SPEECH/MUSIC DISCRIMINATION BASED ON WAVELETS FOR BROADCAST PROGRAMS

E. Didiot, I. Illina, O. Mella, D. Fohr, J.-P. Haton
LORIA-CNRS & INRIA Lorraine
BP 239, 54506 Vandoeuvre-les-Nancy, France

Keywords: Speech/music discrimination, wavelets, static and dynamic parameters, long-term parameters, classifiers fusion.

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 broadcast test corpus, the proposed wavelet approach gives better results than the MFCC one. For instance, we have a significant relative improvements of the error rate of 39% for the speech/music discrimination task.

1 INTRODUCTION

Discrimination between speech and music consists in segmenting an audio stream into acoustically homogeneous segments such as speech, music and speech on music. This segmentation task plays an important role in various multimedia applications. Let us mention several examples. For automatic transcription of broadcast news or programs, non-speech segments must be discarded to avoid high recognition error rate. Audio indexing of multimedia documents requires that music segments have to be labelled. Speech/music discrimination can speed up the task of putting subtitles because by skipping non- speech segments. Automatic real-time captioning of live TV transmissions of events also needs speech/non-speech detection.
Speech/music discrimination requires two steps : parameterization and classification of audio signal.
The parameterization step consists in extracting discriminative features from the audio signal. This article presents a new approach for speech/music discrimination based on the wavelet decomposition of the signal. To our knowledge, a such approach has been never used for this task. Our motivation to apply wavelets to speech/music discrimination is their ability to extract time-frequency features and to deal with non-stationary signals. Moreover, the multiband decomposition made by the dyadic wavelet transform is close to the one made by the human ear (I. Daubechies, 1996). Therefore, we study several features based on wavelet decomposition and test them on some broadcast programs. We also compare their performance with Mel Frequency Cepstral Coefficients (MFCC) because the latter have shown good results in speech/music discrimination (Carey et al., 1999; Logan, 2000), in music modeling (Logan, 2000) and in musical genre classification (Tzanetakis and Cook, 2002). Besides, MFCC features are widely used in speech recognition.
The classification step consists in classifying the audio signal in different categories: speech, music, speech on music. For that two approaches can be considered: either a “class/non-class” approach that builds a classifier for each category or a “competing” approach allowing the competition of several categories in a single classifier.
Moreover, both approaches can use different methods to classify: k-Nearest Neighbours (kNN), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Neural Networks,...
We decide to use the class/non-class approach with intent to obtain the best parameterization for each category. The classification method is based on a Viterbi

Didiot E., Illina I., Mella O., Fohr D. and Haton J. (2006).
SPEECH/MUSIC DISCRIMINATION BASED ON WAVELETS FOR BROADCAST PROGRAMS.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 151-156
DOI: 10.5220/0001572901510156

algorithm using HMM models, because this simultaneously performs classification and segmentation. Besides, in order to decrease the error rate, a classifier fusion is evaluated.
This paper is organized as follows. Section 2 introduces the new features. Section 3 describes our speech/music classification system. Section 4 presents the training and test corpora. Experiments are detailled in section 5: the speech/non-speech and music/non-music discriminations and then the classifier fusion. 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 (Sarikaya and Hansen, 2000; Deviren, 2004).
Discrete Wavelet Transform (DWT) analyses the signal in different frequency bands with various resolutions. Such an analysis allows a simultaneous analysis in time and frequency domains. S. Mallat (Mallat, 1998) has shown that such a decomposition can be obtained by successive low-pass (G) and high-pass (H) filterings of the time domain signal and by downsampling 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 *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.

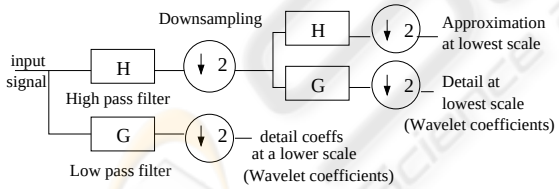

Figure 1: Discrete Wavelet Transform.

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 :

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

- *Logarithm of Teager energy* (**T_E**). The discrete Teager Energy Operator TEO introduced by Kaiser is used (Kaiser, 1990).

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

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 coefficient 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) and the question which may be asked is: can an human ear reliably identify such a short segment as speech or as music? We thus decide to also study some long-term parameters. Firstly, we test 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 (Scheirer and Slaney, 1997). Therefore, the study of this long-term parameter seems interesting.

3.2 System Description

Our classification approach is a "Class/Non-class" one (Pinquier, 2002). 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). Each class is modelled by an HMM model 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, a 0.5s minimal duration is imposed for each recognized segment.

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 Corpus

We carried out experiments on a broadcast corpus composed of three 20-minutes shows (interviews and musical programs). This corpus is considered as quite difficult. Indeed, there are a lot of superimposed 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 corpus allows us to evaluate the proposed parameterization on difficult broadcast programs. Confidence interval is $\pm 1\%$ at the 0.05 level of significance.

5 EXPERIMENTAL RESULTS

5.1 Error Rate Calculation

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

- Global classification error rate:
$$100 * (1 - (n_{SM}^{SM} + n_M^M + n_S^S)/T) \quad (3)$$
- Music/Non-Music classification error rate:
$$100*(1-(n_{SM}^M+n_M^{SM}+n_M^M+n_{SM}^{SM}+n_S^S)/T) \quad (4)$$
- Speech/Non-Speech classification error rate:
$$100*(1-(n_{SM}^S+n_S^{SM}+n_M^M+n_{SM}^{SM}+n_S^S)/T) \quad (5)$$

with n_z^y the number of frames recognized as z when labeled y and T the total number of frames.
Moreover, we consider the 12 MFCC coefficients with their first and second derivatives as the baseline features because they give the best global discrimination error rate compared to other MFCC-based features also evaluated on our test corpus.

Table 1 presents the distribution of recognized frames into speech, music or speech on music categories for the global discrimination task with MFCC parameterization. This Table shows the hardness of the speech/music discrimination, especially for superimposed segments of speech and music.

Table 1: Frames distribution (%) for global discrimination task using 12 MFCC coefficients with their first and second derivatives.

labelled \ recognized	**S**	**SM**	**M**
S	60.9	30.8	8.3
SM	10.1	74.9	15.0
M	2.9	2.5	93.8

5.2 Speech/non-speech Discrimination

After preliminary experiments, we chose two families of wavelets: daubechies wavelet with 4 vanishing moments (*db-4*) and coiflet with 2 vanishing moments (*coif-1*). We used two decomposition levels: 5 and 7, and, computed two energy features on the wavelet coefficients: instantaneous (**E**) and Teager (**T_E**) energies.

Speech/non-speech discrimination results are summarized in Table 2. We can notice that energy features computed on the "coif-1" wavelet parameterization with 5 bands give sligthy better results than the

Table 2: Error rates (%) for speech/non-speech discrimination task using wavelets *db-4* and *coif-1*, 5 and 7 bands.

Wlt	Nb	Param.	Error rate
MFCC+Δ+ΔΔ			*5.8*
Static parameters			
db-4	5	E	5.3
db-4	5	T_E	5.4
db-4	7	E	6.2
db-4	7	T_E	5.4
coif-1	5	E	**4.2**
coif-1	5	T_E	**4.2**
coif-1	7	E	6.8
coif-1	7	T_E	6.1
Dynamic parameters			
coif-1	14	E+Δ	3.4
coif-1	14	T_E+Δ	**2.7**
coif-1	21	E+Δ+ΔΔ	3.1
coif-1	21	T_E+Δ+ΔΔ	2.7
Long-term parameters			
MFCC+Δ+ΔΔ (Var. on 1s)			*4.2*
coif-1	7	E Var 1s	3.5
coif-1	7	T_E Var 1s	**3.2**

Table 3: Error rates (%) for music/non-music discrimination results using wavelets *db-4* and *coif-1*, 5 and 7 bands.

Wlt	Nb	Param.	Error rate
MFCC+Δ+ΔΔ			*23.1*
Static parameters			
db-4	5	E	15.3
db-4	5	T_E	15.1
db-4	7	E	16.1
db-4	7	T_E	16.5
coif-1	5	E	16.5
coif-1	5	T_E	17.0
coif-1	7	E	**14.5**
coif-1	7	T_E	14.6
Dynamic parameters			
coif-1	14	E+Δ	15.2
coif-1	14	T_E+Δ	**15.0**
coif-1	21	E+Δ+ΔΔ	17.4
coif-1	21	T_E+Δ+ΔΔ	17.4
Long-term parameters			
MFCC+Δ+ΔΔ (Var. on 1s)			*23.3*
coif-1	7	E Var 1s	**16.3**
coif-1	7	T_E Var 1s	16.4

MFCC parameters. The addition of dynamic parameters, more precisely first derivatives, gives significantly better performance than MFCC parameters or static wavelet features. Besides, with the same number of parameters (7), the long-term wavelet parameters based on variance computation provide an improvement compared to the static ones.

5.3 Music/non-music Discrimination

For the music/non-music discrimination task, the results are presented in Table 3. Whatever wavelet type, number of bands or energy type, the static wavelet parameters achieve a dramatic decrease of the error rate compared to MFCC parameterization. On the other hand, adding derivative components or using long-term wavelet features is not helpful.

5.4 Global Discrimination

We then conducted some experiments to test different features computed on the "coif-1" wavelet parameterization with 7 bands for the global discrimination task. The results presented in Table 4 confirm the previous obtained conclusions : static wavelet features significantly decrease the error rate compared to MFCC ones. The addition of dynamic coefficients reduces the error rate a little bit more. Finally, variance-based long-term parameters are not very helpful.

Table 4: Error rates (%) for global discrimination task using wavelets *coif-1* and 7 bands.

Param.	Nb	Error rate
MFCC+Δ+ΔΔ	36	*26.2*
Static parameters		
E	7	21.6
T_E	7	18.4
Dynamic parameters		
E+Δ	14	**17.4**
T_E+Δ	14	17.6
Long-term parameters		
E Var 1s	7	18.7
T_E Var 1s	7	18.6

5.5 Fusion of Different Classifiers

In order to improve performance of all the discrimination tasks, we combine the outputs of several class/non-class classifiers. The classifiers differ by the parameterization and features they use. Two types of classifier output fusion were tested.
In the first one, called "fusion A", to outperform the results of the global discrimination task, we combine the outputs of the best speech/non-speech classifier and of the best music/non-music one. For both classifiers, best results are obtained with the 7-band "coif-1" wavelet parameterization. Regarding the energy features computed on this decomposition, the

best speech/non-speech discrimination is achieved with Teager energy and its first derivative and the best music/non-music one with instantaneous energy.
In the second one, called “fusion B”, we choose three parameterizations for each discrimination task (speech/non-speech and music/non-music). Then, the outputs of these classifiers are merged using the majority voting strategy.
We assume that these parameterizations are well performing methods, bring diversity and produce different kinds of mistakes. Combination of such experts should reduce overall classification error and as a consequence emphasize correct outputs.
For every discrimination task, the three parameterizations are chosen as follows: we select the best static feature, the best “dynamic feature” (static components plus derivatives) and the best long-term one. According to our experiments, we obtain:

For speech/non-speech task:

- coif-1 instantaneous energy with 5 bands,
- coif-1 Teager energy with 7 bands with first derivatives,
- variance on 1 second computed on coif-1 Teager energy with 7 bands.

For music/non-music task:

- coif-1 instantaneous energy with 7 bands,
- coif-1 Teager energy with 7 bands with first derivatives,
- variance on 1 second computed on coif-1 instantaneous energy with 7 bands.

Table 5 shows the results of the three discrimination tasks using both fusion approaches. Besides, Table 5 mentions the error rate obtained by the best classifier for the global discrimination task (first line). For fusion A, only the global discrimination error rate must be considered: we can notice a non significant improvement. In the other hand, fusion B slightly improves the speech/non-speech and music/non-music discriminations. Moreover,it provides a significant decrease of the global classification error rate.

To conclude the experimental part, Table 6 shows the classification results using the best fusion of classifiers. Compared to Table 1 (MFCC parameters), a significant reduction of misclassified segments is observed.

Table 5: Error rates (%) for the 3 discrimination tasks using the fusion of classifiers.

Param.	**M/NM**	**S/NS**	**GR**
best feature GR coif-1 7b E+Δ	*15.0*	*3.4*	*17.4*
best feature S/NS coif-1, 7bds, T_E+Δ	–	2.7	Fusion A
best feature M/NM coif-1, 7bds, E	14.5	–	17.0
majority vote with 3 classifiers S/NS	–	**2.5**	Fusion B
majority vote with 3 classifiers M/NM	**14.0**	–	**16.1**

Table 6: Frame distribution (%) for global discrimination task using the best fusion of classifiers.

labelled \ recognized	**S**	**SM**	**M**
S	76.9	22.5	0.5
SM	8.9	86.3	4.6
M	0.2	4.1	94.3

6 CONCLUSION

In this paper, we propose new features based on wavelet decomposition of the audio signal for speech/music discrimination.
These features are obtained by computing different energies on wavelet coefficients. Compared to the MFCC parametrization, the wavelet decomposition gives a non-uniform time resolution for the different frequency bands. Moreover, this parameterization is more robust to signal non-stationarity and allows to obtain a more compact representation of the signal.
We have tested these new features on a difficult real-world corpus composed of broadcast programs with superimposed segments, speech with music or songs with an effect of “fade in-fade out’.
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 40% in error rate. Moreover, Teager energy feature based on coif-1 wavelet seems to be a robust feature for discrimination between speech, music and speech on music.
Another interesting point is that the proposed parameterizations use a reduced number of coefficients to represent the signal compared to MFCC one.
Finally, the fusion between the classifiers using the three best speech/non-speech, music/non-music parameterizations improves the speech/music discrimination results. At last, for the speech/music/speech on music discrimination task, a relative gain of 39% in error rate is obtained, compared to MFCC parameters.

REFERENCES

Carey, M., Parris, E., and Lloyd-Thomas, H. (1999). A Comparison of Features for Speech, Music Discrimination. In *ICASSP-99*.

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

I. Daubechies, S. M. (1996). A Nonlinear Sqeezing of the Continuous Wavelet Transform based on Auditory Nerve Models. In *Wavelets in Medecine and Biology*.

Kaiser, J. (1990). On a Simple Algorithm to Calculate the 'Energy'of a Signal. In *ICASSP-90*.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. In *International Symposium on Music Information Retrieval (ISMIR)*.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press.

Pinquier, J. (2002). Speech and music classification in audio documents. In *ICASSP-02*.

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

Scheirer, E. and Slaney, M. (1997). Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *ICASSP-97*.

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