HAL
open science

A Wavelet-Based Parameterization for Speech/Music Discrimination

E. Didiot, Irina Illina, D. Fohr, O. Mella

To cite this version:

E. Didiot, Irina Illina, D. Fohr, O. Mella. A Wavelet-Based Parameterization for Speech/Music Discrimination. Computer Speech and Language, 2010, 24 (2), pp.341. 10.1016/j.csl.2009.05.003 . hal-00608922

HAL Id: hal-00608922

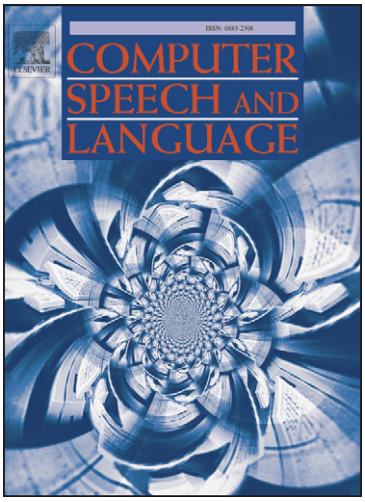
https://hal.science/hal-00608922

Submitted on 16 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



A Wavelet-Based Parameterization for Speech/Music Discrimination

1E. Didiot, I. Illina, D. Fohr, O. Mella

PII: S0885-2308(09)00042-4
DOI: 10.1016/j.csl.2009.05.003
Reference: YCSLA 421

To appear in: *Computer Speech and Language*

Received Date: 27 May 2008
Revised Date: 27 April 2009
Accepted Date: 11 May 2009

Please cite this article as: Didiot, 1E., Illina, I., Fohr, D., Mella, O., A Wavelet-Based Parameterization for Speech/Music Discrimination, *Computer Speech and Language* (2009), doi: 10.1016/j.csl.2009.05.003

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

A Wavelet-Based Parameterization for Speech/Music Discrimination

E. Didiot, I. Illina, D. Fohr, O. Mella [a]

[a] *LORIA-CNRS & INRIA Nancy Grand Est*
Building C, BP 239 54506 Vandoeuvre-les-Nancy, France

Résumé

This paper addresses the problem of parameterization for speech/music discrimination. The current successful parameterization based on cepstral coefficients uses the Fourier transformation (FT), which is well adapted for stationary signals. In order to take into account the non stationarity of music/speech signals, this work proposes to study wavelet-based signal decomposition instead of FT. Three wavelet families and several numbers of vanishing moments have been evaluated. Different types of energy, calculated for each frequency band obtained from wavelet decomposition, are studied. Static, dynamic and long-term parameters were evaluated. The proposed parameterization are integrated into two class/non-class classifiers: one for speech/non-speech, one for music/non-music. Different experiments on realistic corpora, including different styles of speech and music (Broadcast News, Entertainment, Scheirer), illustrate the performance of the proposed parameterization, especially for music/non-music discrimination. Our parameterization yielded a significant reduction of the error rate. More than 30% relative improvement was obtained for the envisaged tasks compared to MFCC parameterization.

Key words: Speech/music discrimination, segmentation, wavelets, static parameters, dynamic parameters, long-term parameters

1 Introduction

This paper addresses the problem of parameterization for speech/music discrimination. We propose to take into account the difference between music and

∗. Corresponding author: I. Illina
Email address: emmanuel.didiot@gmail.com, illina@loria.fr, fohr@loria.fr, mella@loria.fr (E. Didiot, I. Illina, D. Fohr, O. Mella).

ACCEPTED MANUSCRIPT

speech at the parameter level: a combination of time and frequency features that deal with non-stationary signals will be used. The proposed approaches were evaluated on several real-world corpora extracted from radio programs. These corpora contain a lot of superimposed segments, such as speech with music or songs with a "fade-in fade-out" effect.

In real world applications, automatic speech recognition systems (ASRs) are faced with a large diversity of audio signals: speech, music, noise as well as their superimpositions. The performance of standard ASRs usually decreases drastically when they are confronted with this kind of mixed condition. During the automatic speech recognition step, a wide variety of environment adaptation and compensation approaches can be used to treat the differences between training and testing conditions [21]. On the other hand, these techniques are not powerful enough in the case of mixed speech/music, because they only take into account the specificity of speech and are not appropriate for music. In these situations a preprocessing step is necessary before recognition. The basic principle of speech/music discrimination consists in segmenting the signal into homogeneous parts and in classifying each part in predefined categories like speech, music, speech superimposed on music (called *speech over music*). Sometimes more precise categories can be used for music, such as instrumental music, songs, etc. [12], [46]. The music segments are then discarded, to avoid recognition mistakes and the speech over music segments can be used to perform powerful compensation or adaptation. For example, speech/music detection could speed up the process of automatic captioning of TV transmissions by skipping the non-speech segments and avoiding incorrect transcriptions during music, songs or jingle segments. Another realistic application of speech/music discrimination is its ability to give interesting information about the type of music for indexing and retrieval of audio documents. Thus, the development of speech/music discrimination methods has become an important research area.

Speech/music discrimination differs from Voice Activity Detection (VAD). VAD aims to discriminate between noise and speech and not between speech and music. More particularly, VAD is not able to discriminate speech from songs.

Figure 1 illustrates the differences between speech and music signals. A wide variety of parameterization techniques has been used for speech/music discrimination. They can be divided into three classes according to the domain in which they are computed: the time, frequency or mixed (time and frequency) domain.

Time-domain features represent the temporal characteristics of the signal. For example, the zero crossing rate (ZCR) [41], [42], [34] can detect unvoiced parts of the audio signal. During speech there is an alternation of voiced

ACCEPTED MANUSCRIPT

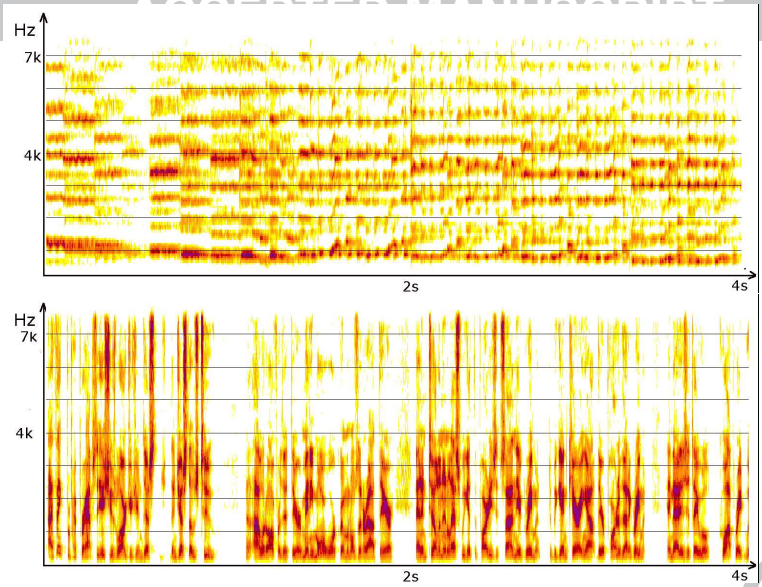


FIG. 1 –. Example of signals: music signal (Vivaldi excerpt, above) and speech signal (below).

and unvoiced segments. ZCR is greater during unvoiced segments than voiced segments. So, peaks occur in the evolution of the ZCR during speech. For music, the variations of the ZCR are smoother.

Frequency-domain features characterize the spectral envelope of the signal. Some examples are spectral centroid [42], harmonic coefficients [6], [49] and spectral peak track [51], [43]. The *Mel Frequency Cepstral Coefficients* parameters (MFCC), which could be classified in this category, are considered as one of the best parameterizations for speech/music discrimination [4], [5], [2], [13], , [16], [15], [18], [44], [19], [29], [37], [38].

Combinations of time and frequency features are for instance the spectral flux [30], [42] or the 4Hz modulation energy [42], [35]. The spectral flux detects the harmonic continuity in music. The high variations of spectral flux are specific for speech. This is due to the alternation of consonants and vowels. The 4Hz modulation energy is more specific for speech than for music, because it corresponds to the syllabic rate.

Concerning the classification step, most systems are based on Gaussian Mixtures Models (GMM) or Hidden Markov Models (HMM). Nevertheless, some systems use other speech/music classifiers, such as Multi-Layer Perceptron [22], [24], Maximum A Posteriori classifier [42], k-Nearest Neighbors [42], and different hybrid systems: MLP/SVM (Support Vector Machine) [14], MLP/HMM [1].

ACCEPTED MANUSCRIPT

This article presents a new parameterization approach for speech/music discrimination based on the wavelet decomposition of the signal. Our goal is not to propose a new wavelet type but to apply the wavelet formalism for speech/music discrimination. Our motivation to apply wavelets to speech/music discriňmination is due to their ability to extract time-frequency features and to deal with non-stationary signals. Kahn and al. Earlier, [22] proposed the wavelet parameterization for speech/music detection. But he used only two values per frame to perform speech/music classification: the mean and the variance of the discrete wavelet transform coefficients. In our work, we use the wavelet coefficients in each frequency band of every frame, so a more accurate analysis can be performed. We study several features based on wavelet decomposition and test them on some broadcast programs. Furthermore, we compare their performance with MFCC because studies [5], [2], [29] have showed that the latter achieve state-of-the-art results in speech/music discrimination. Besides, many automatic news transcription systems use MFCC-based parameterization for speech/music segmentation in different evaluation campaigns, like the DARPA evaluation (1997-2000) or the recent ESTER campaign (2003-2005) [18]. We refer to the systems designed by Cambridge (HTK) [44], LIA [13], LIMSI [16] and LORIA (ANTS, *Automatic News Transcription System*) [4].

To perform the classification we chose a "class/non-class" approach: a speech/non-speech segmentation and a music/non-music segmentation [35]. This approach allows us to determine the best parameters for each task and to increase the accuracy. The classification method is based on the Viterbi algorithm which uses HMM models (HTK toolkit [50]), because it simultaneously performs classification and segmentation.

The paper is organized as follows. First, the wavelet decomposition and the wavelet-based parameters are briefly introduced in section 2. Then, our speech/music discrimination system is presented in section 3. Next, experimental results obtained for speech/music discrimination on various corpora are discussed in section 4, followed by a conclusion in section 5.

2 Wavelet-based Parameters for Speech/Music Discrimination

In this section, we introduce our parameterization method based on wavelet transforms. The signal is first analyzed using the wavelet transform, then different energy parameters are calculated. As the purpose of this article is not wavelet signal analysis but only its use for speech/music discrimination, we shortly introduce the wavelet transforms.

ACCEPTED MANUSCRIPT

For speech/music discrimination, it is essential to deal with non-stationary signals and to achieve variable time and frequency localization of acoustic cues. *Multi-resolution Analysis* (MRA) is a signal analysis, which provides a time-frequency representation of the signal, well suited for non-stationary signals [31], [32]. MRA analysis offers an alternative to the more traditional Short-Time Fourier Transform (STFT). The problem with STFT is that the shorter the analysis window is, the better the time resolution, but the poorer the frequency resolution. This means that STFT is facing the resolution problem, e.g. which window size to use. The solution of this problem is often application dependent. In contrast, MRA analyses the signal at different frequencies with different resolutions and is well adapted for non-stationary signals. Indeed, MRA makes sense especially when the signal has many high frequency components for short durations and low frequency components for long durations, which is often the case for speech and music signals.

In our work, we chose a specific case of MRA: *Discrete Wavelet Transform* (DWT). DWT provides a compact representation of the signal, has a rich set of basis functions and can be implemented very efficiently. Wavelet-based signal analysis has been successfully applied to various problems, such as image size reduction [39], speech denoising [26], automatic speech recognition [7], [40] and audio classification [28], [46].

A DWT can be derived from a Continuous Wavelet Transform (CWT). Given a time signal $x(t)$, the continuous wavelet transform is given by:

$$CWT(r,s) = \frac{1}{\sqrt{|s|}} \int x(t)\Psi^*(\frac{t-r}{s})dt \tag{1}$$

where $*$ is the conjugate operator. $\Psi(t)$ is a time function called "mother wavelet", r ($r \geq 0$) is related to the time location of the analyzing window and s corresponds to scale (scale $s < 1$ dilates the analysis function, scale $s > 1$ compresses the analysis function. By varying r and s, the "mother wavelet" is scaled and shifted. Several "mother wavelets", called wavelet families, have been proposed.
Using the dyadic decomposition ($s = 2^j$, cf. Figure 2), and a discrete signal $x[m]$,$m = 0$,..,$N-1$, a CWT is transformed into a DWT:

$$DWT[n,2^j] = \sum_{m=0}^{N-1} x[m]\Psi^*_{2^j}[m-n] \tag{2}$$

where

$$\Psi_{2^j}[n] = \frac{1}{\sqrt{2^j}}\Psi(\frac{n}{2^j}) \tag{3}$$

ACCEPTED MANUSCRIPT

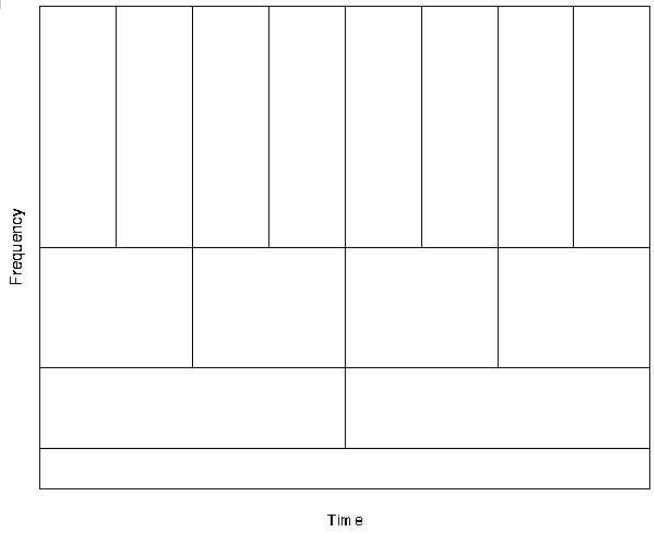


FIG. 2 –. Example of a dyadic time-frequency decomposition.

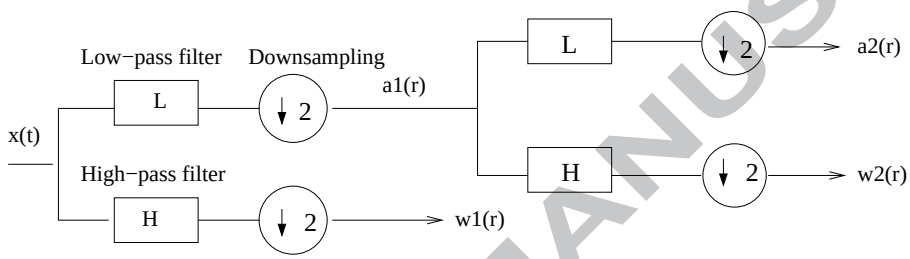


FIG. 3 –. DWT with two decomposition levels. $a_1(r)$,$a_2(r)$ are the approximation coefficients, $w_1(r)$,$w_2(r)$ the wavelet coefficients.

The DWT provides a rough approximation of the Mel scale and can be computed efficiently using a fast, pyramidal algorithm related to a multi-rate filterbank: S. Mallat [32] has shown that frequency band decomposition can be obtained by successive low-pass (L) and high-pass (H) filterings of the signal in the time domain. Figure 3 illustrates a decomposition with two levels. The symbol "$\downarrow$ 2" denotes a down-sampling by 2. This figure illustrates that at each level j, the signal is decomposed into *approximation* coefficients $a_j(r)$ (output of low-pass filter) and *detail* coefficients $w_j(r)$ (output of high-pass filter). Approximation coefficients correspond to local averages of the signal. Detail coefficients, named also *wavelet coefficients*, can be viewed as the differences between two successive local averages, ie. between two successive approximations of the signal [33]. The index j corresponds to the frequency band.

Our work on DWT is based on the Daubechie, Symlet and Coiflet families because these wavelets are some of the best known wavelets and have been successfully used for speech recognition [8], [17]. Daubechie and Symlet wavelet families correspond to FIR filters (L,H). Daubechie and Symlet wavelet

ACCEPTED MANUSCRIPT

families have an interesting property: they have a minimum support[1] for a given number of vanishing moments. Small support size allows better singularity detection. The definition of vanishing moments will be provided in section 4.3.1.

For speech/music discrimination, we propose to use only wavelet coefficients $w_j(r)$ to analyze the acoustic signal, because they can capture the sudden modifications of the signal.

2.2 Energy-based Parameters

The energy distribution in each frequency band is a very relevant acoustic cue. For this reason we employ energy, calculated from DWT, as a speech/music discrimination feature.

Let, as below, $w_j(r)$ denote the wavelet coefficient at time position r and frequency band j. We underline that the frequency band decomposition and time decomposition correspond to the dyadic scale (see Figure 2): time resolution halves while the frequency resolution doubles. If N is the length of the analysis window, $w_j(r)$ has $N_j = N/2^j$ samples[2] and three methods are investigated for extracting the wavelet energies:

– *Instantaneous Energy* (labelled **E** in Tables) gives the energy distribution in each band:

$$f_j^E = \log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}(w_j(r))^2\right) \tag{4}$$

– *Teager Energy* (labelled **T_E** in Tables) was recently applied for speech recognition [36], [11]:

$$f_j^{T_E} = \log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j-1}|(w_j(r))^2 - w_j(r-1) * w_j(r+1)|\right) \tag{5}$$

The discrete Teager Energy Operator (TEO), introduced by Kaiser [23], allows modulation energy tracking and gives a better representation of the formant information in the feature vector compared to MFCC. The Teager energy is a noise robust parameter for speech recognition because the effect of additive noise is attenuated: good results are obtained in presence of car

1. The scaling function is compactly supported if and only if the filter L has a finite support.

2. For instance, using 5 bands on 512 samples window, $N_1 = 256$, $N_2 = 128$, $N_3 = 64$, $N_4 = 32$ and $N_5 = 16$.

ACCEPTED MANUSCRIPT

engine noise [20]. The Instantaneous energy reflects only the amplitude of the signal whereas the Teager energy operator reflects the variations in both amplitude and frequency of the signal [45].
Figure 4 is an example of two spectrograms: one based on wavelet coefficients (Coiflet, 5 bands, Teager energy) and the other based on STFT coefficients for the same signal. The variations of energy in each frequency band are greater for speech than for music. This can be observed for STFT parameters as well as for wavelet parameters.

– *Hierarchical Energy* (labelled **H_E** in Tables), used in automatic speech recognition to parameterize the signal [17], [27]. We wanted to assess the idea presented by Kryze [27]. It provides a hierarchical time resolution and gives more importance to the center of the analysis window:

$$f_j^{H_E} = \log_{10}\left(\frac{1}{N_J}\sum_{r=(N_j-N_J)/2}^{(N_j+N_J)/2}(w_j(r))^2\right) \quad (6)$$

J corresponds to the lowest band.

After energy calculation, we decided not to perform a DCT (*Discrete Cosinus Transform*), like for MFCC, because we want to keep the interpretation of coefficients as frequency band energies.

3 Speech/Music Discrimination System

3.1 System Description

The chosen classification approach is a "class/non-class" one. In other words, class detection is performed by comparing a class model and a non-class model estimated on the same representation space. Two classification systems are implemented: speech/non-speech and music/non-music. By taking the "class/non-class" approach, we will be able to optimize the parameterization separately for each classification system. The decisions of both classification systems are merged and the audio signal is segmented into four categories: speech (S), music (M), speech over music (SM) and silence/noise (N) (cf. Table 1). Figure 5 shows the architecture of our speech/music discrimination system.

According to [42], the choice of classifier (GMM, HMM, NN, etc.) is not important for this kind of discrimination task. Therefore, we decided to choose a stochastic classifier. A GMM model containing between 8 and 64 Gaussians per state is trained to model each class. A frame by frame decision would lead

ACCEPTED MANUSCRIPT

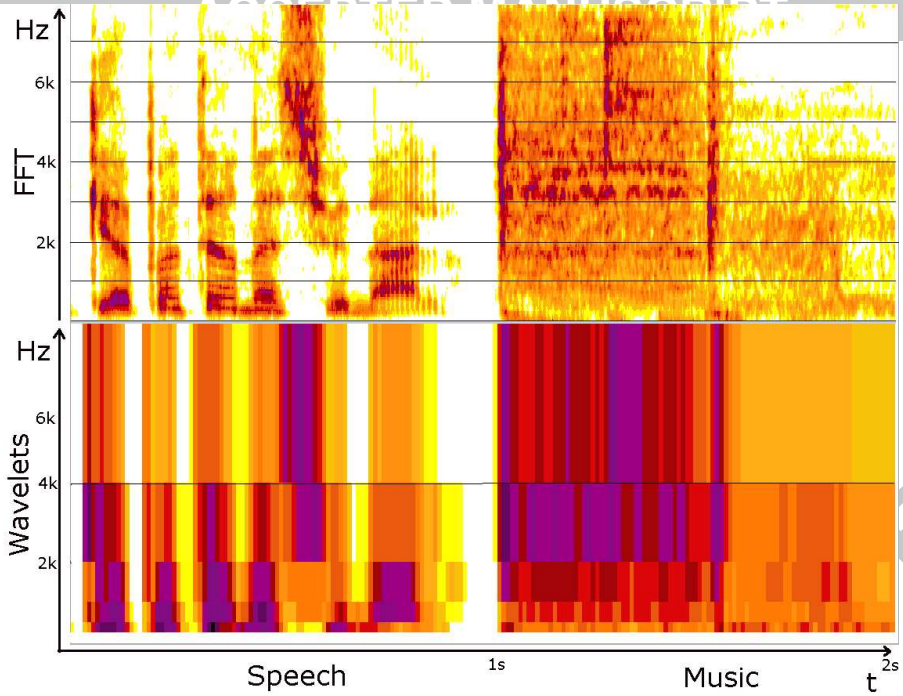


FIG. 4 –. Above: spectrogram based on STFT (128 frequency bands, frame size 32ms), below: spectrogram based on Coiflet, (5 bands, Teager energy), for a 2s signal containing speech during the first part and music during the last one.

S/NS classifier	M/NM classifier	Final decision
Speech	Non-Music	Speech
Speech	Music	Speech over Music
Non-Speech	Music	Music
Non-Speech	Non-Music	Silence/Noise

TAB. 1 –
Final discrimination results for a segment using two classifiers: speech/non-speech and music/non-music.

to unrealistic 10ms segments. To avoid this, for each recognized segment a 0.5s minimal duration is imposed by concatenating 50 GMMs[3]. This gives an HMM model with 50 states. The Viterbi algorithm provides the best model sequence, describing the audio signal.

3. A duration of 0.5 seconds is chosen because we assume that a speech segment contains at least one word and consequently, lasts at least 0.5 seconds.

ACCEPTED MANUSCRIPT

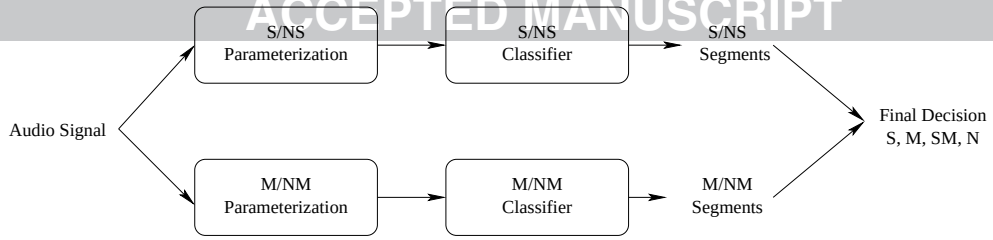


FIG. 5 –. Architecture of our speech/music discrimination system.

3.2 Evaluation

To evaluate our different features, three error rates are computed:

- Music/Non-Music classification error rate (labelled M/NM in the Tables). Music/non-music segmentation could be useful for audio indexing.
- Speech/Non-Speech classification error rate (labelled S/NS in the Tables). Speech/non-speech detection is useful for discarding the non-speech segments when performing the automatic transcription of broadcast programs.
- Global classification error rate (labelled GR in the Tables). Global rate can evaluate the quality of the segmentation system, because this measure takes into account all kinds of segmentation errors. The global error rate corresponds to a more difficult task: we have to segment the audio signal into 4 classes: speech, music, speech over music, other. For S/NS and M/NM tasks there are only 2 kinds of segments, so discrimination is easier and the error rate is smaller. Let n_z^y be the number of frames recognized as z having label y, and T the total number of frames. The global error rate is computed as follow:

$$100 * (1 - (n_{SM}^{SM} + n_M^M + n_S^S + n_N^N)/T)$$

4 Experiments and Results

4.1 Parameterization

The signal is sampled at 16kHz. After pre-emphasis, the following parameters are computed on a 32ms Hamming window with a 10ms shift. 32ms is a commonly used window duration in many ASR systems. We used two types of features:

- *Baseline MFCC features.* 12 MFCC coefficients including C_0 (computed from 24 triangular filters) with their first and second derivatives are computed. This parameterization is the most usual in speech recognition. Finally,

ACCEPTED MANUSCRIPT

a vector of 36 components is obtained. These parameters were chosen as baseline because they have achieved very good performance for speech/music discrimination (cf. section 1).

– *Wavelet-based features.* The energy features, described in section 2.2, are calculated on wavelet coefficients obtained with different wavelet families: Daubechie, Coiflet and Symlet. As previously mentioned, these wavelet families are the most popular ones and have been utilized for speech recognition. Let us point out that we use only detail coefficients. Multi-resolution parameters are computed for different decomposition levels, i.e. for different numbers of frequency bands.

4.2 Database Description

All the following corpora are manually segmented into speech/non-speech and music/non-music. Silence and background noise segments are labelled as non-speech and non-music.

4.2.1 Training Corpus

The training corpus is composed of two parts: "Audio CDs" and "Broadcast programs". The "Audio CDs" corpus (2 hours) is made up of several tracks of instrumental music (jazz, electronic music and classical music) and songs (rock and pop) extracted from CDs. The "Broadcast programs" corpus (4 hours 20mn) contains programs from the French radio: broadcast news as well as interviews and musical programs.

4.2.2 Test Corpora

We carried out test experiments on three entirely different corpora:

– We use only the test part of *Scheirer* corpus built by E. Scheirer and M. Slaney [42]. All audio files are homogeneous and have the same duration of 15 seconds: 20 files of broadband or telephone speech, 21 files of music and 20 files of vocals. Note that this test part does not contain speech with music in background. The audio is recorded from an FM tuner in San Francisco Bay Area using a variety of stations, styles and noise levels. The music styles are more various (jazz, pop, country, etc.) than in the *Entertainment* corpus (see below). Vocals (singing) are labeled as music. This corpus is composed of 32% speech frames and 68% music frames. This corpus allows us to evaluate our new parameterizations on a corpus which has been used in previous studies [42], [48], [3]. We don't exploit the file

ACCEPTED MANUSCRIPT

homogeneity information and our discrimination system can split a file into different segments.

Let us note that compared to [42], the cross-validation testing framework is not used here: only the test part of Scheirer data is used to build this test corpus and our models are trained as explained in 4.2.1. The confidence interval is ±1% at a 0.05 signifiance level for about 5% error rate.

– The *News* corpus consists of three 1-hour files of French radio stations *France-Inter* and *Radio France International* and contains mainly speech or speech over jingles (86% speech, 11% speech over music and 3% music). This corpus is interesting in the way that our speech/music discrimination system can be evaluated on a broadcast news transcription task. The confidence interval is ±0.5% for about 10% error rate.

– The *Entertainment* corpus is composed of three 20-minutes shows (interviews and musical programs). It was recorded and given to us by a French radio station. This corpus is considered as quite difficult. Indeed, there are a lot of superimposed segments, such as speech with music or songs with an effect of "fade-in fade-out". Moreover, it contains an alternation of broadband speech and telephone speech and some interviews are very noisy. It is made up of 52% speech frames, 18% speech over music frames and 30% music frames. The confidence interval is ±1% for about 20% error rate.

As the three test corpora are very different (different kind of radio programs), more often than not, experimental results will be presented corpus by corpus.

4.3 Experimental Results and Discussion

As our goal was to study the relevance of wavelet parameterization for speech/music discrimination, we began our experiments by determining the best wavelets: wavelet type, number of vanishing moments and number of decomposition bands.

We then assessed the performance of the three energy parameters computed from the wavelet coefficients for each segmentation task and we compared these results with the ones obtained by the MFCC baseline segmentation system. Besides, we compared our parameters with 4Hz modulation energy, because according to Scheirer [42] and Pinquier [35] the 4Hz modulation was one of the best parameters for speech/music discrimination.

After evaluating static wavelet parameters, we tested dynamic parameters [10]. Indeed, several studies [42], [47] demonstrated that dynamic features allow to efficiently take into account the specificity of the speech and music structure. The main conclusion of Scheirer's study was that the variance of the parameters give better results than the parameters themselves. [25] also concluded

ACCEPTED MANUSCRIPT

that variance of MFCC parameters is a relevant feature. Indeed, this kind of long-term parameter should capture the rhythm differences between speech and music. For these reasons, we studied the variance of wavelet parameters [9].

4.3.1 Effect of Wavelet Type and the Number of Vanishing Moments

The goal of our first experiment was to study the influence of different families of wavelets (Daubechie, noted as *db* in the Tables, Coiflet, noted as *coif*, Symlet, noted as *sym*) and the number of vanishing moments of the mother wavelets that generated these families. The mother wavelet has p vanishing moments if:

$$\int_{-\infty}^{+\infty} t^k \Psi(t)\ dt = 0, \text{ for } 0 \leq k < p \tag{7}$$

This means that $\Psi(t)$ is orthogonal to any polynomial of degree $p-1$. So, if the signal is well approximated by a Taylor polynomial of degree k, and $k < p$ then the wavelet coefficients at fine scales have a small amplitude [32]. This property is useful to detect abrupt transitions: wavelet coefficients will be larger during a transition.

For this preliminary experiment, we chose to limit our study to static parameters: instantaneous energy and 5 bands. The corresponding frequency limits are [8000-4000], [4000-2000], [2000-1000], [1000-500], [500-250] Hz. To simplify their interpretation, the results are presented on all test corpora together in terms of speech/non-speech and music/non-music error rates.

Table 2 indicates that the best results were obtained with the smallest number of vanishing moments, especially for the music/non-music discrimination task. With a small number of vanishing moments, abrupt transitions give large wavelet coefficients. So the alternation vowel/fricative or vowel/plosive can be better detected and speech/music discrimination is more accurate.

Another conclusion that can be drawn from this Table is that the different wavelet families (Daubechie, Coiflet, Symlet) achieved similar performance when there is a low number of vanishing moments.

As the three wavelet families gave similar performance and in order to reduce the experimental part, we chose to only use Daubechie (*db-2*) and Coiflet (*coif-1*) wavelets in the following experiments.

4.3.2 Static Parameters

In this experiment, static features based on wavelets were studied. More precisely, we evaluated different decomposition levels (number of bands) and

ACCEPTED MANUSCRIPT

Wavelet Type	NbVanishMom	M/NM	S/NS
db-2	2	11.6	4.9
db-4	4	15.8	**4.6**
db-8	8	16.8	5.0
db-12	12	19.0	5.6
coif-1	2	**11.5**	4.8
coif-3	6	16.3	4.9
coif-5	10	19.0	5.6
sym-2	2	11.6	4.9
sym-4	4	16.0	**4.6**
sym-8	8	16.7	5.2

TAB. 2 –
Discrimination results with varying wavelet types and number of vanishing moments. Wavelets with 5 bands and instantaneous energy. Frame error rate in percentages. Scheirer, News *and* Entertainment *corpora.*

different energies: instantaneous (labelled E in the Tables), Teager (labelled T_E) and hierarchical (labelled H_E) energies. As said in the previous section, we used only Daubechie and Coiflet wavelets. Two decomposition levels were evaluated: 5 and 7 because a preliminary study showed that best classification results were achieved with 5 and 7 decomposition bands.

The experimental results for speech/non-speech and music/non-music discrimination for each test corpus are presented in Tables 3 and 4. Several conclusions can be drawn:

– **Wavelets/MFCC**
For speech/non-speech discrimination, the performance of static wavelet features proposed in this paper is comparable to the performance of baseline MFCC features for *Scheirer* and *News* corpora (cf. Table 3). But, wavelet features outperform MFCC features for the most difficult corpus (*Entertainment*) which contains a lot of superimposed segments (speech over music). For the music/non-music discrimination task, wavelet-based parameters are significantly better than MFCC ones (cf. Table 4) for all three corpora. This confirms our hypothesis that wavelet coefficients are better than MFCC for dealing with non-stationary signals.

We can notice that wavelet features have a more compact representation. Indeed, similar or better results are obtained with a 5- or 7-component vector for wavelet parameterization and with 36-component vector for MFCC.

– **Coiflet/Daubechie**

ACCEPTED MANUSCRIPT

Because it is difficult to predict which wavelet family is more suitable for a given task, we evaluated Coiflet and Daubechie for the two tasks. The two wavelet families obtained similar performance.

– **Energies**
For speech/non-speech, Teager Energy features provided slightly better discrimination for all corpora. This can be explained by the fact that Teager Energy has the ability to compensate additive noise [20]. So, speech over music segments can be better classified. On the other hand, for music/non-music, no clear conclusion can be drawn.

– **Number of bands**
For corpora containing a lot of music (*Scheirer*) or speech over music (*Entertainment*) it is better to use 7 bands for the music/non-music discrimination. In the low frequency (7th) band, on average less energy can be found for pure speech compared to music. So, using 7 bands is useful for music/non-music discrimination.

Wavelet	**NbBands**	**NbPar**	**Energy**	**Scheirer**	**News**	**Enter**
MFCC+Δ+ΔΔ		36	–	**2.5**	**2.9**	*5.8*
db-2	5	5	E	3.3 (-32%)	3.6 (-24%)	4.3 (26%)
db-2	5	5	T_E	3.3 (-32%)	3.2 (-10%)	**4.2 (28%)**
db-2	5	5	H_E	3.2 (-28%)	4.6 (-59%)	4.3 (26%)
db-2	7	7	E	3.3 (-32%)	6.5 (-124%)	6.9 (-19%)
db-2	7	7	T_E	3.3 (-32%)	6.4 (-121%)	5.9 (-2%)
db-2	7	7	H_E	3.3 (-32%)	7.6 (-162%)	5.9 (-2%)
coif-1	5	5	E	3.3 (-32%)	3.7 (-28%)	**4.2 (28%)**
coif-1	5	5	T_E	3.3 (-32%)	3.2 (-10%)	**4.2 (28%)**
coif-1	5	5	H_E	3.3 (-32%)	4.4 (-52%)	4.3 (26%)
coif-1	7	7	E	3.3 (-32%)	7.4 (-155%)	6.8 (-17%)
coif-1	7	7	T_E	3.6 (-44%)	6.4 (-121%)	6.1 (-5%)
coif-1	7	7	H_E	3.3 (-32%)	7.6 (-162%)	6.6 (-14%)

Tab. 3 –
Speech/non-speech discrimination results using wavelets db-2 *and* coif-1*, 5 and 7 bands. Frame error rate in percentages. Relative improvement rates compared to MFCC are presented in parentheses.*

In this section, we studied the relevance of static wavelet parameters according to different families, energy features and number of decomposition bands. In accordance with the results presented here, in the following experiments we

ACCEPTED MANUSCRIPT

Wavelet	NbBands	NbPar	Energy	Scheirer	News	Enter
MFCC+Δ+ΔΔ		36	–	*6.5*	*13.1*	*23.1*
db-2	5	5	E	5.3 (18%)	8.3 (37%)	15.9 (31%)
db-2	5	5	T_E	5.4 (17%)	7.9 (40%)	17.0 (26%)
db-2	5	5	H_E	5.1 (22%)	7.2 (45%)	19.2 (17%)
db-2	7	7	E	4.3 (34%)	11.4 (13%)	**13.3 (42%)**
db-2	7	7	T_E	**3.7 (43%)**	10.1 (23%)	14.0 (39%)
db-2	7	7	H_E	**3.7 (43%)**	10.8 (18%)	13.8 (40%)
coif-1	5	5	E	5.3 (18%)	7.8 (40%)	16.5 (29%)
coif-1	5	5	T_E	5.6 (14%)	8.0 (39%)	17.0 (26%)
coif-1	5	5	H_E	5.3 (18%)	**7.0 (47%)**	18.5 (20%)
coif-1	7	7	E	4.3 (34%)	11.4 (13%)	14.5 (37%)
coif-1	7	7	T_E	**3.7 (43%)**	10.1 (23%)	14.6 (37%)
coif-1	7	7	H_E	**3.7 (43%)**	10.9 (16%)	14.8 (36%)

Tab. 4 –
Music/non-music discrimination results using wavelets db-2 *and* coif-1 *with 5 and 7 bands. Frame error rate in percentages. Relative improvement rates compared to MFCC are presented in parentheses.*

restricted the studied parameters to one wavelet family (Coiflet) and to one number of decomposition bands for each task (5 bands for speech/non-speech, 7 bands for music/non-music).

4.3.3 Comparison between the Wavelet-based Parameters and the 4Hz Modulation Parameter

The goal of this section is to compare the performance of the 4Hz modulation parameter and the wavelet-based parameters because 4Hz modulation yielded a good speech/music discrimination. In our work, the 4Hz modulation parameter was computed as follows:

- Speech signal is segmented into 16ms windows without overlapping;
- Mel filter bands are extracted with FFT;
- Each frequency band is filtered with a band-pass filter centered at 4 Hz;
- After this, all filter channels are added and the variance is computed on a 1-second window.

ACCEPTED MANUSCRIPT

Table 5 shows that:

- For speech/non-speech discrimination, wavelet parameters obtain better results than 4Hz modulation parameter;
- For music/non-music task, 4Hz energy works well on the *Scheirer* and *News* corpora but does not obtain good results on *Entertainment* corpus. In this last case the errors are due to the fact that speech over music segments or speech with background noise are misclassified as music segments. An unique parameter (like 4Hz modulation) cannot capture the variability of speech over music or speech with background noise.

Parameterization	**NbPar**	**Scheirer**	**News**	**Enter**
Speech/non-speech				
4Hz modulation	1	5.8	8.4	27.7
coif-1	5	**3.3 (43%)**	**3.7 (127%)**	**4.2 (560%)**
Music/non-music				
4Hz modulation	1	**1.6**	**8.6**	24.2
coif-1	7	4.3 (-63%)	11.4 (-32%)	**14.5 (40%)**

TAB. 5 –
Speech/non-speech and music/non-music discrimination results using wavelet-based (coif-1 E with 5 or 7 bands) and 4Hz modulation parameters. Frame error rate in percentages. Relative improvement rates compared to 4Hz modulation are presented in parentheses.

4.3.4 Dynamic Parameters

In order to study how the discrimination rates depend on the dynamic features, the first (Δ) and second ($\Delta\Delta$) derivatives of the wavelet-based parameters were computed. Tables 6 and 7 present the frame error rate for each corpus separately for dynamic parameters only and for static and dynamic parameters.

Table 6 shows that, for speech/non-speech discrimination, the dynamic coefficients alone are better than the static ones for all corpora and all energy types (except in one case: with Teager energy on the *News* corpus). This means that dynamic parameters are more discriminant than static ones. This is perhaps due to the fact that the variations of speech parameters are specific, for instance, to the alternation vowel-consonant. According to Table 7, for the music/non-music task, the dynamic parameters seem to be more discriminant than static ones on *Scheirer* and *News* corpora. This is not the case for the *Entertainment* corpus. One reason could be the fact that there are more music and speech over music in this corpus than in the other corpora.

ACCEPTED MANUSCRIPT

Tables 6 and 7 also show that the addition of first derivatives (Δ) improves the results compared to static parameters. For instance, using Teager energy (coif-1 with 5 bands) for speech/non-speech discrimination, a relative significant gain of 48% for the *Scheirer* corpus, 16% for the *News* corpus and 31% for *Entertainment* corpus is obtained compared to the static features. For music/non-music discrimination, using Teager energy (coif-1 with 7 bands), a relative significant gain of 51% for the *Scheirer* corpus and 29% for the *News* corpus is obtained compared to the static features. For the *Entertainment* corpus no improvement is observed.

On the contrary, addition of the second derivatives ($\Delta\Delta$) does not improve the results compared to the addition of the first derivatives. We can even see a decrease in the performance for the music/non-music task. We attribute this slight decrease to the nature of $\Delta\Delta$ coefficients. One possible explanation could be that $\Delta\Delta$ coefficients have a high variability and depend on the type of music. So, if the type of music occuring in the test files has not been encountered in the training files, the $\Delta\Delta$ coefficients are not useful and will add "noise" to the models.

In conclusion, the important result of this section is that combining the derivatives with the static wavelet parameters outperforms MFCC results for all corpora and for both segmentation tasks.

4.3.5 Long-Term Parameters

The study of long-term parameters such as variance on a large window (between 1 and 2.5 second duration) seems interesting [42], [47], [48], [25]. We conducted experiments in order to optimize the window duration for the computation of the variance. The best result was obtained for a 1-second window size. We applied this 1-second variance to static coefficients: MFCC and energy features based on the coif-1 wavelet family.

To study the behavior of wavelet variance parameter, we computed the histogram of the variance of the Teager energy computed in the third band, using wavelet coif-1 with 5 bands on the training corpus (cf. Figure 6). For the other bands, the shapes are similar. As expected the variance for speech segments is greater than the variance for music segments because of the alternation of vowel-consonant. The curve corresponding to speech over music segments overlaps the speech and music curves. This explains why it is difficult to discriminate speech over music segments.

Tables 8 and 9 present the discrimination error rates provided only by variance of the parameters and by combining the variance with static parameters. For speech/non-speech discrimination, short term dynamic parameters Δ (cf. Table 6). are better than the long term parameters (cf. Table 8). For

ACCEPTED MANUSCRIPT

Parameters	NbPar	Scheirer	News	Enter
E	5	3.3	3.7	4.2
Δ E	5	**1.7 (48%)**	3.5 (5%)	3.4 (19%)
E+Δ	10	3.0 (9%)	2.7 (27%)	3.0 (29%)
E+Δ+ΔΔ	15	**1.7 (48%)**	**2.6 (30%)**	3.2 (24%)
T_E	5	3.3	3.2	4.2
Δ T_E	5	**1.7 (48%)**	3.8 (-19%)	3.3 (21%)
T_E+Δ	10	**1.7 (48%)**	2.7 (16%)	2.9 (31%)
T_E+Δ+ΔΔ	15	**1.7 (48%)**	2.7 (16%)	**2.8 (33%)**
H_E	5	3.3	4.4	4.3
Δ H_E	5	**1.7 (48%)**	3.2 (27%)	3.4 (21%)
H_E+Δ	10	**1.7 (48%)**	2.8 (36%)	3.2 (26%)
H_E+Δ+ΔΔ	15	**1.7 (48%)**	2.9 (34%)	3.3 (23%)

Tab. 6 –
Speech/non-speech discrimination results using wavelets coif-1 *with 5 bands and dynamic parameters (Δ, ΔΔ). Frame error rate in percentages. Relative improvement rates compared to static parameters are presented in parentheses.*

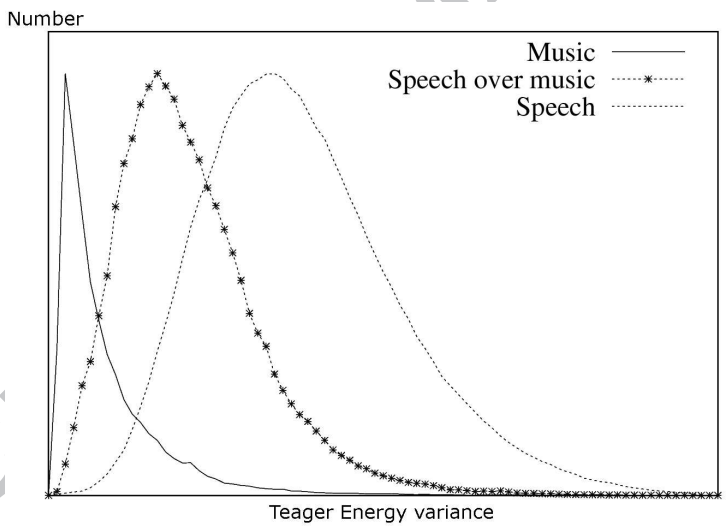


Fig. 6 –. Histogram of 1s variance of the Teager energy of the third band using wavelet coif-1 with 5 bands.

music/non-music task, according to Table 9, the variance parameters give similar results than Δ parameters (cf. Table 7) on the *Scheirer* and *News* corpora, and better results on the *Entertainment* corpus.

ACCEPTED MANUSCRIPT

Parameters	NbPar	Scheirer	News	Enter
E	7	4.3	11.4	**14.5**
Δ E	7	**1.8 (58%)**	8.1 (29%)	18.1 (-25%)
E+Δ	14	**1.8 (58%)**	7.9 (31%)	15.2 (-5%)
E+Δ+ΔΔ	21	**1.8 (58%)**	9.5 (17%)	17.4 (-20%)
T_E	7	3.7	10.1	14.6
Δ T_E	7	3.4 (8%)	**6.3 (38%)**	18.2 (-25%)
T_E+Δ	14	**1.8 (51%)**	7.2 (29%)	15.0 (-3%)
T_E+Δ+ΔΔ	21	**1.8 (51%)**	9.7 (4%)	17.4 (-19%)
H_E	7	3.7	10.9	14.8
Δ H_E	7	3.4 (8%)	8.8 (19%)	20.4 (-38%)
H_E+Δ	14	**1.8 (51%)**	7.2 (34%)	14.8 (0%)
H_E+Δ+ΔΔ	21	**1.8 (51%)**	8.6 (21%)	18.3 (-24%)

TAB. 7 –
Music/non-music discrimination results using wavelets coif-1 *with 7 bands and dynamic parameters (Δ, ΔΔ). Frame error rate in percentages. Relative improvement rates compared to static parameters are presented in parentheses.*

Tables 8 and 9 show that static plus variance parameters do not give any improvement compared to variance parameters. Moreover, for *Entertainment* corpus, a small degradation is observed.

All these results point out that generally Δ parameters are better than long term parameters.

4.3.6 Global Discrimination

This experiment aims to discriminate speech, music, speech over music and silence/noise. As we said previously (see section 3.2) global discrimination is a difficult task and allows to evaluate the quality of the segmentation system, because this measure takes into account all kinds of segmentation errors. This is obtained by performing speech/non-speech discrimination, then music/non-music discrimination, and finally taking into account these results to calculate a global discrimination rate (see section 3.2). For each discrimination task, we used the features giving the best discrimination results in the previous experiments, i.e. *coif-1* with 5 bands for speech/non-speech discrimination and *coif-1* with 7 bands for music/non-music discrimination. In the previous experiments, the three energy types reached almost the same performance.

ACCEPTED MANUSCRIPT

Parameters	NbPar	Scheirer	News	Enter
MFCC+Δ+ΔΔ	36	*2.5*	**2.9**	*5.8*
Var of MFCC	12	2.2 (12%)	4.1 (-41%)	8.1 (-40%)
MFCC+(Var of MFCC)	24	3.4 (-36%)	4.3 (-48%)	10.4 (-79%)
Var of E	5	**1.7 (32%)**	3.9 (-34%)	**3.7 (36%)**
Var of T_E	5	**1.7 (32%)**	4.0 (-38%)	**3.7 (36%)**
Var of H_E	5	**1.7 (32%)**	4.2 (-45%)	4.1 (29%)
E+(Var of E)	10	2.1 (16%)	4.2 (-44%)	4.2 (28%)
T_E+(Var of T_E)	10	**1.7 (32%)**	4.1 (-41%)	4.1 (29%)
H_E+ (Var of H_E)	10	2.1 (16%)	4.5 (-55%)	5.1 (12%)

Tab. 8 –
Speech/non-speech discrimination results using variance on a 1-second window and static with variance coefficients for wavelet coif-1 *and 5 bands. Frame error rate in percentages. Relative improvement rates compared to MFCC are presented in parentheses.*

Parameters	NbPar	Scheirer	News	Enter
MFCC+Δ+ΔΔ	36	*6.5*	*13.1*	*23.1*
Var of MFCC	12	3.1 (52%)	7.7 (41%)	25.1 (-9%)
MFCC+(Var of MFCC)	24	4.7 (28%)	9.4 (28%)	22.5 (3%)
Var of E	7	**1.7 (74%)**	7.5 (43%)	**16.3 (29%)**
Var of T_E	7	1.8 (72%)	**7.1 (46%)**	16.4 (29%)
Var of H_E	7	1.8 (72%)	7.3 (44%)	16.7 (28%)
E + (Var of E)	14	1.8 (72%)	8.3 (37%)	18.4 (20%)
T_E + (Var of T_E)	14	1.8 (72%)	9.2 (30%)	19.2 (17%)
H_E + (Var of H_E)	14	1.8 (72%)	8.6 (34%)	19.1 (17%)

Tab. 9 –
Music/non-music discrimination results using variance on a 1 second window and static with variance coefficients for wavelet coif-1 *and 7 bands. Frame error rate in percentages. Relative improvement rates compared to MFCC are presented in parentheses.*

ACCEPTED MANUSCRIPT

Consequently, we only take into account the Teager energy. We chose to test static parameters plus delta.

Table 10 shows that wavelet-based parameterization gives much better performance than MFCC parameterization for this more difficult task. This improvement is statistically significant and is 58% for *Scheirer* corpus, 40% for *News* corpus and 30% for *Entertainment* corpus compared to MFCC baseline system.

Param.M/NM	**Param.S/NS**	**NbPar**	**Scheirer**	**News**	**Enter**
MFCC+Δ+ΔΔ	*MFCC+Δ+ΔΔ*	36-36	*8.1*	*15.0*	*26.3*
T_E(7bands)+Δ	T_E(5bands)+Δ	10-14	**3.4(58%)**	**9.0(40%)**	**18.4(30%)**

Tab. 10 –
Global discrimination with best features: wavelet coif-1 *with 7 bands and* Δ *for music/non-music discrimination and wavelet* coif-1 *with 5 bands and* Δ *for speech/non-speech discrimination. Frame error rate in percentages. Relative improvement rates compared to MFCC are presented in parentheses.*

5 Conclusion

In this paper we have proposed a new parameterization based on the wavelets for speech/music discrimination. Our goal was not to propose a new wavelet type but to apply the wavelet formalism for speech/music discrimination task.

Compared to MFCC parameters, widely used for this task, wavelet parameters are more compact, allow the extraction of time-frequency features and deal with non-stationary signal. Our discrimination system is based on the GMM class/non-class approach and the Viterbi algorithm performs the classification.

In the experiments, the proposed wavelet features have been compared to MFCC parameters on three various corpora: *Scheirer, News, Entertainment.* Scheirer corpus has been frequently used in previous studies, *News* corpus is a broadcast news corpus. *Entertainment* is considered as quite difficult because it contains a lot of superimposed segments: speech over music. As expected, the classification error rates on this last corpus are higher than on the two other corpora.

The following conclusions have been drawn from these experiments:

– The wavelet parameterization gives better results that MFCC features for all studied discrimination tasks (speech/non-speech, music/non-music and global discrimination) for all three corpora. For instance, compared to MFCC

ACCEPTED MANUSCRIPT

parameters, the wavelet parameterization led to a significant improvement in the error rate for global speech/music discrimination: 58% for *Scheirer*, 40% for *News* and 30% for *Entertainment* corpora.

- The smaller the number of vanishing moments, the better the discrimination results are.
- The choice of the wavelet family has a small effect on the discrimination results.
- As it has been shown in the different studies for other parameterizations [42], dynamic parameters give solid results. Long term parameters achieve slightly worse results.
- Finally, the best results were obtained using wavelet coif-1 Teager energy and Δ: with 7 bands for music/non-music discrimination and 5 bands for speech/non-speech discrimination.

In conclusion, wavelet parameters are well suited for speech/music discrimination, especially when a corpus containing speech over music segments is being used.

6 Acknowledgements

We would like to thank Eric Scheirer and Malcolm Slaney for making their speech/music corpus available for us. We also thank the evaluation project Technolangue EVALDA-ESTER and the CNRS for its support of the RAIVES project.

Références

[1] J. Ajmera, I. McCowan, and H. Bourlard. Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework. *Speech Communication*, 40:351–363, 2003.

[2] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. Lopez-Ferreras. Application of Fisher Linear Discriminant Analysis to Speech/Music Classification. In *IEEE Eurocon*, pages 1666–1669, 2005.

[3] A. Berenzweig and P.W. Ellis. Locating Singing Voice Segments Within Music Signals. *IEEE Workshop on Apps of Sign. Proc. to Acous. and Audio*, 2001.

[4] A. Brun, C. Cerisara, D. Fohr, I. Illina, D. Langlois, O. Mella, and K. Smaili. ANTS : le système de transcription automatique du LORIA. In *Journées d'Etude sur la Parole - JEP'04*, 2004.

[5] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A Comparison of Features for Speech, Music Discrimination. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 149–152, 1999.

ACCEPTED MANUSCRIPT

[6] W. Chou and L. Gu. Robust Singing Detection in Speech/Music Discriminator Design. *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 865–868, 2001.

[7] M. Deviren. *Revisiting speech recognition systems : dynamic Bayesian networks and new computational paradigms*. PhD thesis, Université Henri Poincaré, Nancy, France, 2004.

[8] M. Deviren and K. Daoudi. Frequency Filtering or Wavelet Filtering? *ICANN/ICONIP*, 2003.

[9] E. Didiot, I. Illina, O. Mella, J.-P. Haton, and D. Fohr. A Wavelet-based Parametrization for Speech/Music Segmentation. In *Proc. Int. Conf. on Spoken Language Processing, ICSLP*, pages 653–656, 2006.

[10] E. Didiot, I. Illina, O. Mella, J.-P. Haton, and D. Fohr. Speech/Music Discrimination Based on Wavelets for Broadcast Programs. In *IEEE International Conference on Signal Processing and Multimedia Applications*, pages 151–156, 2006.

[11] D. Dimitriadis, P. Maragos, and A. Potamianos. Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition. *Proc. European Conf. on Speech Communication and Technology*, 2005.

[12] H. Ezzaidi and J. Rouat. Automatic Music Genre Classification Using Second-Order Statistical Measures for the Prescriptive Approach. *Proc. European Conf. on Speech Communication and Technology*, pages 141–144, 2005.

[13] C. Fredouille, D. Matrouf, G. Linares, and P. Nocera. Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ESTER. In *Journées d'Etude sur la Parole - JEP04*, 2004.

[14] A. Ganapathiraju and J. Picone. Hybrid SVM/HMM Architectures for Speech Recognition. In *Neural Information Processing Systems*, 2000.

[15] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1):89–108, 2002.

[16] J.L. Gauvain, L. Lamel, G. Adda, and M. Jardino. The LIMSI 1998 Hub-4E Transcription System. In *Proc. DARPA Broadcast News Transcription Workshop*, pages 99–104, February 1999.

[17] R. Gemello, D. Albesano, L. Moisa, and R. De Mori. Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 121–124, 2001.

[18] G. Gravier, J.F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en francais. In *Journées d'Etude sur la Parole - JEP04*, 2004.

[19] T. Hain and P. Woodland. Segmentation and Classification of Broadcast News Audio. *Proc. Int. Conf. on Spoken Language Processing, ICSLP*, 1998.

[20] F. Jabloun and A. Enis Cetin. The Teager Energy based Feature Parameters for Robust Speech Recognition in Car Noise. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 273–276, 1999.

[21] J.-C. Junqua and Haton J.-P. Robustness in Automatic Speech Recognition: Problems, Issues, and Solutions. *Kluwer*, 1995.

ACCEPTED MANUSCRIPT

[22] M. Kahn, W. Al-Khatib, and M. Moinuddin. Automatic Classification of Speech and Music using Neural Networks. *Proc. ACM Int. Workshop on Multimedia Databases*, pages 94–99, 2004.

[23] J.F. Kaiser. On a Simple Algorithm to Calculate the 'Energy'of a Signal. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 381–384, 1990.

[24] J.S. Keum and H.S. Lee. Speech/Music Discrimination Based on Spectral Peak Analysis and Multi-Layer Perceptron. In *International Conference on Hybrid Information Technology*, volume 2, pages 56–61, 2006.

[25] M. Khan and W.G. Al-Khatib. Machine Learning-Based Classification of Speech and Music. *Multi-Media Systems*, 12:55–67, 2006.

[26] I. J. Kim, S. I. Yang, and Y. Kwon. Speech Enhancement using Adaptive Wavelet Shrinkage. In *ISIE-2001*, volume 1, pages 501–504, 2001.

[27] D. Kryze, L. Rigazio, and J.-C. Junqua. A New Noise-Robust Subband Front-End and its Comparison to PLP. In *Automatic Speech Recognition and Understanding Workshop*, 1999.

[28] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang. Audio Classification and Categorization Based on Wavelets and Support Vector Machine. *IEEE Transactions on Speech and Audio Processing*, 13:644–651, 2005.

[29] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proc. International Symposium on Music Information Retrieval*, 2000.

[30] L. Lu, H-J. Zhang, and H. Jiang. Content Analysis for Audio Classification and Segmentation. In *IEEE Transactions on Speech and Audio Processing*, number 10(7), pages 504–516, 2002.

[31] S. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[32] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[33] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi. Les ondelettes et leurs applications. *Editeur Lavoisier Hermes*, 2003.

[34] C. Panagiotakis and G. Tziritas. A Speech/Music Discriminator Based on RMS and Zero-Crossings. In *IEEE Transaction on Multimedia*, number 7(1), pages 155–166, 2005.

[35] J. Pinquier, C. Senac, and R. Andre-Obrecht. Speech and Music Classification in Audio Documents. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 4164–4167, 2002.

[36] A. Potamianos and P.Maragos. Time-Frequency Distributions for Automatic Speech Recognition. In *IEEE Transactions on Speech and Audio Processing*, pages 196–200, 2001.

[37] J. Razik, D. Fohr, O. Mella, and N. Parlangeau-Vallès. Segmentation Parole/Musique pour la transcription automatique. In *Journées d'Etudes sur la Parole*, 2004.

[38] J. Razik, C. Senac, D. Fohr, O. Mella, and N. Parlangeau-Valles. Comparison of Two Speech/Music Segmentation Systems For Audio Indexing on the Web. *Proc. Multi Conference on Systemics, Cybernetics and Informatics*, 2003.

ACCEPTED MANUSCRIPT

[39] S. Saha. Image Compression from DCT to Wavelets: a Review. *ACM Crossroads*, 6(3):644–651, 2000.

[40] R. Sarikaya and J.H.L. Hansen. High Resolution Speech Feature Parameterization for Monophone-based Stressed Speech Recognition. *IEEE Signal Processing Letters*, 7(7):182–185, 2000.

[41] J. Saunders. Real-Time Discrimination of Broadcast Speech/Music. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 993–996, 1996.

[42] E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 1331–1334, 1997.

[43] T. Taniguchi, M. Tohyama, and S. Katsuhiko. Detection of Speech and Music Based on Spectral Tracking. In *Speech Communication*, number 50, pages 547–563, 2008.

[44] T.Hain, S.E.Johnson, A.Tuerk, P.C. Woodland, and S.J.Young. Segment Generation and Clustering in the HTK Broadcast News Transcription System. In *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 133–137, 1998.

[45] H. Tolba and D. O'Shaughnessy. Automatic Speech Recognition Based on Ceptral Coefficients and a Mel-Based Discrete Energy Operator. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP*, pages 973–976, 1998.

[46] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[47] K. Umapathy, S. Krishnan, and S. Jimaa. Multigroup Classification of Audio Signals Using Time-Frequency Parameters. *IEEE Transaction on Multimedia*, 7(2):308–315, 2005.

[48] G. Williams and D. Ellis. Speech/Music Dscrimination Based on Posterior Probability Features. *Proc. European Conf. on Speech Communication and Technology*, pages 687–690, 1999.

[49] E. Wold, T. Blum, D. Keislar, and J. Wheater. Classification, Search and Retireval of Audio. *CRC Handbook of multimedia computing, CRC Press LLC*, 1999.

[50] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge, England, Entropic Ltd., Microsoft, 1995.

[51] T. Zhang and C.-C. J. Kuo. Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. In *IEEE Transactions on Speech and Audio Processing*, number 9(4), pages 441–457, 2001.

ACCEPTED MANUSCRIPT