14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPEECH/MUSIC DISCRIMINATION FOR RADIO BROADCASTS USING A HYBRID HMM-BAYESIAN NETWORK ARCHITECTURE

Aggelos Pikrakis, Theodoros Giannakopoulos and Sergios Theodoridis

Dept. of Informatics and Telecommunications
University of Athens
Panepistimioupolis, 15784, Athens, Greece
email: {pikrakis,tyiannak,stheodor}@di.uoa.gr

ABSTRACT

This paper presents a speech/music discrimination scheme for radio recordings using a hybrid architecture based on a combination of a Variable Duration Hidden Markov Model (VDHMM) and a Bayesian Network (BN). The proposed scheme models speech and music as states in a VDHMM. A modified Viterbi algorithm for the computation of the observations' probabilities at each state is proposed. This is achieved by embedding a BN, that outputs to the HMM the required probability values. The proposed system has been tested on audio recordings from a variety of radio stations and has exhibited an overall performance close to 95%.

1. INTRODUCTION

The problem of speech/music discrimination is important in a number of audio content characterization applications. Since the first attempts in the mid 90's, a number of speech/music discrimination algorithms have been implemented in various application fields.

In [1], a real-time technique for speech/music discrimination was proposed, focusing on the automatic monitoring of radio channels, using energy and zero-crossing rate (ZCR) as features. In [2], thirteen audio features were used to train different types of multidimensional classifiers, including a Gaussian MAP estimator and a nearest neighbor classifier. In [3], for the purposes of analyzing on-line audiovisual data, energy, ZCR and fundamental frequency were used as features and segmentation/classification was achieved by means of a procedure based on heuristic rules. A framework based on a combination of standard Hidden Markov Models and Multilayer Perceptrons (MLP) was used in [4] for speech/music discrimination of broadcast news. An Adaboost-based algorithm, applied on the spectrogram of the audio samples, was used in [5] for frame-level discrimination of speech and music. In [6], energy and ZCR were employed as features and classification was achieved by means of a set of heuristic criteria that were stemming from the nature of the speech and music signals.

The majority of the above methods deal with the problem of speech/music discrimination in two separate steps. At a first step the audio signal is split into segments by detecting abrupt changes in the signal statistics. At a second step the extracted segments are classified as speech or music by using standard classification schemes. The work in [4] differs in the sense that the two tasks are performed jointly by means of a standard HMM, where a MLP serves as an estimator of continuous observation densities of feature vectors extracted by short-term audio processing.

In this paper, a joint segmentation/classification scheme is employed for speech/music discrimination using a hybrid architecture consisting of a Variable Duration Hidden Markov Model (VDHMM) and a Bayesian Network (BN). The audio recording is first split into a number of short-term frames by means of a short-term processing window and five features are extracted per frame. The resulting sequence of feature vectors is subsequently fed as input to a VDHMM, where state duration is modeled explicitly, in contrast to the previously used standard HMM modeling, where self-transitions to states impose an exponentially decaying state duration probability. In the present paper, a BN is employed as a probability estimator. When a sequence of feature vectors is emitted by a state, the joint probability of these observations is determined by the BN, which has been trained as a binary classifier (speech and music being the two classes), and returns the posterior class probability. The optimal cost of the VDHMM is computed in the Viterbi algorithm sense, where the required state observation probabilities are computed by the BN.

The novelty of our approach is twofold: a) a VDHMM is used for speech/music modeling. Music/speech classes correspond to HMM states and the use of state duration provides the means of an explicit modeling of the time the model spends at each state; b) a modification of the Viterbi algorithm is proposed, so that a BN can serve as a probability estimator for sequences of observations emitted by the HMM states. The use of the BN provides the means of overcoming the assumption of independence that underlies the Viterbi algorithm.

The paper is organized as follows: Section 2 is the feature extraction stage, Section 3 describes the VDHMM architecture and related issues and Section 4 presents the adopted BN. Results and experiments are presented in Section 5 and conclusions are drawn in Section 6.

2. FEATURE EXTRACTION

At a first step, the audio recording is broken into a sequence of non-overlapping short-term frames (50ms long) and five features, popular in speech/music discrimination, are extracted per frame, namely energy, zero-crossing rate, spectral entropy and the first two Mel-Frequency Cepstrum Coefficients (MFCCs). Therefore, at the end of this stage, the audio recording is represented by a sequence $\mathbf{F}$ of five-dimensional feature vectors, i.e.,

$$\mathbf{F} = \{\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_T\}$$

where T is the number of short-term frames. This choice of features was the result of extensive experimentation. It has to be noted that this is not an optimal feature set in any sense and other choices may also be possible.

If $\{x(0), x(1), \dots, x(N-1)\}$ is the short-term frame, then the adopted features are given by:

1. **Energy:** This is a popular feature used in speech/music discrimination systems and is defined as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n) \tag{1}$$

2. **Zero Crossing Rate (ZCR):** It measures the number of time-domain zero crossings (divided by the frame's length).

$$ZCR = \frac{1}{N} \sum_{n=1}^{N-1} \frac{|sgn\{x(n)\} - sgn\{x(n-1)\}|}{2} \tag{2}$$

where $sgn(.)$ stands for the sign function, i.e., $sgn\{x(n)\} = +1$ if $x(n) \geq 0$ and -1 if $x(n) < 0$.

3. **Spectral Entropy**. Entropy is a measure of the uncertainty or disorder in a given distribution [7]. In order to compute spectral entropy [8], the spectrum of the short-term frame is first divided into L sub-bands (bins). The energy X_i of the i-th sub-band, $i = 0,\ldots,L-1$, is then normalized by the total spectral energy, yielding $n_i = \frac{X_i}{\sum_{i=0}^{L-1} X_i}$, $i = 0,\ldots,L-1$. The entropy of the normalized spectral energy is then computed by the equation:

$$H = -\sum_{i=0}^{L-1} n_i \cdot log_2(n_i) \qquad (3)$$

4. **The first two Mel Frequency Cepstral Coefficients (MFCCs)**. The filter bank used for the computation of the MFCCs consists of 40 triangular bandpass filters, with bandwidth and spacing determined by a constant mel-frequency interval. More specifically, the first 13 filters are linearly-spaced with 133.33Hz between center frequencies and are followed by 27 log-spaced filters, whose filter centers are separated by a factor of 1.0711703 in frequency. The adopted filter bank covers the frequency range $0-8$kHz, suggesting a sampling rate of 16kHz. If $\widetilde{S_k}$, $k = 1,\ldots,40$ is the output of the k-th filter, then the first two MFCCs are given by the equation

$$\widetilde{c_n} = \sum_{k=1}^{40} (\log \widetilde{S_k}) \cos[n(k-\frac{1}{2})\frac{\pi}{40}], \;\; n = 1,2$$

3. HIDDEN MARKOV MODEL

The extracted feature sequence, $\mathbf{F}$, is subsequently fed as input to the VDHMM of figure 1. This is a two state model, where state M models music and state S models speech. The basic idea behind this approach is that states M and S are allowed to emit sequences (a number of successive frames) of feature vectors (*audio segments*). The length of each sequence varies, hence the term variable duration HMM (this also justifies the fact that self-transitions to these states are not possible). As a result, a sequence of successive states corresponds to a sequence of audio segments. Therefore, if the Viterbi algorithm is used, then the best-state sequence corresponds to the optimal sequence of segments in the Viterbi sense.

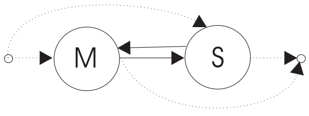

Figure 1: Basic VDHMM architecture

Translated in the HMM terminology, let $\mathscr{H}$ be the resulting VDHMM, π_{2x1} the vector of initial probabilities, $A_{2\times 2}$ the state transition matrix and $p_j(\tau)$ the state duration probability, i.e. the probability that the model emits a segment of τ successive observations (feature vectors) at state j, $j = 1,2$. For such a model, the "forward variable" [9] is defined as

$$a_t(j) = P(\mathbf{O}_1\mathbf{O}_2\ldots\mathbf{O}_t,\, state\ j\ ends\ at\ t|\mathscr{H}), j = 1,2$$

$a_t(j)$ stands for the probability that the model finds itself in the j-th state after the first t feature vectors have been emitted. It can be shown that ([10]),

$$\begin{aligned} a_t(j) &= \max_{D_{j_{min}} \le \tau \le D_{j_{max}}, 1\le i \le 2, i \ne j} [\delta_t(i,\tau,j)] && (4) \\ \delta_t(i,\tau,j) &= a_{t-\tau}(i) A_{ij} p_j(\tau) \prod_{s=t-\tau+1}^{t} b_j(\mathbf{O_s}) && (5) \end{aligned}$$

where $D_{j_{min}}$ and $D_{j_{max}}$ is the minimum and maximum allowable duration at state j. In addition, $b_j(.)$ is the continuous observation density function for state j (for the case of continuous-valued features, as in our approach). Usually, $b_j(.)$ is modeled as a mixture of log-concave or elliptically symmetric densities (e.g. Gaussians). It is important to observe that in equation (5), it has been assumed that the feature vectors to be emitted from state j are *statistically independent*, hence the term $\prod_{s=t-\tau+1}^{t} b_j(\mathbf{O_s})$.

3.1 Embedding a BN in the Viterbi algorithm

In this paper a modification of equation (5) is proposed. Instead of assuming statistical independence (which is hardly the case in practice), we treat the sequence of feature vectors to be emitted from state j as a joint entity (segment). To compute the joint probability of a segment, a BN is used. We assume that the BN has been trained as a binary classifier. Given a segment, the BN classifier decides in favor of either the speech or music class. The BN outputs the respective posterior probability, p, of the winning class. Obviously, the probability that the segment does not belong to the winning class is $1-p$. As a consequence, the term $\prod_{s=t-\tau+1}^{t} b_j(\mathbf{O_s})$ in (5) can be replaced by the probability returned by the BN for class j, $j = 1,2$. Let $pBN_j(\mathbf{O_{t-\tau+1}},\ldots,\mathbf{O_t})$ be the BN probability that the segment $\{\mathbf{O_{t-\tau+1}},\ldots,\mathbf{O_t}\}$ is emitted by state j, i.e., for state M ($j = 1$) this is equivalent to the probability that this is a music segment and for state S ($j = 2$) this stands for the probability that this is a speech segment. As a result, equations (4) and (5) become

$$\begin{aligned} a_t(j) &= \max_{D_{j_{min}} \le \tau \le D_{j_{max}}, 1\le i \le 2, i \ne j} [\delta_t(i,\tau,j)] && (6) \\ \delta_t(i,\tau,j) &= a_{t-\tau}(i) A_{ij} p_j(\tau) pBN_j(\mathbf{O}_{t-\tau+1},\ldots,\mathbf{O}_t) && (7) \end{aligned}$$

If $a_t(j)$ is computed for all time instants, then a standard backtracking procedure will extract the best-state sequence, which will correspond to a *sequence of segments*. The class of each segment (i.e., speech/music) coincides with the label of the state from which the respective observations have been emitted. The above suggests that speech/music discrimination becomes a joint segmentation/classification task, performed by a hybrid VDHMM-BN architecture.

The remaining parameters of the VDHMM, i.e., π, A and $p_j(\tau)$, $j = 1,2$ can be set to predefined values according to the following rationale:

- π stands for the probability that the first segment is music or speech. As both cases are a priori equiprobable, it makes sense to set $\pi = [0.5 \;\; 0.5]^T$.
- Due to the fact that self transitions to states are not allowed by the definition of the variable duration HMM, the state transition matrix A becomes
$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
- As is always the case with VDHMMs, the minimum and maximum allowable state duration is an issue that needs careful consideration. In real world audio signals from radio broadcasts, a speech or music segment can be arbitrarily long. Equivalently, feature sequences (segments) emitted by states can be arbitrarily long. As can be seen from equations (6) and (7) this results in a significant increase in computational cost, due to the large number of δ arguments over which maximization takes place when $\alpha_t(j)$ is computed. To remedy this problem, led us to modify the VDHMM architecture of figure 1, by introducing in the VDHMM a so-called non-emitting state [10], as shown in figure 2. The use of the non-emitting state, N, allows state sequences to consist of a succession of music (speech) states. This is in a way equivalent to permitting self-transitions to states. As a result, in the best-state sequence a long music (speech) segment will appear as a sequence of shorter music (speech) segments. This is not a restriction, because, once the best-state sequence is extracted, a subsequence of segments belonging to the same class can be merged to form a longer segment by means of a simple post-processing algorithm. It is now possible to set the limits of state durations to smaller values, thus solving the problem

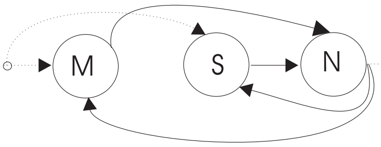

Figure 2: Modified VDHMM architecture

of increased computational complexity. In our experiments, the minimum and maximum state duration, for both classes, were set equal to 1 and 3 seconds respectively (20 to 60 observations, given a short-term processing step of 50*ms*). For the new model, which has three states, $\pi = [0.5\ \ 0.5\ \ 0]^T$ and A now becomes:

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

It also has to be noted, for the sake of completeness, that the forward variable is also computed for state N at time instant t, after $a_t(j)$ is calculated for the first two states. This is achieved by the following equation

$$a_t(j=3) = max\{a_t(j=1), a_t(j=2)\} \quad (8)$$

- The last set of parameters to deal with, is the set of state duration probabilities. The physical meaning is to account for how frequently in real-world signals, a segment of specific length is likely to be encountered. Given the nature of radio broadcasts and diversity of radio stations, it is hard to decide on the values of the state duration probabilities for the two states that model music and speech. Therefore it makes sense to assume a uniform distribution for both states, for segments whose length lies in the range $[1\text{secs}, 3\text{secs}]$. The fact that we choose the same duration limits for the two states, leads to identical state duration probabilities for the two states and as a result, the term $p_j(\tau)$ can be omitted from equation (7). Of course, if a priori knowledge of the audio broadcasts to be segmented permits certain hypotheses to be adopted for the state duration probabilities (and/or for the allowable state durations), then this type of knowledge can be straightforwardly accommodated in equation (7).

3.2 Treating zero-energy segments as a third class

The above architecture does not treat zero-energy (silent) segments as a distinct class, which is also the case with a number of proposed systems. It may be desirable to enhance the above VDHMM by permitting the possibility of zero-energy segments as a separate class (state). This can be achieved by resorting to the HMM of figure 3, where we have added a new state, labeled E. This state is also allowed to emit sequences of feature vectors, as the states modeling music and speech. However, in this case, the joint probability of emitted observations is computed differently, i.e., zero-energy is not included as a class in the BN. More specifically, the joint probability that observations $\{\mathbf{O_{t-\tau+1}}, \ldots, \mathbf{O_t}\}$ form a zero-energy segment, is given by the number of observations whose energy (first feature) is below a predefined threshold divided by τ, the length of the emitted sequence of observations from state E. Clearly, this fraction takes values in the range $[0,1]$ and can thus be interpreted as a probability. Experimentation led us to set the value of this threshold to be equal to $\frac{1}{8}$ of the mean value of the energy across the whole audio recording.

Given the fact that the new model has four states, $\pi = [\frac{1}{3}\ \ \frac{1}{3}\ \ \frac{1}{3}\ \ 0]^T$ and A becomes

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}$$

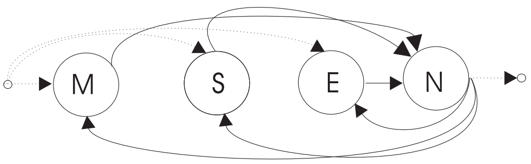

Figure 3: A third state has been added to the VDHMM in order to model zero-energy segments

4. BAYESIAN NETWORK ARCHITECTURE

As explained in Section 3, the classification probability returned by a BN has been embedded in the Viterbi algorithm for the calculation of the forward variable. To this end, the BN has been trained as a binary classifier over speech and music data. In other words, given a segment (feature sequence), the BN returns the respective class label along with the classification probability, by *combining* the results of five individual classifiers, each of which operates on a separate feature. The idea for such a classification scheme was to use for each feature a very simple linear classifier, and then to combine the results via a BN. This has the advantage of using very simple classifiers and then exploit a BN as a combiner to boost the performance. Such an architecture has the advantage that the required (by the HMM) conditional probability (given the segment) is naturally provided by the BN, whose function is known to be tailored for such a task.

4.1 Individual Classifiers

At a first step, a separate statistic is calculated for each feature dimension (distinct feature) of the feature vector sequence to be classified. The statistics that we use are shown in Table 1.

Feature	Statistic
Energy	$\frac{\sigma^2}{\mu^2}$
ZCR	$\frac{max}{\mu}$
Spectral En.	σ^2
MFCC 2	σ^2
MFCC 1	μ

Table 1: Statistics for each one the five features that have been used

The choice of statistics was motivated by the nature of the audio signals under study. For example, the standard deviation of the second MFCC, computed over a number of short-term frames, exhibits higher values for speech segments. This can be observed in the histogram of Figure 4. Similar observations hold for the other statistics. In the literature, similar features have been frequently used (e.g., [6]). For each extracted statistic, an individual binary classifier is used, i.e., a separate classification decision is taken. The nature of the specific task dictates to optimize the classifiers during training, by minimizing a risk function and not the overall classification error. This is equivalent to a relative shift of the decision threshold. In order to determine the value of the threshold, a training set is used per classifier, containing segments of both classes. The histogram of the statistic is calculated per class, and the threshold is chosen to be equal to the value for which the error rates of the two classes are equal. If the minimum overall error rate is chosen as a criterion instead, this may lead to different error rates per class, with the risk of ending up with a threshold biased towards one of the classes, especially in the case of insufficient training sets (see Figures 4, 5).

4.2 Bayesian Network Classifier

Each one of the above individual binary decisions is fed as input to a BN, which combines the results and takes the final decision. BNs

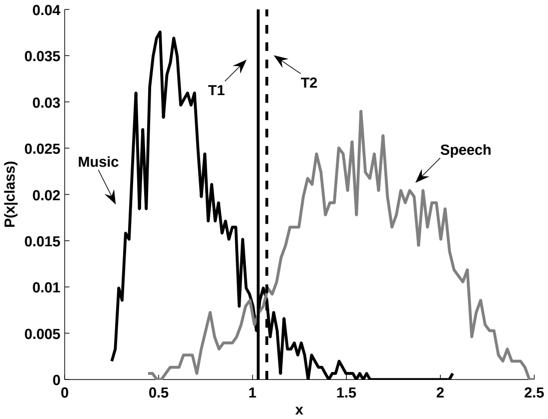

Figure 4: Histograms of music and speech data for the standard deviation of the second MFCC. T_1 is the adopted threshold, whereas T_2 is the threshold that minimizes the overall classification error

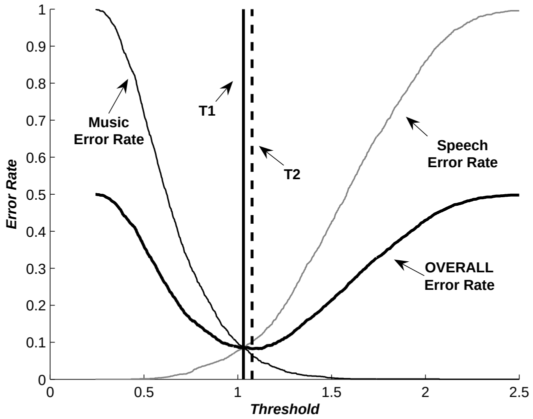

Figure 5: Classification errors (overall and per class) for the classifier operating on the standard deviation of the second MFCC, plotted over various threshold values. T_1 is the adopted threshold, whereas T_2 is the threshold that minimizes the overall classification error, which however yields different error rates per class

are directed acyclic graphs (DAGs) *that encode conditional probabilities* between a set of random variables. In the case of discrete random variables, for each node (random variable) A, with parents B_1, ..., B_k a conditional probability table (CPT) $P(A|B_1,...,B_k)$ is defined. In this paper, the BN architecture shown in figure 6, has been used as a scheme for combining the decisions of the individual classifiers described above [11]. We will refer to this type of BN as the BNC (Bayesian Network Classifier). Nodes h_1, ..., h_n (also called hypotheses, rules, attributes or clauses) correspond to the binary decisions of the individual classifiers, while node Y is the output node and corresponds to the true class label. In the BN training step, the CPTs of the BN are learned according to the set:

$$S = \{(h_1(1),\ldots,h_n(1),s(1)),\ldots,(h_1(m),\ldots,h_n(m),s(m)))\} \quad (9)$$

where $h_j(i)$ is the result of classifier $j = 1,\ldots,n$ for input x_i^j, where x_i^j is the feature value presented to the j-th classifier representing the i-th input pattern, $s(i)$ is the *true label* for $x_i^j, j = 1,\ldots,n$ and m is the total number of training samples. Set S is generated by validating each individual classifier with a test set of length m, in our case a set of m audio segments with known true class label. The CPTs of the BN are learned according to the Maximum Likelihood principle ([12]).

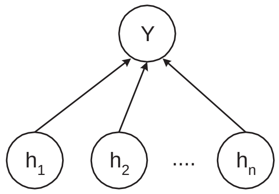

Figure 6: BNC architecture

The BN makes the final decision, based on the conditional probability $P_{dec} = P(Y|h_1,...,h_n)$. The process of calculating P_{dec} is called *inference* and it is in general a very time consuming task. However, for the adopted BNC architecture no actual inference algorithm is needed, since the required conditional probability is given by the CPT itself, that has been learned in the training phase. Another advantage of the specific architecture *is that no assumption of conditional independence between the input nodes is made*, like e.g. in the Naive Bayesian Networks.

To summarize, in the current work, a BN that has been trained as a binary classifier, has been embedded in the mechanism of a VDHMM (see Section 3). Therefore, the joint probability of a segment of successive observations that are emitted by a state of the HMM, has been replaced by the classification probability of the segment, returned by the BN. This classification probability is computed in a three-step process, namely:

1. For any segment, the values of the five statistics are calculated, i.e., x_j, $j = 1,\ldots,5$.
2. x_j is fed as input to the j-th classifier. Therefore, five binary decisions h_j are extracted.
3. $P_{dec} = P(Y|h_1,...,h_5)$ is calculated by inferring in the trained BN.

If the label of the state of the HMM coincides with the class predicted by the BN, then the probability of the segment emitted by the state is set equal to P_{dec}. If, on the other hand, they are different, then the probability of the segment is set equal to $1 - P_{dec}$.

5. EXPERIMENTS

5.1 BN-related training and testing issues

The hybrid HMM-BN scheme that is proposed in this paper suggests that a BN has to be trained before it is employed by the VDHMM. To this end, a dataset (referred to D_1 in the sequel) consisting of 340 minutes of BBC on-line radio recordings has been compiled in order to train and evaluate both the individual classifiers and the BN. More than 5000 audio segments were extracted from D_1 and were manually labeled as speech or music. The duration of those segments varied in the range $[0.5\text{secs}, 5\text{secs}]$. At a first step the five individual classifiers were trained, i.e., the five decision thresholds were set. For this purpose, 20% of the segments in D_1 were used, generating approximately 500 segments for each class. Table 2 presents the threshold value for each individual binary classifier.

Classifier	En.	ZCR	Sp. En.	MFCC2	MFCC1
Threshold	-0.58	-2.84	-0.36	-1.06	-15.49

Table 2: Threshold values for the individual classifiers

At a second stage, the BN was trained, as explained in section 4.2. For this purpose, 60% of the segments in D_1 were used, resulting into 1600 audio segments for each class.

Finally, the remaining 20% of the segments in D_1 were used for testing the classification performance of both the BN and the individual classifiers. In Table 3, the error rates of the individual classifiers and the BNC are presented. The classification error rate for

Classifier	E.	ZCR	Sp. E.	Mfcc2	Mfcc1	BNC
% Error	17.2	18.1	16.6	6.9	12.9	5.2

Table 3: Classification results

the BNC is 5.2%, while the error rate for the best classifier (i.e. the classifier corresponding to the 2nd MFCC) is 6.9%. This amounts to approximately a 25% error reduction with respect to the performance of the best individual classifier.

5.2 Overall system testing

In order to measure the overall segmentation/classification accuracy, a dataset, D_2, consisting of 11 uninterrupted radio recordings was compiled. The length of the recordings varied in the range $5-45$ minutes, yielding a total duration of 2.5 hours. D_2 was compiled from 7 different on-line BBC radio stations, covering a wide range of broadcasts. In order to estimate the overall discrimination error, all recordings were manually segmented and labeled as speech, music or silence. This manual procedure revealed that 79.89% of D_2 was music, 19.16% was speech and 0.95% was silence.

The proposed HMM-BN scheme classified 78.58% of the data as music, 19.89% as speech and 1.53% as silence. Table 4 presents the corresponding confusion matrix. Each element, C_{ij}, of the matrix, corresponds to the percentage of data whose true class label was i and was classified to class j. The sum of values at the i-th row is equal with the percentage of data in D_2 whose true class label is i. Similarly, the sum of values at the j-th column is equal with the percentage of data which were classified in class j.

	Music	Speech	Silence
Music	76.88	2.12	0.89
Speech	1.63	17.47	0.06
Silence	0.06	0.30	0.59

Table 4: Confusion matrix of the proposed discrimination scheme

From the confusion matrix we can directly extract the following measures for each class:

1. **Recall (R_i):** This measure is defined as $R_i = \frac{C_{ii}}{\sum_{j=1}^{3} C_{ij}}$. R_i stands for the proportion of data with true class label i, that were correctly classified in that class.
2. **Precision (P_i):** P_i is defined as $P_i = \frac{C_{ii}}{\sum_{j=1}^{3} C_{ji}}$. P_i is the proportion of data classified in class i, whose true class label is indeed i.

Recall and precision values for each class are presented in Table 5. The confusion matrix along with the recall and precision values, reveal the fact that, the proposed method achieved high classification rates for speech and music and a lower classification rate for silence. This is because a large number of silent segments were attached to the endpoints of neighboring speech segments. Given the fact that speech/music discrimination is usually a preprocessing stage in various application fields, the low classification rate for silence is not necessarily a restriction.

Furthermore, it can be observed that 0.89% of D_2 (i.e. C_{13}) were music data which were misclassified as silence. This mainly stems from music data with a very low energy value, which results in high probabilities being generated by the state of the VDHMM that models silence.

As a final remark, the overall accuracy of the proposed method is 94.95%, which stands for the percentage of correctly classified data. The overall accuracy is equal with the sum of values in the diagonal of the confusion matrix.

6. CONCLUSIONS

In this paper, we presented a joint segmentation/classification scheme for speech/music discrimination using a hybrid architecture consisting of a Variable Duration Hidden Markov Model (VDHMM) and a Baysian Network (BN). The novelty of our approach lies in the facts that : a) a variable duration HMM is used for speech/music modeling and b) a modification of the Viterbi algorithm is proposed, so that the BN can serve as a probability estimator for sequences of observations emitted by the HMM states. The system was tested on approximately 3 hours of audio recordings of on-line radio broadcasts from various radio stations. Approximately 95% of the audio data was correctly segmented and classified. In the future, given the fact that the performance of the BN is crucial to the system, we will experiment with more features that have been proposed in the literature and different BN architectures. Furthermore, we will investigate the possibility to model silence more efficiently and include, as a class in the BN, the case of speech-over-music segments, which in the current approach are given a speech label.

	Music	Speech	Silence
Recall	96.24	91.21	61.97
Precision	97.84	87.83	38.59

Table 5: Recall and precision values for each class

REFERENCES

[1] J. Saunders, ”Real-time discrimination of broadcast speech/music”, in *Proc. ICASSP 1996,* vol 2, pages 993-996, Atlanta, May 1996.

[2] E. Scheirer and M. Slaney, ”Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator”, in *Proc. ICASSP 1997*, pages 1331–1334, Munich, Germany

[3] Tong Zhang and C.-C. Jay Kuo, ”Audio Content Analysis for Online Audiovisual Data Segmentation and Classification”, in *IEEE Transactions On Speech And Audio Processing*, Vol. 9, No. 4, MAY 2001

[4] Jitendra Ajmera, Iain McCowan and Herve Bourlard, ”Speech/music segmentation using entropy and dynamism features in a HMM classification framework”, in *Speech Communication 40 (2003) 351363,*

[5] N. Casagrande, D. Eck, and B. Kigl. ”Frame-level audio feature extraction using AdaBoost.”, In *Proc. ISMIR 2005*, London, UK, 2005.

[6] C. Panagiotakis and G. Tziritas, “A Speech/Music Discriminator Based on RMS and Zero-Crossings,” *IEEE Trans. Multimedia*, vol. 7(1), pp. 155–166, Feb. 2005.

[7] A. Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stohastic Processes, 4th edition,* McGraw-Hill, NY, 2001.

[8] Hemant Misra, Shajith Ikbal, Herve Bourlard, and Hynek Hermansky, ”Spectral entropy based feature for robust ASR ,” in *Proc. of ICASSP*, Montreal, Canada, 2004.

[9] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, 3d edition*. Academic Press, 2005.

[10] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of the IEEE*, Vol. 77, No. 2, 1989

[11] A. Garg, V. Pavlovic and T.S. Huang, “Bayesian Networks as Ensemble of Classifiers”, Proceedings of the IEEE International Conference on Pattern Recognition, pp. 779-784, Quebec City, Canada, August 2002.

[12] D. Heckerman, “A Tutorial on Learning With Bayesian Networks”, Microsoft Research, MSR-TR-95-06, Mar. 1995