# ROBUST HMM-BASED SPEECH/MUSIC SEGMENTATION

*Jitendra Ajmera* *Iain A. McCowan* *Hervé Bourlard*

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P. O. Box 592, CH-1920 Martigny, Switzerland
{jitendra, mccowan, bourlard}@idiap.ch

**ABSTRACT**

In this paper we present a new approach towards high performance speech/music segmentation on realistic tasks related to the automatic transcription of broadcast news. In the approach presented here, the local probability density function (PDF) estimators trained on clean microphone speech are used as a channel model at the output of which the entropy and "dynamism" will be measured and integrated over time through a 2-state (speech and and non-speech) hidden Markov model (HMM) with minimum duration constraints. The parameters of the HMM are trained using the EM algorithm in a completely unsupervised manner. Different experiments, including a variety of speech and music styles, as well as different segment durations of speech and music signals (real data distribution, mostly speech, or mostly music), will illustrate the robustness of the approach, which in each case achieves a frame-level accuracy greater than 94%.

## 1. INTRODUCTION

The problem of speech/music discrimination has become increasingly important as automatic speech recognition systems (ASR) are applied to more real world applications. For example, in the task of automatic speech recognition of broadcast news and sports videos, the data typically contains clean speech interspersed with segments of other sounds (generally music). In order to transcribe the speech content in an audio stream of this nature, it is necessary to first segment the stream into homogenous regions of "recognizable" speech and music so that the ASR can only be applied to the valid speech segments.

In recent years, many systems have been proposed for real-time discrimination of speech/music signals [1, 2, 3, 4]. Most of these systems are based on acoustic features that attempt to capture the temporal and spectral structures of the signals. These features include, among others, zero-crossing information, energy, pitch, cepstral coefficients, line spectral frequencies (LSF), 4 Hz modulation energy, amplitude and perceptual features like timbre and rhythm. Recently, a different approach was considered by Williams and Ellis [5], who investigated the use of features based on the phonetic posterior probabilities generated in a speech recognition system.

Besides the selection of appropriate acoustic features, another issue is the classification algorithm. Different classifiers such as the Bayesian information criterion (BIC) [4], Gaussian likelihod ratio (GLR) [1, 2, 6], quadratic Gaussian classifier (QGC) [7], nearest neighbourhood classifier [6, 7] and hidden Markov model (HMM) [5] have been used for this purpose.

In this work, we use posterior probability based features introduced in [5], namely, *entropy* and *dynamism*, and confirm that they indeed exhibit discriminatory properties for speech/music signals. In our proposed system, the features are then integrated over time through a 2-state HMM with minimum duration constraints, enabling the audio stream to be classified into a sequence of speech and music segments. The minimum duration constraint eliminates short, ambiguous sounds, which we assume to not carry significant information. The HMM is trained in an unsupervised fashion, leading to a system which can be easily adapted to new conditions, and could be extended to further sound classes. The proposed system thus has the advantages of

1. using more appropriate, higher-level features based on the quality of the classifier, considered here as a channel model,
2. being a threshold-free, global decision making strategy,
3. allowing completely unsupervised training of the HMM parameters, removing the need for labeled training data, and
4. facilitating unsupervised adaptation for different applications.

The system is tested with four audio streams having speech and music distributions that vary in both duration and style (for example, male and female speech, different types of music). The results obtained across this database demonstrate the robustness of the proposed approach.

## 2. SPEECH/MUSIC DISCRIMINATION SYSTEM

The complete block diagram of the proposed speech/music discrimination system is shown in Figure 1. Individual blocks are described in the following subsections.

### 2.1. Multilayer Perceptron (MLP)

The MLP in the proposed system is the same as the one used in a hybrid HMM/MLP ASR system, where its role is to estimate the posterior probabilities of the speech phonetic classes given the acoustic feature vectors. We can consider such an MLP to be a channel trained to process speech. If the input to this channel is indeed a speech signal, we can expect certain behaviour at the channel output. In contrast, if the input is non-speech, the channel output will not display this characteristic behaviour. In this way, careful examination of the channel output should enable us to infer whether the input signal is speech or not.

The MLP estimates the posterior probabilities of the speech phonemes given acoustic feature vectors corresponding to a temporal contextual window of a certain duration, that is, $P(q_k|x_n)$ where $q_k$ is the phoneme (with $k = 1, \ldots, K$, where $K$ is the total number of phonemes) and $x_n$ is the acoustic feature vector at

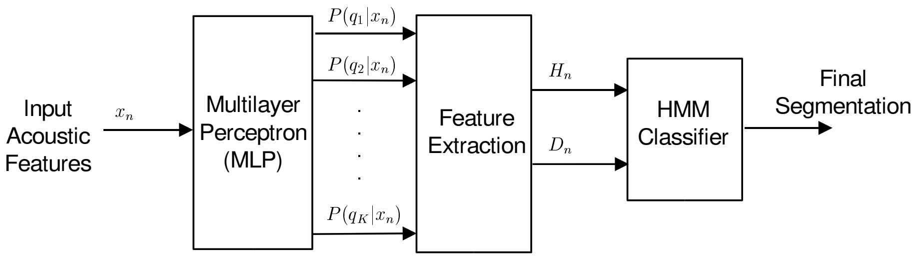


**Fig. 1. Block diagram of the proposed system**

time $n$. According to the channel model, these probability values should exhibit certain behaviour if the input signal is speech. In order to examine the behaviour of these probabilities, two features are extracted, as discussed in the next section.

### 2.2. Feature Extraction

In order to examine both the intra-frame and inter-frame (temporal) behaviour of these probabilities, we calculate two features, namely entropy $H_n$ and dynamism $D_n$.

Entropy is a measure of the distribution of the probability values within a frame, and is calculated as

$$H_n = -\frac{1}{N} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} \sum_{k=1}^{K} P(q_k|x_m) \log_2 P(q_k|x_m) \qquad (1)$$

where the frame values are averaged over a temporal window of size $N$ for smoothing. In the case of speech, the probability of a particular phoneme (the recognized phoneme) will generally be much higher than other phonemes. This means that the value of the entropy will ideally tend to zero for speech input. In the case when a music or non-speech signal is passed through the MLP, the probabilities will be more uniformly distributed (as there is no "true" phoneme present), resulting in a higher entropy.

Dynamism measures the temporal variations in the probability values between frames, and is calculated as

$$D_n = \frac{1}{N} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} \sum_{k=1}^{K} [P(q_k|x_m) - P(q_k|x_{m+1})]^2 \qquad (2)$$

As speech is a highly time-varying signal, it can be expected that the probability values will change rapidly from one frame to another, leading to a high value of dynamism. Conversely, music is a harmonic signal with slower time variations, and hence we would expect a lower value of dynamism for such an input signal.

In the proposed system, every acoustic frame $x_n$ of the input signal is converted into a two-dimensional feature vector $y_n = [H_n \; D_n]^T$. To demonstrate the appropriateness of these features for speech/music discrimination, Figures 2 and 3 respectively show Gaussian mixture models (GMM's) for entropy and dynamism features extracted from real speech and music data. It can be seen that the features occupy well-separated regions in the feature space, and thus make good discriminatory features. It is also clear from the figures that the time-averaging of the local feature values (over the window of $N$ frames) makes them better Gaussian processes, and increases the distinction between the classes.

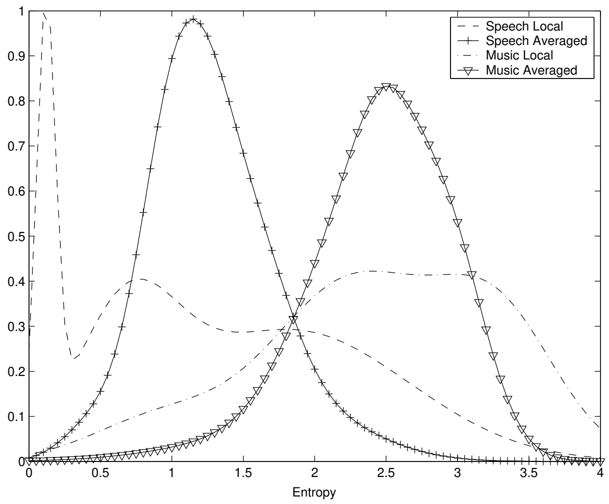


**Fig. 2**. Distribution of local and average entropy for speech and music. As expected, the average entropy is usually lower for speech than for music.

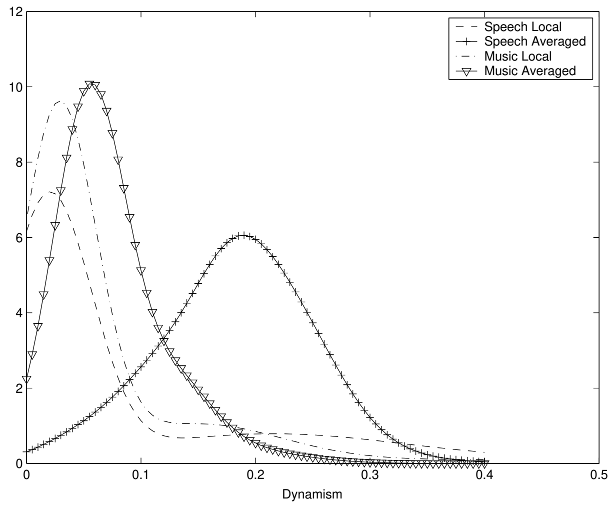


**Fig. 3**. Distribution of local and average dynamism for speech and music. As expected, the speech average dynamism is usually higher than the average music average dynamism.

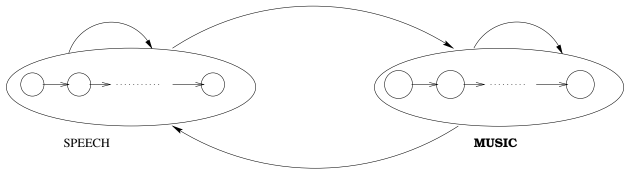


**Fig. 4**. HMM topology for the proposed system

### 2.3. HMM Classifier

The HMM topology for the proposed system is shown in Figure 4. This is a fully connected 2-class HMM with each class having several identical states in cascade to impose the minimum duration constraint. The emission probabilities of the states belonging to the same class are the same and are modeled using a GMM.

The parameters of the HMM are trained in a completely unsupervised manner, eliminating the need for labeled training data. The model was trained using the Baum-Welch algorithm to calculate the GMM parameters for the two classes. The transition probabilities are set manually to favour staying in the same class once the minimum duration is reached. The initial probabilities are also set manually to make the classes equally likely at the beginning of the stream.

At the time of segmentation, given the observation sequence $y_n$, the local likelihood of each class is calculated at every frame $n$. The Viterbi algorithm is then used to find the best possible state sequence which could have emitted this observation sequence, according to the maximum likelihood (ML) criterion.

In our experiments, the back-tracking part of the Viterbi algorithm is performed after reaching the end of each audio sequence. This gives the sound segment (speech/music) sequence resulting in maximum likelihood. However, for large audio databases, it may be necessary to break the data into chunks of manageable size and then perform Viterbi decoding.

Using such an HMM classifier to perform the segmentation has a number of advantages. Most existing techniques for speech/music segmentation make local, threshold-based decisions. Integrating the entropy and dynamism features within an HMM system allows for a global segmentation of the audio sequence which is optimal in the maximum likelihood sense. In addition, the minimum duration constraint can be easily imposed in the HMM system design by simply configuring the number of cascaded states within each class. Indeed, in this way it could be possible to impose different minimum duration constraints for the different acoustic classes. Finally, the use of unsupervised training enables the system to be trained without explicitly labeled data, facilitating both the initial training of the system, and also any subsequent adaptation to new conditions. Such a system could also be easily extended to additional acoustic classes - for example, in automatic transcription of sports videos it may be desirable to distinguish between the three classes of speech, music and cheering.

## 3. EVALUATION EXPERIMENTS

### 3.1. Implementation

For the posterior probability calculation, we use a (9x13)-2000-42 MLP with a softmax output layer trained via back-propagation to a minimum-cross-entropy criterion. The input acoustic features to this MLP are the first 13 cepstra of a $12^{th}$-order PLP filter to the spectrum of the 16 Khz sampled data, using a 32 ms window and a 16 ms frame shift. No delta, double delta, or explicit energy terms are used. Nine successive feature frames are presented to the neural network at a time.

For the purpose of feature calculation, the number of phonemes $K$ is 42 and the size of averaging window $N$ is 40.

Approximately 2.5 hours of audio data was used for the unsupervised HMM training. Five mixture GMM's were used for all states. The number of states used to impose the minimum duration constraint in the HMM was fixed to 180, meaning that in our case a recognised speech or music segment is never shorter than 2.88 seconds (16ms $\times$ 180).

### 3.2. Evaluation

The system was evaluated using 4 labeled test data sets, each 10 minutes long. These data sets contain a variety of speech and music. For example, they contain speech from a number of both male and female speakers, as well as different types of music, such as jazz, pop, and country. In addition, to observe the effect of varying segment durations, the test data was mixed together with different speech and music durations as follows

1. Test Set 1: Alternate speech and music segments of equal duration (15 seconds each).
2. Test set 2: Varying length of alternate speech and music segments. These segments include very short (2 seconds) as well as long (14 seconds) durations.
3. Test Set 3: This test set contains mostly speech data. 15 second segments of speech data are interleaved with shorter music segments.
4. Test Set 4: This test set contains mostly speech data. 15 second segments of music data are interleaved with shorter segments of speech.

The results on these four data sets are shown in Table 1. These results are obtained in terms of frame-level accuracy. We calculate three different statistics in each case: the percentage of true speech frames identified as speech, the percentage of true music frames indetified as music, and the overall percentage of speech and music frames identified correctly.

| $Test\ Set$ | $Speech$ | $Music$ | $Total$ |
|---|---|---|---|
| 1 | 92.9 | 99.1 | 95.7 |
| 2 | 91.2 | 98.8 | 94.8 |
| 3 | 96.6 | 92.8 | 95.0 |
| 4 | 91.9 | 97.1 | 94.9 |

**Table 1**. Classification results

### 3.3. Discussion

On the basis of these results, we observe the following

- Overall, entropy and dynamism make very good discriminatory features for speech/music discrimination. However, the performance of these features is better for music segments than for speech. This can be explained by observing that the dynamism for frames within a speech segment can sometimes be as low as for a music segment, particularly for low speaking rates and if considerable silence periods occur within the speech. As a result, some speech frames may be confused as music, but the converse will not be the case - the dynamism for music frames will not behave like speech.

- Test set 3 represents a situation similar to that found in broadcast news applications, where large speech segments are interleaved by short non-speech (typically music) segments. High performance of the system, especially for the speech segments, ensures a good pre-processing for transcription using a speech recognition system. In these situations, additional care can be taken by appending short margins at the left and right of the detected speech segments to avoid abrupt beginnings and ends.

- Test sets 3 and 4 contain several segments of very short duration, that is, less than our minimum duration constraint. Due to the design constraint on minimum duration, we would expect these short segments to be discarded by the system, and this expectation is reflected in our reference transcriptions when calculating the frame accuracy rate. The results for these portions indicate that the system has indeed been successful in discarding these short segments.

- Figures 2 and 3 show the feature distribution calculated using labeled data. The HMM parameters are not set using these, but trained in an unsupervised manner. The high performance of the system on all four data sets verifies the validity of the unsupervised training process. A major advantage of such training is that the HMM parameters can be adapted in an unsupervised manner depending on the application. For example, in applications where the non-speech sound class is not purely 'music', but also consists of other noises such as clapping and cheering, unsupervised adaption of the HMM parameters would improve segmentation.

- In the proposed system we do not rely on the setting of a hard threshold value, as would be the case for schemes based on the GLR. In such systems, a threshold value is generally estimated on the basis of experiments and is then used later for the segmentation, making for a less robust system. Also, unlike the BIC and GLR which tend to make decisions every frame, global decisions are made in this case, allowing for a more robust segmentation of a continuous audio stream.

As an aside, we note that some of the error may be attributed to the inherent latency of the system. At a first level, a high amount of averaging is done in the pre-processing stages in order to extract the speech recognition features and contextual information for input to the MLP. This is followed by another level of averaging to obtain the average entropy and average dynamism features. Due to these factors, the features will not change abruptly as the signal makes a transition from speech to music and vice-versa. Another level of latency is introduced by the minimum duration constraint within the HMM, where a specified minimum amount of time is required to decide whether the signal is really a music or a speech signal. The combined effect of these factors will mean that perfect 100% accuracy at the frame level is unlikely to be achieved in practice.

## 4. CONCLUSION

In this paper we have proposed a new approach for speech/music discrimination. Posterior probability based *entropy* and *dynamism* features are integrated over time through a 2-class HMM with minimum duration constraints. The HMM parameters are trained in a completely unsupervised manner eliminating the need for labeled training data, and potentially facilitating online adaption of the HMM parameters. The overall system has the benefit of making global, threshold-free segmentation decisions.

The system was tested with different speech and music styles, as well as different mixtures of speech and music segment durations. The results of these tests illustrate the robustness of the approach, with the system achieving consistent frame accuracies of approximately 95% across a variety of realistic test scenarios. From these results, we conclude that using these features within an HMM classification framework with minimum duration constraints makes a powerful system for speech/music segmentation.

In the framework of the ASSAVID project for automatic sports video annotation, the technique was extended to calculate a confidence measure of the speech and music segments. Such a measure is particularly useful when both speech and music are present simultaneously. A demonstration of this work can be found at `http://www.idiap.ch/~jitendra/speech-music`.

## 5. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation grant 2100-65067.01 (AudioSkim), and the European Commission ASSAVID project (IST-1999-13082).

## 6. REFERENCES

[1] E. Sheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/ music discriminator," April 1997, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1331–1334.

[2] J. Saunders, "Real-time discrimination of broadcast speech/ music," May 1996, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 993–996.

[3] T. Zhang and J. Kuo, "Hierarchical classifiaction of audio data for archiving and retrieving," March 1999, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 3001–3004.

[4] E. S. Parris M. J. Carey and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," March 1999, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing.

[5] G. Williams and D. Ellis, "Speech/ music discrimination based on posterior probabilites," Sept. 1999, Proc. European Conf. on Speech Commun. and Technology, pp. 687–690.

[6] S. S. Chen and P. S. Gopalkrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *IBM Technical Journal*, 1998.

[7] G. Petrucci K. El-Maleh, M. Klein and P. Kabal, "Speech/ music discrimination for multimedia application," June 2000, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 2445–2448.