A Wavelet-Based Parameterization for Speech/Music Segmentation

E. Didiot, I. Illina, O. Mella, D. Fohr, J.-P. Haton

LORIA-CNRS & INRIA Lorraine
BP 239, 54506 Vandoeuvre-les-Nancy, France
{didiot,illina,mella,fohr,jph}@loria.fr

Abstract

The problem of speech/music discrimination is a challenging research problem which significantly impacts Automatic Speech Recognition (ASR) performance. This paper proposes new features for the Speech/Music discrimination task. We propose to use a decomposition of the audio signal based on wavelets, which allows a good analysis of non stationary signal like speech or music. We compute different energy types in each frequency band obtained from wavelet decomposition. Two class/non-class classifiers are used : one for speech/non-speech, one for music/non-music. On the different test corpora, the proposed wavelet approach gives better results than the MFCC one. For instance, we have a significant relative improvements of the error rate of 58.0% on the "Scheirer" corpus for the speech/music discrimination task.
Index Terms: speech/music discrimination, wavelets, static and dynamic parameters, long-term parameter.

1. Introduction

Discrimination between speech and music consists in segmenting an audio stream in acoustically homogeneous segments such as speech, music and speech on music. This is an important research area, for example, for audio content indexing and for automatic transcription of audio broadcasts. Regarding the transcription task, separation between speech and music allows to eliminate spurious segments containing only music and thus to decrease the number of recognition mistakes.
Speech/music segmentation needs two steps : parameterization and classification of audio signal. The parameterization step extracts from the signal discriminative features for speech and music. Various parameterizations have been proposed in the state of the art of the domain. They can be classified in :

- frequency features : spectral centroid, spectral flux, spectral rolloff point, etc. [1, 2];
- temporal features : energy, Zero Crossing Rate, etc. [1, 2];
- mixed features : 4 Hz modulation energy [2], low frequency modulation amplitude features [3], percentage of low energy frames [1];
- cepstral domain features : for instance MFCC [4, 5, 6].

This article presents a new approach for speech/music discrimination based on the use of wavelet decomposition of the signal. To our knowledge, a such approach was never used for this task. Our motivation to use wavelets is their ability to measure the time variations of the spectral components and to extract time-frequency features. Moreover, the multi-band decomposition made by the dyadic wavelet transform is close to the one made by the human ear [7]. Compared to MFCC coefficients, wavelets have a good time-frequency resolution, allow a more compact representation, have a rich set of basis functions and are more robust to the non-stationarity and the distortions of the signal.
In this work, we study energy-based features extracted from different wavelet decompositions of the audio signal. We also evaluate short and long-term variations by taking into account first and second derivatives and variance of extracted features. To validate the proposed approach, experiments are made on three different broadcast corpora that cover various real-world cases. We compare the wavelet's performance with the MFCC one, because MFCC features show a good results in speech recognition, in music modeling [5], in speech/music discrimination [4, 5] and in musical genre classification [6].
This paper is organized as follows. Sections 2 and 3 introduce the new features and describe our speech/music discrimination system. Section 4 presents the corpora. Our experiments are detailed in Section 5. Finally, section 6 gives some conclusions.

2. Wavelet-based parameters

Wavelet-based signal processing has been successfully used for various problems : for example, in denoising task or, recently, in automatic speech recognition [8, 9].
Discrete Wavelet Transform (DWT) analyses the signal in different frequency bands with various resolutions. Such analysis allows a simultaneous analysis in time and frequency domains. S. Mallat [10] has shown that such decomposition can be obtained by successive low-pass (G) and high-pass (H) filtering of the time domain signal and by down-sampling the signal by 2 after each filtering. This process is repeated on the results of the low-pass filtering until the required number of frequency bands is obtained. Figure 1 shows a two-level decomposition where the symbol $\downarrow 2$ denotes a down-sampling by 2. The signal is decomposed into

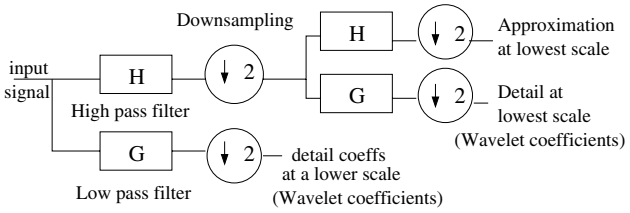

Figure 1: Discrete Wavelet Transform

approximation coefficients and *detail* coefficients. Approximation coefficients correspond to local averages of the signal. Detail coefficients, named "wavelet coefficients", depict the differences between two successive local averages, ie. between two successive

approximations of the signal.
For speech/music discrimination task, we propose to use only wavelet coefficients to parameterize the acoustic signal. The use of wavelet coefficients allows to capture the sudden modifications of the signal. Indeed, the wavelet coefficients have high values during such events. In our study, we compute dyadic wavelet transform corresponding to octave-band filter banks. The dyadic wavelet transform performs a non-uniform bandwidth decomposition of the signal, and thus permits to obtain a decreasing frequency resolution when frequency increases. So this wavelet decomposition gives a multi-resolution analysis of the signal : a fine time resolution and a coarse frequency resolution at high frequencies and inversely at low frequencies.
Several features based on energy are computed on wavelet coefficients in each frequency band. In the following, w_k^j denotes the wavelet coefficient at position k and band j. N_j denotes the number of coefficients at band j, and f_j the feature vector for band j. We compute :

- *Logarithm of energy* (**E**). The instantaneous energy :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j}(w_k^j)^2\right)$$

- *Logarithm of Teager energy* (**T_E**). The discrete Teager Energy Operator TEO introduced by Kaiser is used [11].

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j-1}|(w_k^j)^2 - w_{k-1}^j w_{k+1}^j|\right)$$

- *Logarithm of hierarchic energy* (**H_E**). Hierarchic energy corresponds to the computation of energy at the center of the analysis window with the same number of coefficients in every band :

$$f_j = \log_{10}\left(\frac{1}{N_J}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2}(w_k^j)^2\right)$$

J corresponds to the lowest band. This energy was successfully used in automatic speech recognition to parameterize the signal [12].

3. Speech/music discrimination system

3.1. Parameterization

The signal is sampled at 16kHz. After pre-emphasis, we use a 32ms Hamming window with a 10ms shift. Our parameters are :
Baseline MFCC features: 12 MFCC coefficients with their first and second derivatives. Finally, a 36 coefficients vector is obtained.
Wavelet based features: The above-described energy features are calculated on wavelet coefficients obtained with two wavelet families : daubechies wavelet and coiflet. Multiresolution parameters are computed for two decomposition levels, i.e. for different number of bands (5 and 7).
Our static features are computed on a very short time duration (32ms). We decided to study also some long-term parameters. Firstly, we use the first and second derivatives of the energy parameters. Secondly, Scheirer and Slaney have shown that the use of variance computed on a one-second window improves the results in speech/music discrimination [2]. The study of this long-term parameter seems interesting.

3.2. System description

Our classification approach is a "Class/Non-class" one [13]. In other words, class detection is performed by comparing a class model and a non-class model estimated on the same representation space. Two subsystems are implemented : speech/non-speech and music/non-music.
The decisions of both classifiers are merged and the audio signal is classified into three categories: speech (S), music (M), and speech on music (SM). To model each class, HMM model is used (with between 8 and 64 gaussians per state). The Viterbi algorithm is used to provide the best sequence of models, describing the audio signal. A frame by frame decision would lead to unrealistic 10ms-length segments. To avoid this, for each recognized segment a minimal duration (0.5s) is imposed.

4. Corpora

4.1. Training corpus

The HMM models were trained on two databases : "Audio CDs" and "Broadcast programs". The "Audio CDs" corpus (120 mn) is made up of several tracks of instrumental music and songs extracted from CDs. The "Broadcast programs" corpus (976 mn) contains programs from French radios: broadcast news as well as interviews and musical programs.

4.2. Test corpora

We carried out experiments on three corpora that are totally diverse :

- The *Scheirer* part is the English test corpus built and used by E. Scheirer and M. Slaney [2]. All the audio files are homogeneous and have the same duration, 15 seconds: 20 files of broadband or telephone speech and 41 files of music or vocals. The music styles are more various (jazz, pop, country, etc.) than in the *Entertainment* part. This part does not contain speech with music in background. It is composed of 32% of speech frames and 68% of music frames. This corpus allows to evaluate our new parameterizations on a often used in previous studies corpus. We don't use file homogeneity information and our discrimination system segments and classifies the files. Confidence interval is $\pm 2\%$ at the 0.05 level of significance.
- The *News* part consists of three 1-hour files of French radios "France-Inter" and "Radio France International" and contains mainly speech or speech on jingles (86% of speech, 11% of speech on music and 3% of music). This corpus is interesting in the way that our speech/music discrimination system can be evaluated on a broadcast news transcription task. Confidence interval is $\pm 0.5\%$.
- The *Entertainment* part is composed of three 20-minutes shows (interviews and musical programs). This corpus is considered as quite difficult. Indeed, there are a lot of superposed segments, speech with music or songs with an effect of "fade in-fade out". Moreover, this part contains an alternation of broad-band speech and telephone speech and some interviews are very noisy. It is made of 52% of speech frames, 18% of speech on music frames and 30% of music frames. Thus, this last corpus allows us to evaluate the proposed parameterization on difficult broadcast programs. Confidence interval is $\pm 1\%$.

5. Experimental results

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

- Global classification error rate :
$$100 * (1 - (n_{SM}^{SM} + n_M^M + n_S^S)/T)$$
- Music/Non-Music classification error rate :
$$100 * (1 - (n_{SM}^M + n_M^{SM} + n_M^M + n_{SM}^{SM} + n_S^S)/T)$$
- Speech/Non-Speech classification error rate :
$$100 * (1 - (n_{SM}^S + n_S^{SM} + n_M^M + n_{SM}^{SM} + n_S^S)/T)$$

with n_z^y the number of frames recognized as z when labeled y and T the total number of frames.

As baseline feature, the 12 MFCC with their first and second derivatives are used because for the test corpora they give the best global discrimination error rate compared to other MFCC features.

5.1. Static parameters

In this experiment, static features based on wavelets are evaluated. After preliminary experiments, the following wavelets have been chosen: daubechies wavelet with 4 vanishing moments (*db-4*) and coiflet with 2 vanishing moments (*coif-1*). We use two decomposition levels : 5 and 7. Instantaneous (**E**), Teager (**T_E**) and hierarchic (**H_E**) energy, calculated on wavelet coefficients, are studied.

Table 1: *Speech/non-speech discrimination results using wavelets* db-4 *and* coif-1*, 5 and 7 bands. Error rate (%).*

Wlt	Bds	Ener.	Scheirer	News	Entertain.
MFCC+Δ+ΔΔ			*2.5*	*2.9*	*5.8*
db-4	5	E	3.3	3.9	5.3
db-4	5	T_E	3.3	3.6	5.4
db-4	5	H_E	3.3	4.5	5.7
db-4	7	E	2.9	5.5	6.2
db-4	7	T_E	**2.9**	4.4	5.4
db-4	7	H_E	3.3	5.6	6.2
coif-1	5	E	3.3	3.7	**4.2**
coif-1	5	T_E	3.3	**3.2**	4.2
coif-1	5	H_E	3.3	4.4	4.3
coif-1	7	E	3.3	7.4	6.8
coif-1	7	T_E	3.6	6.4	6.1
coif-1	7	H_E	3.3	7.6	6.6

5.1.1. Speech/non-speech discrimination

Speech/non-speech segmentation results are summarized in table 1. This table shows that static features proposed in this paper are comparable to the baseline MFCC features but have a more compact representation. Indeed, similar results are obtained with a 5 components features vector for wavelet parametrisation and with 36 components features vector for MFCC.

Nevertheless, the Teager energy on 5 bands with "coif-1" wavelet gives an absolute gain of 2.6% for the "Entertainment" corpus.

5.1.2. Music/non-music discrimination

For the music/non-music discrimination task, the results are given in table 2. The Teager energy feature gives good results for every wavelet decomposition. For the most difficult corpus ("Entertainment") all the wavelet parameters achieve better results than MFCC parameters.
So, for the music/non-music discrimination task, wavelet based parameters are better than MFCC. We decided to carry on our experiments for the music/non-music discrimination task.

Table 2: *Music/non-music discrimination results using wavelets* db-4 *and* coif-1 *with 5 and 7 bands. Error rate (%).*

Wlt	Bds	Ener.	Scheirer	News	Entertain.
MFCC+Δ+ΔΔ			*6.5*	*13.1*	*23.1*
db-4	5	E	5.1	15.1	15.3
db-4	5	T_E	5.1	13.1	15.1
db-4	5	H_E	5.1	14.3	19.2
db-4	7	E	5.0	17.8	16.1
db-4	7	T_E	5.0	17.0	16.5
db-4	7	H_E	5.1	15.7	16.4
coif-1	5	E	5.3	7.8	16.5
coif-1	5	T_E	5.6	8.0	17.0
coif-1	5	H_E	5.3	**7.0**	18.5
coif-1	7	E	4.3	11.4	14.5
coif-1	7	T_E	**3.7**	10.1	14.6
coif-1	7	H_E	3.7	10.9	**13.8**

5.2. Dynamic parameters

To study the influence of dynamic features, the first (Δ) and second derivatives (ΔΔ) are added to our wavelet-based parameters. We chose as static features "coif-1" wavelet with 7 bands, because for music/non-music classification task and for every corpus they give on the whole good results.
For the music/non-music segmentation task, results are summarized in Table 3. In this table, Nb corresponds to the feature vector size.

For the "Scheirer" and the "News" corpora, the addition of first derivatives (Δ) improves the results compared to static parameters. A relative gain of 72.3% for the "Scheirer" and 45.0% for the "News" is obtained compared to the MFCC features. On the contrary, addition of the second derivatives (ΔΔ) does not improve the results compared to the addition of the first derivatives. We can even see a degradation of the results.

5.3. Long-term parameters

The study of long-term parameters such as variance seems interesting [2]. Therefore, variance of the static features, based on the "coif-1" wavelet and 7 bands, is computed on a one-second window. The one-second variance is also calculated on baseline MFCC features for comparison. Results for music/non-music discrimination are given in table 4.
The variance based features give results similar to the features based on the addition of the first derivatives (Δ) (see table 4). The advantage of the variance-based features is the compactness of the

Table 3: *Music/non-music discrimination results using wavelets* coif-1 *with 7 bands and dynamic parameters (Δ, ΔΔ). Error rate (%).*

Param.	Nb	Scheirer	News	Entertain.
MFCC+Δ+ΔΔ	36	*6.5*	*13.1*	*23.1*
E+Δ	14	1.8	7.9	15.2
T_E+Δ	14	**1.8**	**7.2**	15.0
H_E+Δ	14	1.8	7.2	**14.8**
E+Δ+ΔΔ	21	1.8	9.5	17.4
T_E+Δ+ΔΔ	21	1.8	9.7	17.4
H_E+Δ+ΔΔ	21	1.8	8.6	18.3

Table 4: *Music/non-music discrimination results using variance on a one-second window with wavelet* coif-1 *and 7 bands. Error rate (%).*

Param.	Nb	Scheirer	News	Entertain.
MFCC+Δ+ΔΔ	36	*6.5*	*13.1*	*23.1*
MFCC+Δ+ΔΔ (Variance on 1s)	36	*3.4*	*9.1*	*23.3*
E Var 1s	7	**1.7**	7.5	**16.3**
T_E Var 1s	7	1.8	**7.1**	16.4
H_E Var 1s	7	1.8	7.3	16.7

signal representation : the features are only composed of 7 components.

5.4. Global discrimination

The last experiment corresponds to simultaneously discriminate speech, music and speech with music. For the test, the same features are used for speech/non-speech discrimination and for music/non-music discrimination. The previous best features are used here. Table 5 gives the global discrimination error rate computed on the different corpora.

Table 5: *Global discrimination with best features : Δ with wavelet* coif-1 *and 7 bands. Error rate (%).*

Param.	Nb	Scheirer	News	Entertain.
MFCC+Δ+ΔΔ	36	*8.1*	*15.0*	*26.3*
E+Δ	14	3.4	9.6	17.4
T_E+Δ	14	**3.4**	**8.8**	17.6
H_E+Δ	14	3.4	9.3	**17.1**

We have an important improvement for each corpus compared to MFCC. For the "Scheirer" corpus, the significant relative gain is 58.0%. For the "News" corpus, the gain is 41.3%. For the "Entertainment" corpus, the gain is 35.0%. Table 5 shows that Teager energy seems to be a robust feature to discriminate speech, music and speech on music.

6. Conclusion

In this paper, we propose new features based on wavelet decomposition of the audio signal. These features are obtained by computing different energies on wavelet coefficients. Compared to MFCC, wavelet decomposition gives a non-uniform time resolution for the different frequency bands. Moreover, this parameterization allows to obtain a more compact representation of the signal and is more robust to signal non-stationarity. The proposed parameterization is used for speech/music discrimination task.
The new parameterization gives better results than MFCC based one for speech/music discrimination. Best improvements are obtained for the music/non-music discrimination task, with a relative gain of 72.3% in error rate for the "Scheirer" corpus, 45.0% for the "News" corpus and 40.3% for the "Entertainment" corpus. Moreover, Teager energy feature based on coif-1 wavelet seems to be robust features for discrimination between speech, music and speech on music. Finally, the proposed parameterizations use a reduced number of coefficients to represent the signal compared to MFCC parameterization.

7. Acknowledgements

We would like to thank Eric Scheirer and Malcolm Slaney for making available to us their speech/music corpus.

8. References

[1] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music," in *ICASSP-96*, 1996.

[2] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *ICASSP-97*, 1997.

[3] S. Karneback, "Discrimination between Speech and Music based on a Low Frequency Modulation Feature," in *European Conf. on Speech Comm. and Technology*, 2001.

[4] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A Comparison of Features for Speech, Music Discrimination," in *ICASSP-99*, 1999.

[5] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in *International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[6] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transaction on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[7] S. Maes I. Daubechies, "A Nonlinear Sqeezing of the Continuous Wavelet Transform based on Auditory Nerve Models," in *Wavelets in Medecine and Biology*, 1996.

[8] R. Sarikaya and J.H.L. Hansen, "High Resolution Speech Feature Parameterization for Monophone-based Stressed Speech Recognition," *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 182–185, 2000.

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

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

[11] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy'of a Signal," in *ICASSP-90*, 1990.

[12] R. Gemello, D. Albesano, L. Moisa, and R. De Mori, "Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System," in *ICASSP-01*, 2001.

[13] J. Pinquier, "Speech and music classification in audio documents," in *ICASSP-02*, 2002.