Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 628570, 13 pages
doi:10.1155/2009/628570

Research Article

Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy

Marko Kos, Matej Grašič, and Zdravko Kačič

Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ul. 17, 2000 Maribor, Slovenia

Correspondence should be addressed to Marko Kos, marko.kos@uni-mb.si

Received 6 March 2009; Revised 4 June 2009; Accepted 2 September 2009

Recommended by Aggelos Pikrakis

This paper presents a novel feature for online speech/music segmentation based on the variance mean of filter bank energy (VMFBE). The idea that encouraged the feature's construction is energy variation in a narrow frequency sub-band. The energy varies more rapidly, and to a greater extent for speech than for music. Therefore, an energy variance in such a sub-band is greater for speech than for music. The radio broadcast database and the BNSI broadcast news database were used for feature discrimination and segmentation ability evaluation. The calculation procedure of the VMFBE feature has 4 out of 6 steps in common with the MFCC feature calculation procedure. Therefore, it is a very convenient speech/music discriminator for use in real-time automatic speech recognition systems based on MFCC features, because valuable processing time can be saved, and computation load is only slightly increased. Analysis of the feature's speech/music discriminative ability shows an average error rate below 10% for radio broadcast material and it outperforms other features used for comparison, by more than 8%. The proposed feature as a stand-alone speech/music discriminator in a segmentation system achieves an overall accuracy of over 94% on radio broadcast material.

1. Introduction

Segmentation of audio data has become a very important procedure in audio processing systems. It is especially significant in applications such as automatic speech recognition (ASR), where only the speech segments of an input audio stream are led to the system's input, and nonspeech segments are discarded [1, 2]. In this way, the speed and accuracy of an ASR system can be improved, and the computation load is also reduced. The prior segmentation of audio data is also very important for applications such as broadcast news transcription [3], where the speech is typically interspersed with music and background noise. With the development of the internet, content-based indexing [4–6] has emerged, because there is a lot of audio data that is not indexed by web search engines. In such systems, audio segmentation is part of the indexing task. Segmentation is also used in systems for audio and speaker diarization [7–9], retrieval of audio-visual data [10, 11], and so forth.

One of the more often used acoustic segmentation types is speech/music segmentation. This is not surprising, because speech and music are two of the most important and often present acoustic classes within the audio domain. In the spectrum of typical speech and music signals, we can see the difference between these two acoustic classes. Typically different kinds of speech have certain common features, for example, most of speech energy is in the lower part of the frequency spectrum (below 1 kHz). Depending on the type of music the music frequency spectrum can be quite different. There are many different music genres: rock, pop, rap, classical, hip-hop, electronic, latin, jazz, country, dance, and so forth. In some music genres (e.g., rap music) music can also be quite similar to speech. Because new domains for segmentation are constantly emerging, speech/music discrimination and segmentation is an active field of research.

To date, a lot of research effort has been put into speech/music segmentation. Many different systems for segmentation have been introduced and many different features proposed (some of the features are compared in [12]), such as zero-crossing rate (ZCR), low-frequency modulation (4 Hz typically), root mean square (RMS), spectral roll-off point (SR), spectral centroid (SC), spectral flux (SF, also known as delta spectrum magnitude), percentage of "low energy" frames (PLEFs), line spectral frequencies, perceptual features such as timbre and rhythm, Mel-Frequency Cepstral Coefficients (MFCCs), entropy and dynamism features, and so forth. Some of the above-mentioned features are more successful when their variance values are used (e.g., zero-crossing rate and spectral flux). Frameworks, such as neural networks, Gaussian Mixture Models (GMMs), support vector machines, Hidden Markov Models (HMMs), and the nearest-neighbour, have been used for classification. Although some frameworks perform better than others, features are still one of the main factors for final performance.

Several approaches for speech/music discrimination have been proposed in the past. Saunders [13] proposed a method for real-time automatic monitoring of radio channels. His system was based on using zero-crossing rate and energy as features, extracted in a 2.4 second window. The author reported an accuracy of 98%.

In their work, Scheirer and Slaney [14] used 13 features to characterize the distinct properties of speech and music signals. They also examined three classification schemes, that is, the GMM classifier, the multidimensional MAP Gaussian classifier, and the nearest-neighbour classifier. The best classifier accuracy was reported at over 94%, and when integrated into long segments of sound (2.4 seconds), it achieved accuracy over 98%.

The authors in [15] proposed a method based on the entropy and dynamism features within an HMM classification framework. In their approach, an artificial neural network trained on clean speech only is used as a channel model, at the output of which entropy and dynamism are measured every 10 milliseconds. These features are then integrated over time through an ergodic 2-state HMM, using minimum duration constraints. Different experiments, including different music styles, as well as different temporal distributions of speech and music signals, have been conducted with a reported accuracy of over 90%. These authors also noted that this method can be easily adapted to other speech/nonspeech discrimination applications.

Two methods for speech/music classification for multimedia applications were compared in [16]. The first method is based on a zero-crossing rate and Bayesian classification. This method is very simple from the computational point of view, and gives good results in the case of pure music or speech, but some performance degradation arises when the music segment also contains speech superimposed on music, or strong rhythmic components. In order to overcome these problems, the authors proposed a second method, which is based on neural networks. It is reported that this method performs better at the expense of a limited growth in computational complexity. In practice, real-time implementation is possible, even if using low-cost embedded systems.

The authors in [17] investigated several audio features that have not been previously used in speech/music classification. Three different classification frameworks have also been studied, and tests have shown that multilayer perceptron neural networks achieve the best performance.

A classification method based on sinusoidal trajectories is introduced in [18]. Sinusoidal trajectories represent the temporal characteristics of each sound category, such as speech, singing voice, and a musical instrument. Twenty temporal features are extracted from trajectories and used to classify sound segments into categories, by using statistical classifiers. The authors developed an optimal spectral tracking algorithm with low computational complexity, in order to handle the temporal overlapping of sounds.

The author in [19] presented a method for performing automatic segmentation based on features relating to rhythm, timbre, and harmony. A comparison was made between features only, and between the features and manual segmentation of a 48-song database. Standard information retrieval performance measures were used for measuring performance. Results show that the timbre-related features perform best.

In [20], the authors performed speech/music classification based on one-class support vector machines. The experimental results show that the classification method, which can be easily implemented, performs better than the other methods implemented on the same database.

The authors in [21] proposed a method for speech/music discrimination based on RMS and zero crossings. Experimental results show good efficiency and performance. The segmentation and classification algorithms were benchmarked on a large dataset, with a classification accuracy of about 95% and a segmentation accuracy of about 97%.

A computationally-efficient speech/music discriminator for radio recordings was presented in [22]. It is an offline system based on a region growing technique operating on a single feature (chromatic entropy). This system was tested on recorded internet radio broadcast material and achieved an average discrimination accuracy of 93.38%.

The authors in [23] presented a robust and computationally efficient speech/music discriminator. Their approach is based on the extraction of four features (zero-crossing rate, spectral roll-off, loudness, and fundamental frequencies). The feature values are combined linearly into a unique parameter. This method has achieved very good accuracy, even for severely degraded and noisy signals, and it is also remarkably robust in unknown situations. The low computational complexity of the method makes it appropriate for applications that demand real-time operation.

In [24], the authors proposed two novel features for speech/music discrimination, called Average Pitch Density and Relative Tonal Power Density. The features were compared to RMS, ZCR, variance of SF, and PLEF features. The two novel features have proved to be more robust under noisy conditions.

A method based on a low-frequency modulation feature is presented in [25]. The low-frequency modulation amplitudes calculated over 20 critical bands, and their standard deviations were found to be good features for speech/music discrimination and were also discovered to be less sensitive to channel quality and model size than MFCC features.

The authors in [26] introduced an evolutionary speech/music discrimination method for audio coding improvement. In order to discriminate between speech and music, a fuzzy rules-based system is incorporated into the decision stage of a traditional speech/music discrimination system. Experimental results demonstrated the robustness of the proposed system and a classification accuracy of about 94% was obtained over a wide-range of audio samples.

In [27], the authors presented a fast and robust speech/music discrimination approach, based on a Modified Low Energy Ratio feature (MLER). The feature is extracted from each window-level segment as the only feature. A novel context-based postdecision method was designed to refine the classification results. The proposed method was evaluated on various audio data, containing clean and noisy speech from various speakers, as well as a wide range of musical content. A classification accuracy of 97% was achieved despite the low complexity of the method.

In this paper, we propose a novel feature for speech/music discrimination. The main idea for the feature construction is that energy in a narrow frequency sub-band varies more rapidly, and to a greater extent for speech than for music. The energy variance in such a sub-band is, therefore, greater for speech than for music.

The remainder of this paper is organized in the following way. In Section 2, we present the most commonly used set of features for speech/music discrimination, which we will use for comparison. Our proposed feature for speech/music discrimination is presented in Section 3. Section 4 contains an analysis of the feature's discrimination ability. Section 5 presents the experimental setup for speech/music segmentation. Descriptions of experiments and results are given in Section 6. Conclusions and findings are drawn together in Section 7.

2. Speech/Music Discrimination Features

Many standard features are available for the task of discriminating between speech and music signals. For the purpose of comparison with the proposed novel feature, we implemented the following set of standard features.

(i) Zero-crossing rate (ZCR) is a member of the time-domain features, and is the number of zero-crossings of a signal within a predefined window. Zero-crossing occurs when successive samples have different algebraic signs [28]. ZCR can be computed as

$$\mathrm{ZCR} = \frac{1}{2N}\sum_{n=1}^{N-1} \left|\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]\right|, \tag{1}$$

where N is the number of samples in one window, $x(n)$ represents the samples of the input window, and $\operatorname{sgn}[x(n)]$ is ± 1 as $x(n)$ is positive or negative, respectively. ZCR is widely used in practice and is also a strong measure for discerning fricatives from voiced speech. The sampling rate of a signal should be high enough to detect any crossing through zero. It is also very important that the signal is normalized, so that the amplitude average of the signal is equal to zero [29]. The ZCR of music is usually higher than that of speech, because ZCR is proportional to the dominant frequency (music has higher average dominant frequency [30]).

(ii) Spectral roll-off (SR) is the measure of skewness of the signal's frequency spectrum. It is the value of the frequency under which usually 95% of the signal's power resides. It is a good measure for distinguishing between voiced and unvoiced speech. It is expected that speech has a lower value of spectral roll-off, because it has most of the energy in the lower part of the frequency spectrum. The mathematical expression used, is

$$\sum_{k=1}^{R} X(k) = 0.95\sum_{k=1}^{M} X(k), \tag{2}$$

where k is the frequency bin index, M is the total number of frequency bins, $X(k)$ is the amplitude of the corresponding frequency bin, and R is the spectral roll-off number.

(iii) Spectral centroid (SC) is defined as the centre of a signal's spectrum power distribution. Like spectral roll-off, spectral centroid is also a measure of spectral shape. Music signals have high spectral centroid values because of the high frequency noise and percussive sounds. On the other hand, speech signals have a narrower range, where pitch stays at fairly low values. It has different values for voiced and unvoiced speech, and can be calculated as

$$\mathrm{SC} = \frac{\sum_{k=1}^{M} k \cdot X(k)}{\sum_{k=1}^{M} X(k)}, \tag{3}$$

where k is the frequency bin index, M is the total number of frequency bins, and $X(k)$ is the amplitude of the corresponding frequency bin. Higher values mean "brighter" sound with higher frequencies.

(iv) The percentage of low energy frames (PLEF) is a percentage measure of low energy frames, and is also known as Low Short Time Energy Ratio (LSTER) [31]. PLEF is defined as the proportion of frames, with RMS power of less than 50% of the mean RMS power within a specific window (usually 1 second). A higher value for speech is expected, because it contains more silent moments than music. The mathematical expression for PLEF feature calculation is

$$\mathrm{PLEF} = \frac{1}{2N}\sum_{n=0}^{N-1}\left[\operatorname{sgn}(0.5 \cdot \mathrm{STE}_{\mathrm{AV}} - \mathrm{STE}(n)) + 1\right], \tag{4}$$

where n is the index of a current frame, N is the total number of frames in a window, $\mathrm{STE}(n)$ is the short-time energy of a current frame, and $\mathrm{STE}_{\mathrm{AV}}$ is the average short-time energy in a window, and can be calculated as

$$\mathrm{STE}_{\mathrm{AV}} = \sum_{n=0}^{N-1} \mathrm{STE}(n). \tag{5}$$

(v) Spectral Flux (SF) [32]: spectral flux, also known as delta spectrum magnitude, is a measure which characterizes the change in the shape of the signal's spectrum. The rate of change in spectral shape is higher for music and, therefore, this value is higher for music than for speech. Spectrum flux can be calculated as the ordinary Euclidean norm of the delta spectrum magnitude:

$$\mathrm{SF} = \frac{1}{M}\sqrt{\sum_{k=1}^{M}\left(X_i(k) - X_{i-1}(k)\right)^2}, \tag{6}$$

where M is the total number of frequency bins, i is the frame index, k is the frequency bin index, and X_i and X_{i-1} are the spectrum magnitude vectors of frames i and i-1, respectively. It is known that speech alternates between transient and nonperiodic speech to short-time stationary and periodic speech, due to phoneme transitions (e.g., consonant to vowel). On the other hand, music and environmental sounds can be periodic or monotonic, and have more constant rates of change versus that typical for speech. This means the variance of SF for speech is larger than for music and most environmental sounds.

3. Variance Mean of Filter Bank Energy

In this section, we will describe the motivation for constructing the newly-proposed Variance Mean of Filter Bank Energy (VMFBE) feature.

Our goal was to analyze the possibilities of constructing a good discriminator between speech and singing voice with instrumental accompaniment. As can be seen in Figure 1, spectral representations of speech and music can be very different, despite the fact that there is a human voice present in both cases. It is typical for speech that the speaker's pitch can have values between 50 Hz and 400 Hz, and can vary by as much as 160 Hz, especially if the speaker is excited or surprised [33, 34]. Also, the duration of the phonemes is shorter for speech (40–200 milliseconds) than for the singing voice (600–1200 milliseconds) [35]. This can be seen in Figure 1, where changes in individual speech harmonics are more rapid for speech than for music. Furthermore, Figure 1 also shows that speech harmonics in music tend to have steadier values during longer periods of time than with speech. If we exploit this fact and divide the signal's spectrum into several sub-bands, narrow enough to catch the variation of pitch and higher harmonics, we can expect the energy of an individual sub-band to go through more drastic and rapid changes during speech than music. Thus, the variance in energy of such a sub-band should be higher for speech than for music. With this idea in mind, we now define the VMFBE feature calculation procedure.

3.1. Calculation of a VMFBE Feature. The first three steps of VMFBE feature calculation are sampling, windowing, and DFT calculation. These steps will be described in detail later (in Section 5), when the experimental framework is presented because they are also common to some other features. Now, we will focus solely on VMFBE feature calculation.

After calculating the signal's spectrum magnitude, we filter the spectrum with a set of triangular filters. These filters are distributed evenly on the melodic scale (mel-scale) frequency axis with 50% overlap. The equation for transforming frequency from linear scale (f_{lin}) to mel-scale (f_{mel}) [36] is

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(\frac{f_{\text{lin}}}{700} + 1\right). \qquad (7)$$

The centre frequencies on linear scale (f_C) of mel-distributed triangular filters [36] are calculated according to

$$f_C(l) = 700 \cdot \left(10^{(\text{STF}+l\cdot(\text{SAF}/2-\text{STF})/(L+1))/2595} - 1\right), \qquad (8)$$

where STF is the start frequency of the first (lowest) sub-band (32 Hz in our case), SAF is the sampling frequency, L is the number of filters, and l is the sub-band filter index.

Every DFT magnitude coefficient is multiplied by the corresponding sub-band filter channel gain and the logarithmic energy in that sub-band [36] is calculated as

$$E_{l,n} = \log \sum_{k=1}^{M} (X_n[k] \cdot F_1[k])^2, \quad n = 1 \cdots W, \qquad (9)$$

where $l = 1 \cdots L$ is the filter number index, k is the frequency bin index, M is the number of frequency bins, n is the frame index in the variance calculation window, W is the number of frames in the variance calculation window, $E_{l,n}$ is the logarithmic energy coefficient of the corresponding filter, X_n is the spectrum magnitude vector, and F_l is the corresponding filter channel gain function.

After filter bank energy calculation, we calculate the energy variance for each filter channel. The variance can be calculated over different time windows. The variance calculation window has to be long enough to capture enough energy dynamism but, at the same time, it must not be too long, because time resolution for the segment border calculation will be low. The variance of an individual filter channel can be expressed as

$$V_l = \text{var}(E_l), \quad E_l = [E_{l,1}, \dots, E_{l,W}], \quad l = 1 \cdots L, \qquad (10)$$

where V_l is the variance of the lth corresponding filter channel, l is the filter number index, L is the number of filters, and E_l is the energy vector of the lth filter channel. Visual representation of the filter bank energy variance coefficients can be seen in Figure 2. However, 200 filters were used for better visual representation. In the figure there are three explicit regions. The left region represents speech (0–8 seconds), the middle region represents silence (8–10 seconds), and the right region represents music. It can be seen that the energy filter bank variance coefficients representing speech attain higher values than those representing music, especially in the lower half of the frequency band. This is due to the fact that the majority of energy in speech resides in the lower half of the frequency spectrum.

In order to obtain only one feature from a number of filter bank energy variance coefficients, we calculate the mean of those coefficients. Mathematically, we formulate this as

$$\text{VMFBE} = \frac{1}{L}\sum_{l=1}^{L} V_l, \qquad (11)$$

where l is the filter bank number index, L is the number of filters, V_l is the variance value of the corresponding filter's energy, and VMFBE is the calculated feature value. After the

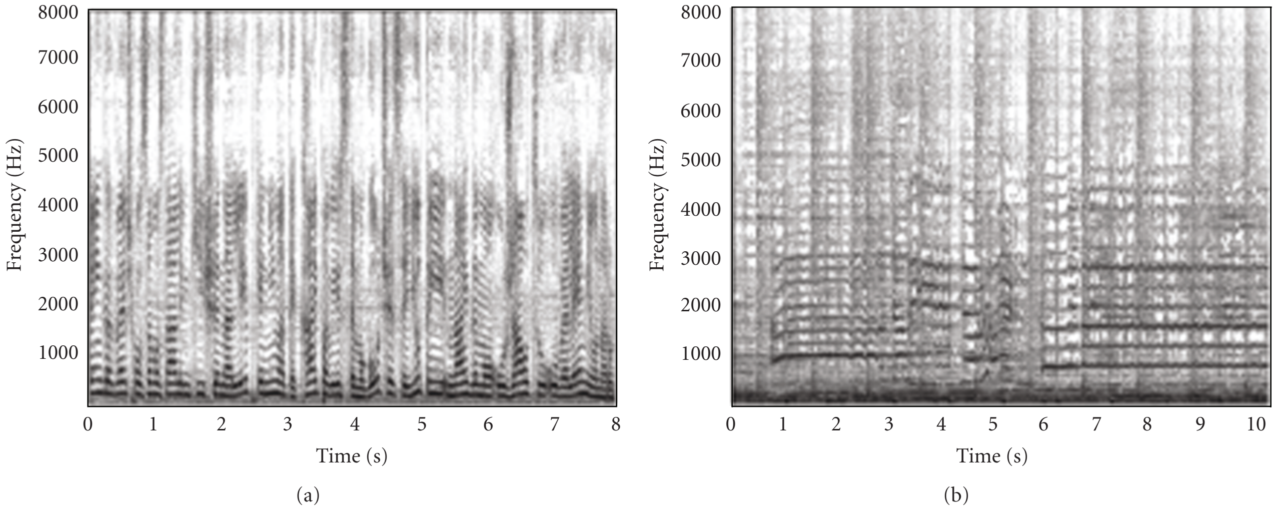


Figure 1: Spectrograms of (a) Speech of a female radio speaker, (b) music including vocals, recorded from public broadcast radio station.

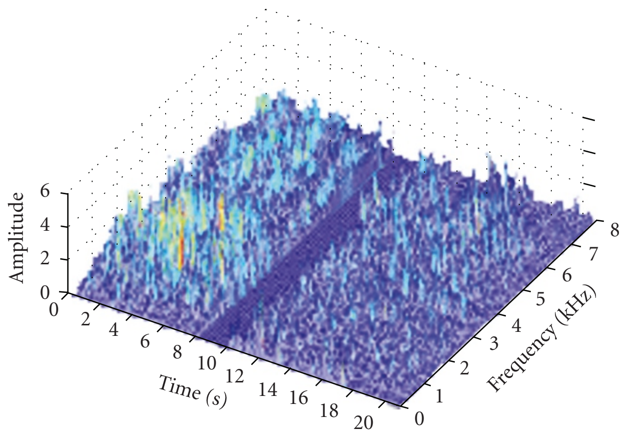


Figure 2: Visual representation of the energy filter bank variance coefficients (0–8 seconds speech, 8–10 seconds silence and 10–20 seconds music).

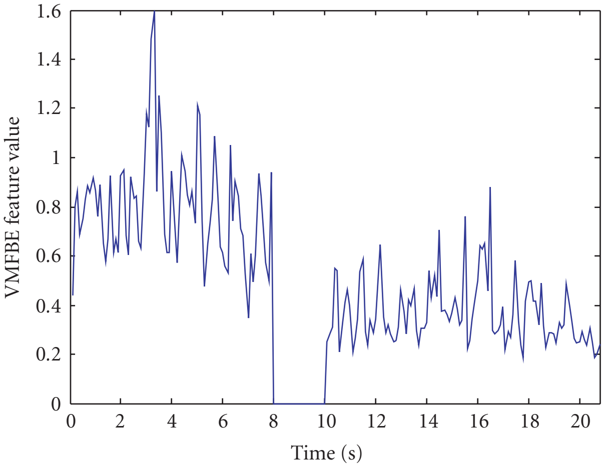


Figure 3: Visual representation of the VMFBE feature (value for speech is higher than for music; 0–8 seconds speech, 8–10 seconds silence, and 10–20 seconds music).

calculation of mean, we get the average variance value for our window of observation. The values of the VMFBE feature for our demo recording are shown in Figure 3. This figure shows that the VMFBE feature achieves higher values for speech, than for music, as was anticipated.

If we compare the calculation steps of the VMFBE feature calculation with the calculation procedure of the MFCC features, we notice that 4 out of 6 steps are in common (windowing, DFT calculation, mel-filtering, and filter log-energy calculation) [36]. By taking this into account, only two additional steps (sub-band filter energy variance calculation and energy variance mean calculation) are required to implement a speech/music discriminator in an ASR system, based on MFCC features. For example, if speech/music discriminator would be implemented using variance of SF, only 2 out of 5 steps would be in common with MFCC feature calculation procedure.

4. Discrimination Ability Analysis of the Proposed Feature

Analysis of the proposed VMFBE features' discrimination ability, together with other features mentioned in Section 2 (standard used features for speech/music discrimination), is performed on two different databases. The first database is composed of television news shows and late-night news shows, recorded mostly in a studio environment. Recordings of lower quality are usually telephone recordings or news reports from the field. The second database is the database of radio broadcast recordings from several Slovenian radio stations.

4.1. BNSI Broadcast News Database. The Slovenian BNSI Broadcast News database was collected in a cooperation between the University of Maribor and the Slovenian public

Table 1: Focus conditions of the BNSI database.

FC	Description	Perc. (%)
F0	Read studio speech	36.6
F1	Spontaneous studio speech	16.2
F2	Clean telephone speech	1.65
F3	Speech with music background	6.0
F4	Read or spontaneous speech with background other than music	37.6
F5	Speech of nonnative speakers	0.1
Fx	Unclassified	1.85

broadcast company RTV SLO [37]. Two different types of news shows are incorporated into the BNSI database. The first is the evening news, and the other the late-night news shows, where more detailed analysis of major daily events is given. The audio format is 16 bit 16 kHz wav. The speech corpus consists of 42 news shows which include 36 hours of speech. Also, 30-hour material is used as a train set, 3 hours as development, and 3 hours as evaluation set. The database material is divided into 7 focus conditions (FC). The focus conditions are presented in more detail in Table 1.

As can be seen from the table, the two most frequent conditions are F0 (read studio speech) and F4 (read or spontaneous speech with a background other than music). The database contains both male and female speakers with approximately equal shares. Speech represents 88% of the database material and 12% is nonspeech material. The nonspeech part of the database is composed of silence, music, and noise. More than 70% of nonspeech material is music, and it is composed of jingles, intros, and so forth, the music is mostly instrumental and electronic (no singing). Manual transcriptions are available for the database. Transcriptions also include information about a speaker's gender, background, sound fidelity, channel bandwidth, commercials, and so forth.

4.2. Radio Broadcast Database. For the purpose of our analysis a radio broadcast database was built. The database contains radio broadcast material collected from several public radio stations. The idea was to collect more diverse material from as many music genres as possible. The material was collected by sampling an FM tuner connected to a desktop PC. Recordings were sampled at 16 KHz using a 16-bit resolution, single channel. The database contains both male and female speakers, with in studio and telephonic channel conditions. Background conditions sometimes vary (silence, noise, and often silent music, which is typical for a radio broadcast material). Many different performers (local and international) and music genres are represented in the database: pop, jazz, various types of rock, dance, hip-hop, rap, and so forth. Altogether 48 hours of sound material was compiled, where music represents 65%, speech 25%, and 10% other radio broadcast material (commercials, intros, etc.). Two subsets were defined within the structure of the database. The first subset is a train set and is assigned for training purposes. It is compiled from 2 hours of randomly chosen speech and music material from the database (60% music, 40% speech). The test set also includes 2 hours of randomly chosen speech and music material from the database. The material from the train and test sets are selected in such a way that they do not overlap with each other. Manual annotations were also created for the database. These annotations contain marks for speech segments, music segments, and other (commercials, intros, etc.). Transcriptions for the database were not created because the database is only used for speech/music segmentation evaluation.

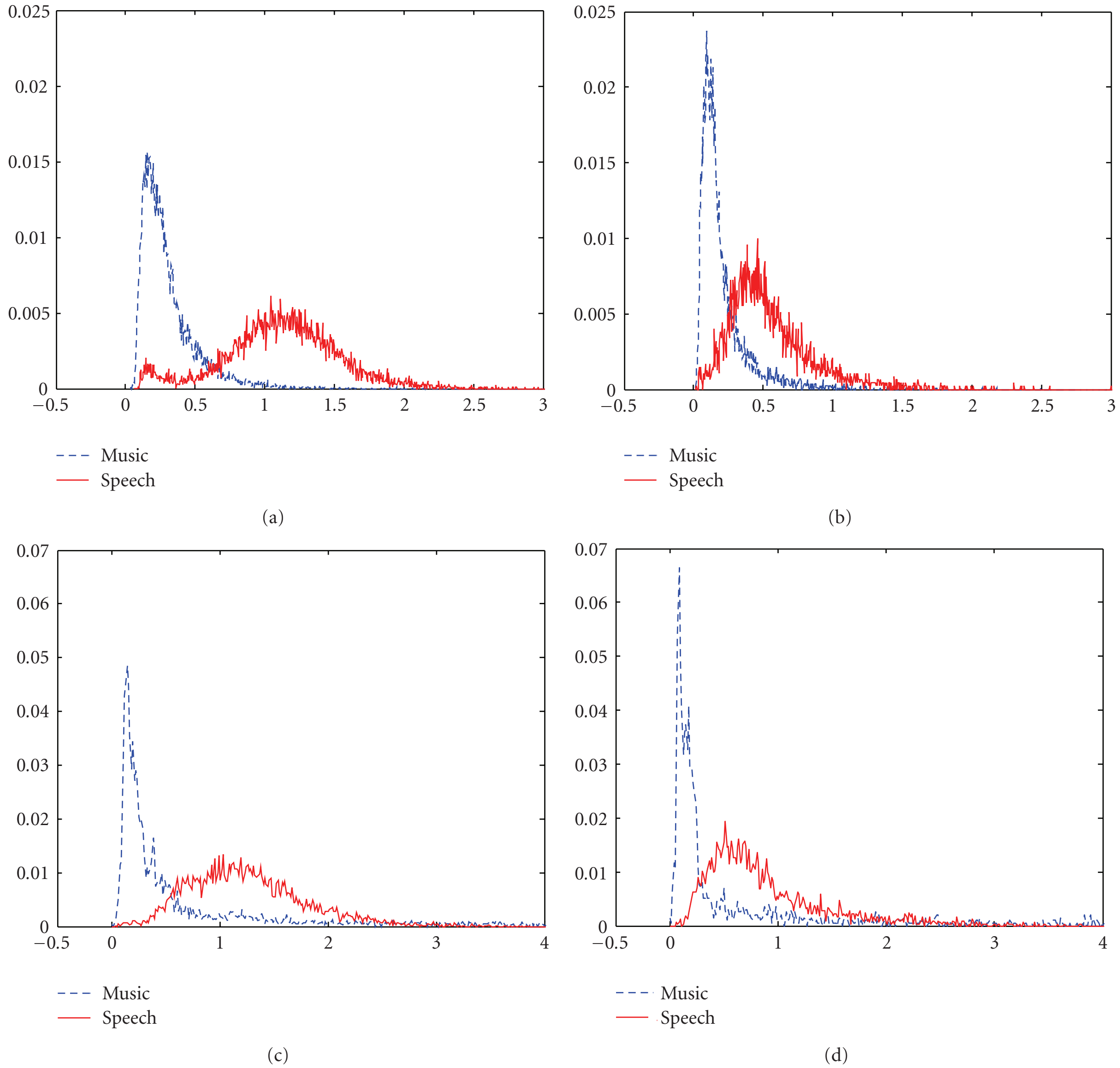
4.3. Discrimination Ability Estimation. This section analyses the ability of features to discriminate between speech and music audio class. This method for estimating classification error is based on estimating the probability density function (PDF) of each class using histograms. Analyses were carried out on both databases. An example of histograms for the VMFBE feature, and the variance of SF feature for both databases, is shown in Figure 4.

The ZCR feature was calculated over a 20 millisecond frame with a 10 millisecond frame shift. The signal was first normalized to obtain the correct result. PLEF was another feature calculated within the time-domain. Short time energy (STE) was calculated within a 20 millisecond frame with a 10 millisecond frame shift. The ratio of frames with STE lower than 50% of the average STE was calculated over a time period of 1 second. SR, SC, and SF are members of the frequency domain features. All three were calculated using 32 milliseconds long frames with a frame shift of 10 milliseconds. A Hamming window was used for windowing, and a DFT of the order 512 was calculated. The SR feature was calculated with a roll-off coefficient of 0.95. The VMFBE feature was calculated using the same front-end setup as for the SC, SR, and SF features. The order of DFT was also 512. The magnitude of the frequency spectrum was then filtered using 24 triangular filters, evenly distributed on the melodic scale.

Results for the speech/music discrimination abilities of all the presented features can be seen in Tables 2 and 3. Table 2 shows the results of tests conducted on the radio broadcast database, and Table 3 shows the results of tests conducted on the BNSI database. In both cases, the train sets of the databases were used.

In regard to features ZCR, SC, SR, and SF, a variance version (in tables marked as "Var. of") was also calculated and compared to other features, in addition to their basic version. The variance of a particular feature was calculated within a 1 second window.

As expected, from the list of standard speech/music discriminative features, the variance of SF performed the best. On the radio broadcast database (see Table 2) it outperformed the second rated standard feature (PLEF) by 4% absolute average. The proposed VMFBE feature proved to be the most effective. Results obtained on the radio broadcast database show that the VMFBE feature has more than 8% average better discriminative ability than



(a)

(b)

(c)

(d)

Figure 4: Histograms of (a) VMFBE feature on radio broadcast database, (b) Var. of SF feature on radio broadcast database, (c) VMFBE feature on BNSI database and (d) Var. of SF feature on BNSI database.

the second-rated feature (variance of SF). The same as in previous research work [14], the variance of SF proved to be a good discriminator between speech and music, and performed best from the list of standard speech/music discriminators used in this discriminative ability test. For this reason, we chose this feature to compare its performances to the VMFBE feature in the speech/music classification and segmentation task.

Results obtained on the BNSI database (see Table 3) show a minor advantage of the VMFBE feature over the variance of SF. Only 2.3% higher-average discrimination ability was achieved. The BNSI database has very low portion of music material and the material is mostly instrumental (jingles, intros, etc.). Therefore, results obtained on the radio broadcast database show more realistic performance abilities of features for speech/music discrimination.

5. Description of the Experimental Framework

The performance of the speech/music segmentation and classification was evaluated on the databases, as described in Sections 4.1 and 4.2.

5.1. Basic Structure of the Speech/Music Segmentation and Classification Framework. The block scheme of the defined speech/music segmentation and classification framework is shown in Figure 5. The input audio signal is sampled at 16 KHz with 16 bit resolution. The Hann window is used for windowing, with a length of 512 samples, which is equal to 32 milliseconds at a sampling frequency of 16 KHz. Window shift is 10 milliseconds (160 samples). Over each window, the 512-order discrete Fourier transformation (DFT) is applied, which is followed by feature calculation. When calculating

Table 2: Speech/Music discrimination ability of features. Experiments were performed on radio broadcast database.

Feature	Music	Speech	Average
		Error (in %)	
ZCR	30.92	32.50	31.71
Var. of ZCR	18.61	37.34	27.97
SC	25.71	28.06	26.88
Var. of SC	13.37	43.98	28.67
SR	18.72	35.11	26.91
Var. of SR	14.41	34.50	24.45
PLEF	29.87	13.50	21.68
SF	25.70	45.46	35.58
Var. of SF	18.05	16.26	17.15
VMFBE	7.25	10.61	8.93

Table 3: Speech/Music discrimination ability of features. Experiments were performed on BNSI database.

Feature	Music	Speech	Average
		Error (in %)	
ZCR	28.68	39.42	34.05
Var. of ZCR	14.24	26.90	20.57
SC	25.61	43.31	34.46
Var. of SC	09.18	26.15	17.66
SR	26.66	43.15	34.90
Var. of SR	13.49	21.36	17.42
PLEF	43.12	06.94	25.03
SF	30.37	56.09	43.23
Var. of SF	26.03	7.64	16.83
VMFBE	21.35	7.71	14.53

Table 4: Results for discrimination ability with a shorter variance calculation window (200 milliseconds).

Feature	Music	Speech	Average
	Error (in %) on radio broadcast database		
Var. of SF	24.82	26.67	25.74
VMFBE	13.82	21.58	17.70
	Error (in %) on BNSI database		
Var. of SF	24.86	20.53	22.69
VMFBE	15.87	17.63	16.75

the variance of SF, we calculate the variance over a period of 200 milliseconds with 100 millisecond overlap. If we were to calculate variance in a 1 second window, the time resolution would be too inexact for successful determination the beginning or the end of a segment. Variance in the VMFBE feature calculation procedure was also calculated over 200 milliseconds. We tested the basic discrimination ability again, because the length or the variance calculation window changed. The results are shown in Table 4.

As can be comprehended from Table 4, the discrimination abilities of both features dropped when using a shorter time window for variance calculations. Nevertheless, the VMFBE feature still outperforms the variance of SF by 7% on average.

After the feature calculation step, feature classification and segmentation procedures are performed. The result of segmentation and classification is written into the output segmentation text file (SEG). The output segmentation file is used later during the evaluation process.

5.2. Classification Step. The feature classification procedure is executed after the feature calculation step, using a GMM classifier. Classification using GMM uses a likelihood estimate for each model, which measures how well the calculated feature is modelled by the trained Gaussian clusters. A feature is assigned to whichever class is the best model of that feature. On the basis of the likelihood values, the frames are classified and in the segmentation step they are grouped into segments according to minimum segment duration rules.

The train sets of the databases were used for training the GMM models. We trained one model for each acoustic class (one for music and one for speech) using five Gaussian mixtures per class. The number of mixtures was defined empirically, by considering the achieved speech/music discrimination accuracy. The speech training material in both databases included male and female speakers under different environmental circumstances (studio recording, recording over a telephone line, etc.). Radio broadcast database speech material typically includes quite a large portion of speech with quiet music in the background. On the other hand, the music material of the databases used differs a great deal. The BNSI database has low portion of music material. This music material is mostly instrumental (jingles, intros, etc.), with no singing voice present. As mentioned earlier in this article, the radio broadcast database contains a wide variety of music. In the database different genres of music and different performers (local and international) can be found.

5.3. Segmentation Step. After the classification procedure, frames are grouped into segments according to the classification tag (whether a frame was classified as a speech frame or as a music frame). The classification result is smoothed out using mean filter, which filters out any glitches during the classification step. The segments are created according to the minimum speech and music segments' duration rules. An example of the segmentation procedure is shown in Figure 6. When the transition from speech to music (or vice versa) is detected (e.g., t1 mark in Figure 6), the algorithm marks the transition point as the potential beginning of the music segment and, at the same time, the measurement of segment duration begins. The algorithm then waits for the transition from music to speech (t2), and if the segment duration S1 is long enough, the segment is confirmed, otherwise the segment is refuted. In Figure 6, the segment S1 is refuted because the segment does not fulfill the minimum segment duration rule. The segment S2 is verified, because it fulfills the rule. The segment t2-t3 was not compared to the minimum speech size because music segment S1 (t1-t2) was not confirmed as a valid music segment.

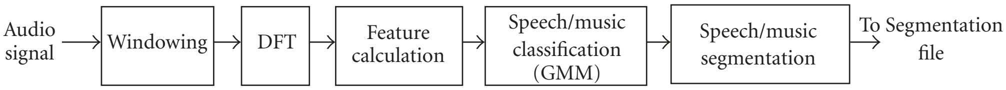


Figure 5: Block scheme of the proposed speech/music segmentation and classification framework.

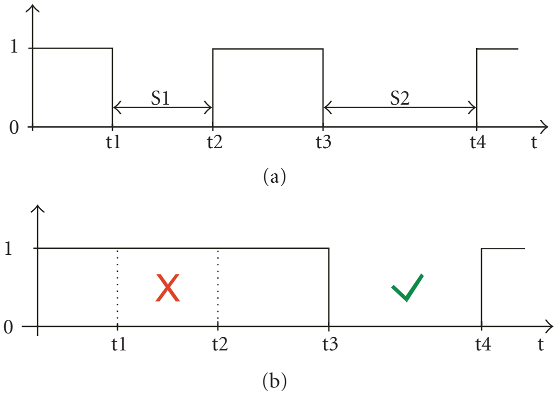


Figure 6: An example of segmentation procedure; (a) marking the potential segments, (b) refuting the unsuitable and confirming the suitable segments (speech = 1, music = 0).

The minimum speech segment duration rules for radio broadcast database are the same for both acoustic classes (minimum segment duration is set at 3 seconds). In the database there are almost no labelled segments shorter than that. Minimum segment duration rules are different for the BNSI database. Minimum nonspeech duration is set at 1500 milliseconds, because the transcription rules for the BNSI database instructs that 1500 milliseconds is the minimum nonspeech section duration. The minimum speech duration is set at 600 milliseconds. Many speech segments in the BNSI database begin with the greeting of the news anchor, followed by a short pause (nonspeech). The duration of this short speech segment is around 600–700 milliseconds and the duration of the short pause segment is around 300 milliseconds. By setting the minimum speech duration to the mentioned value we can, in such a case, successfully determine the beginning of the speech segment.

The whole framework works online. The delay of the system depends on the longest minimum segment duration, set by the segment duration rules (1.5 seconds for BNSI database and 3 seconds for radio broadcast database).

6. Evaluation and Results

6.1. Evaluation. We used the percentage of frame-level accuracy measure for the evaluation metric. We calculated three different frame-level accuracies: speech, music, and overall frame-level accuracy. Speech frame-level accuracy is defined as a percentage of the true speech frames classified as speech, the music frame-level accuracy is defined as a percentage of the true music frames classified as music, and the overall accuracy is defined as a percentage of correctly classified speech and music frames. The mathematical formula for overall accuracy is

$$\text{Overall acc.} = \frac{\text{FS} + \text{FM}}{\text{TS} + \text{TM}} \cdot 100\%, \qquad (12)$$

where FS and FM stand for found speech and music frames, and TS and TM stand for true speech and music frames. Commercials were discarded from the evaluation. The reason for discarding commercials is that they are labelled as homogenous segments and in order to use them in the evaluation procedure, they should be labelled in more detail (which part is speech and which is music). It should be noted that when one class dominates the other (as in BNSI, speech class dominates the nonspeech), overall accuracy mostly depends on the accuracy of that dominant class. In such a case, the overall accuracy itself does not provide enough information; therefore, all three accuracies need to be presented. Transcriptions in the BNSI database and the evaluation tool (ELIS-SEG; developed during the COST278 project campaign) do not explicitly support speech/music segmentation evaluation, but only speech/nonspeech. Because music in the BNSI database represents more than 70% of all non speech material (the rest is mostly silence), we used the speech/nonspeech evaluation procedure to evaluate our speech/music segmentation framework on the BNSI database. Regular speech/music segmentation evaluation was performed on the radio broadcast database.

Table 5: Accuracy results for speech/music segmentation, performed on BNSI and radio broadcast database.

Feature	Music	Speech	Overall
	Accuracy (in %; radio broadcast database)		
Var. of SF	90.49	91.12	90.70
VMFBE	94.24	92.78	94.05
	Accuracy (in %; BNSI database)		
Var. of SF	84.01	95.76	94.51
VMFBE	95.85	95.90	95.90

6.2. Results. Speech/music segmentation performance for both the variance of SF and VMFBE features was tested on test sets of the BNSI and radio broadcast databases. The results are shown in Table 5. Performance was measured as frame-level accuracy of speech, music, and overall.

As given in Table 5, the performance of the VMFBE feature compares favourably with the performance of the SF feature's variance, on both databases. The results obtained on the BNSI database show 1.4% better performance although the VMFBE feature outperformed the variance of SF by more

Table 6: Accuracy results (in %) for speech/music segmentation for multifeature discriminator.

Database	Music	Speech	Overall
Radio broadcast	96.26	93.37	95.91
BNSI	95.69	98.17	97.87

than 10%, regarding music accuracy. The reason why this difference contributes so little to the overall performance difference is in the fact that speech is the dominant class in the BNSI database, and music represents only 10% of the test material in the evaluation set of the BNSI database. It is typical that television broadcast news databases include a fairly small amount of music material (it is similar for other broadcast news databases collected within the COST278 project [38]).

The results obtained on the radio broadcast database show a bigger advantage over the VMFBE feature regarding the variance of SF in segmentation performance. In contrast to the BNSI database, music is the dominant class in this database and represents almost 65% of the database evaluation set. The VMFBE feature shows a 3.35% segmentation performance gain according to the variance of SF feature. The results obtained on this database are more representative, because this database contains more diverse music material than the BNSI database.

After reviewing the segmentational files of the VMFBE feature segmentation procedure, we noticed that errors mostly occurred regarding rap music material. This happens quite often for this type of music as it is closest to natural human speech, although it has a strong beat, but it does not have such a distinctive melody, like some other music genres. This characteristic makes rap music harder to discern from speech than some other genres.

We also tested the joint-discriminating ability of the presented features, by joining variance versions of all features and a PLEF feature, into a vector. In this way, we obtained a feature vector with 6 feature coefficients (VMFBE, var. of SF, var. of SC, var. of SR, PLEF, and var. of ZCR). We trained the GMM model on the same data as before. However, 30 Gaussian distributions were used for individual acoustic classes. Variance of features was, as in the previous example, calculated over a period of 200 milliseconds with 100 millisecond overlapping. The segmentation process and minimum segment duration rules were the same as before. The results obtained for multi-feature speech/music classification and segmentation framework, are shown in Table 6.

The results in Table 6 show almost 2% overall accuracy gain for the speech/music segmentation task, performed on the BNSI and radio broadcast databases, comparing to the results in Table 5. The results for radio broadcast database show a bigger accuracy gain for music (2.02%) than for speech (0.59%). For the BNSI database it is the opposite case. Speech shows a 2.27% accuracy gain, while music accuracy decreased by 0.16%. Because the segmentation accuracy of the dominant acoustic class improved quite noticeably, overall speech/music segmentation accuracy also essentially increased.

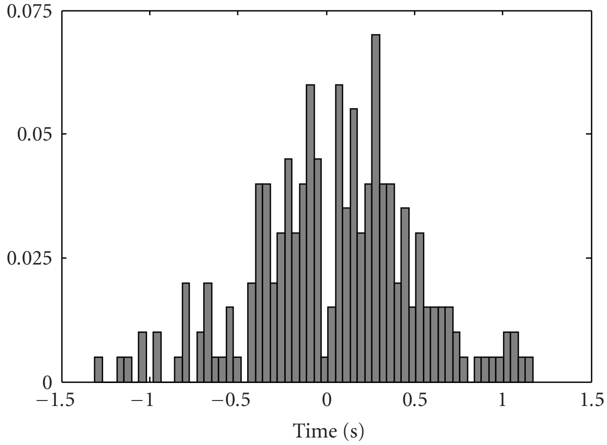


Figure 7: Histogram of speech/music segmentation time-error analysis.

Table 7: Segmentation performance of MFCC features, performed on BNSI and radio broadcast database.

Features	Music	Speech	Overall
	Accuracy (in %; radio broadcast database)		
MFCC 128 mix.	94.12	97.53	94.56
MFCC 256 mix.	94.25	97.95	94.73
	Accuracy (in %; BNSI database)		
MFCC 128 mix.	91.23	97.54	96.91
MFCC 256 mix.	91.72	97.76	97.15

In order to evaluate the time accuracy of the speech/music segmentation, we performed the time-error analysis. We measured the time differences between hand-labeled reference borders and automatically calculated borders. The results of time-error analysis are shown in Figure 7 in the form of a histogram. Results of the analysis show that the average segmentation time-error is +0.36 seconds. This means that, on average, the segment borders are set later than they actually occur.

For the purpose of comparison with the VMFBE feature, we tested the ability of MFCC features to discriminate between the speech and music acoustic classes. As in other experiments, these experiments were also performed on both databases. We used the same training material to train the GMM models, as used for other features. We used two different model complexities; one model with 128 Gaussian mixtures and the other with 256 Gaussian mixtures. We calculated 12 standard MFCC features extended by the log-energy. The feature vector, therefore, has 13 coefficients. The results for speech/music segmentation accuracy for MFCC features, are shown in Table 7.

The results from Table 7 show that there is only a slight difference between models with 128 mixtures and 256 mixtures, therefore, tests with higher model complexities are not needed. If we compare the overall results from Table 7 with the overall results from Table 5, we can see that MFCC

Table 8: Cross-test segmentation performance of VMFBE and MFCC features on BNSI and radio broadcast database.

Features	Music	Speech	Overall
	Accuracy (in %; on radio broadcast database with BNSI GMM models)		
MFCC 256 mix.	90.68	93.10	90.99
VMFBE	94.11	92.81	93.94
	Accuracy (in %; on BNSI with radio broadcast GMM)		
MFCC 256 mix.	71.86	94.33	92.23
VMFBE	96.37	94.95	94.89

features perform slightly better than VMFBE feature (0.7% on radio broadcast database and 1.2% on BNSI database), mainly because of the better performance for the speech acoustic class. The performances of MFCC and VMFBE features for the music acoustic class are very similar on the radio broadcast database. The VMFBE feature slightly outperforms the MFCC features over a 128 mixture experiment, while the performance with 256 mixture experiment is practically the same. On the BNSI database (regarding the music acoustic class), the VMFBE feature outperforms the MFCC features by more than 4%. Although MFCC features show small overall performance advantage (less than 1%, on average), the VMFBE feature shows solid performance when discriminating the music acoustic class. The VMFBE feature also has the advantage of having a faster feature extraction algorithm than MFCC features and, therefore, represents a smaller computation load for a system. The tests made on the BNSI evaluation set show that the VMFBE feature extraction algorithm is 19% faster than the MFCC feature extraction algorithm. For 3 hours, 4 minutes, and 58 seconds of material, the VMFBE feature extraction algorithm needed 85 minutes and 55 seconds to complete, and the MFCC feature extraction algorithm needed 104 minutes and 28 seconds. For classification the VMFBE feature classification procedure needed only 0.47 seconds and the MFCC feature classification procedure needed 4 minutes and 20 seconds. The main reason for such a big difference in time needed for classification is in the different model complexities. The VMFBE feature has only a 5 mixture model, whereas the MFCC features have a 256 mixtures model. In the overall (feature extraction and classification step), the VMFBE feature represents a 22% smaller computation load than the MFCC features. The time complexities of the VMFBE and MFCC features were measured on an Intel Core2duo 3.0 GHz with 4 GB of RAM within Linux environment.

To evaluate the robustness of the VMFBE feature, we cross-tested the speech/music segmentation performance with BNSI GMM models on the radio broadcast database, and vice versa. For the comparison we also cross-tested the segmentation performance of the MFCC features. The results are shown in Table 8.

As the results in Table 8 show, the VMBFE feature proves to be more robust, because it achieves higher cross-test segmentation accuracy (3.05% higher on the radio broadcast database and 2.66% higher on the BNSI database), than MFCC features, and also the drop in segmentation performance is smaller. Comparing with the results from Table 7, the MFCC features achieved 3.7% lower accuracy on the radio broadcast database, and 4.9% on the BNSI database. Comparing the results from Table 5, the VMFBE feature achieved only 0.11% lower segmentation accuracy on the radio broadcast database, and 1.01% on the BNSI database.

It is always a difficult task to compare the proposed methods against methods presented in the past by other authors. The main reason for this are the diverse datasets, used by different authors, which are often not available to others. There are also differences, if the proposed systems work online or offline. Thus, it is only reasonable to compare the proposed method to methods evaluated on similar databases (same kind of material, e.g., broadcast news), and similar circumstances. The results of our system obtained on the radio broadcast database can be compared to the system performance in [14]. The authors in that paper used 40 minutes of recorded radio broadcast material with a wide-variety of music genres and different speakers. Also, 36 minutes of material was used for classifier training and 4 minutes for testing (we used 2 hours of material for training and analysis and 2 hours for testing). The same as in our case, in [14] the variance of SF proved to be the best standard feature for speech/music discrimination. When testing the features' discriminatory ability on their database, the variance of the SF feature produced a 13% error-rate. The same feature produced a 17.15% error rate on our database, while the feature proposed in this paper (VMFBE) produced an 8.93% error-rate. The results for multifeature speech/music segmentation cannot be directly compared, because different sets of features were used in individual systems.

In adition, the authors in [39] trained and evaluated their system on a database containing radio broadcast material. Their database was composed of various male and female speakers, and various music genres. The amount of speech material was 9.3 minutes and the amount of music was 10.7 minutes. They used four features for speech/music discrimination: line spectral frequencies (LSFs), differential line spectral frequencies (DLSFs), line spectral frequencies with higher order crossings (LSF-HOCs), and line spectral frequencies with linear prediction zero-crossing ratio (LSF-ZCR). The authors implemented segment-level classification by making decisions over 50 frames (1 second). An accuracy of 95.9% was reported. The authors also tested the performance of a speech/music segmentation system proposed in [14] on their database, and 93.2% accuracy was achieved. If we indirectly compare the accuracy performance of our multifeature method with those results, performance is at a similar level. Therefore, we can say that our database has a similar structure as the databases used in [14, 39], and the results obtained on our database show the true advantage of the VMFBE feature over other features used and tested in this article. We also used more training and testing material (2 hours each set), than the authors in [14, 39] (40 minutes and 20 minutes resp.).

7. Conclusions

This paper presents a novel feature (VMFBE) for speech/music discrimination. Discrimination ability analyses and comparative tests were performed on the BNSI broadcast news database, and the radio broadcast database. This feature was compared to several other standard speech/music discrimination features (zero-cross rate (ZCR), spectral centroid (SC), spectral roll-off (SR), spectral flux (SF), and percentage of low energy frames (PLEF)). Variance versions of the features were also calculated and compared. The results show more than 8% better average discrimination ability in a 1 second window than the second rated feature (variance of SF). On the radio broadcast material, 3.3% accuracy gain is achieved for speech/music segmentation with the VMFBE feature over the variance of SF. As a standalone discriminator in a speech/music segmentation system, the VMFBE feature achieves an overall accuracy of over 94%. On the BNSI database, where music material is not as diverse, it achieves almost 96% overall accuracy. The experiment based on all 6 features, presented in this article, was also performed, and an overall accuracy of 95.91% was achieved on the radio broadcast material. Although MFCC features perform slightly better regarding overall segmentation results, the VMFBE feature shows better discrimination abilities for the music acoustic class and also 22% lower computation cost, than the MFCC features. The cross-test results also show that the VMFBE feature proves to be more robust to new conditions and unseen data than the MFCC features.

The VMFBE feature is a very convenient speech/music discriminator for automatic speech recognition systems, where MFCC features are used for recognition. When calculating the VMFBE feature, 4 out of 6 steps are in common with the MFCC features calculation procedure. Valuable processing time can be saved and almost no additional computational cost is needed to implement the speech/music discrimination procedure into a real-time ASR system based on MFCC features.

References

[1] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 432–435, Hong Kong, March 2003.

[2] H. K. Maganti, P. Motlicek, and D. Gatica-Perez, "Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 1037–1040, Honolulu, Hawaii, USA, April 2007.

[3] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, and S. J. Young, "Experiments in broadcast news transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 2, pp. 909–912, Seattle, Wash, USA, May 1998.

[4] L. Zhu and H. Qian, "Content-based indexing and retrieval-by-example in audio," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '00)*, vol. 2, pp. 877–880, New York, NY, USA, July 2000.

[5] J. Razik, C. Sénac, D. Fohr, O. Mella, and N. Parlangeau-Vallès, "Comparison of two speech/music segmentation systems for audio indexing on the Web," in *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (WMSCI '03)*, Orlando, Fla, USA, July 2003.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[7] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 5, pp. 953–956, Philadelphia, Pa, USA, March 2005.

[8] J. M. Pardo and X. Anguera, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[9] X. Anguera Miro, *Robust speaker diarization for meetings*, Ph.D. thesis, Universitat Politécnica de Catalunya, Barcelona, Spain, 2006.

[10] T. Zang and C. C. Jay Kuo, *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*, Kluwer Academic Publishers, Dordrecht, The Nederlands, 2001.

[11] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

[12] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proceedings of the IEEE Transactions on Audio and Speech Processing (ICASP '99)*, vol. 1, pp. 1432–1435, Phoenix, Ariz, USA, March 1999.

[13] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 2, pp. 993–996, Atlanta, Ga, USA, May 1996.

[14] E. Scheirer and 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 (ICASSP '97)*, vol. 2, pp. 1331–1334, April 1997.

[15] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[16] A. Bugatti, A. Flammini, and P. Migliorati, "Audio classification in speech and music: a comparison between a statistical and a neural approach," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 4, pp. 372–378, 2002.

[17] M. K. S. Khan and W. G. Al-Khatib, "Machine-learning based classification of speech and music," *Multimedia Systems*, vol. 12, no. 1, pp. 55–67, 2006.

[18] T. Taniguchi, M. Tohyama, and K. Shirai, "Detection of speech and music based on spectral tracking," *Speech Communication*, vol. 50, no. 7, pp. 547–563, 2008.

[19] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 73205, 11 pages, 2007.

[20] S. O. Sadjadi, S. M. Ahadi, and O. Hazrati, "Unsupervised speech/music classification using one-class support vector machines," in *Proceedings of the 6th International Conference on Information, Communications and Signal Processing (ICICS '07)*, pp. 1–5, Singapore, December 2007.

[21] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, 2005.

[22] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A computationally efficient speech/music discriminator for radio recordings," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR '06)*, Victoria, Canada, October 2006.

[23] J. G. A. Barbedo and A. Lopes, "A robust and computationally efficient speech/music discriminator," *Journal of the Audio Engineering Society*, vol. 53, no. 7-8, pp. 571–588, 2006.

[24] Z. H. Fu and J. F. Wang, "Robust features for effective speech and music discrimination," in *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing (ROCLING '08)*, Taipei, Taiwan, September 2008.

[25] S. Karneback, "Discrimination between speech and music based on a low frequency modulation feature," in *Proceedings of the 7th European Conference on Speech Communication and Technology (ISCA EUROSPEECH '01)*, pp. 1891–1894, Aalborg, Denmark, September 2001.

[26] J. E. M Exposito, S. G. Galan, N. R. Reyes, and P. V. Candeas, "Audio coding improvement using evolutionary speech/music discrimination," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 1–6, London, UK, July 2007.

[27] W. Q. Wang, W. Gao, and D. W. Ying, "A fast and robust speech/music discrimination approach," in *Proceedings of the 10th IEEE International Conference on Communication Systems (ICCS '06)*, vol. 3, pp. 1325–1329, Singapore, December 2003.

[28] B. Kotnik, D. Vlaj, and B. Horvat, "Efficient noise robust feature extraction algorithms for distributed speech recognition (DSR) systems," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 205–219, 2003.

[29] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, 1986.

[30] J. Bakus, *The Acoustical Foundations of Music*, W. W. Norton & Company, Pennsylvania, Pa, USA, 2nd edition, 1997.

[31] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[32] H. Rongqing and J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 907–919, 2006.

[33] N. Jhanwar and A. K. Raina, "Pitch correlogram clustering for fast speaker identification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 17, pp. 2640–2649, 2004.

[34] A. A. Razak, M. I. Z. Abidin, and R. Komiya, "Emotion pitch variation analysis in Malay and English voice samples," in *Proceedings of the Asia Pacific Conference on Communication (APCC '03)*, vol. 1, pp. 108–112, Penang, Malaysia, September 2003.

[35] A. I. Al-Shosan, "Speech and music classification and separation: a review," *Journal of King Saud University; Engineering Sciences*, vol. 19, no. 1, pp. 95–133, 2006.

[36] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '01)*, vol. 1, pp. 73–76, May 2001.

[37] A. Žgank, D. Verdonik, A. M. Zögling, and Z. Kačič, "BNSI Slovenian Broadcast News database—speech and text corpus," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1537–1540, Lisbon, Portugal, September 2005.

[38] J. Žibert, H. Meinedo, J. Neto, et al., "The COST278 broadcast news segmentation and speaker clustering evaluation—overview, methodology, systems, results," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 629–932, Lisbon, Portugal, September 2005.

[39] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 6, pp. 2445–2448, Istanbul, Turkey, June 2000.