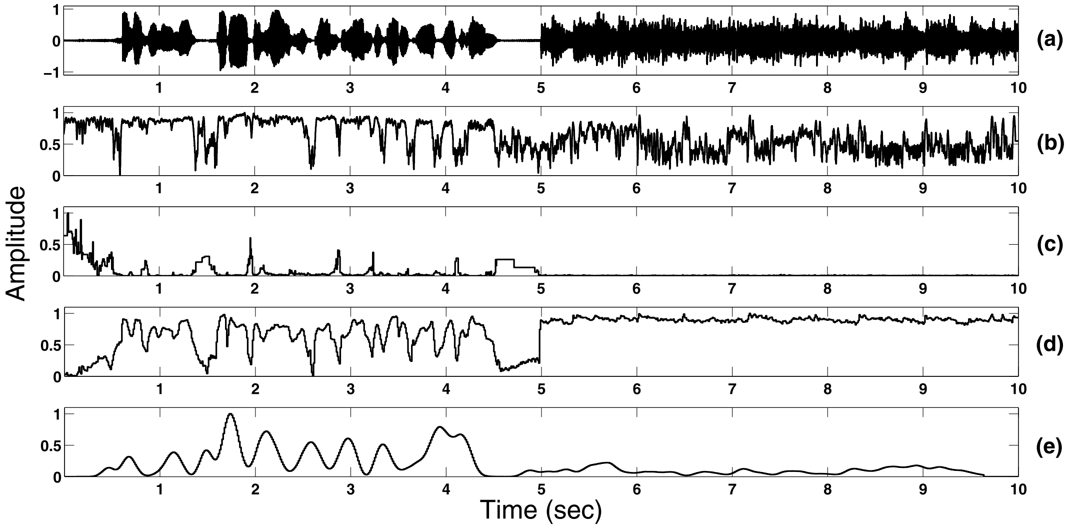

Fig. 6. (a) Audio signal, where the first 5 s correspond to speech and the next 5 s correspond to music, (b) NAPS of ZFFS, (c) PSR of HE of LP residual, (d) Log mel energy, (e) 4 Hz Modulation spectrum energy.

3.1. Speech / music classification by non-linear mapping and combining

The speech / music classification task involves assigning a particular label to speech and music. In this work, speech is given a label as *one* and music as *zero*. With this objective in mind, smoothing and non-linear mapping of the features is performed. Ideally the value of the feature is mapped to one for speech and zero for music, hence performing the classification task. It can be seen in Fig. 6, the NAPS of ZFFS, PSR of HE of LP residual, and modulation spectrum have mostly high values in the speech regions compared to music regions. However, there are some speech regions in which the feature values may be lower than the music regions which are categorized to be spurious. This spurious can be reduced by smoothing the features. It may be noted that the above features have been computed for a window size of 30 ms with a shift of 1 ms. Before the smoothing process, interpolation of the missing samples is required. The interpolation process is performed by duplicating the single value obtained for every frame to the missing samples caused by the frame shift to the next frame. A small shift has been chosen to reduce the number of samples required for interpolation. If a larger shift is chosen, more samples need to be interpolated and may affect the accuracy of the smoothing process and thereby reduce the overall accuracy. The mean over 1 s frame and every sample shift is computed for the NAPS, PSR, and modulation spectrum and the smoothed values are shown in Fig. 7. The variation of the interpolation method does not significantly change the final smoothed value of the features. Even the standard linear interpolation method results in a very similar smoothing effect on the features as the interpolation method mentioned earlier. However, the linear interpolation method is not done here since it is computationally more intensive than the interpolation method followed in this work.

For the log mel spectrum energy, the variance over 1 s frame for every sample shift is computed for smoothing. It can be observed from Fig. 6 (d), the variation of the feature in the speech regions is very high compared to the music regions. The window size for computing smoothed mean and variance contours is experimentally chosen for the best values. It was observed that there is not much difference in the mean and variance contours while smoothing with window sizes in the range of 500 ms to 1200 ms.

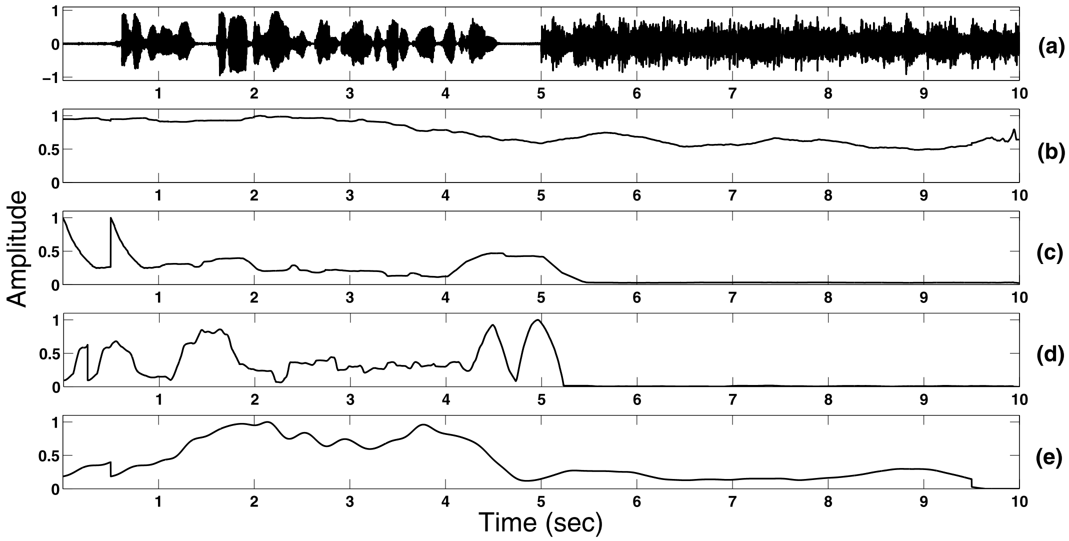

Fig. 7. (a) Audio signal, Smoothed, (b) NAPS of ZFFS, (c) PSR of HE of LP residual, (d) Log mel energy and, (e) 4 Hz Modulation spectrum energy. The smoothed contours have been computed from the features in Fig. 6.

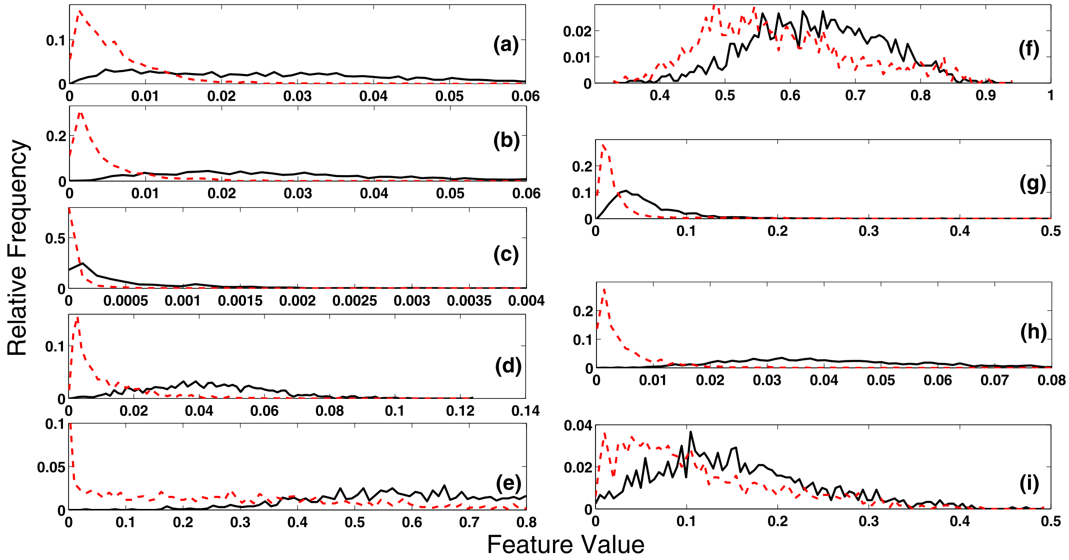

Fig. 8. Histogram plot for (a) ZCR variance, (b) Spectral centroid variance, (c) Spectral flux variance, (d) Spectral roll-off variance, (e) Percentage of low energy frames, (f) NAPS of ZFFS Mean, (g) PSR of HE of LP residual mean, (h) Log mel energy variance, (i) 4 Hz Modulation spectrum energy mean. Note that the continuous line represents speech and the dashed line represents music.

If the window size is chosen beyond this range, severe degradation in the smoothed contours is observed. It can be seen that the features having values lower in the speech regions compared to the music regions shown in Fig. 6, are now having their values smoothed to their nearest higher values as shown in Fig. 7, thus reducing spurious.

The smoothed evidences are then mapped using the non-linear mapping function given by

$$P_m = \frac{1}{1+e^{-(P_s-\Theta)/\tau}} + \alpha \qquad (13)$$

where, P_m is non-linearly mapped value, P_s is the smoothed evidence value, Θ, τ are the slope parameters and α is the offset which is the minimum value of the function. The values of τ and α are set to 0.001 and 0, respectively, since the binary mapping of either 0 or 1 is required. The main tunable parameter is Θ which is set experimentally, and this value is kept in the range of 0.3 to 0.6 based on the experiments performed on the databases. The overall performance does not change significantly if the value of Θ is varied in this range. The non-linearly mapped plots are shown in Fig. 9 (b)–(e), where it is seen that nearly all the speech regions have a value of one and the music regions have a value of zero.

A classification framework is presented, wherein these non-linear mapped values are combined by summing them together to produce an evidence for the speech / music classification task. The summed evidence is next non-linearly mapped using the same mapping function and a fixed threshold of 0.4. After that, a 1 s window with a non-overlapping 1 s shift of the audio segment is considered and the mean is computed. This window size is chosen in most of the segment based speech / music classification task [2, 9]. If the value of the mean is greater than a threshold which is 0.08 (this value is not crucial for the task), the segment is classified as speech, otherwise it is classified as music. The final result of the classification is shown in Fig. 9 (f) where all the 5 speech segments are classified correctly whereas only 4 out of 5 music segments are classified correctly. Misclassification of 1 music segment is observed in the figure.

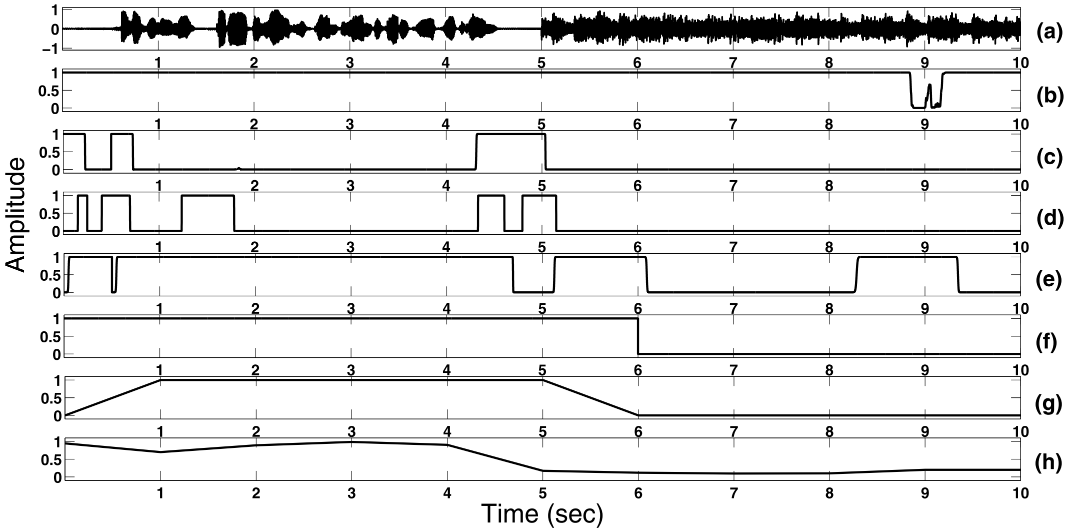

Fig. 9. (a) Audio signal, non-linear mapped value of smoothed, (b) NAPS of ZFFS, (c) PSR of HE of LP residual, (d) Log mel energy, (e) 4 Hz Modulation spectrum energy. The non-linear mapped values have been computed from the smoothed contours in Fig. 7. Classification result using (f) non-linear mapping, (g) Gaussian Mixture Models (GMM), (h) Support Vector Machines (SVM).

It can be seen that the non-linear mapping technique requires thresholds to be defined. However, the advantage of this method is that there is no requirement of training data as required by the classifiers. This method can also be described as the signal processing approach for classification, where the accuracy will be decided by the linear separability of the speech-specific features. This kind of signal processing approach has been followed in other tasks like voiced / unvoiced detection [26]. Since those tasks are similar to the speech / music classification task, we have employed this kind of approach for classification in this work.

The non-linear mapping technique gives an overall improved accuracy. To illustrate this, the classification is directly performed on the smoothed values. The smoothed log mel spectrum energy value of Fig. 7(d) is taken as an example and its mean is computed for 1 s window with a non-overlapping 1 s shift. If the mean is greater than a threshold of 0.5, the segment is classified as speech, otherwise it is classified as music. Similarly, the classification on the non-linear mapped value of this feature is performed by computing the mean of the non-linear mapped value ($\Theta = 0.5$) of the log mel spectrum energy shown in Fig. 9(d), compared the mean to a threshold of 0.08 as earlier, and the same kind of segment classification is performed. An overall accuracies of 68.46% and 74.34%, respectively, are obtained for the two cases, on the Broadcast News database (to be described later), indicating the significance of using the non-linear mapping technique.

3.2. Speech / music classification using gaussian mixture models and support vector machines

Gaussian mixture models (GMMs) have been explored earlier for the speech / music classification task [2]. The models are trained for the speech and music signals by using the expectation maximization algorithm. A new feature is assigned to a particular model which has a higher likelihood estimate. Diagonal covariances for the GMM have been used in this work.

Support vector machines (SVMs) are well suited for binary classification tasks and have shown considerable success in a variety of domains. The use of SVMs for speech / music classification has been explored in [27,28]. All the experiments using SVM in this work, were carried out using the libSVM [29] with a radial-basis function (RBF) kernel of the form,

$$K(x, y) = exp(-Y||x - y||^2) \tag{14}$$

The classification result using GMM and SVM for the audio file in Fig. 9 (a) is shown in Fig. 9 (g) and (h), respectively. The statistics of the raw speech specific features shown in Fig. 6 are computed for a window size of 1 s with a non-overlapping shift of 1 s [2,9]. These statistics of the features are concatenated to form a feature vector. This feature vector is given as input to the classifiers. It can be seen that using GMM, the first segment of speech has been misclassified and all the other remaining segments are classified correctly whereas using SVM all the speech and music segments have been classified correctly. The classifiers have been trained with a particular database and the details of the training process is given in the next section.

4. Results and discussion

The proposed method for speech / music classification is first evaluated on a database which has been recorded at random from the radio during the summer of 1996 by Scheirer and Slaney [2] which will be referred to as Scheirer and Slaney (S&S) database. The training examples taken from this database include 80 files of speech and 80 files of music without vocals which are of 15 seconds each. Another database evaluated in this work is the GTZAN database which has been explored in [9,30]. This database contains 64 files of speech and 64 files of non-vocal music each of length 30 seconds. The audio data in both of the databases has a sampling frequency of 22 050 Hz and has been down-sampled to 8000 Hz for the task. The evaluation is also done on the database containing audio data recorded from the Indian broadcast news channels which has a total of 104 files of speech and 104 files of non-vocal music each length of 5 seconds with 8000 Hz sampling frequency.

4.1. Non-linear mapping and combining

First, the results using non-linear mapping of the speech-specific features are shown in Table 1. For the GTZAN database, the performance in music is higher, however for the case of S&S and Broadcast news database, the performance in speech is higher. This depends on the threshold. However, varying the threshold does not change the overall performance for the three databases to larger extent. The results shown in the table are evaluated for a threshold (θ of the non-linear mapping function) of 0.5 for all the speech-specific features.

Table 1
Results using non-linear mapping in terms of classification accuracy (%).

Features →	Speech-specific features		
Database ↓	Speech	Music	Overall
S&S	84.33	64.58	74.45
GTZAN	70.26	78.80	74.53
Broadcast news	96.34	60.96	78.65

4.2. Classifiers

The use of thresholds for the task may not give optimal performance. Hence classifiers like GMM and SVM are used for the classification task on the speech-specific features. The width parameter Y and the cost parameter c of the SVM as well as the mixture k of GMM is varied to achieve optimal performances. The cost parameter c is set to 1 and the width parameter Y is set to 3 in this work. The number of mixtures for the GMM has been set to $k = 8$. The SVM parameters $(c = 1, Y = 3)$ and the number of mixtures of the GMM $(k = 8)$ have been fixed at their optimal values based on the results for the test data across the different databases. These parameters have been fixed for all the features and across different databases to show the impact of the speech-specific features for the classification task. The existing features like the ZCR, spectral centroid, spectral flux and spectral roll-off as well as the percentage of low energy frames are also considered for evaluation in order to compare the performances of the speech-specific features. The statistics of speech-specific features as well as the existing features are computed using a window size of 1 s with a non-overlapping shift of 1 s. As mentioned in Section 14, the statistics are computed on the raw features and not on the smoothed features. The individual features are evaluated and their performances are shown in Table 2. A 4-fold cross-validated scheme was used for evaluation with separate files in training and testing datasets. It can be observed that the variance of log mel spectrum energy feature, which is the speech-specific feature representing the vocal tract system shows superior performance on all the three databases. The existing features like the variance of spectral centroid, variance of spectral roll-off and percentage of low energy frames also show good performances.

It is interesting to see that individually, most of the existing features show better performances than the source and modulation spectrum features which belong to the category of speech-specific features. However, on combining the speech-specific features, we get a better performance than combining the existing features. This is reflected in the 4th and 5th row of Table 3, where the feature combination is performed by concatenating the features together to form a feature vector and fed as input to the classifiers. The reason for the better performance of the combined speech-specific features could be due to the capturing of complementary information by each of the speech-specific features, since these features represent the different aspects of speech production. The existing features, being general audio features may not be able to characterize the speech regions as much as the speech-specific features.

It can also be seen from Tables 1 and 3, that the overall performance of speech-specific features increased gradually from 74.45% when using threshold based approach to 95.12% using GMM and finally 95.87% using SVM on the S&S database. Similar trends are observed for the GTZAN and Indian broadcast news database. On combining the existing features with the speech-specific features, the best performances are obtained and are shown in the last row of Table 3. The best overall performance obtained is for the S&S database, which is 96.75% using SVM. The earlier work [2] on this database showed the best overall performance of 94.2%. Similarly for the GTZAN database, the best performance obtained is 92.03% which is comparable to the best performance reported in [9] which is 93.5%.

Table 2
Performance in terms of classification accuracy (%) using the different individual features on the Scheirer and Slaney (S&S) database, the GTZAN database and the Broadcast News (BN) database. In the table, the abbreviations, GMM indicates the Gaussian mixture model classifier and SVM indicates support vector machines. For the different features, the statistics are computed on the raw features and not on the smoothed features.

Database →	S&S database						GTZAN database						BN database					
Classifier →	GMM			SVM			GMM			SVM			GMM			SVM		
Features ↓	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall
ZCR var.	74.25	90.25	82.25	70.00	93.91	81.95	75.78	72.81	74.29	58.12	85.62	71.87	51.15	84.42	67.78	60.00	79.80	69.90
Spec. centroid var.	89.91	87.25	88.58	86.00	91.50	88.75	84.94	80.15	82.55	72.03	89.68	80.85	89.03	76.53	82.78	77.30	84.42	80.86
Spec. flux var.	80.91	80.66	80.79	55.91	92.00	73.95	68.33	70.36	69.34	48.59	85.00	66.79	67.30	69.61	68.46	57.50	81.53	69.51
Spec. roll-off var.	85.75	85.58	85.66	85.50	85.91	85.70	78.38	79.53	78.95	74.16	83.75	78.95	89.03	85.76	87.40	87.11	88.26	87.69
Percent. of low energy frames	83.75	77.75	80.75	88.58	72.25	80.41	78.17	77.08	77.63	83.54	72.23	77.89	90.38	80.57	85.48	90.38	80.00	85.19
NAPS mean	69.08	56.00	62.54	73.75	52.83	63.29	61.77	56.09	58.93	62.39	58.07	60.23	69.23	73.07	71.15	72.69	71.15	71.92
PSR mean	82.50	72.16	77.33	74.66	80.00	77.33	74.37	67.55	70.96	57.39	80.05	68.72	81.53	78.07	79.80	75.76	85.38	80.57
Log mel spec. energy var.	94.75	94.58	94.66	94.08	94.91	94.50	85.62	85.46	85.54	82.96	88.33	85.65	92.50	82.69	87.59	88.65	85.19	86.92
Modulation spec. energy mean	71.58	53.66	62.62	69.25	56.58	62.91	45.78	63.33	54.55	29.94	83.33	56.64	77.69	59.42	68.55	65.00	70.57	67.78

Table 3
Performance in terms of classification accuracy (%) using the existing, speech-specific and combined set of features on the Scheirer and Slaney (S&S) database, the GTZAN database and the Broadcast News (BN) database. In the table, the abbreviations, GMM indicates the Gaussian mixture model classifier and SVM indicates support vector machines.

Database →	S&S database						GTZAN database						BN database					
Classifier →	GMM			SVM			GMM			SVM			GMM			SVM		
Features ↓	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall	Speech	Music	Overall
Existing	91.58	87.16	89.37	90.50	89.83	90.16	87.70	81.66	84.68	87.29	88.43	87.86	90.76	81.53	86.15	92.11	89.42	90.76
Speech-specific	95.08	95.16	95.12	95.08	96.66	95.87	89.68	84.01	86.84	88.64	87.55	88.09	89.61	85.38	87.50	92.88	88.07	90.48
Combined	96.91	94.66	95.79	96.91	96.58	96.75	91.77	87.08	89.42	93.48	90.57	92.03	91.02	90.00	90.51	93.26	91.34	92.30

From Table 3 it can be observed that the difference of the performances in speech and music is higher, for the existing features compared to the speech-specific features when using GMM. This reflects the inability of the existing features to reduce the confusion between speech and music. The speech signal is more controlled in terms of its production and hence the signal characteristics of speech is similar. The music signal has a complex nature which may be produced by different instruments and some characteristics of the music signal may be similar to speech. The existing features are able to characterize the speech segments to a certain extent due to the similar signal characteristics of speech. However, since the existing features are not derived from the speech-specific knowledge, they are not able to discriminate the speech like music segments and tend to describe these segments as speech. On the other hand, the speech-specific features which represent the source, vocal tract system and syllabic rate aspects of speech are able to capture the speech segments of the audio signal successfully. For those music segments which have a nature similar to speech, the speech-specific features deviate significantly from their normal behavior since those speech like music segments may not be completely described by the speech-specific features, thereby reducing the confusion between speech and music.

4.3. Canonical correlation analysis (CCA)

In order to measure the correlation of each speech-specific feature to the combined cases, canonical correlation analysis (CCA) is performed, initially between each speech-specific feature with the existing features consisting of spectral flux, spectral centroid, spectral roll-off, zero crossing rate and percentage of low energy frames. Next, CCA is performed with the other speech-specific features. Finally, CCA is performed with the overall set of features consisting of the existing and the speech-specific features. The result of this analysis is shown in Table 4. This analysis was performed for the features computed on the S&S database. In the Table 2, it shows that the performance of the log mel energy variance feature is best individually than the other speech-specific features. However, CCA analysis shows that its value is greater than the other speech-specific features. This means that it is more correlated to the overall combined set of features than the other speech-specific features. CCA analysis also shows that the NAPS of ZFFS mean feature is the most uncorrelated feature to the combined set of features, followed by the PSR of HE of LP residual mean and the modulation spectrum mean features. This shows that the speech-specific features are mostly uncorrelated and combine effectively for the speech / music classification task.

4.4. Feature selection

An experiment is performed to find the minimum subset of features having performances close to the combined case in Table 3. Table 5 shows the result of the subset of those features. Since the log mel energy variance feature performs the best, this is used as the base feature. For the S&S database, the addition of NAPS mean provides the best additive improvement compared to the other features followed by the addition of the spectral roll-off variance feature. The performance saturates after the addition of the third feature. Similarly, for the other databases, the subset of features giving good performances are shown in Table 5.

4.5. Mismatched training and testing data

In order to study the performances of the mismatched training and testing data cases, an experiment is performed which involves one database as the training set and the other database

Table 4
Level of canonical correlation.

NAPS of ZFFS			PSR of HE of LP		
Existing	Speech-specific excluding NAPS of ZFFS	All excluding NAPS of ZFFS	Existing	Speech-specific excluding PSR of HE of LP	All excluding PSR of HE of LP
0.1318	0.1331	0.2040	0.5508	0.5090	0.5895
Log Mel			Modulation spectrum energy		
Existing	Speech-specific excluding log mel	All excluding log mel	Existing	Speech-specific excluding modulation spectrum energy	All excluding modulation spectrum energy
0.7773	0.5165	0.7952	0.2797	0.3267	0.4003

Table 5
Performance in terms of classification accuracy (%) of first three features on the three databases using SVM classifier.

Database →	S&S database		
Features ↓	Speech	Music	Overall
Log mel+NAPS mean	95.33	96.75	96.04
Log mel+NAPS mean+spec. roll-off var.	96.25	97.25	96.75
Database →	GTZAN database		
Features ↓	Speech	Music	Overall
Log mel+spec. centroid var.	85.93	90.10	88.02
Log mel+spec. centroid var.+PSR mean	91.04	87.44	89.24
Database →	BN database		
Features ↓	Speech	Music	Overall
Log mel+spec. roll-off var.	93.46	88.46	90.96
Log Mel+Spec. Roll-off Var.+ NAPS Mean	92.69	90.19	91.44

as the testing set. Table 6 shows the results of testing the broadcast news data on models trained on the GTZAN as well as the S&S database. The performances of the combined speech-specific features are better than the combined existing features. When the speech-specific and the existing features are combined, the best performance is obtained. Hence similar trends in the results are obtained even for the mismatched training and testing data.

4.6. Analysis on vocal music

An analysis of the behavior of the speech-specific features for the vocal music is briefly discussed here. The vocal music considered here involves singing mixed in with musical instruments. Fig. 10 shows the behavior of the speech-specific features for an audio signal which consists of speech for the first 5 s and vocal music for the next 5 s. It can be seen that there is some kind of discrimination between speech and vocal music especially for the NAPS of ZFFS, PSR of HE of LP residual, and log mel spectrum energy. The present work mostly focuses on the discrimination be-

Table 6
Performance in terms of classification accuracy (%) on the Broadcast News database using the models trained on the GTZAN database and S&S database.

Broadcast news test	Classifier →	GMM			SVM		
Models trained on	Features ↓	Speech	Music	Overall	Speech	Music	Overall
GTZAN database	Existing	94.23	70.19	82.21	85.76	75.57	80.67
GTZAN database	Speech-specific	93.65	79.03	86.34	95.00	80.57	87.78
GTZAN database	Combined	97.30	77.11	87.21	87.11	82.30	84.71
S&S database	Existing	93.65	75.00	84.32	92.11	83.07	87.59
S&S database	Speech-specific	94.03	81.73	87.88	93.46	84.23	88.84
S&S database	Combined	95.96	80.19	88.07	94.42	84.23	89.32

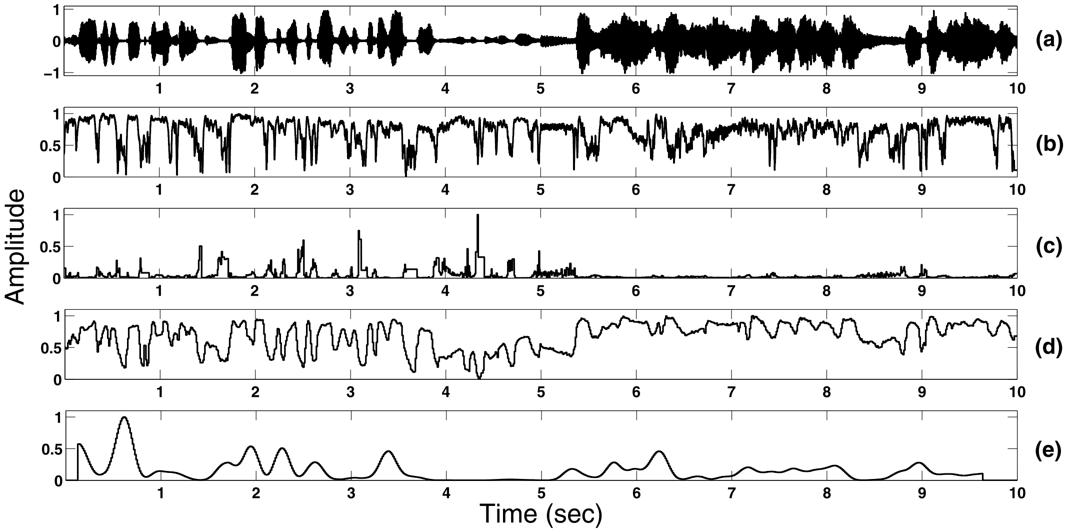

Fig. 10. (a) Audio signal, where the first 5 s correspond to speech and the next 5 s correspond to vocal music, (b) NAPS of ZFFS, (c) PSR of HE of LP residual, (d) Log mel energy, (e) 4 Hz Modulation spectrum energy.

tween speech and non-vocal music to understand the behavior of the speech-specific features and their discrimination for speech and non-vocal music regions. The work can be extended to the task of discriminating speech against vocal music. In particular, the behavior of the features for the vocal music segments which contain singing (with or without the mixing of musical instruments) can be explored in detail. Fig. 10 shows that there is potential for exploration of this case in the future.

5. Summary and conclusion

The use of the speech-specific features for the task of speech / music classification is explored. The NAPS of ZFFS, PSR of HE of LP residual, log mel spectrum energy, and modulation spectrum energy are considered as speech-specific features. The behavior of each feature is studied independently to demonstrate its potential for speech / music classification. Non-linear mapping of the features is done initially for the speech / music classification task. The speech-specific features are then classified using GMM and SVM. Their performances on the S&S database, GTZAN database, and the Indian broadcast news database are tabulated. The performance of the combined speech-specific features has been compared to the combined existing features where it is observed that the performance of speech-specific features is better. The existing features are then combined with the speech-specific features for the speech / music classification task and the best performance is achieved with this feature combination. Similar trends in the performances were obtained when testing the broadcast news database on the models trained with either the S&S or the GTZAN database.

Acknowledgments

This work is part of the project titled *Multi-modal Broadcast Analytics: Structured Evidence Visualization for Events of Security Concern* funded by the e-Security division of the Department of Electronics & Information Technology (DeitY), Govt. of India. The authors would like to thank Professor Dan Ellis of Columbia University for providing the Scheirer & Slaney database.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.dsp.2015.09.005.

References

[1] J. Saunders, Real-time discrimination of broadcast speech/music, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996, pp. 993–996.

[2] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, pp. 1331–1334.

[3] G. Williams, D.P.W. Ellis, Speech/music discrimination based on posterior probability features, in: Proceedings of the 6th European Conference on Speech Communication and Technology, EUROSPEECH '99, 1999, pp. 687–690.

[4] J. Ajmera, I. McCowan, H. Bourlard, Speech/music segmentation using entropy and dynamism features in a hmm classification framework, Speech Commun. 40 (2003) 351–363.

[5] C. Panagiotakis, G. Tziritas, A speech/music discriminator based on rms and zero-crossings, IEEE Trans. Multimed. 7 (2005) 155–166.

[6] Y. Lavner, D. Ruinskiy, A decision-tree-based algorithm for speech/music classification and segmentation, EURASIP J. Audio Speech Music Process. 2009 (2009).

[7] J. Shirazi, S. Ghaemmaghami, Improvement to speech–music discrimination using sinusoidal model based features, Multimed. Tools Appl. 50 (2010) 415–435.

[8] M. Kos, Z. Kačič, D. Vlaj, Acoustic classification and segmentation using modified spectral roll-off and variance-based features, Digit. Signal Process. 23 (2013) 659–674.

[9] G. Sell, P. Clark, Music tonality features for speech/music discrimination, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2014, pp. 2489–2493.

[10] K. Deepak, B.D. Sarma, S.M. Prasanna, Foreground speech segmentation using zero frequency filtered signal, in: Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[11] N. Adiga, S.R.M. Prasanna, Detection of glottal activity using different attributes of source information, IEEE Signal Process. Lett. 22 (2015) 2107–2111, http://dx.doi.org/10.1109/LSP.2015.2461008.

[12] V.C. Raykar, B. Yegnanarayana, S.M. Prasanna, R. Duraiswami, Speaker localization using excitation source information in speech, IEEE Trans. Audio Speech Lang. Process. 13 (2005) 751–761.

[13] C.H. Lee, J.L. Shih, K.M. Yu, H.S. Lin, Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features, IEEE Trans. Multimed. 11 (2009) 670–682.

[14] K.S.R. Murthy, B. Yegnanarayana, Epoch extraction from speech signals, IEEE Trans. Audio Speech Lang. Process. 16 (2008) 1602–1613.

[15] A. De Cheveigné, H. Kawahara, Yin, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111 (2002) 1917–1930.

[16] P. Krishnamoorthy, S.R.M. Prasanna, Reverberant speech enhancement by temporal and spectral processing, IEEE Trans. Audio Speech Lang. Process. 17 (2009) 253–266.

[17] S.M. Prasanna, B.S. Reddy, P. Krishnamoorthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies, IEEE Trans. Audio Speech Lang. Process. 17 (2009) 556–565.

[18] J. Makhoul, Linear prediction: a tutorial review, in: Proc. IEEE, 1975, pp. 561–580.

[19] T. Ananthapadmanabha, B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval, IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (1979) 309–319.

[20] A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, India, N. Delhi, India, 1975.

[21] S. Greenberg, B.E.D. Kingsbury, The modulation spectrogram: in pursuit of an invariant representation of speech, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, 1997, pp. 1647–1650.

[22] C.L. Smith, C.P. Browman, R.S. McGowan, B. Kay, Extracting dynamic parameters from speech movement data, J. Acoust. Soc. Am. 93 (1993) 1580–1588.

[23] H. Dudley, Remaking speech, J. Acoust. Soc. Am. 11 (1939) 169–177.

[24] R. Drullman, J.M. Festen, R. Plomp, Effect of temporally envelope smearing on speech reception, J. Acoust. Soc. Am. 95 (1994) 1053–1064.

[25] B.E.D. Kingsbury, N. Morgan, S. Greenberg, Robust speech recognition using the modulation spectrogram, Speech Commun. 25 (1998) 117–132.

[26] N. Dhananjaya, B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs, IEEE Signal Process. Lett. 17 (2010) 273–276.

[27] K. Sang-Kyun, J.H. Chang, Speech/music classification enhancement for 3gpp2 SMV codec based on support vector machine, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 92 (2009) 630–632.

[28] C. Lim, J.H. Chang, Efficient implementation techniques of an SVM-based speech/music classifier in SMV, Multimed. Tools Appl. (2014) 1–26.

[29] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[30] G. Tzanetakis, P. Cook, Sound analysis using mpeg compressed audio, in: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00, 2000, pp. II761–II764.

Banriskhem K. Khonglah was born in India in 1987. He received the B.E. degree in Electronics and Communication Engineering from Sri Jayachamarajendra College of Engineering, Visvesvaraya Technological University, Mysore, India, in 2010 and the M.Tech. degree in Microelectronics and VLSI from the National Institute of Technology, Silchar, India, in 2012. He is currently pursuing the Ph.D. degree in Electronics and Electrical Engineering at the Indian Institute of Technology Guwahati, India. His research interests include speech processing, analysis and recognition.

S.R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in Electronics Engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994. He received the M.Tech. degree in Industrial Electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently a Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati. His research interests are in speech and signal processing.