Comparison OF Wavelet Based Feature Extraction Methods for Speech/Music Discrimination

Timur DÜZENLİ[1,2]Nalan ÖZKURT[1,3]
[1]Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Izmir, Turkey.
[2]Izmir University of Economics, Department of Electronics And Telecommunications Engineering,
Faculty of Engineering and Computer Sciences,35330,Balçova, İzmir, Turkey
Tel: +90-232-488-8272, Fax: +90-232-488-8475
[3]Yaşar University, Department of Electrical and Electronics Engineering,
Faculty of Engineering, 35100,İzmir, Turkey
Tel: +90-232-411-5000, Fax: +90-232-374-5474

timur.duzenli@ieu.edu.tr, nalan.ozkurt@yasar.edu.tr

The speech/music discrimination systems have gaining importance in several intelligent audio retrieval algorithms due to the increasing size of the multimedia sources in our daily lives. This study aims to propose a speech/music discrimination system which utilizes the advantages of the wavelet transform. Also, the performance of the discrete wavelet transform and the dual- tree wavelet transform has been compared with the conventional time, frequency and cepstral domain features used in speech/music discrimination. The speech and music samples collected from common databases, CD recording and internet radios have been classified with artificial neural networks with different feature sets. The principal component analysis has been applied to eliminate the correlated features before classification stage. Considering the number of vanishing moments and orthogonality, the best performance has been obtained with Daubechies8 wavelet among the other members of the Daubechies family. According to the results, the proposed feature set outperforms the traditional ones.

Keywords: *Speech/music discrimination, Discrete wavelet transform, Dual-tree wavelet transform, Daubechies mother wavelet.*

1. Introduction

The multimedia sources have a dominant role in our lives and due to the increasing size of these sources, the importance of the intelligent search and retrieval algorithms have also increased. The speech / music discrimination (SMD) systems are commonly used in preprocessing stage in audio coding [1,2], automatic speech recognition [3] and content based multimedia retrieval [4].

There are several methods and several different segmentation schemes for speech/music classification task in the literature. In [5], one of the pioneering studies, the distribution of zero crossing rate and lopsidedness of this distribution have been used for feature extraction.

Another important study proposes a feature set constructed from 13 features such as 4Hz modulation energy, percentage of low energy frames, spectral roll-off points, spectral centroid, spectral flux and zero-crossings rate [6]. The effects of the features to the classification performance has been considered for GMM, k-NN and k-d spatial classifiers in this study. Following these researches, there have been growths in the proposed features and classification methods. The entropy and dynamism features have been used for hidden Markov model classification in [7]. In [8], a system which uses zero crossings and average frequency is proposed to find transitions between music and speech.

Despite the wide spread of the proposed features and segmentation procedures, finding discriminative features and accurate classification schemes for an effective SMD is still a hot topic. The time-frequency domain based tools especially wavelets have also been used for feature extraction [9–12].

Received on: 06.04.2011
Accepted on: 17.06.2011

The wavelets are suitable tools for SMD because they have ability to deal with non-stationary signals such as music and speech, analyze the signals in different scales and achieve variable time-frequency localization [13].

In this study, three different types of wavelet transform based features have been extracted. The statistical measures of Discrete Wavelet Transform (DWT) coefficients, DWT based energy features and Complex Wavelet Transform (*C*WT). The use of *C*WT which eliminates some shortcomings of DWT is a novel technique in SMD. Also, it has been shown that the shorter windows accomplish successful classification. The paper is organized as follows: a brief introduction of the previous and proposed features; data set and classification algorithm; the result and discussion and finally the conclusions and future works.

2. Features for Speech/ Music Discrimination

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

2.1. Conventional features

The time and frequency domain features such as number of zero crossings, 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 including mean and variances of these time / frequency based features has been constructed. Length of the feature vector for this method is 21.

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

$$Z_i = \frac{1}{2}\sum_{n=2}^{N}\left|sgn\big(x(n)\big) - sgn\big(x(n-1)\big)\right| \qquad (1)$$

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

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

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

where X$_k$ is the magnitude of kth frequency component in the frame. Since the energy distribution for speech is more left-skewed than for music, this measure is higher for speech.

Spectral Centroid is the measure of the center of mass in the frequency spectrum and calculated as

$$SC = \frac{\sum_{k=1}^{K} f_k x_k}{\sum_{k=1}^{K} x_k}$$
(3)

where X$_k$ is the magnitude of the component of the frequency bin f$_k$ [6].

Spectral Roll-off is important in determining the shape of the frequency spectrum. The spectral roll-off point R$_k$ is the frequency value where the 95% of the spectral power lies below as summarized in

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

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 [6].

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

$$SF_h = \sum_{k=1}^{K}\left(X_k^h - X_k^{h-1}\right)^2 \qquad (5)$$

where X_k^h is the kth frequency component of the hth frame. Then the average of the all frames is calculated. The music has a higher rate of change than the speech except consonant-vowel boundaries, thus this value is usually higher for music [6].

Mel Frequency Cepstrum Coefficients (MFCC) are calculated from the Mel frequency spectrum which is defined as 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 where the frequency bands are not linearly spaced. Thus the sound is represented better.

2.2 Discrete Wavelet Transform

The Multi-Resolution Analysis (MRA) provides time-frequency representation and also well suited for non-stationary signals. MRA decomposes and analyses the signal at different frequencies and different resolutions. Discrete wavelet transform (DWT), which is a specific case of MRA, has been used widely in all areas of signal processing and has established an impressive reputation especially in image and speech processing. There is a rich set of basis functions for DWT and it is possible to get a compact representation of signal using this transform. The continuous wavelet transform of a signal x(t) is a representation in time-scale space by projecting the signal on a space spanned by scaled and translated wavelets as

$$\mathrm{CWT}(s,r) = \frac{1}{\sqrt{s}}\int x(t)\psi^*\left(\frac{t-r}{s}\right) \qquad (6)$$

where ψ(t) is the mother wavelet, s and r are the scale and translation coefficients, respectively [13]. For computational issues, the scale and translation coefficients are discretized and Discrete Wavelet transform (DWT) is obtained. In practical applications, DWT is applied in sampled signal x[n]; n = 1... N as

$$DWT[n,2^j] = \sum_{m=0}^{N-1} x[m]\psi^*_{2^j}[m-n] \qquad (7)$$

where

$$\psi_{2^j}[n] = \frac{1}{\sqrt{2^j}}\psi\left(\frac{n}{2^j}\right) \qquad (8)$$

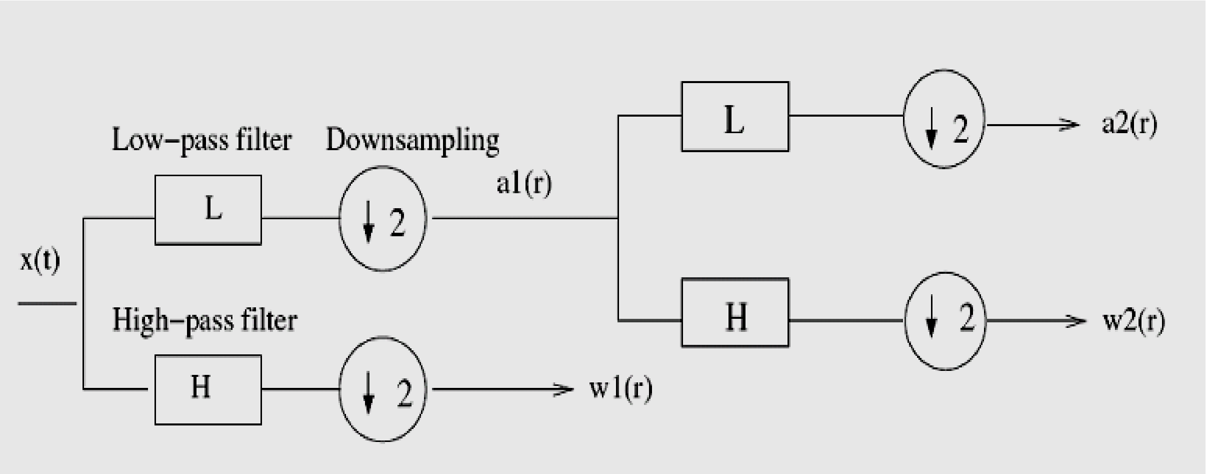

Figure 1. DWT with two composition levels [10]

DWT is implemented using a pyramidal algorithm related to a multi-rate filterbank approach for multi-resolution 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 Fig.1 [10], [13].

After each filtering stage i, the outputs are down-sampled by 2, where the outputs $a_i(r)$ are approximation and $w_i(r)$ are detail coefficients. Approximation coefficients give local averages of the signal. On the other hand, detail coefficients show the differences between local averages. In this work, Daubechies family is used as the mother wavelet which is known as one of the best families for speech processing applications [10],[11]. For feature extraction, 12-level decomposition has been considered which covers the analyzed frequency range in detail. Therefore, 1 approximation and 12 detail signals are obtained for each frame.

The length for feature vector which is constructed from the statistical measures including mean, standard deviation and ratios of the decomposed signals is 38.

2.3 Wavelet Transform Based Energy Features

In [10], it has been talked about the energy based features which are calculated using wavelet transform. 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.

The Instaneous Energy is a feature which gives the energy distribution in each band and given as:

$$f_j^E = \log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}(w_j(r))^2\right) \qquad (9)$$

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.

Teager Energy has been recently applied for speech recognition and given as:

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

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 MFCC as described in [10]. 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.

In this method, only detail coefficients have been used at the feature extraction stage. The decomposition has been performed for 5 levels of subbands and two energy parameters; instantaneous and teager energy, have been obtained for each band. In this way, length of the feature vector for each sample is 10.

2.4 Complex Wavelet Transform

Conventional discrete wavelet transform suffers from some fundamental short-comings although 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. Second problem of DWT is the lack of directionality for two and higher dimension signals. Complex Wavelet Transform (*C*WT) has been proposed inspiring by the Fourier Transform which does not suffer from these types of problems. A complex wavelet is defined as

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \qquad (11)$$

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 if 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 $2^j\psi_c(2^j t - n)$, the complex wavelet coefficients are obtained as

$$d_c(j,n) = d_r(j,n) + jd_i(j,n) \qquad (12)$$

Complex Wavelet Transform can be performed in two class. In 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. Resulting *C*WT 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 prefferred. Dual tree approach is based on two filterbank trees and it has two bases. 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 [15].

Dual-Tree Complex Wavelet Transform (DT-*C*WT) was first introduced by Kingsbury in 1998 [16]. The dual tree implements an analytic wavelet transform by using two real discrete wavelet transform; the first DWT gives the real and the second one gives the the imaginary part of the *C*WT. Analysis and synthesis filter banks can be illustrated in the Figure 2 where $h_0(n)$ and $h_1(n)$ denote the low-pass / high-pass filter pair for the upper filterbank 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.

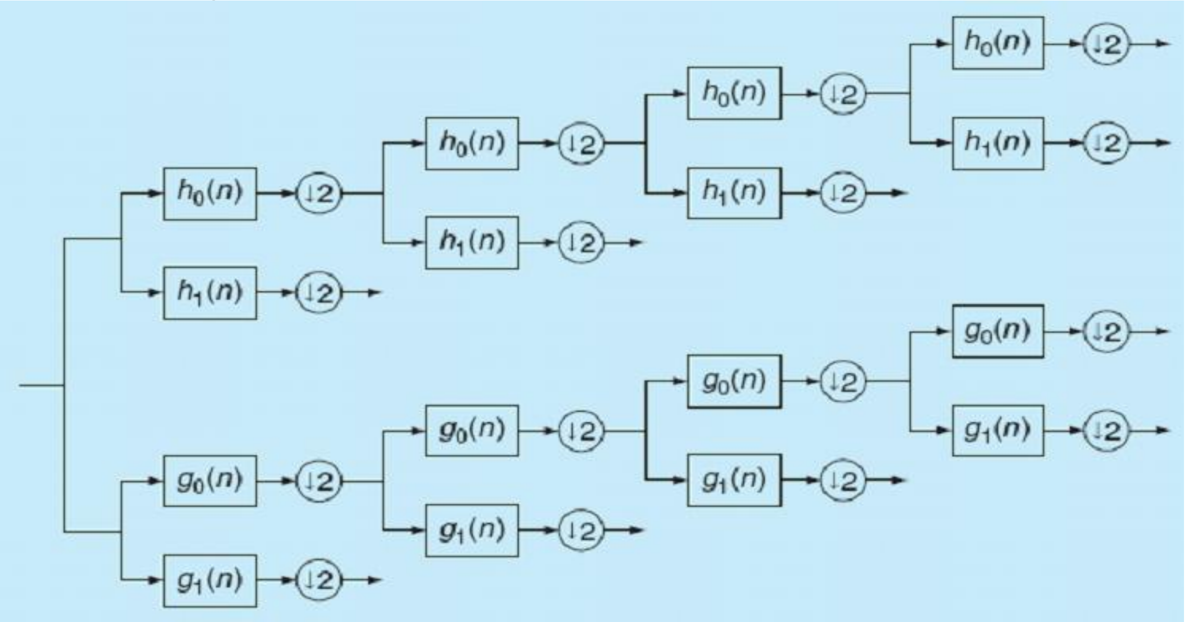

Figure 2. Analysis filter bank for the dual tree CWT [15]

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

$$\sum_n h_0(n) h_0(n+2k) = \delta(k) \tag{13}$$

$$h_1(n) = (-1)^n h_0(M-n) \tag{14}$$

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

$$g_0(n) = h_0(n-0.5) \rightarrow \psi_h(t) = \mathcal{H}\{\psi_g(t)\} \tag{15}$$

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

$$|G_0(e^{jw})| = |H_0(e^{jw})| \tag{16}$$

$$\angle G_0(e^{jw}) = \angle H_0(e^{jw} - 0.5w) \tag{17}$$

There are various methods for design of filters for DT-*C*WT. In this paper, Q-Shift and Common Factor solutions have been used at filter design stage [15].

While constructing feature vector, the decomposition has been made for 5 and 7 levels in order to avoid increasing length. The feature vectors have been constructed from mean, variance and median of the magnitudes of complex wavelet coefficients at each band instead of using all coefficients. In this way, the length of feature vectors for each sample is given as 25 or 35 for 5-level and 7-level decomposition, respectively.

3. Dataset and Classification Algorithm

The two different data sets have been utilized in the study and the features have been extracted separately for these two different datasets. In Dataset1, TIMIT database has been used for speech and several CD recordings with various musical genres have been used for music database. To obtain Dataset2, radio broadcasts containing music and speech were recorded. The sampling frequency was set as 44100 Hz. However, since the data taken from TIMIT database is sampled with 16000 Hz, they have been interpolated in the pre-processing stage in order to set sampling frequency to 44100 Hz. The segmentation has been performed for a frame of 4196 samples with 512 samples overlapping which corresponds to a frame length of 95 ms since use of shorter window lengths may limit the discriminative characteristics of window. Both datasets used in the study contain samples with length of 0.5 sec. [17]. A detailed representation for dataset1 and dataset2 is given in Table 1.

Table 1. Content of datasets

	Overall Database		Train Set		Test Set	
	Speech	Music	Speech	Music	Speech	Music
Dataset1	4620	4290	3080	2860	1540	1430
Dataset2	2624	2190	1749	1460	875	730

Before classification stage, the features that are highly correlated with the other features have been eliminated using principal component analysis (PCA) to reduce length of feature vectors. The principal components that contribute less than 0.05% to the total variation in the data set have been eliminated.

In classification stage, the feedforward artificial neural networks with the scaled conjugate gradient (SCG) backpropagation algorithm in MATLAB's Neural Networks Toolbox which belongs to class of the conjugate gradient algorithms have been used. SCG algorithm uses step size scaling instead of line-search per learning iteration and this makes it faster than other second order algorithms [18]. 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, heuristically.

4. Results and Discussion

The performance has been given as the accuracy of the classification which can be formulated as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (18)$$

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

When Table 2 is taken into consideration, it can be seen that wavelet based parameters have higher classification results than traditional features. In general, all methods are successful in classification of samples in Dataset1, which indicates that the TIMIT speech data and CD recordings are separable. However, it is not possible to say same thing for Dataset2 since the samples in Dataset2 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 performs better than conventional features 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.

When these feature extraction methods are considered in terms of their calculation times, DWT based energy features emerge as the fastest algorithm in terms of feature extraction since it contains only ten parameters in feature vector. On the other hand, DWT based energy features have the lowest classification performance among the considered methods according to results. In Table 3, the computation times for feature extraction stage for all methods used in the study are given.

Table 2. General classification results

Performance (%)	Dataset1	Dataset2
Conventional	99.72	94.27
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
DWT_Energy (db8)	89.02	91.21
DWT_Energy (coif1)	82.93	77.45
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

According to average computation times in Table 3, a sorting among the feature extraction methods can be made as:

$$t_C > t_{DWT} > t_{CWT} > t_{DWTE}$$

where t_C, t_{DWT}, t_{CWT} and t_{DWTE} show the computation time for the methods based on conventional, DWT, *C*WT and DWT based energy features.

The calculation time of DWT based feature extraction methods depends on the used mother wavelet. The wavelet families having large number of vanishing moments such as db15 and db20 spend more time for computation of coefficients comparing to other wavelet families since they have more filter coefficients. 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.

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

Conventional analysis is the slowest method since it performs many computations in time and frequency domain.

Table 3. Average computation times

	Speech (ms)	Music (ms)
Conventional	0.2768	0.2745
Haar	0.0357	0.0382
db2	0.0401	0.04
db8	0.0485	0.0462
db15	0.1035	0.1034
db20	0.155	0.1547
DWT_Energy (db8)	0.0216	0.0217
DWT_Energy (coif1)	**0.0176**	**0.0176**
Q-shift (5 Levels)	0.0298	0.0296
CFS (5 Levels)	0.0301	0.03

5. Conclusion & Future Works

In this study, the speech and music samples with length of 0.5 sec. have been used at feature extraction and classification stages. Although longer segments are used in the literature generally, it has been shown that 0.5 sec. length is enough to get high performance in classification of speech and music. The proposed algorithm used in this work is computationally efficient (average running time for proposed method is <50 msn) and this allows the use of method for online implementation. As mentioned before, a fast running speech / music discrimination system with high accuracy can be designed by using suggested method as a preprocessing stage for several applications.

Since the SMD is an hot topic for multimedia applications, the studies can be extended in several directions. One of them might be the research on adaptive filter design methods to reveal more advantages of CWT on DWT in speech / music discrimination. Therefore, the parameters for the SMD tasks can be determined automatically according to the problem at hand.

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 performance of CWT based features can be further examined to construct a more discriminated feature space.

6. References

[1] Ambikairajah, O. M. E., Epps, J., "Novel features for effective speech and music discrimination," in Proc. IEEE Int. Conf. on Engineering of Intelligent Systems, pp. 1–5, 2006.

[2] Exposito, N. R. J.E.M., Galan, S.G., Candeas, P., "Audio coding improvement using evolutionary speech/music discrimination," in Proc. IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE), pp. 1–6, 2007.

[3] El-Maleh, K., Petrucci, M. G., Kabal, P., "Speech/music discrimination for multimedia applications," in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 2445–2448, 2000.

[4] Gedik, A., Bozkurt, B., "Pitch frequency histogram based music information retrieval for turkish music," Signal Processing, vol. 10, pp. 1049–1063, 2010.

[5] Saunders, J., "Real time discrimination of broadcast speech/music," in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 993–996, 1996.

[6] Scheier, E., Slaney, M., "Construction and evaluation of a robust multifeature speech/music discriminator," in Proc. IEEE Int. Conf. On Acoustics, Speech, and Signal Processing, ICASSP'97, pp. 1331–1334, 1997

[7] Ajmera, I. M. J., Bourlard, H., "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," Speech Communication, vol. 40, pp. 351–363, 2003.

[8] Panagiotakis, C., Tziritas, G., "A speech/music discriminator based on RMS and zero-crossings," IEEE Trans. Multimedia, vol. 7, pp. 155–166, 2005.

[9] Tzanetakis, G. E. G., Cook, P., "Audio analysis using the discrete wavelet transform," in Proc. Conf. in Acoustics and Music Theory Applications. WSES, pp. 318–323, 2001.

[10] Didiot E., Illina, I., Fohr, D., Mella, O., "A wavelet-based parameterization for speech/music discrimination," Computer Speech and Language, vol. 24, pp. 341–357, 2010.

[11] Ntalampiras, S., Fakotakis, N., "Speech /music discrimination based on discrete wavelet transform," in Proc. of 5th Hell. Conf. On Art.Int., SETN'08, LNAI 5138, Greece, Oct. 2008, pp. 205–211, 2008

[12] Khan, M., Al-Khatib, W., "Machine-learning based classification of speech and music," ACM Jour. on Multimedia Systems, vol. 12, pp. 55–67, 2006.

[13] Mallat, S., *A wavelet tour of signal processing*. Academic Press, 1999

[14] Zheng, F., Zhang, G., Song, Z., "Comparison of different implemantations of mfcc," Arch. Rat. Mech. Anal., vol. 16, pp. 582–589, 2001.

[15] Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.G. "The Dual-Tree ComplexWavelet Transform", IEEE Sig.Proc. Mag. 22, pp. 123–151, 2005.

[16] Kingsbury, N.G., "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters", Proc. of the IEEE Digital Signal Processing Workshop, 1998.

[17] Düzenli, T., (2010). Classification of Speech and Musical Signals Using Wavelet Domain Features, MSc. Thesis submitted to Dokuz Eylül University, Graduate School Of Natural And Applied Sciences.

[18] Charalambous, C., Conjugate gradient algorithm for efficient training of artificial neural networks. IEEE Proceedings-G on Circuit Devices and System, 139 (3), pp. 301-310, 1992

[21] A. Toker, S. Özcan, H. Kuntman, O. Çiçekoğlu, "Supplementary all-pass sections with reduced number of passive elements using a single current conveyor", Int J of Electronics, vol.88, pp.969-976,2001.

[22] U. Çam, O. Çiçekoğlu, M. Gülsoy, H. Kuntman, "New voltage and current mode first-order all-pass filters using single FTFN", Frequenz, vol.7-8, pp.177-179,2000.

[23] R. Schauman, M. E. Valkenburg, "Design of analog filters", Oxford University Press, New York, 2001.

Nalan Özkurt received her B.S., M.S. and Ph.D. degree in Electrical and Electronics Engineering from the Dokuz Eylul University, in 1994, 1998 and 2004, respectively. She is currently an assistant professor in the Department of Electrical and Electronics Engineering at Yaşar University. Her research interests are wavelets, nonlinear static and dynamical systems, chaos. She is a member of Association of Electrical and Electronic Engineers of Turkey.

Timur Düzenli received his B.S. in 2007 and his M.S. in 2010, both in Electrical and Electronics Engineering, from Dokuz Eylul University. He is currently a Ph.D. student at the same department. Her research interests are wavelets, time-frequency analysis, and digital communication systems.