International Scholarly Research Network
ISRN Signal Processing
Volume 2011, Article ID 269361, 10 pages
doi:10.5402/2011/269361

Research Article

Discrete and Dual Tree Wavelet Features for Real-Time Speech/Music Discrimination

Timur Düzenli[1,2] and Nalan Özkurt[1,3]

[1] *Graduate School of Natural and Applied Sciences, Dokuz Eylul University, 35160 Buca, İzmir, Turkey*
[2] *Department of Electronics and Telecommunications Engineering, Izmir University of Economics, 35330 Balçova, İzmir, Turkey*
[3] *Department of Electrical and Electronics Engineering, Yaşar University, 35100 Bornova, İzmir, Turkey*

Correspondence should be addressed to Nalan Özkurt, nalan.ozkurt@deu.edu.tr

Received 5 January 2011; Accepted 1 March 2011

Academic Editor: P. C. Yuen

The performance of wavelet transform-based features for the speech/music discrimination task has been investigated. In order to extract wavelet domain features, discrete and complex orthogonal wavelet transforms have been used. The performance of the proposed feature set has been compared with a feature set constructed from the most common time, frequency and cepstral domain features such as number of zero crossings, spectral centroid, spectral flux, and Mel cepstral coefficients. The artificial neural networks have been used as classification tool. The principal component analysis has been applied to eliminate the correlated features before the classification stage. For discrete wavelet transform, considering the number of vanishing moments and orthogonality, the best performance is obtained with Daubechies8 wavelet among the other members of the Daubechies family. The dual tree wavelet transform has also demonstrated a successful performance both in terms of accuracy and time consumption. Finally, a real-time discrimination system has been implemented using the Daubhecies8 wavelet which has the best accuracy.

1. Introduction

The discrimination of music and speech has been an important task in multimedia signal processing with the increasing role of the multimedia sources in our life. The speech/music discrimination (SMD) systems can be used in the development of the efficient coding algorithms for audio decoders [1]. Thus, if the speech can be separated, it can be coded with a speech coder which needs less bandwidth relative to the audio coder. Also, speech/music discriminator is important in automatic speech recognition when the recordings include music such as radio broadcasts [2]. The content-based multimedia retrieval [3] and automatic channel selector design for radios are other emerging applications with a growing interest for speech/music discriminators. There are several features which are used to classify music and speech such as number of zero crossings [4], spectral centroid, and Mel frequency cepstral coefficients [5, 6]. The entropy and dynamism features have been used for hidden Markov model classification in [7]. In [8], a system which looks for the transition between music and speech using two characteristics of signals, which are RMS-based average density of zero crossings and average frequency, has been proposed. In [9], the harmonic features are used to discriminate speech and music using a hierarchical oblique decision tree. An SMD system designed for radio broadcasts that has been proposed in [10] uses a three-stage structure, which determines the speech and music segments that are separable at first glance with spectral entropy and region growing. Recently, a segmentation method has been proposed in [11] which uses the property that the mean and the variances of the filter bank changes more rapidly and reaches higher value for speech than music.

Since it is important to deal with nonstationary signals and to achieve variable time and frequency localization of acoustic data, the wavelet-based parameters became an important tool in speech/music discrimination [12–15]. However, there are some important choices to make for reconstruction of the features from the wavelet coefficients. In some studies, only one level of wavelet decomposition is

employed, and the wavelet transform is applied to 20 or 30 ms windows. To capture the nonstationary information from a time series, a longer window should be used. The windows in order of seconds are used in [12]; however, this is too long for a real-time application. Also, decomposing signal into several frequency bands gives us more discriminative features. In another recent work, Didiot et al. employed energy-based features extracted from the 5 and 7 bands of discrete wavelet components [15]. For audio analysis including SMD and genre classification, Tzanetakis and Cook employed mean and variances of 12 bands of wavelet coefficients and the ratio of the mean values for adjacent bands using Daubechies 4 wavelets in 3 sec windows [12]. The accuracy of the method is approximately 90%. Along with several advantages, the discrete wavelet transform has some disadvantages such as lack of time invariance and oscillatory behavior. In order to cope with these problems, complex wavelet transform has been proposed [16]. The dual tree complex wavelet transform, which is a specific case of complex wavelet transform, has been introduced in [17].

Therefore, the aim of this study is both to consider the feature extraction for SMD with DWT to further improve performance for real-time applications and to examine the features obtained from the statistical parameters of the dual tree wavelet coefficients. In order to compare the performance of the proposed feature sets, the previously used time, frequency, and discrete wavelet features have been used to classify the same data set. The database has been constructed from the speech samples taken from TIMIT database and the music samples which are recorded from the audio CDs and radios with different genres such as classical, jazz, and pop. The artificial neural networks (ANNs) have been used for the comparison of the effectiveness of the feature extraction algorithms.

The paper is organized as follows: a brief introduction of the previous and proposed feature extraction methods are explained in Section 2; the introduction of the data set and classification algorithm with the results are given in Section 3; the conclusions and future works are discussed in Section 4.

2. Features for Speech/Music Discrimination

In this section, the related theoretical background on the features used for speech/music discrimination systems will be given briefly.

2.1. Common Features. The time-domain features such as number of zero crossings and frequency-domain features such as low energy ratio, spectral centroid, spectral roll-off and spectral flux are commonly used for speech/music discrimination. Also, Mel frequency cepstrum coefficients are shown to be successful in speech/music classification and recognition applications. For comparison, a feature vector constructed from these features have been used for classification in the first method of this study.

2.1.1. Number of Zero Crossings. It is the time-domain feature which represents the number of zero crossing in a frame. It is a useful feature in music and speech discrimination, since it is a measure of the dominant frequency in the signal [5, 6]. The number of zero crossings are calculated as

$$Z_t = 0.5 * \sum_{n=1}^{N} \left| \operatorname{sgn}(x(n)) - \operatorname{sgn}(x(n-1)) \right|, \quad (1)$$

where $x(n)$ is the nth component of the frame of length N.

2.1.2. Low Energy Ratio. This feature gives the number of the frames of which the effective or root mean square (RMS) energy is less than the average energy. The RMS energy for each frame is determined as

$$X_{\mathrm{RMS}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} X_k^2}, \quad (2)$$

where X_k is the magnitude of kth frequency component in the frame. Since the energy distribution is more left-skewed than for music, this measure will be higher for speech [6].

2.1.3. Spectral Centroid. This is the measure of the center of mass of the frequency spectrum and calculated as

$$\mathrm{SC} = \frac{\sum_{k=1}^{K} f_k X_k}{\sum_{k=1}^{K} X_k}, \quad (3)$$

where X_k is the magnitude of the component in the frequency band f_k [5, 6].

2.1.4. Spectral Roll-off. This feature is important in determining the shape of the frequency spectrum. The spectral roll-off point R_k is the frequency where the 95% of the spectral power lies below as summarized in

$$\sum_{k=1}^{R_k} X_k^2 = 0.95 \sum_{k=1}^{K} X_k^2, \quad (4)$$

where X_k is the magnitude of the component of the kth frequency. Since the most of the energy is in the lower frequencies for speech signals, R_k has lower values for speech [5, 6].

2.1.5. Spectral Flux. It represents the spectral changes between adjacent frames and calculated as

$$\mathrm{SF}_t = \sum_{k=1}^{K} \left(X_k^t - X_k^{t-1} \right)^2, \quad (5)$$

where X_k^t is the kth frequency component of the tth frame. Then the average of the all frames are calculated. The music has a higher rate of changes than the speech does, thus this value is higher for music [5, 6].

2.1.6. Mel Frequency Cepstrum Coefficients (MFCCs). The Mel frequency spectrum is the linear cosine transform of

a log power spectrum on a nonlinear Mel scale of frequency [18]. The Mel scale is inspired from the human auditory system in which the frequency bands are not linearly spaced. Thus, the sound is represented better. The calculation of the MFCCs includes the following steps.

(i) The discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain, and the short-term power spectrum $P(f)$ is obtained.

(ii) The spectrum $P(f)$ is warped along its frequency axis f (in hertz) into the mel-frequency axis as $P(M)$, where M is the Mel frequency using

$$M(f) = 2595 * \log\left(1 + \frac{f}{700}\right). \tag{6}$$

(iii) The resulted warped power spectrum is then convolved with the triangular band-pass filter $P(M)$ into $\theta(M)$. The convolution with the relatively broad critical-band masking curves $\theta(M)$ significantly reduces the spectral resolution of $\theta(M)$ in comparison with the original $P(f)$, which allows for the downsampling of $\theta(M)$.

$$\theta(M_k) = \sum_{M} P(M - M_k)\psi(M), \quad k = 1, \ldots, K. \tag{7}$$

Then, K outputs $X(k) = \ln(\theta(M_k)), k = (1, \ldots, K)$ are obtained. In the implementation, $\theta(M_k)$ is the average instead of the sum.

(iv) The MFCCs are computed as

$$\mathrm{MFCC}(d) = \sum_{k=1}^{K} X_k \cos\left[d\left((k - 0.5)\frac{\pi}{K}\right)\right], \quad k = 1, \ldots, D. \tag{8}$$

2.2. Discrete Wavelet Transform. The multiresolution analysis (MRA) provides a time-frequency representation well suited for nonstationary signals. MRA decomposes and analyzes the signal at different frequencies and different resolutions in the space spanned by wavelets and scaling functions. The continuous wavelet transform of a signal $x(t)$ is the projection of the signal on this space as

$$WT(s, r) = \frac{1}{\sqrt{s}} \int x(t)\psi^*\left(\frac{t - r}{s}\right), \tag{9}$$

where $\psi(t)$ is the mother wavelet s and r are the scale and translation coefficients, respectively [19]. For computational issues, discrete wavelet transform (DWT) is obtained. There is a rich set of basis functions for DWT and it is possible to get a compact representation of signal using this transform. In practical applications, DWT is applied in sampled signal $x[n]$; $n = 1, \ldots, N$ as

$$\mathrm{DWT}\left[n, 2^j\right] = \sum_{m=0}^{N-1} x[m]\psi_{2^j}^*[m - n], \tag{10}$$

where $\psi_{2^j}[n] = (1/\sqrt{2^j})\psi(n/2^j)$.

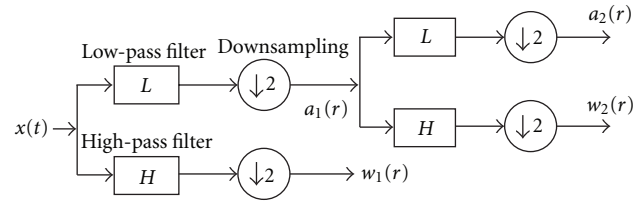

Figure 1: DWT with two composition levels [15].

DWT is implemented using a pyramidal algorithm related to a multirate filter-bank approach for multiresolution analysis. Mallat has shown that it is possible to make frequency band decomposition by using successive low-pass (L) and high-pass (H) filters in the time domain as shown in Figure 1 [15].

After each filtering stage i, the outputs are downsampled by 2, and the outputs are $a_i(r)$ approximation and $w_i(r)$ detail or wavelet coefficients for low- and high-pass filters, respectively. Approximation coefficients give local averages of the signal. On the other hand, detail coefficients show the differences between local averages. In this work, Haar wavelet and Daubechies family, which is known as one of the best families for speech processing applications, is used as the mother wavelet [9, 10].

2.3. Discrete Wavelet Transform Based Energy Features. In study of Didiot et al. [15], the energy-based features which are calculated using wavelet transform have been proposed. According to study, the energy distribution in each frequency band is a very relevant acoustic cue and energy, calculated from DWT, can be used as a speech/music discrimination feature. In our study, these energy based parameters have also been used in order to make comparison among different feature extraction methods.

2.3.1. Instantaneous Energy. This is a feature which gives the energy distribution in each band and given as

$$f_j^E = \log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}\left(w_j(r)\right)^2\right), \tag{11}$$

where $w_j(r)$ is the wavelet coefficient at time position r and frequency band j and N is the length of the analysis window.

2.3.2. Teager Energy. Teager energy has been recently applied for speech recognition and given as

$$f_j^{T_E} = \log_{10}\left(\left|\frac{1}{N_j}\sum_{r=1}^{N_j-1}\left(w_j(r)\right)^2 - \left(w_j(r-1) * w_j(r+1)\right)\right|\right). \tag{12}$$

It is said that the discrete teager energy operator (TEO) allows modulation energy tracking and gives a better representation of the formant information in the feature vector compared to MFCCs in [15]. It is also pointed out that the

Teager energy is a noise robust parameter for speech recognition, because the effect of additive noise is attenuated.

2.4. Complex Wavelet Transform. Conventional discrete wavelet transform suffers from some fundamental shortcomings despite its compact representation and efficient computational algorithm. The first shortcoming of DWT is that a small shift of the signal in time domain yields distortion in the wavelet coefficient oscillation pattern around singularities. The second problem of DWT is the lack of directionality for two and higher dimension signals. Complex wavelet transform (CWT) has been proposed inspiring by the Fourier transform which does not suffer from these types of problems. CWT is defined with a complex-valued scaling function and complex-valued wavelet

$$\psi_c(t) = \psi_r(t) + i\psi_i(t), \tag{13}$$

where $\psi_r(t)$ and $\psi_i(t)$ are real and imaginary parts. If these functions are 90° out of phase with each other, that is they form a Hilbert transform pair, then $\psi_c(t)$ is an analytic signal and it has a one-sided spectrum. Projecting the signal onto $2^j\psi_c(t)(2^jt - n)$, the complex wavelet coefficients are obtained as

$$d_c(j,n) = d_r(j,n) + jd_i(j,n). \tag{14}$$

Complex Wavelet Transform can be performed by two schemes. In the first one, a complex wavelet $\psi_c(t)$ that forms an orthonormal or biorthogonal basis is searched. The second method seeks a redundant representation, and it searches $\psi_r(t)$ and $\psi_i(t)$ that provide orthonormal and biorthogonal bases individually. The resulting CWT has $2x$ redundancy in 1-D and has power to overcome the shortcomings of DWT. In this study, the dual-tree approach for performing complex wavelet transform which is a natural approach to second, redundant type has been preferred.

2.4.1. Dual-Tree Complex Wavelet Transform (DT-CWT). Dual-tree complex wavelet transform was first introduced by Kingsbury in 1998 [20]. The dual tree implements an analytic wavelet transform by using two real discrete wavelet transform with two filter-bank trees; the first DWT gives the real, and the second one gives the imaginary part of the CWT. Analysis and synthesis filter-banks can be illustrated as in Figure 2, where $h_0(n)$ and $h_1(n)$ denote the low-pass/high-pass filter pair for the upper filter-bank which implements WT for real part. In the same way, $g_0(n)$ and $g_1(n)$ denote the low-pass/high-pass filter pair for the lower filterbank for imaginary part. In this approach, the key challenge is joint design of two filterbanks to get complex wavelet and scaling function as close as possible to analytic [20].

The filters used for real and imaginary parts of the transform must satisfy the perfect reconstruction condition given as

$$\begin{aligned} &\sum_n h_0(n)h_0(n+2k) = \delta(k), \\ &h_1(n) = (-1)^n h_0(M-n). \end{aligned} \tag{15}$$

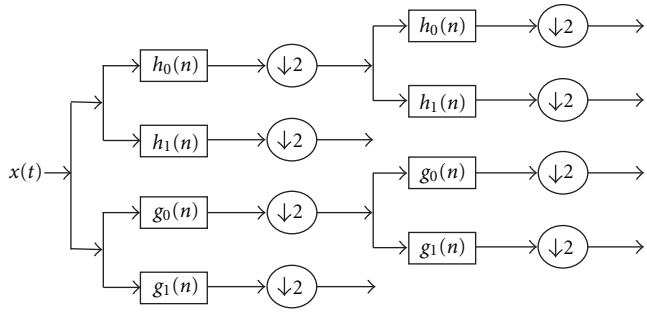

Figure 2: Analysis filter bank for the dual tree CWT [21].

Two low-pass filters of dual tree $h_0(n)$ and $g_0(n)$ satisfying a very simple property make corresponding wavelets to form an approximate Hilbert Transform pair: one of them must be approximately a half-sample shift of the other [21]

$$g_0(n) = h_0(n - 0.5) \Longrightarrow \psi_g(t) = H\{\psi_h(t)\}. \tag{16}$$

Since $h_0(n)$ and $g_0(n)$ are defined only on integers, it will be useful to rewrite the half-sample delay condition in terms of magnitude and phase functions separately in frequency domain to make the statement rigorous

$$\begin{aligned} \left|G_0\left(e^{jw}\right)\right| &= \left|H_0\left(e^{jw}\right)\right| \\ \angle G_0\left(e^{jw}\right) &= \angle H_0\left(e^{jw} - 0.5w\right). \end{aligned} \tag{17}$$

There are two popular methods for design of filters for DT-CWT [20].

Q-Shift Solution. According to q-shift solution, $g_0(n)$ must be selected as

$$g_0(n) = h_0(N - 1 - n), \tag{18}$$

where N is the length of filter $h_0(n)$ and is even. In this case, the magnitude condition is satisfied but not the phase condition as shown in (19) in the frequency domain

$$\begin{aligned} \left|G_0\left(e^{jw}\right)\right| &= \left|H_0\left(e^{jw}\right)\right|, \\ \angle G_0\left(e^{jw}\right) &\neq \angle H_0\left(e^{jw} - 0.5w\right), \end{aligned} \tag{19}$$

The quarter-shift (q-shift) solution has an interesting property that causes to take its name: when you ask that $g_0(n)$ and $h_0(n)$ be related as $g_0(n) = h_0(N - 1 - n)$ and also that they approximately satisfy $\angle G_0(e^{jw}) = \angle H_0(e^{jw} - 0.5w)$, then it turns out that the frequency response of $h_0(n)$ has approximately linear phase. This is verified by writing $g_0(n) = h_0(N - 1 - n)$ in terms of Fourier transforms

$$G_0\left(e^{jw}\right) = H_0^*\left(e^{jw}\right)e^{-j(N-1)w}, \tag{20}$$

where the $*$ represents complex conjugation. This implies that the phases satisfy

$$\angle G_0\left(e^{jw}\right) = -\angle H_0\left(e^{jw}\right) - (N-1)w. \tag{21}$$

If the two filters satisfy the phase condition approximately, it can be written that

$$\angle H_0\left(e^{jw}\right) - 0.5w = -\angle H_0\left(e^{jw}\right) - (N-1)w. \tag{22}$$

and we have the equation

$$\angle H_0\left(e^{jw}\right) \approx -0.5(N-1)w + 0.25w. \tag{23}$$

As it can be seen, $h_0(n)$ is an approximately linear-phase filter. This means that $h_0(n)$ is approximately symmetric around the point $n = 0.5(N-1) - 0.25$. This is one quarter away from the natural point of symmetry, and solutions of this kind were introduced as q-shift dual-tree filters for this reason [20].

2.4.2. Common Factor Solution (CFS). Another method for filter design stage named as *common factor solution (CFS)* can be used to design both orthonormal and biorthogonal solutions for the dual tree CWT [20].

In this approach, the filters, h_0 and g_0 are set as

$$\begin{aligned} h_0(n) &= f(n) * d(n), \\ g_0(n) &= f(n) * d(L-n), \end{aligned} \tag{24}$$

where $d(n)$ is supported on $0 \le n \le L$ and $*$ represents the discrete time convolution. In terms of Z-transform, we have

$$\begin{aligned} H_0(z) &= F(z)D(z), \\ G_0(z) &= F(z)z^{-L}D\left(\frac{1}{z}\right). \end{aligned} \tag{25}$$

For this solution, the magnitude part of half-sample delay condition is satisfied; however, the phase part is not exactly satisfied as in q-shift solution [20]

$$\begin{aligned} \left|G_0\left(e^{jw}\right)\right| &= \left|H_0\left(e^{jw}\right)\right|, \\ \angle G_0\left(e^{jw}\right) &\neq \angle H_0\left(e^{jw}\right) - 0.5w. \end{aligned} \tag{26}$$

So, the filters must be designed so that the phase condition is approximately satisfied.

3. Experimental Study and Results

3.1. Dataset and Preprocessing. The two different data sets have been utilized, and the features have been extracted separately for these two different datasets. In the first dataset, TIMIT database has been used for speech and several CD recordings with various musical genres have been used for music database. To obtain the second dataset, radio broadcasts were recorded containing music and speech. The sampling frequency was set as 44100 Hz in every stage of study. However, since the data taken from TIMIT database is sampled with 16000 Hz, they have been interpolated in the preprocessing stage in order to set sampling frequency to 44100 Hz. For the common features, the segmentation has been performed for a frame of 4196 samples with 512 samples overlapping which corresponds to a frame length of 95 ms. Our experiments showed that the use of shorter window lengths limits the discriminative characteristics of window.

Table 1: Content of datasets.

	Overall database		Train set		Test set	
	Speech	Music	Speech	Music	Speech	Music
Dataset 1	4620	4290	3080	2860	1540	1430
Dataset 2	2624	2190	1749	1460	875	730

Both datasets used contain samples with length of 0.5 sec. While the first dataset includes 4290 music and 4620 speech samples, the second dataset contains 2190 music and 2624 speech samples entirely derived from radio broadcasts in contrary to the first dataset. In the rest of the context, the first and second data sets will be named as Dataset 1 and Dataset 2, respectively. For the performance evaluation, the data sets have been divided into two groups as training and test sets. A detailed representation for dataset 1 and dataset 2 is given in Table 1.

The four different feature extraction methods have been employed in this study. The first method has a parameter vector which contains time-frequency-based features and Mel cepstrum coefficients with length of 21. The second and third methods use DWT-based features. The second method contains DWT-based energy parameters Teager and instantaneous energy as described in Section 2.3. The length of feature vector for third method is 10.

The novel methods proposed in this study are the last two feature extraction schemes. In the third method, 12-level DWT decomposition has been performed to cover the analysis frequency range in detail using several types of mother wavelets of Daubhecies family. The length of feature vector constructed from the statistical measures of the coefficients, and ratios between the adjacent subbands is 38. Usage of ratio parameters and the selection of the size of the analysis window are the important aspect which improves the performance of the classification in this study.

The last method is based on complex wavelet transform (CWT), and two different filter design strategies including common factor solution and Q_shift solution have been used at feature extraction stage. CWT has been performed for 5-level and 7-level to avoid further increase in the length of the feature vector which results in feature vectors with length of 25 and 35 for 5 and 7 bands, respectively.

Before classification stage, the features that are highly correlated with the other features have been eliminated using principal component analysis (PCA) to reduce the length of feature vectors. The principal components that contribute less than 0.05% to the total variation in the data set have been eliminated. Table 2 shows the length of the feature vectors before and after PCA. It has been observed from Table 2 that the DWT-based energy features include discriminative information. Similarly, approximately half of preliminary feature vectors are correlated with each other especially for CWT-based features.

Table 2: Lengths of the feature vectors before and after PCA.

		Original length	After PCA
Conventional	Common features	21	20
DWT based Energy	DWT Energy (db8)	10	5
	DWT Energy (coif1)	10	3
DWT based	Haar	38	19
	db2	38	19
	db8	38	22
	db15	38	21
	db20	38	21
CWT based	CFS (5 Levels)	25	11
	Q_Shift (5 Levels)	25	11
	CFS (7 Levels)	35	15
	Q_Shift (7 Levels)	35	14

Table 3: General classification results.

	Performance (%)	Dataset 1	Dataset 2
Conventional	Common features	99.72	94.27
DWT based Energy	DWT energy (db8)	89.02	91.21
	DWT energy (coif1)	82.93	77.45
DWT based	Haar	99.9	96.51
	db2	99.93	97.69
	db8	**99.97**	**99.19**
	db15	99.83	98.63
	db20	99.9	98.69
CWT based	CFS (5 Levels)	99.12	98.13
	Q_shift (5 Levels)	99.93	97.95
	CFS (7 Levels)	99.87	97.82
	Q_shift (7 Levels)	99.93	97.57

The feedforward artificial neural networks with the scaled conjugate gradient (SCG) backpropagation algorithm in MATLAB's neural networks toolbox which belongs to the class of the conjugate gradient algorithms have been used for classification. SCG algorithm uses step size scaling instead of line-search per learning iteration, and this makes it faster than other second-order algorithms [22]. This algorithm performs well for networks with a large number of weights, where it is as fast as the Levenberg-Marquardt and resilient backpropagation algorithms; its performance does not degrade quickly. Also, the conjugate gradient algorithms have relatively modest memory requirements. The number of hidden neurons has been preferred as 40, and the target mean square error has been defined as 0.001, as a result of extensive simulations.

All codes and programs in this study have been written in MATLAB. The codes for common features, DWT-based statistical and energy features have been written by the authors. For DWT-based analysis, wavelet toolbox of MATLAB has been used. For CWT-based analysis, the codes are taken from the study of Selesnick [21] for common factor solution-based filter design and the programs written by two students under supervision of I. Selesnick have been used for Q-shift filter based analysis [23].

In the following section, the general classification results will be given for feature vectors. The performance has been given as the accuracy of the classification which can be formulated as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}, \tag{27}$$

where TP, TN, FP, and FN represent the number of speech samples labeled as speech, the number of music samples labeled as music, the number of music samples labeled as speech and the number of speech samples labeled as music, respectively.

3.2. General Classification Performance. The results of the experiments are summarized as general performance in Table 3.

When Table 3 is taken into consideration, it can be seen that wavelet-based parameters have higher classification results than traditional methods. In general, all methods are successful in the classification of samples in Dataset 1, which indicates that the TIMIT speech data and CD recordings are separable. However, it is not possible to say same thing for Dataset 2, since the samples in Dataset 2 reflects a more realistic case, where samples are recorded from radio broadcast. The best performance has been obtained with db8 wavelet. The complex wavelet-based features perform better than conventional methods and wavelets with fewer vanishing moments. However, they are not as successful as the db8. The similarity of the mother wavelet with the analyzed waveforms is an important criterion for the wavelet analysis which may be the cause of this performance difference. Therefore, the accuracy for different databases may differ drastically.

The feature extraction methods are considered in terms of their calculation times, and the average computation times for feature-extraction stage for all methods used in this study are given in Table 4.

According to Table 4, a sorting among the feature extraction methods can be made as

$$t_C > t_{\text{DWT}} > t_{\text{CWT}} > t_{\text{DWTE}}, \tag{28}$$

where t_C, t_{DWT}, t_{CWT}, and t_{DWTE} show the computation time for the methods based on conventional, DWT, CWT, and DWT-based energy features. As expected, the conventional methods take more time, since they include time domain calculations. Although energy-based features are the fastest, they have poor classification performance.

The calculation times for DWT-based statistical feature extraction show differences according to the used wavelet in the analysis. Wavelet families including more vanishing moments such as db15 and db20 spend more time for computation comparing to other wavelet families, since they have longer filters. It is encountered that the db8 families as the optimum wavelet for DWT-based analysis since it shows highest performance in classification of speech and music and it has acceptable calculation time.

Table 4: Average computation times for used feature extraction methods.

		Speech (msn.)	Music (msn.)
Conventional	Common features	0.2768	0.2745
DWT based Energy	Daubechies8-based energy features	0.0216	0.0217
	Coiflet1-based energy features	0.0176	0.0176
DWT based	Haar	0.0357	0.0382
	Daubechies2	0.0401	0.04
	Daubechies8	0.0485	0.0462
	Daubechies15	0.1035	0.1034
	Daubechies20	0.155	0.1547
CWT based	Q-shift-based CWT features	0.0298	0.0296
	CFS-based CWT features	0.0301	0.03

CWT-based method is faster than DWT-based analysis, and it shows performance results close to DWT. In this perspective, CWT-based features can be used for online implementation as well.

It should be noted that the silent parts of samples could be determined quicker than the feature-extraction methods during online implementation, since only an energy threshold value is considered to make a decision if the segment is silence or not.

3.3. Graphical User Interface (GUI) Design for Speech/Music Discrimination. A graphical user interface has been designed as well in order to perform speech music discrimination visually. An online labelling module has been also embedded to the interface and observation of performance for real-time classification has become possible with this tool.

3.3.1. Main Module. Main module can be used to see classification results obtained by methods. In Figure 3, the GUI designed for main module is shown. Using "load file" button, the file to be analyzed is selected, and the "play" button plots and plays the signal at the same time. Since time-/frequency-based features are also used at the classification stage, it is important to see the general structure of spectrogram. For this aim, there is a button named as "spectrogram of signal" in the module to plot time/frequency properties. In the DWT-based features part of module, 12-level decomposition is performed using selected wavelet from pop-up menu. It is also possible to see shape of wavelet and wavelet coefficients using "show wavelet" and "plot wavelet coefficients" buttons, respectively. For CWT-based feature, there is also a pop-up menu which you can select the filter design method for analysis. It can be seen the existing complex wavelet with its real and imagined parts using "plot filter Coefficients" button in CWT-based feature extraction part of the module. For time/frequency based feature extraction, there is also a button and the values of parameters such as spectral centroid, spectral rolloff, spectral flux, the number of zero crossings, and low energy ratio of loaded file exist in the blanks when this button is pushed. It is possible to get the classification results of four-feature extraction methods simultaneously using "classification results" button. If "online labelling module" button is pushed, the online labelling module will appear in a new window.

3.3.2. Real-Time Speech/Music Discrimination Module. This module has been designed to observe speech/music classification performance for online implementations. In the given graphical user interface in Figure 4 , a sample file is fetched using "open" button and the online labelling process is started using "start" button. "pause" button makes it possible to stop the process temporarily, and using "continue" button the labeling process can be continued from where it is stopped. The "stop" button interrupts the program and ends the label assignment process.

In the module, the red letters under the signal graph shows the preassigned labels for data and "*S*", "*M*", and "_" are used to indicate speech, music, and silence parts of data, respectively. Online classification results are shown with blue color under preassigned labels, as it can be seen from Figure 4. These labels are assigned for segments which have the length of 0.5 sec. In online label assigning, the features are extracted using 12-level DWT with db8 wavelet for each sample, since it has given the most accurate results in experiments and a previously trained artificial neural network is used to determine the labels.

4. Conclusions and Future Work

There has been a growing interest in speech/music discrimination systems which is usually used in several applications as a preprocessing stage of several audio systems. However, still improvements for SMD systems are required to develop effective real-time systems. Therefore, in this study, two feature-extraction schemes based on discrete and dual-tree wavelet transforms have been proposed which are suitable for real-time applications. After comparing the proposed features with the conventional features and recently proposed wavelet-based energy features, an online labeling module has been implemented with the Daubhecies8 wavelet.

Regarding the experiments and results given in the previous section, it has been observed that conventional features are not very effective in discrimination of speech/music samples when they are used alone. However, if they are used together, the accuracy tends to slightly increase.

The wavelet transform methods except the energy-based ones have shown higher performance for Database 1 than the results for Database 2, because the second database consists of the recordings from radio broadcasts which reflects a more realistic case.

The selection of the analysis window length which specifies the content of the nonstationary signal and the speed of implementation is an important choice for SMD. The selection of a short window order of milliseconds as in literature will not give the necessary information on time-varying frequency content, since the signal can be assumed

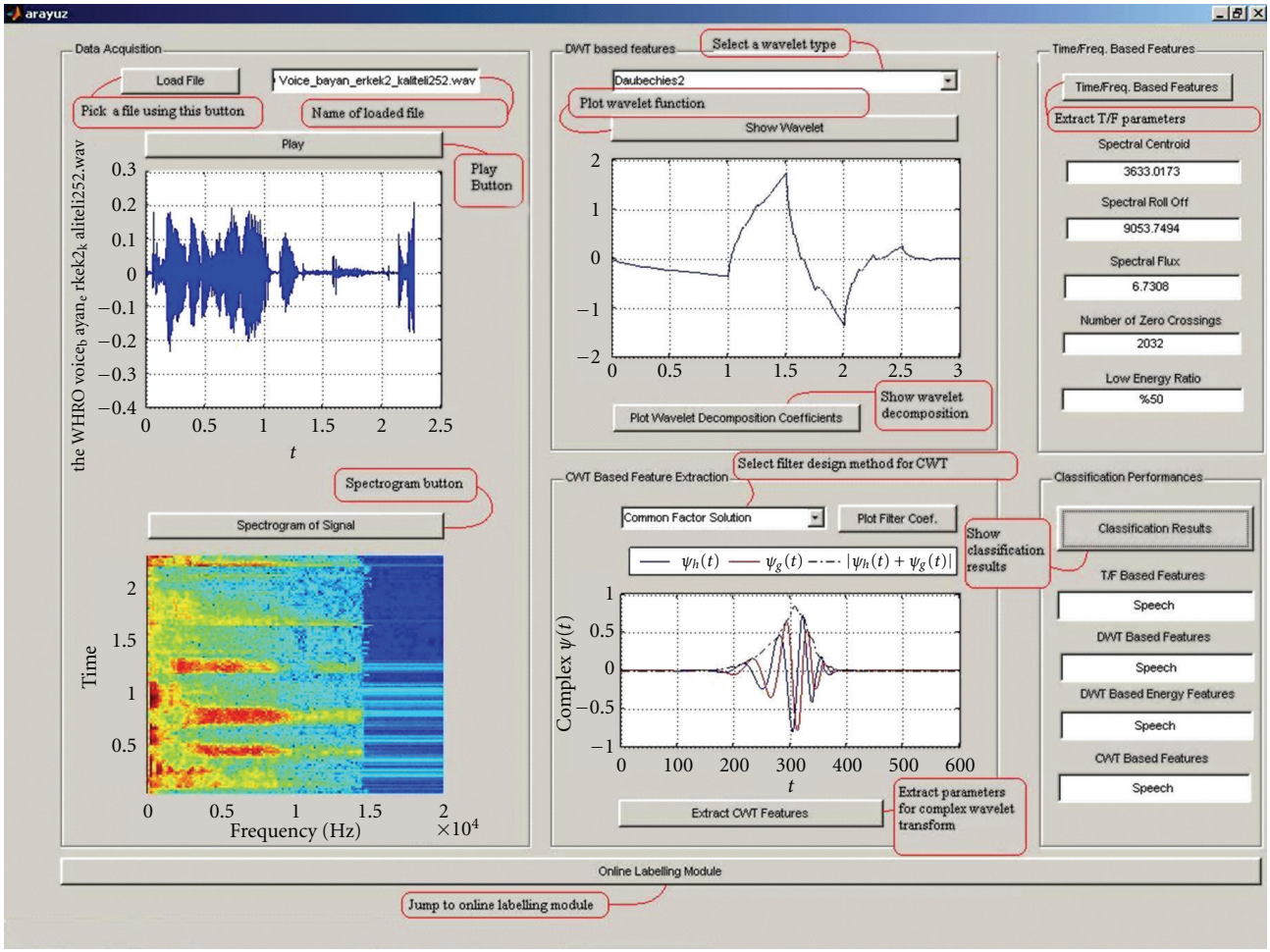

Figure 3: Graphical user interface for main module.

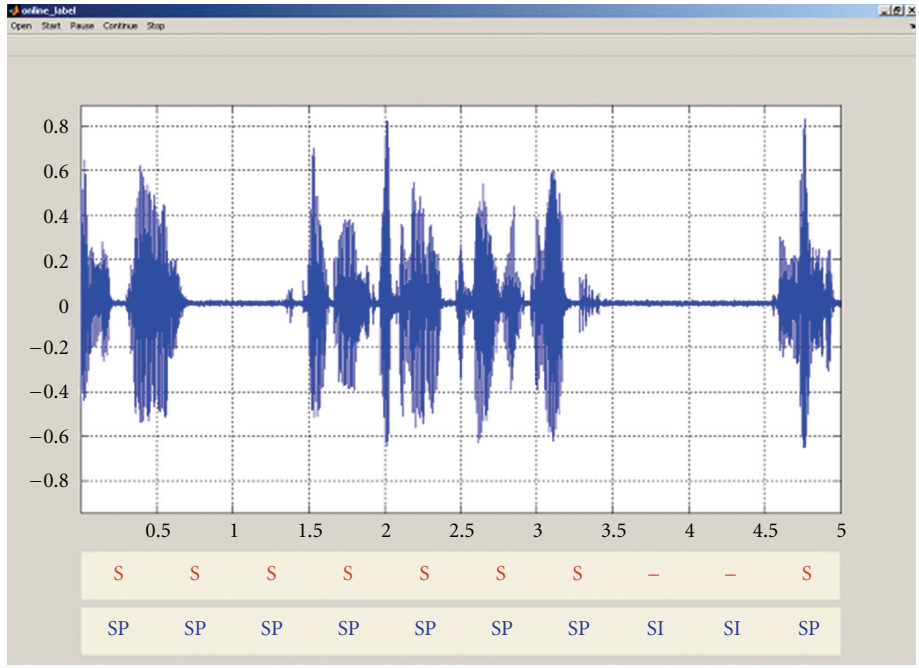

Figure 4: Graphical user interface for online labelling module.

as stationary in this interval. On the other hand, the usage of long window order of seconds, which is reported as successful, limits the online application of the algorithms. The proposed algorithm is computationally efficient (average running time for the proposed method is <50 msn), and this allows the use of method for real-time implementation.

The classification performance of DWT-based feature vector varies depending on the mother wavelet used in the feature-extraction stage. When the number of vanishing moments is increased, the wavelet becomes smoother. These smooth wavelets produce large coefficients for slowly changing signals like music, while it produces relatively small coefficients for speech signals. This can be used as a discriminative property for SMD. The Haar and db2 wavelets have a few vanishing moment this may prevent the good representation of signal in frequency domain. On the contrary, db15 and db20 have much more filter coefficients and vanishing moments, but this increases the complexity in the feature space and also requires longer computations. In this way, db8 has emerged as the most ideal wavelet type.

Among the CWT-and DWT-based features, DWT feature extracted with Daubechies8 wavelet has demonstrated the highest classification performance, while the CWT-based classification has shown results as 99.93% for Database1 and 98.13% for Database2. When all features are concerned, we see that Daubechies8-based parameters have superior discrimination features in terms of classification of speech and music.

In the study, the contribution of the ratio parameters to the discrimination performance have also been examined for DWT- and CWT-based features. It has been observed that ratio parameters provide a contribution about 1–1.5% to the overall performance for DWT-based parameters. However, the results were inconclusive for CWT.

In the classification stage, artificial neural networks have been used as classification tool. The number of hidden neurons has been preferred as 40, and the target mean square error has been defined as 0.001, heuristically. Conjugate gradient algorithms have been selected as learning algorithm, since they have advantages according to other methods. Also, principal component analysis has been performed before the classification stage to represent signals more efficiently and to decrease the dimension of feature vector.

The observed SMD performance of CWT-based features was less than the DWT-based ones. To obtain a better performance, a feature set which reflects the most powerful properties of CWT must be constructed. The filter structure used in CWT-based parameterization has the possibility of presence of unsuitable characteristics in terms of speech/music discrimination, and as a result, the accuracy is observed as lower than performance of DWT-based features. In this manner, adaptive filter design is required to get more successful results. The dataset can be expanded to include mixed speech-music samples. In this way, a multiclass classification can be performed instead of binary classification for future studies.

The proposed feature extraction is also suitable for a hardware implementation using digital signal processors to have a faster SMD system.

References

[1] O. M. Mubarak, E. Ambikairajah, and J. Epps, "Novel features for effective speech and music discrimination," in *Proceedings of the IEEE International Conference on Engineering of Intelligent Systems (ICEIS '06)*, Sayfa, Islamabad, Pakistan, April 2006.

[2] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, pp. 2445–2448, June 2000.

[3] A. C. Gedik and B. Bozkurt, "Pitch-frequency histogram-based music information retrieval for Turkish music," *Signal Processing*, vol. 90, no. 4, pp. 1049–1063, 2010.

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

[5] 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)*, pp. 1331–1334, April 1997.

[6] E. M. Saad, M. I. El-Adawy, M. E. Abu-El-Wafa, and A. A. Wahba, "A multifeature speech/music discrimination system," in *Proceedings of the 19th National Radio Science Conference (NRSC '02)*, pp. 208–213, 2002.

[7] 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.

[8] 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.

[9] J. Wang, Q. Wu, H. Deng, and Q. Yan, "Real-time speech/music classification with a hierarchical oblique decision tree," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 2033–2036, April 2008.

[10] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A speech/music discriminator of radio recordings based on dynamic programming and Bayesian networks," *IEEE Transactions on Multimedia*, vol. 10, no. 5, Article ID 4540196, pp. 846–857, 2008.

[11] M. Kos, M. Grašič, and Z. Kačič, "Online speech/music segmentation based on the variance mean of filter bank energy," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 628570, pp. 1–13, 2009.

[12] G. E. G. Tzanetakis and P. Cook, "Audio analysis using the discrete wavelet transform," in *Mathematics and Simulation with Biological, Economical and Musicoacoustical Applications*, pp. 318–323, WSES, 2001.

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

[14] S. Ntalampiras and N. Fakotakis, "Speech /music discrimination based on discrete wavelet transform," in *Proceedings of the 5th Hellenic Conference on Artificial Intelligence (SETN '08)*, vol. 5138 of *Lecture Notes in Artificial Intelligence*, pp. 205–211, Syros, Greece, October 2008.

[15] E. Didiot, I. Illina, D. Fohr, and O. Mella, "A wavelet-based parameterization for speech/music discrimination," *Computer Speech and Language*, vol. 24, no. 2, pp. 341–357, 2010.

[16] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters,"

in *Proceedings of the IEEE Digital Signal Processing Workshop*, 1998.

[17] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.

[18] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[19] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[20] N. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proceedings of the IEEE Digital Signal Processing Workshop*, pp. 1–4, 1998.

[21] I. W. Selesnick, "The design of Hilbert transform pairs of wavelet bases," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 170–173, 2001.

[22] C. Charalambous, "Conjugate gradient algorithm for efficient training of artificial neural networks," *IEEE Proceedings-G on Circuit Devices and System*, vol. 139, no. 3, pp. 301–310, 1992.

[23] S. Cai and K. Li, "Matlab Implementation of Wavelet Transforms," 2002.