Speech/Music Discrimination Based on Discrete Wavelet Transform

Stavros Ntalampiras and Nikos Fakotakis

Electrical and Computer Engineering Department, Wire Communication Laboratory,
University of Patras, 26500 Rio - Patras, Greece
sntalampiras@upatras.gr, fakotaki@wcl.ee.upatras.gr

Abstract. In this paper we present an effective approach which addresses the issue of speech/music discrimination. Our architecture focuses on the matter from the scope of improving the performance of a speech recognition system by excluding the processing of information which is not speech. Multiresolution analysis is applied to the input signal while the most significant statistical features are calculated over a predefined texture size. These characteristics are then modeled using a state of the art technique for probability density function estimation, Gaussian mixture models (GMM). A classification scheme consisting of a conventional maximum likelihood decision methodology constitutes the next step of our implementation. Despite the fact that our system is based solely on wavelet signal processing, it demonstrated very good performance achieving 91.8% recognition rate.

Keywords: Computer audition, content-based audio classification, discrete wavelet transform, Gaussian mixture model.

1 Introduction

Over the past decades a lot of work has been conducted in the area of speech processing (SP) and especially in the field of automatic speech recognition. In order to achieve better performance we need the recognizer component to elaborate only on speech data and nothing else. Here enters the idea of speech/music discrimination. Having a system that identifies which part of a sound includes speech or music, we can activate or not the component that processes speech resulting this way to a better accuracy. Afterwards a speech enhancement algorithm may be employed, and in combination with a voice activity detector, have the signal recognized.

Another important scientific domain, audio processing, is getting more and more attention lately. Usually the same techniques as in SP are applied here as well, which shows that there is still much work to be done to attain better performance in such systems. A brief review of the area of speech/music discrimination is following. Slaney and Scheirer [1] utilize signal processing techniques in time and frequency domain to extract 13 features. Their experiments are made in different classification schemes with overall performance of 5.8% error rate. El-Maleh et al present a feature set which combines the line spectral frequencies (LSFs) and zero-crossing-based

J. Darzentas et al. (Eds.): SETN 2008, LNAI 5138, pp. 205–211, 2008.
© Springer-Verlag Berlin Heidelberg 2008

features for speech/music differentiation [2]. For feature evaluation Bayes error rate with empirical estimation was used while a quadratic Gaussian classifier categorizes the input signal resulting in 95.9% accuracy. In [3] a fusion of two speech/music classification approaches is presented consisting of two different subsystems. The utilization of GMM and spectral features is shown to provide 94% and 90% accuracy for speech and music detection respectively. Tzanetakis et al [4] propose a framework for audio analysis using the Discrete Wavelet transform (DWT) with 12 different subbands while a comparison of Mel Frequency Cepstral Coefficients (MFCC) and Short Time Fourier transform coefficients was made. The application of a Gaussian classifier indicated that features derived from the wavelet transform have similar performance to the other two feature sets. In [5] a multi-layer perceptron forms the basis to classify speech and music and the performance of several features is investigated including mean and variance of DWT. Seven features comprise the final feature set and 96.6% accuracy with ten neurons is obtained. Didiot et al [6] combined energy-based features from the wavelet coefficient of each frequency band and 12 MFCCs in order to train Hidden Markov models. Finally, a class/non-class strategy is employed to distinguish speech and music.

This work is contributing to the field of automatic acoustic analysis for the purpose of "understanding" the surrounding environment by exploiting *only* the perceived auditory information in the way humans exhibit quite effortless. We explore the usage of a feature set based solely on multiresolution processing to achieve efficient speech/music discrimination. The basis of our implementation is the wavelet transform (Fig. 1). Wavelets are usually used in data compression with the well known paradigm of the image compression algorithm, JPEG 2000 proposed by the Joint Photographic Experts Group. We are going to show that the application of the DWT can be of great importance in classification tasks which involve audio processing. The transform has the property of treating with great accuracy the lower frequencies of the signal in contrast to the higher ones. The fundamental property of the Fourier transform is the usage of sinusoids with infinite duration. While sinusoids are smooth and predictable, wavelets tend to be irregular and asymmetric. This is the principal property of the wavelet representation and will be discussed next.

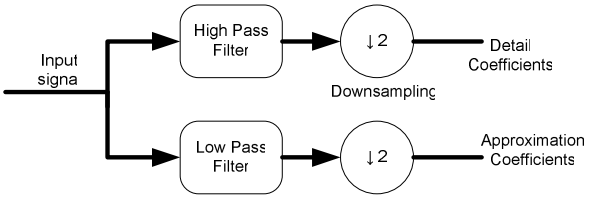

Fig. 1. Discrete Wavelet Transform

2 Wavelet Analysis

During the past years, wavelets techniques have become a common tool in digital signal processing. This kind of analysis has been used in many different researching areas including denoising of signals and applications in geophysics (tropical convention, the

dispersion of ocean waves etc) [7]. One can conclude that this emerging type of signal analysis is adequate to provide strong solutions in many and completely different researching areas. The main advantage of the wavelet transform is that it can process time series, which include non stationary power at many different frequencies (Daubechies 1990). Wavelet comprises a dynamic windowing technique which can treat with different precision low and high frequency information. The first step of the DWT is the choice of the original (or mother) wavelet and by utilizing this function, the transformation breaks up the signal into shifted and scaled versions of it. There are many types of mother wavelet functions and in this work the next four are investigated: haar (or db1), db 4, symlet 2 and a biorthogonal function (bior 3.7). A function must have zero mean and be localized in both time and frequency in order to form a mother wavelet (Farge 1992). Several experiments took place before the decision of these functions was made. Although these functions differ a lot (Fig. 2), we will see that the results are pretty much the same. The application of the DWT with these four different original wavelets consists of one-stage filtering of the audio signals as we can see in Fig. 1. Subsequently the data series are downsampled due to the Nyquist theorem in order not to end up having twice as many data comparing to the ones that we started with. In this paper we take under consideration only the *Approximation* coefficients which contain the low frequency information of the input sound, which is considered to contain the most important information as regards to human perception (inspired by the field of image processing).

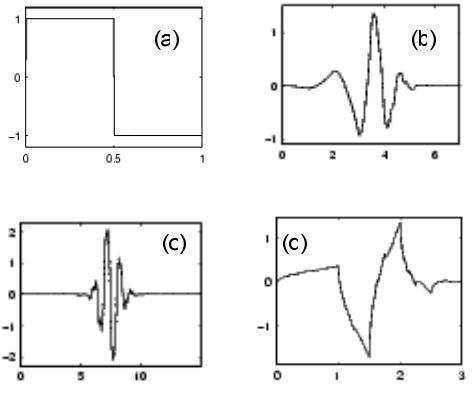

Fig. 2. Shapes of the mother wavelets that were employed (a) Haar (b) Daubechies 4 (c) Biorthogonal 3.7 (d) Symlet 2

2.1 The Feature Set

In our implementation, speech/music discrimination is based on six statistical measurements taken from the low frequency information of the signal. At the primary stage the DWT coefficient is cut into equal chunks of data using a texture size of 480 samples (30 ms), which was determined after extensive experimentations (see Fig. 3). It should be noted that no preemphasis is applied and the analysis is performed onto non overlapping chunks without considering the incomplete ones at the end of each file (in case there is one). For all the experiments the standard wavelet toolbox of the MATLAB^{TM} framework was used.

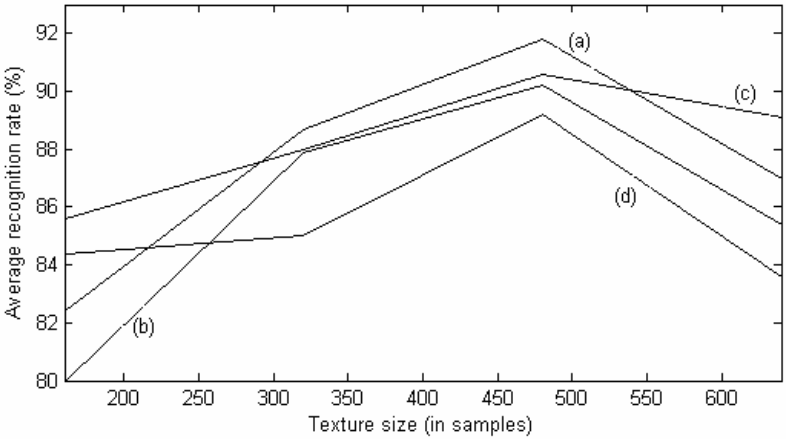

Fig. 3. Average recognition rates against different texture sizes achieved by the following mother wavelet functions: (a) Haar, (b) Daubechies 4, (c) Symlets 2 and (d) Biorthogonal 3.7

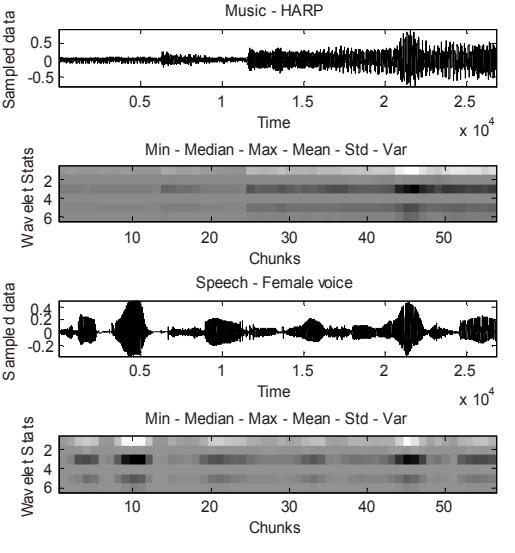

Fig. 4. Feature values of sound samples belonging to both categories

The hierarchical role of the feature extraction process is to capture the characteristics that distinct these two audio classes. Moreover, it is a technique of data reduction and in this case the suppression of the input data is huge having an average proportion of 4.1/100 (feature vector bytes/initial audio signal bytes). We sustain only a small amount of discriminative information of the audio signal using only the following six statistical features measured over the texture size (see also Fig. 4): (i) mean, (ii) variance, (iii) minimum value, (iv) maximum value, (v) standard deviation and (vi) the median. Afterwards we used speech and music data to train probabilistic models

(GMMs), which are described in the next section. Both male and female speech obtained from the TIMIT database, and an EBU music collection [8-9] which incorporates a large variety of musical instruments and compositions were employed to build up speech and music models. All the sounds were sampled at 16 KHz with 16 bit analysis while the average duration was 5.6 seconds.

3 Experimental Setup

The recognition process is consisted of probability computations of two Gaussian mixtures which represent the a priori knowledge that the system includes (Fig. 5). K-means initialization algorithm along with a standard version of the Expectation Maximization (EM) process was used for the training of eight components for each category. Furthermore it should be mentioned that diagonal covariance matrices are utilized for the construction of each mode.

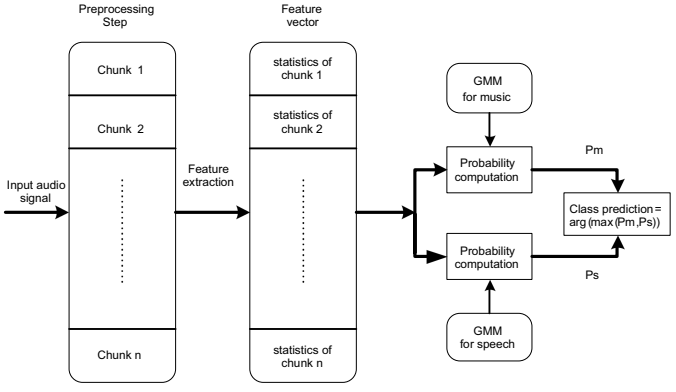

Fig. 5. Overall system architecture

During the testing phase, ten-fold cross validation was employed in order to obtain reliable results. Class decisions are made per frame while care has been taken not to have parts of the same file in both train and test sets simultaneously. As we tabulate in Table 1, the recognition rates achieved by our system are relatively high, concerning the small number of descriptors facilitated by our methodology.

Table 1. Recognition Rates (%)

Mother Wavelet	Music	Speech	Overall
Haar	89.4	94.2	**91.8**
Daubechies 4	89	91.4	**90.2**
Symlets 2	88.5	92.7	**90.6**
Biorthogonal 3.7	88.3	90.1	**89.2**

The four different original wavelets that were examined produced almost equal recognition rates. Despite its simplicity, haar function is proved to have the best performance while the biorthogonal one provides the worst results. Thus, we conclude that *Daubechies 1* mother wavelet function should be used in the task of speech/music discrimination.

4 Conclusions and Future Work

In this paper we explained a wavelet-based architecture for the purpose of efficient and simple speech/music discrimination under the scope of enhancing the performance of a speech recognition system. Our approach utilizes a limited amount of information produced by a well-known multiresolution technique. A comparison between three different wavelet families was conducted and the slight superiority of haar mother wavelet function was made clear.

We conclude that the specific type of analysis provides a strong basis for the implementation of a system with low computational needs as regards the specific classification task. The present contribution proposes a new feature set consisting only by six dimensions, while it reaches very high recognition accuracy. This work completes an initial step towards building a robust system for automatic speech/music discrimination. Our future work includes blind signal separation (BSS) to discriminate overlapping signals, incorporation of a silence detection algorithm and non-redundant fusion of the presented group of descriptors with well known sets, such as MFCC and MPEG-7 low level descriptors.

References

1. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 1331–1334 (1997)
2. El-Maleh, K., Klein, M., Petrucci, G., Kabal, P.: Speech/music discrimination for multimedia applications. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, pp. 2445–2448 (2000)
3. Pinquier, J., Rouas, J.-L., Andre-Obrecht, R.: A fusion study in speech/music classification. In: International Conference in Multimedia and Expo., vol. 2, pp. 409–412 (2003)
4. Tzanetakis, G., Essl, G., Cook, P.R.: Audio analysis using the wavelet transform. In: International Conference on Acoustics and Music: Theory and Applications (AMTA) (2001)
5. Kashif Saeed Khan, M., Al-Khatib, W.G., Moinuddin, M.: Automatic classification of speech and music using neural networks. In: 2nd ACM international workshop on Multimedia databases, pp. 94–99 (2004)
6. Didiot, E., Illina, I., Mela, O., Haton, J.P., Fohr, D.: A wavelet-based parameterization for speech/music segmentation. In: International Conference on Spoken Language Processing, pp. 653–656 (2006)
7. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bulletin of the American Meteorological society 79, 61–78 (1998)
8. EBU.: SQAM - CD: Sound quality assessment material, Polygram Cat. No 422 204-2, European Broadcasting Union (EBU) (1988)

9. EBU.: Sound quality assessment material, Recordings for subjective tests – Users/ handbook for the EBU-SQAM Compact Disc, Tech 3253, European Broadcasting Union (EBU) (1988)
10. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2, 674–693 (1989)
11. Foote, J.: An overview of audio information retrieval. ACM Multimedia Systems 7, 2–10 (1999)
12. Nabney, I.: Netlab: Algorithms for Pattern Recognition. Springer, London (2002)