IDIAP RESEARCH REPORT

IDIAP
Martigny - Valais - Suisse

USING POSTERIOR PROBABILITIES FOR SPEECH/MUSIC DISCRIMINATION

Maja Popović

IDIAP–RR 01-08

MARCH 2001

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 01-08

USING POSTERIOR PROBABILITIES FOR SPEECH/MUSIC DISCRIMINATION

Maja Popović

MARCH 2001

Abstract. Automatic speech/music discrimination has been receiving importance recently, for example when large multimedia documents have to be processed by an ASR system, or for indexing and retrieval of such documents. This work presents using outputs of a speech recognition acoustic classifier (neural network) for determining if the signal is speech or something else. We describe two posterior probability measures, entropy and dynamism [9] and test them on a databases of clean speech and music files, as well as on two broadcast news files containing one speech and one music segment. Likelihood ratio classification is performed on a frame-level, with entropy and dynamism calculated over N frames. The higher value of N (longer segments), the less the error, but classification is good enough up to N=40 frames. Acoustic change detection via the Bayesian Information Criterion [2] using those entropy and dynamism instead of usual set of acoustic features is also applied to the same two news files. This approach seems to be effective, although it has to be evaluated on more data.

1 INTRODUCTION

The problem of speech/music discrimination is receiving more and more attention [1, 6, 7, 9]. Typical use of such discrimination is the extraction of speech segments from multimedia documents (broadcast news, etc.) for further processing by ASR (Automatic Speech Recognition system). If the data contains both speech and music, recognizer will generate some meaningless word sequences for a non-speech parts thus increasing WER. And decoder will expend considerable computation effort for this. So, it's preferable to avoid attempting to recognize non-speech segments if it's possible to distinguish them quickly from the speech ones.

The original motivation for the experiments in this work was to make the segmentation for the sports BBC news data in the framework of the ASSAVID project (Automatic Segmentation and Semantic Anotation of Sports Videos). In summary, this project has objectives to develop techniques for automatic segmentation, classification and annotation of sports videos to faciliate access to this multimedia content for indexing and retrieval. The segmentation and classification should be performed hierarchically down to the finest level, so the first step of audio segmentation would be discrimination between speech and music segments.

There are lot of possible approaches to this problem. Quite a big difference between speech and music is noticeable even just looking at the spectrogram [1], but their efficient discrimination is still an opened problem.

Many acoustic (spectral and temporal) features could be used in solving it, for example zero-crossing rate features [6, 7], energy distribution features [7], modulation spectrum [1, 7], etc. Those features are specifically selected to represent the differences between speech and music, and then used to build the distribution models for data classification.

But it is also possible to determine if the input signal is speech or music using a speech recognizer [9]. In this case, the acoustic features on the input of a classifier are specificaly developed to represent speech, but the probabilities on the output of a classifier behave differently if the incoming signal is something else. Experiments with four measures derived from the speech recognizer neural network posterior probabilities [9] shown that they could be highly effective in distinguishing speech segments from those containing music. Those experiments were carried out on the audio already broken into consistent 15 sec and 2.5 sec long segments of speech and nonspeech, with a features calculated all over the segment.

In this work we have experimented with two of those four features calculated every frame over the N-frame segments, where N is taking several different values from 10 to 100, and we have used them for segmentation of audio containing both music and speech. The features, entropy and dynamism, are described in Section 2, and in Section 3 experiments and results with different segment lengths, and for two hypotheses testing methods, likelihood ratio and Bayesian Information Criterion (BIC) [2].

2 DIFFERENT APPROACHES

There are a lot of different approaches to the speech and music discrimination problem, and variety of features and classification techniques have been examinated.

Saunders [6] proposed a real-time speech/music discriminator using four features based on zero-crossing rate, using the derivative variance, third central moment, threshold, and skewness measure.

Scheirer and Slaney [7] experimented with a multidimensional classifier for speech and music data recorded from radio stations. They experimented with several classification methods, but performance differences between the classifiers were insignificant. They examined 13 features to measure distinctive properties of speech and music signals, and concluded that not all the features are necessary to perform accurate classification. The best-three-feature subset they proposed contains 4 Hz modulation, variance of spectral flux, and pulse metric.

An interesting speech/music discrimination experiment based on the modulation energy only [1] showed good results for 3.56 sec long clean speech and music segments.

The other set of acoustic features for hierarchical audio classification is presented in [10]. They distinguish two types of audio features: physical, which refer to mathematical measures computed directly from the sound wave, and perceptual, which are related to the preception of sounds by human beings. First step, the coarse-level classification of four sound classes, silence, music, speech and environmental sounds, has been done using the set of physical features (energy, zero-crossing rate and fundamental frequency) and simple threshold decisions. Then, the fine-level classification is performed using HMM's based on two perceptual features (timbre and rhythm), for distinguishing more specific classes within each audio type.

New set of features for robust narrowband speech/music discrimination system is proposed in [3]. They have combined the line spectral frequencies (LSF's) with zero-crossing-rate features, and obtained good classification results using short audio segments with frame-delay of 20 ms.

Speech/music discrimination in [9] is based on the quite different approach. The idea is not to develop a set of acoustic features which have distinctive properties for speech and music, but to use an existing speech recognizer. Experiments with four measures derived from posterior probabilities of the speech recognizer neural network showed that this approach is very effective for classification of a long sound segments (15 sec).

In this work, we are interested to examine this approach using short audio segments to make it suitable for real-time applications on multimedia audio files.

3 POSTERIOR PROBABILITY FEATURES

As shown in [9], one possible approach for speech/music discrimination is using the speech recognizer. A hybrid connectionist HMM speech recognizer uses a neural network acoustic classifier [5]. This net is trained to estimate the posterior probabilities $p(q_k|x_n)$ of phonemes q_k ,k=1,...,K, given the vector of acoustic features at the time n, x_n. Those acoustic features are developed to represent speech, and to hide and remove other irrelevant informations such as speaker identity, environment, etc. So, probabilities on the output of the net are specific for speech, and not for the other audio. However, they behave very differently when the input of neural network is not speech [9]. In this work, we made experiments with two of four posterior probability measures presented in [9], entropy and dynamism.

3.1 Entropy

According to information theory, entropy is the measure of unpredictability of the procces. If we are transmitting the signal through the channel, the output of the channel depends not only on the input, but also on the characteristics of the channel. If input signal corresponds to the nature of the channel, the process is predictable and the entropy is low. But if not, the process becomes less predictable, so the entropy becomes higher.

If we consider the speech recognizer neural network as the channel tuned for speech signals, entropy of phoneme posterior probabilities on the output is interpreted like this:

- when the incoming signal is speech, it corresponds to the channel, and the entropy is low;
- when the signal is music, it is not tuned to the channel, so the entropy is high.

Entropy at the time n (n-th frame) is defined as:

$$H(n) = -\sum_{k=1}^{K} p(q_k|x_n) log_2(p(q_k|x_n)) \tag{1}$$

In our experiments we were using the average entropy over N frames:

$$H(n,N) = \frac{1}{N}\sum_{j=n}^{n+N} H(j) \tag{2}$$

where n is the frame number (time index), N is the number of frames in the segment, and $p(q_k|x_n)$ is the posterior probability of phoneme q_k for the given acoustic vector x_n.

Histograms of the average 40-frame entropy calculated for about 20000 speech and music samples (Figure 1) show that those two distributions are quite exclusive, so the discrimination between speech and music using average entropy can be effective.

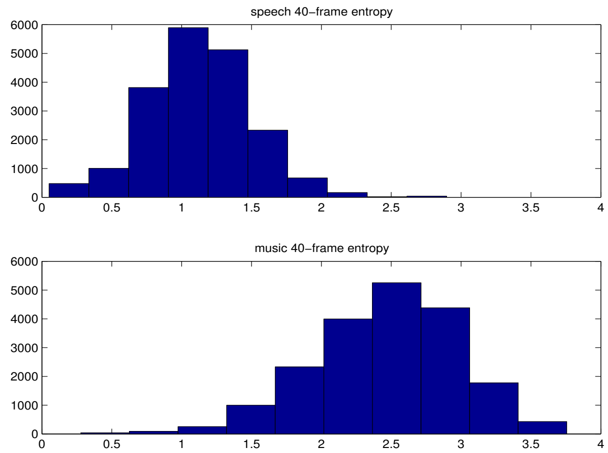

Figure 1: *Histograms of 40-frame entropy values for speech and music class*

3.2 Dynamism

Squared first-order difference, or so-called dynamism [9] is the measure based on the observation that the probabilities estimated for speech signals are changing quite abruptly and frequently, because in the speech signals the phonemes are changing every few tens of miliseconds. Non-speech signals, by contrast, are crossing those phone boundaries less frequently and more gradually.

- posteriors of speech segments are changing frequently and abruptly, so the dynamism is high;
- music segments have posteriors that change gradually and less often, and the dynamism is therefore low.

Dynamism at time n is defined as:

$$d(n) = \sum_{k=1}^{K} (p(q_k|x_n) - p(q_k|x_{n-1}))^2 \tag{3}$$

Average dynamism over N frames in our experiments was calculated like:

$$d(n,N) = \frac{1}{N}\sum_{j=1}^{n+N} d(j) \tag{4}$$

Histograms in Figure 2 show that dynamism distributions for speech and music class are also quite distinctive.

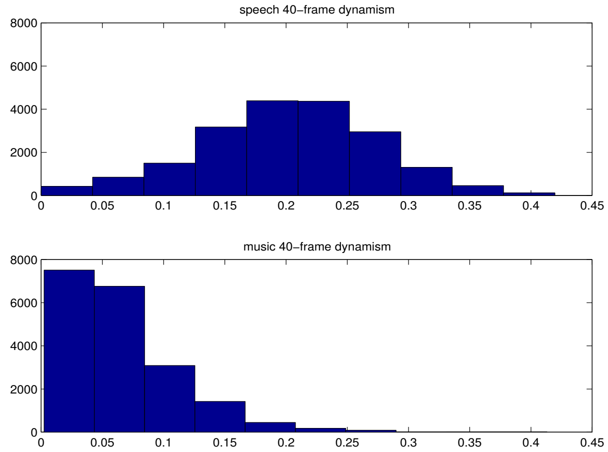

Figure 2: *Histograms of 40-frame dynamism values for speech and music class*

4 HYPOTHESES TESTING

As the histograms on Figure 1 and Figure 2 show, distributions of entropy and dynamism for speech and music class are not really Gaussian, but can be well assumed as Gaussians.

Assuming that they are Gaussians, we have experimented with three hypotheses testing methods:

- Gaussian likelihood ratio test, for the local (frame-level) decision whether is signal at time n speech or music;
- Symmetric Kullback Leibler (KL2) distance, which also gives local decision [8]; however, in our experiments we calculated this metric only for an illustration of distance between speech and music class, we did not use it for testing itself.
- Bayesian Information Criterion (BIC), for the decision if one changing point is present in the signal [2]; this decision is however not local, but segment-level, since it integrates all the data in the segment.

4.1 Gaussian likelihood ratio test

Hypotheses testing using Gaussian likelihood ratio is based on the likelihood ratio discrimination function.

In our experiments, for the decisions based on entropy, this function is defined as:

$$h(H(n,N)) = \frac{(H(n,N) - m_{H_S})^2}{2\sigma_{H_S}} - \frac{(H(n,N) - m_{H_M})^2}{2\sigma_{H_M}} + \frac{1}{2} ln \frac{|\sigma_{H_S}|}{|\sigma_{H_M}|} \tag{5}$$

where $H(n,N)$ is the N-frame entropy measured at the time n (equation (2)) m_{H_S} and m_{H_M} are mean values for speech and music class respectively, and σ_{H_S}, σ_{H_M} their standard deviations.

Similarly, for the dynamism we have:

$$h(d(n,N)) = \frac{(d(n,N) - m_{d_S})^2}{2\sigma_{d_S}} - \frac{(d(n,N) - m_{d_M})^2}{2\sigma_{d_M}} + \frac{1}{2} ln \frac{|\sigma_{d_S}|}{|\sigma_{d_M}|}. \tag{6}$$

For each frame n, the values of the functions (5) and (6) were calculated and then compared with corresponding threshold t; if $h(n) < t$, decision is speech, and if $h(n) > t$, music.

So in this experiment we are making only the local (frame-level) decision. These decisions should be then integrated over time, for example using HMM's for speech and music class (this is planned for the future work).

4.2 KL2-distance

Kullback Leibler distance (or Relative Cross Entropy) between two random variables A nad B is formulated as:

$$KL(A,B) = E_A(log(P_A) - log(P_B)) \tag{7}$$

where E_A is expectation operation performed with respect of probability distribution function of the class A, and P_A, P_B are distributions of classes A and B respectively.

The greater this value, the greater distance between classes A and B, but since this expression is not symmetric, it is not actually a distance metric.

Therefore, the other metric called symmetric KL distance, or KL2 distance, was introduced in [8]:

$$KL2(A,B) = KL(A,B) + KL(B,A) \tag{8}$$

and when both classes have Gaussian distributions, equation (8) becomes:

$$KL2(A,B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (m_A - m_B)^2(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}) \tag{9}$$

where m_A, m_B are mean values and σ_A, σ_B standard deviations of class A and class B respectively. The greater this value, the greater the distance between the classes.

In our experiments, assuming that entropy and dynamism distributions for speech and music class are Gaussians, we calculated KL2 distances for entropy and dynamism according to the equation (9) but just for the illustration of the distance between speech and music class. We did not use this metric for hypotheses testing, but it might be investigated in some future experiments.

4.3 Bayesian Information Criterion

Bayesian Information Criterion (BIC) is a likelihood criterion penalized by the complexity of the model, i.e. number of parameters in the model. If we have data set $X = \{x_i, i = 1, ..., I\}$ which we want to model, and set of possible models $M = \{m_i, j = 1, ..., J\}$, the BIC criterion for the model m_i is defined as:

$$BIC(m_i) = ln(L(X, m_i)) - \frac{\lambda}{2}\#(m_i)ln(N) \tag{10}$$

where $L(X, m_i)$ is the likelihood function for the model m_i, $\#(m_i)$ is number of parameters in the model, and λ is the penalty weight.

The BIC procedure is choosing the model m_i from the set of models M for which the BIC criterion is maximized.

In [2] is shown that BIC criterion can be used for acoustic change detection in the audio stream assuming that the sequence of acoustic feature vectors (set of data to be modeled) is a Gaussian process, and this approach is also used in the other experiments [4].

The simplest problem is to find one changing point in the segment, i.e. to test the hypothesis that there is a change which occures at time n versus the hypothesis that there are no changes in the segment.

Assuming that there is change in the segment at time n, the likelihood ratio is:

$$R(n) = Nln|\Sigma| - N_1 ln|\Sigma_1| - N_2 ln|\Sigma_2| \tag{11}$$

where Σ, Σ_1 and Σ_2 are the sample covariance matrices for all data, for first N_1 samples $\{x_1, ..., x_n\}$, and for the rest N_2 samples $\{x_{n+1}, ..., x_N\}$ respectively, (x_n is the vector of features at time n), and $N = N_1 + N_2$ is the total number of frames in the segment.

The change occures than at

$$n_{cp} = argmax_n(R(n)). \tag{12}$$

The other hypothesis is that there is no changes in the segment, i.e. the segment is acoustically homogeneous. So we are comparing two models: one which models data as two Gaussians, and the other which models data as just one Gaussian. Difference between the BIC functions for those two models at time n is:

$$BIC(n) = R(n) - \lambda P \tag{13}$$

where $R(n)$ is the likelihood ratio defined in (11), penalty P is defined as:

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))lnN \tag{14}$$

where d is dimension of the space, and λ is the penalty weight.
Therefore, if (13) is positive, the model of two Gaussian is favored.

The decision of existing changing point at time n is thus made if

$$max_n(BIC(n)) > 0 \tag{15}$$

or expressing changing point itself as

$$n_{cp} = argmax_n(BIC(n)). \tag{16}$$

In our experiments we tried to use BIC criterion function for this simplest problem, but using entropy and dynamism instead of usually used acoustic features like in [2, 4].

5 EXPERIMENTS

5.1 Methods

Speech/music discrimination in our experiments is based on the following general ideas:

- calculation of posterior probabilities using neural network;
- calculating average entropy (2) and dynamism (4) for speech and music segments;
- estimation of entropy and dynamism distribution values for speech and music;
- decision between speech and music based on Gaussian likelihood ratio test for entropy (5) and for dynamism (6).

We also made one preliminary experiment of acoustic change detection in the segment using Bayesian Criterion Information, which is based on:

- calculation of posterior probabilities of the segment;
- calculating average entropies (2) and dynamismes (4) at each frame along the segment;

- calculating Bayesian Information Criterion difference function at each frame (13); instead of usually used spectral feature vectors, the vectors in our experiments contain the values of entropy and dynamism;
- finding the changing point as the frame index of positive BIC function maximum (16).

5.2 Acoustic model

For the posterior probabilities calculation we used a (9x13)-2000-42 MLP with a softmax output layer trained via back-propagation to a minimum-cross-entropy criteria (i.e. a standard quicknet training).

The input features are the first 13 cepstra of a 12th order PLP fit to the spectrum, based on 16 kHz sampled data, using a 32 ms window and a 16 ms hop size; 9 successive feature frames (thus, 144 ms) are presented to the network at a time. No deltas, double deltas, or explicit energy term is used (however, c0 amounts to an energy term).

The features are normalized with separate online estimates of mean and variance within each dimension, using simple first-order recursive estimator with a time constant of 200 frames (3.2 seconds).

The training set was 45 hours of BBC radio and TV news broadcasts, recorded between 1996 and 1998, manually transcribed at the word level, then labeled via forced alignment over several earlier generations of the acoustic model.

The label set is a 42 phone British-English set from a BEEP42 dictionary, based on BEEP45 but omitting the dipthongs /ea/, /ia/ and /ua/, which are replaced by pairs of monophones.

5.3 Data

For our experiments, we used speech files from BBC news broadcast database recorded between 1996 and 1998, and music files from the THISL music database which contains different types of music (rock, pop, classical, hard, etc.) recorded from different swiss radio stations in 1997.

- For estimation of entropy and dynamism distribution values (so-called training), we used 200 clean speech and 200 music 15 sec long files.
- For tests, in the first set of experiments we used another 100 speech and 100 music 15 sec files (Data 1).

  In the other set of experiments, we used two manually labeled files from BBC news database, each containing one speech and one music segment, i.e. one changing point (Data 2).

  - First test file (file 1) has 50 sec, and speech-to-music changing point after 31 sec;
  - Second test file (file 2) has 40 sec, and music-to-speech changing point after 16 sec.

6 RESULTS

6.1 Data 1

In this preliminary experiment, average entropy (2) and average dynamism (4) were calculated over the whole 15 sec segment (927 frames), and distribution parameters (means and variances) are estimated for both classes and both posterior measures. Gaussian likelihood ratio test was performed on each test file, for both entropy (5) and dynamism (6).

Classification accuracy is calculated for both classes and both measures as a percent of speech frames classified as speech, i.e. percent of music frames classified as music.

The KL2 distance (9) between speech and music class was also calculated for both measures.

Classification accuracy results as well as KL2 distances for both classes and both measures are reported in Table 1.

feature	speech	music	kl2-distance
entropy	98%	100%	80.7
dynamism	96%	99%	48.1

Table 1. *Classification accuracy and KL2 distance measure for 15 sec segments*

Table 1 shows that, as expected [9], classification accuracy is very high for both classes, as well as distance between them. However, class distance for dynamism is lower than the one for entropy, unlike in [9].

Still, classification using such a long segments presumes availability of audio data already broken into speech and non-speech segments, but for segmentation itself this approach would be not good.

6.2 Data 2

In this experiment we tried speech/music discrimination based on short segment features.

This time, each 16 ms (one frame) average entropy (2) and dynamism (4) along N frames were calculated, for several different values of N between 10 and 100 (i.e. 0.16 - 1.6 sec long segments). For the hypothesis testing, we also used Gaussian likelihood ratio functions (5) and (6).

All results are reported in Table 2.

	entropy					dynamism				
	speech		music		kl2	speech		music		kl2
N	file 1	file 2	file 1	file 2		file 1	file 2	file 1	file 2	
10	84.1%	81.6%	59.5%	58.5%	10.9	19.7%	25.1%	99.8%	100%	9.5
20	91.2%	84.8%	61.6%	60.6%	14.8	24.4%	31.4%	100%	100%	11.6
40	95.6%	90.9%	70.8%	59.8%	21.6	19.4%	28.7%	100%	100%	14.5
50	95.5%	94.4%	72.7%	59.1%	24.7	47.9%	63.5%	100%	100%	14.1
60	96.0%	94.6%	72.9%	60.2%	23.5	46.8%	64.2%	100%	100%	19.3
100	98.3%	97.3%	79.1%	58.8%	39	74.3%	88.4%	100%	100%	25.6

Table 2. *Classification accuracy and KL2 distance measure for N-frame segment test*

The experimental results reported in Table 2 showed that in general, longer segments give higher distance between the classes, and higher classification accuracy.

For 10 or 20 frames in segment, the results are not very good. For 100 frames, they are much better. For the other values of N (40, 50, 60), results are very good, but longer segments do not provide necessarely better results.

So, we believe that with choosing N=40, segments would be quite short (0.64 sec), and classification still would be effective enough.

Table 2 also showes that for entropy, classification accuracy for both classes is less than in first experiment, but still high enough.

But for dynamism, music classification accuracy is extremely high, while for speech it is very low.

So this showes that the threshold should not be zero as we assumed. Optimal threshold value is usually obtained by optimizing EER, but in this experiment we did it just looking at dynamism likelihood ratio function (6) of two test files (Figure 2) which shows that threshold could be changed to higher value instead of 0.

We have tried then with the new threshold value, $t = 0.25$ instead of $t = 0$, and results reported in Table 3 show that classification acurracy for both classes became good enough.

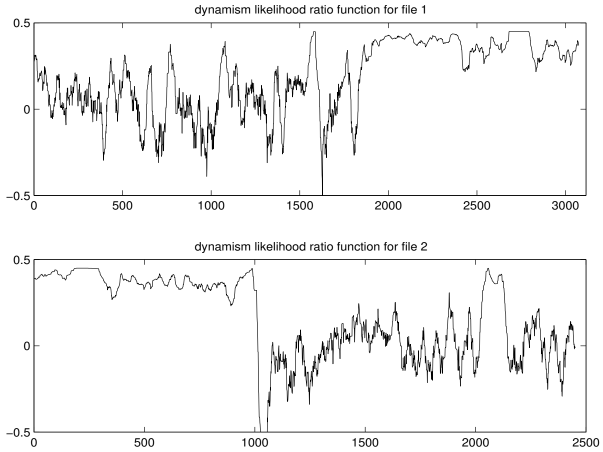

Figure 3: *Dynamism Likelihood Ratio Function for two news files; file 1 has 50 sec, and speech-to-music changing point after 31 sec (1928 frames); file 2 has 40 sec, and music-to-speech changing point after 16 sec (991 frames)*

	entropy					dynamism				
	speech		music		kl2	speech		music		kl2
N	file 1	file 2	file 1	file 2		file 1	file 2	file 1	file 2	
10	84.1%	81.6%	59.5%	58.5%	10.9	59.2%	68.4%	95.8%	97.8%	9.5
20	91.2%	84.8%	61.6%	60.6%	14.8	75.3%	82.8%	96.5%	97.5%	11.6
40	95.6%	90.9%	70.8%	59.8%	21.6	87.4%	91.4%	96%	98.7%	14.5
50	95.5%	94.4%	72.7%	59.1%	24.7	94.4%	93.2%	87.8%	93.9%	14.1
60	96.0%	94.6%	72.9%	60.2%	23.5	93%	93.6%	96.4%	99.9%	19.3
100	98.3%	97.3%	79.1%	58.8%	39	95.8%	96.5%	91.2%	96.1%	25.6

Table 3. *Classification accuracy for N-frame segments with changed treshold value for dynamism (t=0.25 instead of t=0)*

6.3 BIC changing point detection on Data 2

Our fourth experiment is based on the idea of finding one changing point in the segment using Bayesian Information Criterion, because this approach has been shown to be robust and threshold free [2].

So we made the preliminary experiment using two-dimensional entropy-dynamism space instead of usual spectral feature vectors [2, 4]. Since our vectors contain two elements, entropy and dynamism, we have the dimension space $d = 2$. For this experiment, calculation of average posterior measures was carried out over N=40 frames.

As for penalty weight λ, we simply used $\lambda = 1$ according to the BIC theory, although this parameter can be tuned to obtain better segmentation [2].

The BIC function (16) for two news test files is shown in Figure 4.

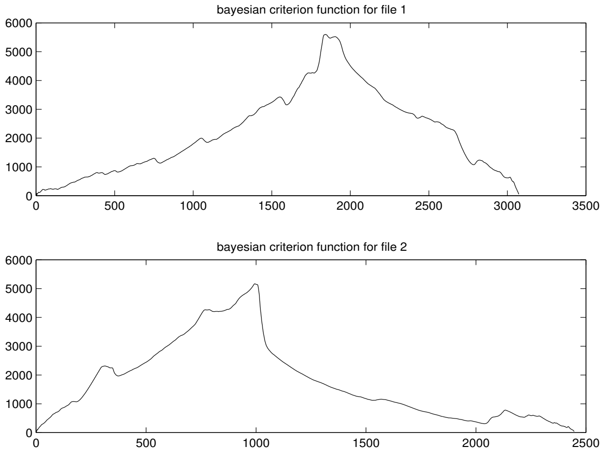

Figure 4: *Bayesian Information Criterion function for two news files; file 1 has 50 sec, and speech-to-music changing point after 31 sec (1928 frames); file 2 has 40 sec, and music-to-speech changing point after 16 sec (991 frames)*

Figure 4 shows that BIC-function based on entropy and dynamism behaves in the same way like the one based on acoustic feature vectors, i.e. for audio segments containing one changing point has one very distinctive maximum. So it's necessarely just to detect if this maximum is positive, and if yes, to find out at which time (frame) it occured. So using BIC with posterior measures space for the speech/music segmentation is apparently possible, but this results are based only on one first preliminary experiment. So it has to be tested on much more data.

7 CONCLUSION

Those experiments have confirmed the idea that the measures derived from phone posterior probabilities can be quite effective in distinguishing speech segments from music ones [9].

With the experiment on 15 sec files, we basically confirmed the results in [9].

With tests on N-frame average entropy and dynamism, we shown that frame-level speech/music discrimination can be also effective using short segments. But since it is only local decision, in the further work we plan to integrate this decisions in time using HMM's.

Our last experiment showed that those two posterior measures might be also used for changing point detection via BIC, although for some more reliable results it should be tested on more data.

In future work, we will perform segment-level speech/nonspeech segmentation of large audio files, HMM's as sound classes models, and try BIC based segmentation on this data.

References

[1] P. Balabko, "Speech and music discrimination based on signal modulation spectrum", research report, 1999

[2] S. S. Chen, P. S. Gopalakrishnan, "Speaker, environment and channel change detection via the Bayesian Information Criterion", DARPA Speech Recognition Workshop, 1998

[3] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/music discrimination for multimedia applications", ICASSP 2000

[4] J. Ferreiros Lopez, D. P. W. Ellis, "Using acoustic condition clustering to improve acoustic change detection on broadcast news", ICSLP 2000

[5] S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, "Connectionist probability estimators in HMM speech recognition", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 1, Part II, 1994

[6] J. Saunders, "Real time discrimination of broadcast speech/music", proc. of ICASSP 1997, pp 993-996

[7] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", Proc. ICASSP, Munchen 1997

[8] M. Siegler, U. Jain, B. Raj, R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio", Proc. DARPA Speech Recognition Workshop, 1997

[9] G. Williams, D. Ellis, "Speech/music discrimination based on posterior probability features", Eurospeech, Budapest 1999

[10] T. Zhang, C.-C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving", Proc IEEE International Conference on Acoustic, Speech and Signal Processing, pp 605-608, 1999