Speech/Music Segmentation using Entropy and Dynamism Features in a HMM Classification Framework

Jitendra Ajmera[a,b], Iain McCowan[a], Hervé Bourlard[a,b]
[a]IDIAP, Martigny, Switzerland
[b]EPFL, Lausanne, Switzerland
email: [jitendra,mccowan,bourlard]@idiap.ch

11th June 2002

Corresponding Author
Jitendra Ajmera
IDIAP
case postal 592
Rue du Simplon 4
CH-1920 Martigny
Switzerland
email: jitendra@idiap.ch
Phone: +41-27-721.77.48
Fax: +41-27-721.77.12

List of abbreviations

HMM: Hidden Markov Model
GMM: Gaussian Mixture Model
ANN: Artificial Neural Network
MLP: Multi-Layer Perceptron
MFCC: Mel Frequency Cepstral Coefficients
GLR: Gaussian Likelihood Ratio
BIC: Bayesian Information Criterion
QGC: Quadratic Gaussian Classifier
LSF: Line Spectral Frequencies
PDF: Probability Density Function
EM: Expectation Maximisation
ML: Maximum Likelihood

Number of tables: 5
Number of figures: 4
NUmber of pages:25

Keywords: speech/music discrimination, audio segmentation, entropy, dynamism, HMM, GMM, MLP.

List of Figures

List of Tables

Abstract

In this paper, we present a new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news. In the approach presented here, an artificial neural network (ANN) trained on clean speech only (as used in a standard large vocabulary speech recognition system) is used as a channel model at the output of which the entropy and "dynamism" will be measured every 10 ms. These features are then integrated over time through an ergodic 2-state (speech and and non-speech) hidden Markov model (HMM) with minimum duration constraints on each HMM state. For instance, in the case of entropy, it is indeed clear (and observed in practice) that, on average, the entropy at the output of the ANN will be larger for non-speech segments than speech segments presented at their input. In our case, the ANN acoustic model was a multilayer perceptron (MLP, as often used in hybrid HMM/ANN systems) generating at its output estimators of the phonetic posterior probabilities based on the acoustic vectors at its input. It is from these outputs, thus from "real" probabilities, that the entropy and "dynamism" are estimated. The 2-state speech/non-speech HMM will take these two dimensional features (entropy and "dynamism") whose distributions will be modeled through multi-Gaussian densities or a secondary MLP. The parameters of this HMM are trained in a supervised manner using Viterbi algorithm.

Although the proposed method can easily be adapted to other speech/non-speech discrimination applications, the present paper only focuses on speech/music segmentation. Different experiments, including different speech and music styles, as well as different temporal distributions of the speech and music signals (real data distribution, mostly speech, or mostly music), illustrate the robustness of the approach, always resulting in a correct segmentation performance higher than 90%. Finally, we will show how a confidence measure can be used to further improve the segmentation results, and also discuss how this may be used to extend the technique to the case of speech/music mixtures.

Résumé

Dans cet article, nous présentons une nouvelle approche particulièrement performante de discrimination parole/musique dans le cadre d'applications réelles de transcription de nouvelles diffusées. Dans cette approche, un réseau de neurones artificiels (ANN) entraîné exclusivement sur de la parole claire (provenant d'un système standard de reconnaissance de la parole grand vocabulaire) est utilisé comme modèle de canal à la sortie duquel nous mesurons toutes les 10 ms l'entropie et le "dynamisme". Ces caractéristiques sont alors intégrées dans le temps à l'aide d'un modèles de Markov caché (HMM) ergodique à deux états (parole et non-parole) incluant également des contraintes de durée minimum sur chaque état. Par exemple, dans le cas de l'entropie, il est effectivement clair (et observé en pratique) que l'entropie à la sortie du ANN sera en moyenne plus élevée pour des segments non-parole que des segments de parole présentés à son entrée. Dans notre cas, le modèle acoustique ANN est un perceptron multicouche (MLP, comme souvent utilisé dans les systèmes hybrides HMM/ANN) générant à sa sortie des estimateurs de probabilités a posteriori de phonèmes étant donné les vecteurs acoustiques d'entrée. C'est à partir de ces sorties, et donc de "vraies" probabilités que l'entropie et le "dynamisme" sont estimés. Le modèle HMM parole/musique à deux états prends ensuite ces deux caractéristiques (entropie et "dynamisme") dont les distributions sont modélisées par des densités multi-gaussiennes ou par un second MLP. Les paramètres de ce modèle HMM sont entraînés par un Viterbi supervisé.

Bien que l'approche proposée ici puisse être facilement adaptée à d'autres applications de discrimination parole/non-parole, nous nous focalisons ici sur le problème de segmentation parole/musique. Différentes expériences, incluant différents styles de parole et musique, ainsi que différentes distributions temporelles des signaux de parole et musique (distributions réelles, surtout parole, ou surtout musique), illustrent la robustesse de l'approche qui résulte toujours en des performances de segmentation correcte supérieure à 90%. Finalement, nous montrons comment l'utilisation d'un niveau de confiance peut améliorer les résultats de segmentation, et comment ceci peut être utilisé pour traiter les cas de mélanges de parole et musique.

1 Introduction

The problem of distinguishing speech signals from other audio signals (e.g., music) has become increasingly important as automatic speech recognition (ASR) systems are applied to more real-world multimedia domains, such as the automatic transcription of broadcast news, in which speech is typically interspersed with segments of music and other background noise. Standard speech recognizers attempting to perform recognition on all input frames will naturally produce high error rates with such a mixed input signal. Therefore, a pre-processing stage that segments the signal into periods of speech and non-speech is invaluable in improving recognition accuracy. This also has the benefit of reducing overall computational load, as the full speech recognition system is only enabled for speech segments.

Another application of speech/music discrimination is low bit-rate audio coding. Traditionally, separate codec designs are used to digitally encode speech and music signals. An effective speech/music discrimination decision will enable these to be merged in a universal coding scheme capable of reproducing well both speech and music.

More generally, audio segmentation (which could be performed by using a more appropriate feature set and generalizing the speech/music discrimination approach presented in the present paper) could allow the use of ASR acoustic models trained on particular acoustic conditions, such as wide bandwidth (high quality microphone input) versus telephone narrow bandwidth, male speaker versus female speaker, etc., thus improving overall performance of the resulting system. Finally, this segmentation could also be designed to provide additional interesting information, such as the division into speaker turns and the speaker identities (allowing, e.g., for an automatic indexing and retrieval of all occurrences of a same speaker), as well as 'syntactical information' (such as end of sentences, punctuation marks, etc).

One of the issues in the design of a signal classifier is the selection of an appropriate feature set that captures the temporal and spectral structure of the signals. Many such features for speech/music discrimination have been suggested in the literature, including zero-crossing information, energy, pitch, cepstral coefficients, line spectral frequencies (LSF), 4 Hz modulation energy, amplitude, and perceptual features like timbre and rhythm [1, 2, 3, 4, 5]. In this work, we use posterior probability based features introduced in [6], namely entropy and dynamism. As we will show, these features indeed exhibit nice discriminant properties yielding to high performance speech/music segmentation.

Another issue in the system design is the selection of a classification algorithm. Different classifiers like the Bayesian Information Criterion (BIC) [7], Gaussian likelihood ratio (GLR) [2, 6, 1, 5], quadratic Gaussian classifier (QGC) [4], nearest neighborhood classifier [1, 4] and hidden Markov model (HMM) [8] have been used for this purpose.

Nowadays, an algorithm based on the BIC [7] is perhaps the most commonly used technique for audio segmentation. It assumes that the sequence of acoustic feature vectors is a Gaussian process, and measures the likelihood that two consecutive acoustic frames were generated by two processes rather than a single process. The BIC technique is useful for general audio change detection, as it does not require any *a priori* information about the particular acoustic classes present. However, in the case that the number and type of acoustic classes is known, it should be advantageous to explicitly incorporate this information into the design of the segmentation system. In real applications, the BIC technique also poses a number of practical problems, such as high computational complexity and the need to tune a threshold parameter (λ) to optimise performance.

In this work, we use the entropy and dynamism features estimated at the output of the multi-layer perceptron (MLP, referred to as primary MLP) used in a regular hybrid HMM/MLP large vocabulary continuous speech recognition system. Depending on the data presented at the input of the primary MLP, these features will exhibit different properties and can be used in a secondary 2-state (speech/non-speech) HMM system, where the state probability densities are estimated by either Gaussian mixture models (GMM) or a secondary MLP. This approach has two advantages, i.e.:

1. Using features that have been shown to have discriminant properties for speech and music classes, and

2. Being a threshold-free, global decision making strategy.

In the same framework, we also investigate the use of a confidence measure to improve the performance and application of the discrimination system. This measure can be used to improve the discrimination accuracy by removing short, low confidence segments. In addition, such a confidence measure could be used in the framework of speech/music mixtures, where it is desirable to determine the 'amount' of speech or music present in the audio signal, rather than simply providing hard segmentation boundaries.

2 Posterior Probability Based Features

According to information theory, a channel designed for a particular type of signal will exhibit characteristic behaviour at its output when that signal is passed through the channel. Conversely, the presence of a different type of signal will result in uncharacteristic behavior at the channel output. In the case where the channel is an MLP trained to emit posterior probabilities for speech recognition [9], it should therefore be possible to distinguish between speech and non-speech signals by examining the behaviour of these probabilities. Following Williams and Ellis [6], we base our speech/music decision on statistics of the output of an acoustic model intended originally for discriminating the phonemes of speech. Specifically, we use their *entropy* and *dynamism* features as defined below. While in this work we focus on application to speech/music discrimination, we note that these features should essentially distinguish between 'recognisable speech' and other signals.

2.1 Entropy

Entropy is a measure of the uncertainty or disorder in a given distribution [10]. In the case of a primary MLP trained to emit posterior probabilities for K output classes (usually associated with speech phones or HMM states q_k, $k = 1, \ldots, K$), the *instantaneous entropy* h_n at a specific time frame n is defined as:

$$h_n = -\sum_{k=1}^{K} P(q_k|x_n) \log_2 P(q_k|x_n) \tag{1}$$

where x_n represents the acoustic vector at time n, q_k the k-th primary MLP output class, and $P(q_k|x_n)$ the posterior probability of class (phone) q_k given x_n at the input.

The posterior probabilities at a given time represent a true PDF, and the entropy of that PDF (the expected value of the log probability) is a measure of the goodness-of-fit of the current observation to the acoustic model (channel). Generally, in the case of speech, the value of the posterior probability for a particular phoneme (the 'recognized' phoneme) is much higher than other phonemes. This means that the value of the entropy will be close to zero, indicating that little information will be gained by knowing its actual value, or, equivalently, that there is little uncertainty over the unknown segments. In the case when a music signal is passed through the primary MLP, the values of probabilities will be more uniformly distributed, resulting in a higher value for entropy.

Equation (1) gives the instantaneous value of the entropy at frame n. As we will see in the subsequent discussion, and to perform a first smoothing, it is advantageous to average this instantaneous entropy over a window of several frames, resulting in the *averaged entropy* at time n:

$$H_n = \frac{1}{N} \sum_{t=n-N/2}^{n+N/2} h_t \tag{2}$$

where n is the index of the current acoustic frame and N is the size of the averaging window.

2.2 Dynamism

Dynamism is a measure of the rate of change of a quantity. In this case, and using the same notation as above, the *instantaneous dynamism* at time n is defined as:

$$d_n = \sum_{k=1}^{K} [P(q_k|x_n) - P(q_k|x_{n+1})]^2 \tag{3}$$

This feature captures the dynamic behaviour of the probability values. As speech involves more transitions through the speech-specific primary feature space, the phoneme posteriors will exhibit more abrupt changes than other acoustic signals such as music, resulting in higher dynamism.

Similar to the case of entropy, it can be beneficial to average the instantaneous values of dynamism over a certain number of frames, resulting in the *average dynamism* at time n:

$$D_n = \frac{1}{N} \sum_{t=n-N/2}^{n+N/2} d_t \tag{4}$$

where N is the size of the averaging window.

3 Speech/Music Segmentation System

The complete block diagram of the proposed speech/music segmentation system is shown in Figure 1. We describe the individual blocks in following subsections.

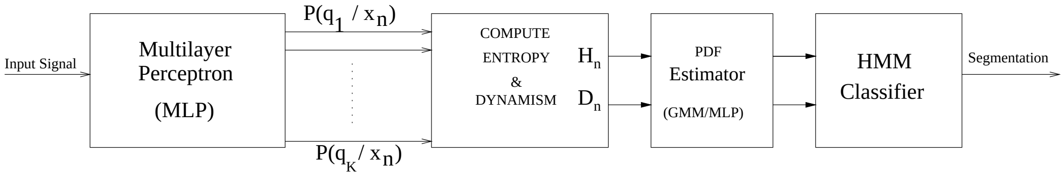

Figure 1: Block diagram of the proposed system where PDF estimator is a GMM or a secondary MLP

3.1 Multilayer Perceptron (MLP)

The primary 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.

In practice, the primary MLP estimates the posterior probabilities of the output classes (in our case, phones) given feature vectors corresponding to a temporal contextual window of a certain duration (typically 9 acoustic frames of 16-ms), i.e., $P(q_k|x_n)$ where q_k is the phonetic class (with $k = 1, \ldots, K$, where K is the total number of output classes) and x_n is the feature vector at time n. Careful observation of these probabilities shows a marked distinction between segments consisting of clean speech and other segments, such as music or very noisy speech. If it is decided that these posterior probabilities correspond to speech segments, they can then be converted to likelihoods and passed to a Viterbi decoder for word recognition, as in a standard hybrid ASR system.

3.2 Feature Computation

The output of the primary MLP is a set of K posterior probabilities, i.e., $P(q_k|x_n)$. For every acoustic frame (16 ms in our case), we calculate the average entropy H_n and average dynamism D_n according to (2) and (4). These values are combined to form a two-dimensional vector, $y_n = (H_n, D_n)^T$, which is then used as the HMM observation vector.

3.3 Probability Density Function Estimator

For every acoustic frame x_n of the input signal, the feature vector y_n is thus constructed and sent to the PDF estimator. The role of this block is to estimate the emission probabilities of the HMM states given the observation vector y_n. We investigate two estimators for this purpose: namely, the GMM and a secondary MLP.

3.3.1 Gaussian Mixture Model (GMM)

The parameters of GMMs for the two classes can be trained by using standard (supervised or unsupervised) *expectation maximization* (EM) algorithm. At the time of segmentaion, these GMMs estimate the likelihood of each class given feature vector y_n, i.e. $p(y_n|C)$. In this case, y_n is a two-dimensional vector composed of the average entropy H_n and average dynamism D_n, as defined earlier.

The individual PDFs of entropy and dynamism are shown in Figures 2 and 3, respectively. The figures show distributions for both the instantaneous (local) as well as the averaged values of entropy and dynamism. The need for averaging the instantaneous values of these features is evident from these figures. This can be further understood by considering the within-phone regions of speech segments. These are regions of very low transitions, yielding low dynamism (like music). It is only at the phoneme boundaries that the dynamism would become high. Similarly, the entropy is low during the phoneme, then temporarily peaks during transitions. Thus, averaging over several phoneme durations is very important.

3.3.2 Multilayer Perceptron (MLP)

The GMM can be substituted by another MLP (referred to as the secondary MLP) for estimating the emission probabilities of the HMM states. The secondary MLP is trained with the observation vector y_n defined earlier at its input, along with several context frames. At the time of segmentation, y_n is presented along with the context frames at the input of the secondary MLP, and the output is obtained as the set of posterior probabilities $P(C|y_n)$ for the two classes (speech and music). Using Bayes rule, these posterior probabilities can be turned into scaled likelihoods that can be used as HMM emission probabilities:

$$\frac{p(y_n|C)}{P(y_n)} = \frac{P(C|y_n)}{P(C)} \tag{5}$$

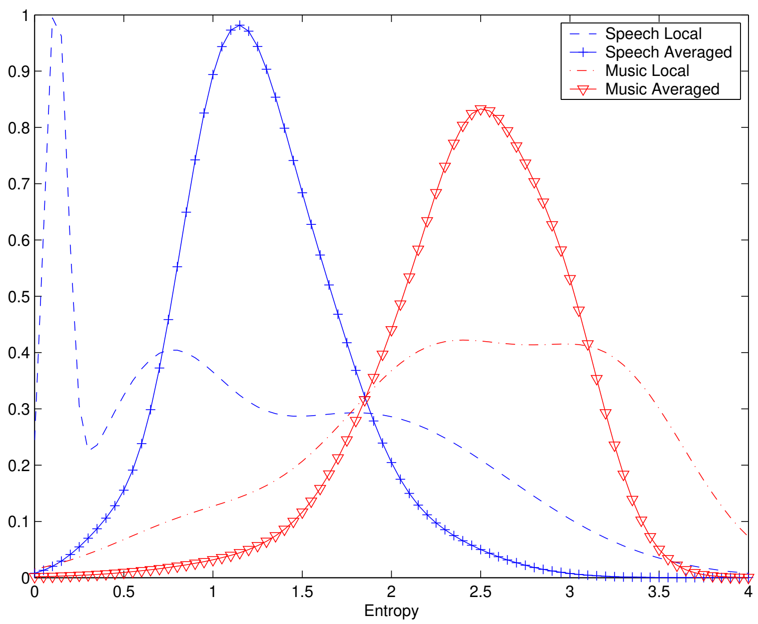

Figure 2: Distribution of local and average entropy for speech and music. As expected, the average entropy is usually higher for music than for speech.

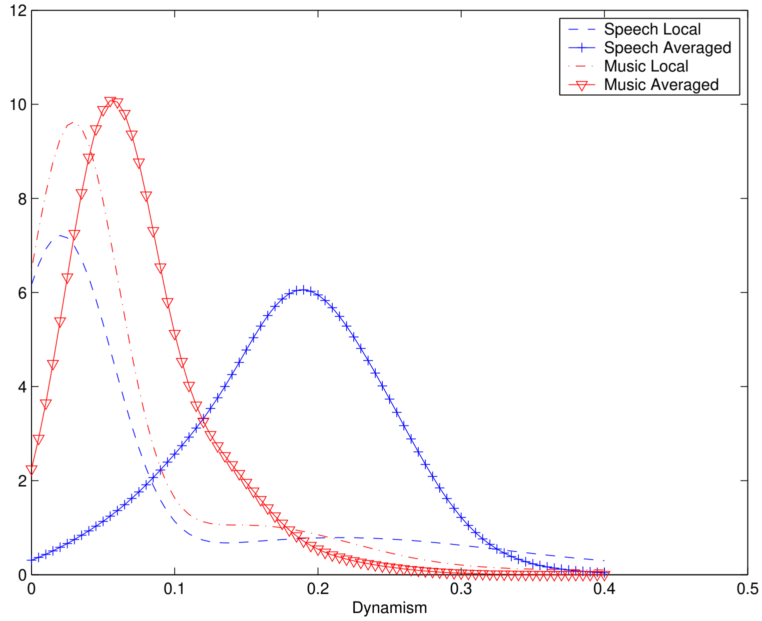

Figure 3: Distribution of local and average dynamism for speech and music. As expected, the speech average dynamism is usually higher than the average music average dynamism.

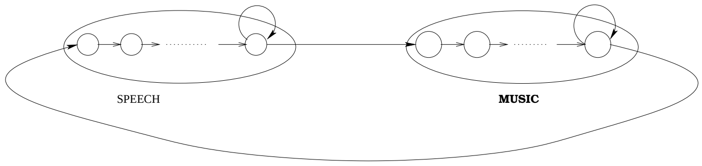

Figure 4: HMM topology for the proposed system

where $P(C)$ is the prior probability of the class C, as estimated on the training data, and $P(y_n)$ is independent of the class and simply appears as a constant scaling factor. In our case, the training data was characterized by equal speech and music priors, so it was not necessary to divide by $P(C)$.

3.4 HMM Classifier

The HMM topology for the proposed system is shown in Figure 4. In the case of speech/music discrimination, this HMM is a 2-state fully connected model, where a minimum duration is imposed for each state. This is achieved by simply concatenating internal states associated with the same PDF.

In some preliminary experiments, we observed that the values of the transition and initial probabilities of the HMM do not greatly affect the results, provided the within-class transitions are favoured (self loop probability for the last state of each class > 0.5). So, these values were set manually to favour remaining in the current state. Similarly, initial probabilities are set manually to make speech and music segments equally likely in the beginning. The emission probabilities for the HMM states are estimated by either a GMM or secondary MLP expert.

The parameters of secondary MLP are trained via the error back propagation (EBP) algorithm. Equal amounts of labeled clean speech and music data are used for training the secondary MLP. The feature vectors $y_n = (H_n, D_n)$ from the training data are presented at the input layer of the secondary MLP along with several context frames. The parameters of the GMM are trained (using the same data) in a supervised manner using standard EM algorithm. In subsequent work [11], we have shown that this training can be done in an unsupervised way eliminating the need for labeled training data.

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

In this case, the backtracking part of the Viterbi algorithm is performed after reaching the end of the audio sequence. This gives the sound (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. These chunks may also be made to overlap to measure the confidence of segments at the boundaries.

There are several advantages of using this classification strategy. First, it eliminates the need for a hard threshold value. In schemes like the BIC and GLR, for practical applications, a threshold value is calculated on the basis of experiments and is used for making a decision at the time of segmentation. Sometimes, this threshold value can be imprecise and misleading. The HMM approach allows for a more principled parameter selection based on a training data set. Second, with the HMM it is possible to easily impose a constraint on the minimum segment duration. If any sound (speech or music) lasts less than a minimum duration, we consider that it does not carry any useful information. We impose this constraint by having several states belonging to the same class in cascade as shown in Figure 4. Also, unlike the BIC and GLR schemes which tend to make independent decision every frame, global decisions over this minimum duration are made in the case of the proposed system. Another important advantage of the system is the low computational complexity of the HMM classifier compared with other decision making strategies. The Viterbi HMM scheme has a complexity of approximately order KN (where K is the state space size and N is the number of input frames), while the BIC system is an order N^2 algorithm.

4 Evaluation Experiments

4.1 Implementation

For the posterior probability calculation, we use a (9x13)-2000-42 MLP [1] with a softmax output layer trained via back-propagation to a minimum-cross-entropy criterion. The input features 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 training the GMM and secondary MLP experts. The GMMs for both speech and music have 5 Gaussian distributions and were trained using the EM algorithm as described earlier. These Gaussians have diagonal covariance matrices, meaning that the two features are not correlated in a two-dimensional feature space. The secondary MLP is a (9x2)-5-2 structure with a softmax output layer trained via the back-propagation algorithm.

The number of states used to impose the minimum duration constraint in the HMM was fixed to 180, thus assuming in our case that any speech or music segment is never shorter than 2.88 seconds (16ms $\times$ 180). The self loop probabilities were set to 0.9 for the last state of each class.

[1]This MLP was trained by our colleagues at ICSI Berkeley

To assess the effectiveness of the entropy and dynamism features, a baseline system using 24 dimensional Mel Frequency Cepstral Coefficients (MFCC) (no delta or acceleration terms were used) in a GMM/HMM framework (using the same topology in Figure 4), was also included in the evaluation for comparison purposes.

4.2 Evaluation

We evaluated the system using 4 labeled data sets, each 10 minutes long. Each data set was constructed by concatenating speech and music segments taken from real broadcast audio data. The purpose of having 4 different test sets was to assess the performance of the system on different degrees of mixing of these sounds, that is, to observe the effect of duration of these sounds. Moreover, segments having wide variety of speech and music were chosen. For example, they contain speech from a variety of both male and female speakers, as well as different types of music, such as jazz, pop, and country.

Results were obtained in terms of the percentage 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 identified as music, and the overall percentage of speech and music frames identified correctly.

4.3 Results

4.3.1 Test Set 1

This is a 10 minute audio stream having alternate speech and music segments of equal (15 seconds) duration. The classification results are shown in Table 1.

Table 1 to be placed here

In this case, both entropy and dynamism features are capable of identifying the speech segments with a high degree of accuracy, as is required in ASR applications. However, with dynamism the frame level accuracy for music segments is low. We found that some of these segments had rap music, with less instrumental music in the background. Still, entropy performs well in discriminating these segments from speech. It is also clear that the performance of these two features is better than that of the 24-dimensional MFCC features. The performance of the GMM and MLP experts are comparable in this case.

4.3.2 Test Set 2

The second test set represents a more realistic task as it consists of varying lengths (including very short and long durations) of alternate speech and music segments. These classification results are shown in Table 2.

Table 2 to be placed here

The performance of different features and experts for this data set follows the same trend as in Test Set 1. This indicates the robustness of the proposed system to the sound durations. To test the significance of the minimum duration constraint, some segments shorter that the minimum duration were included in this test set, with the expectation that they should be filtered out. Examination of the output transcripts shows that these undesired short segments are correctly ignored in the case of the entropy and dynamism features. In contrast, due to the slower and less abrupt variations in the features, these segments appear in the MFCC system output, borrowing frames from neighbouring segments to respect the minimum duration.

4.3.3 Test Set 3

The third test set consists of a 10 minute audio stream comprising mainly of speech data. In this case, 15 second segments of speech data are interleaved with short segments of music. This represents a more likely scenario for the case when the speech/music discrimination is being used as a pre-processing step to speech recognition, as the audio signal will be predominantly speech. These classification results are shown in Table 3.

Table 3 to be placed here

The advantage of combining the two features becomes evident from the results of this test set, clearly improving the total performance and the performance over music segments.

4.3.4 Test Set 4

The final test set contains a 10 minute audio stream consisting mostly of music data. In this case, 15 second music segments are interleaved with short segments of speech. The classification results are shown in Table 4.

Table 4 to be placed here

We note that the speech segments are detected with a high degree of accuracy, despite having been interleaved with large segments of music. This ensures no loss of information when speech/music discrimination is carried out as a pre-processing stage for speech recognition.

4.4 Discussion

The observations from these four data sets can be summarized as follows:

- The 2-dimensional entropy and dynamism feature vector shows better performance (overall 95.2%) in discriminating between speech and music classes compared to standard 24-dimensional MFCC features (92.9%).

- Overall, entropy is a better discriminatory feature than dynamism, especially during music segments. Both features individually are capable of detecting the speech frames with a high degree of accuracy. The combination of the two features, in some cases, significantly improves the results, showing that the two features have complementary information. This can be attributed to the fact that entropy captures the frame-level behaviour, whereas dynamism captures the temporal behaviour, of the posterior probabilities.

- Dynamism fails to detect music frames correctly, especially if the music is composed of more vocal sounds than instrumental music, as in the case of rap music. However, entropy still performs adequately in this situation. This shows that the output of the primary MLP is still music-like in nature (high entropy, probabilities uniformly distributed) within a frame, but changes rapidly between frames, giving higher values of dynamism.

- In general, the relative behaviour of entropy and dynamism does not change in the GMM and MLP frameworks. The performance of the two experts (GMM/MLP) is also comparable in almost all cases.

- In the framework of speech recognition, an important advantage of audio segmentation is the saving in computation time for non-speech segments. When the proposed system is used in conjunction with a hybrid HMM/MLP speech recognition system, computation is reduced, as music segments are not passed to the Viterbi decoder, which is the most computationally intensive element of the hybrid recogniser.

- In [11], it is shown that the parameters of the GMM for the two classes (speech and music in this case) can be trained in an unsupervised manner. This eliminates the need for labeled training data. Also, it makes the online adaptation of the system easier. This way, the system can easily be adapted to more general speech/non-speech classification problem.

5 Confidence Measure

In many situations it is desirable to not only have the segmentation information, but also a measure of the confidence that we have in the segmentation decision. In this section, we first discuss *mean posterior confidence measure* (MPCM) [12] and then briefly discuss its use for two different purposes. As the secondary MLP expert outputs real posterior probabilities, it offers a more convenient framework for the development of such a confidence measure. For this reason, the following discussion focuses on the system employing the secondary MLP.

5.1 Definition of Confidence Measure

In the context of the secondary MLP system, we obtain the posterior probabilities for the speech $P(S|y_n)$ and music $P(M|y_n)$ $(= 1 - P(S|y_n))$ classes for each input frame. For a segment of multiple frames $(N_1 < n < N_2)$, we can define a measure of the confidence of the speech and music classes from the arithmetic mean of these frame probabilities. We adopt the arithmetic mean in this case so that the segmental confidence measure is not unduly biased by the probability estimates of a single frame (it is evident that use of the geometric mean would result in an average confidence of 0 if only one of the frames gave a probability of 0). Thus, the MPCM is defined as:

$$R_C(N_1, N_2) = \frac{1}{N_2 - N_1} \sum_{n=N_1}^{N_2} P(C|y_n) \qquad (6)$$

where C represents either the speech or music class. This confidence measure is convenient as it is has a range of $0 \leq R_C \leq 1$ and obeys the constraint that $R_S + R_M = 1$.

5.2 Improving Speech/Music Discrimination Accuracy

In the experiments reported in Section 4.3, the segmentation resulting from the 2-state HMM has alternate speech and music segments with minimum durations of 2.88 seconds. However, it was observed that, sometimes, a short speech (music) segment may be recognised between two large music (speech) segments. While in some cases these segments may be valid, they could also be attributed to several factors, such as long pauses during speech, rap music, etc. In such cases, we require a strategy to excise these unwanted, incorrect segments.

To this end, we investigated the use of a simple heuristic algorithm in which low confidence segments are merged with the neighbouring segments if

1. the confidence of a segment falls below a threshold, and
2. the confidences of the neighbouring segments are above this threshold value.

We set a confidence threshold at 0.65 and use the above algorithm on the results (secondary MLP system only, using both entropy and dynamism features) of Section 4.3. The results are shown in Table 5.

Table 5 to be placed here

These results demonstrate two important points. First, we can achieve a reduction in error rate by removing low confidence segments. In this case we see the overall error rate decrease from 5.2% to 4.3%, corresponding to a relative error rate reduction of approximately 17%. Second, from the fact that the error rate does not increase noticeably in any case, we can also conclude that, for these test sets, all of the correct speech and music segments are being recognised with a high confidence greater than 0.65. This is as we would hope, as the segments used in these test sets are all 'pure' speech or music segments, and thus the discrimination system should have high confidence in making correct segmentation decisions.

5.3 Speech/Music Mixtures

In the present paper, we have concentrated on the problem of segmenting an audio file consisting of pure speech or music portions. A related problem, and a natural extension of the technique, is determining the 'amount' of speech present in a signal containing a mixture of both speech and music at the same time. In the previous sub-section, we have seen that such pure speech or music segments are recognised with high confidence (above 0.65 for this test data). In the case of speech and music mixtures, it would also be of interest to use the confidence measure as an indication of the relative levels of speech and music present at a given time.

Such a measure would have applications, for example, in the context of a multi-modal fusion application in which the speech/music discrimination information, and indeed the speech recognition output, are simply input cues (or features) for higher-level processing decisions combining cues from different modalities. Such a technique for classifying speech/music mixtures has been applied in the framework of the European ASSAVID project, which is concerned with automatic indexing of sports videos, and a demonstration of initial results of the scheme on a sample audio segment is available at `http://www.idiap.ch/~jitendra/speech-music`. The demonstration consists of an MPEG file which plots the value of the speech confidence measure calculated over segments as the audio signal is played. While this remains the topic of ongoing research, these initial results give a (subjective) indication of the potential of the confidence measure for speech/music mixtures.

6 Conclusion

In this paper, we have presented a new approach for speech/music discrimination. *Entropy* and *dynamism* features based on posterior probabilities of speech phonetic classes (as obtained at the ouptut of an ANN, as used in HMM/ANN large vocabulary continuous speech recognition system) used to form a two-dimensional observation vector sequence which is used in a HMM classification framework. We compare the use of both GMM and secondary MLP experts to estimate the probability density functions of the HMM states. The relative performances of entropy/dynamism and GMM/MLP are demonstrated and discussed in the context of an experimental evaluation.

The system was tested with different speech and music styles, as well as different distributions of speech and music signals. The results of these tests illustrate the robustness of the approach, with the system achieving consistent frame accuracies from 93% to 96% across a variety of realistic test scenarios. From these results, we conclude that entropy and dynamism together make a powerful feature set for speech/music discrimination. In more general terms, by using features based on the phonetic posterior probabilities, the system allows us to locate speech segments within the audio signal that can be well recognised by the speech recognition system.

While the overall performance of the GMM and MLP systems is comparable, the MLP system outputs a set of real probabilities, which may make this system preferable if further confidence statistics are to be calculated. Such a confidence measure was proposed and investigated for the purpose of removing short low-confidence segments, further improving the frame accuracy over the baseline system. The potential use of such a confidence measure in the context of speech/music mixtures was also briefly discussed.

In summary, the proposed speech/music discrimination system provides a powerful, robust technique for reliable segmentation of audio streams.

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). We also thank the anonymous reviewers for their constructive comments, which helped improving the quality of the paper.

References

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

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

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

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

[5] E. S. Parris, M. J. Carey, and H. Lloyd-Thomas, "Feature fusion for music detection," *European Conference on Speech Communication and Technology*, pp. 2191–2194, Sept. 1999.

[6] G. Williams and D. Ellis, "Speech/ music discrimination based on posterior probabilities," *European Conference on Speech Communication and Technology*, pp. 687–690, Sept. 1999.

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

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

[9] N. Morgan and H. Bourlard, "An introduction to the hybrid hmm/ connectionist aprroach," *IEEE Signal Proc. Magazine*, pp. 25–42, May 1995.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 3 ed., 1991.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Robust HMM based speech/music segmentation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 297–300, May 2002.

[12] G. Bernardis and H. Bourlard, "Improving posterior based confidence measures in hybrid hmm/ann speech recognition systems," *International Conference on Spoken Language Processing*, vol. 3, pp. 775–778, 1998.

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	98.7	92.7	95.6
GMM	Dynamism	98.9	64.9	81.8
GMM	**Both**	98.8	93.9	**96.2**
MLP	Entropy	98.8	93.7	96.2
MLP	Dynamism	95.0	74.8	84.8
MLP	**Both**	97.4	93.7	**95.5**
GMM	**MFCC**	94.3	91.6	**92.8**

Table 1: Classification results for Test Set 1

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	97.5	91.4	94.2
GMM	Dynamism	99.6	79.4	89.0
GMM	**Both**	98.1	91.6	**94.5**
MLP	Entropy	97.4	92.4	94.6
MLP	Dynamism	93.0	84.8	88.6
MLP	**Both**	98.6	94.6	**96.3**
GMM	**MFCC**	93.9	92.2	**92.9**

Table 2: Classification results for Test Set 2

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	96.9	82.8	91.5
GMM	Dynamism	91.7	82.8	88.3
GMM	**Both**	97.0	93.6	**95.6**
MLP	Entropy	94.8	87.2	91.8
MLP	Dynamism	83.5	91.0	86.2
MLP	**Both**	91.6	96.4	**93.2**
GMM	**MFCC**	95.3	87.8	**92.4**

Table 3: Classification results for Test Set 3

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	93.3	91.1	91.8
GMM	Dynamism	98.3	70.8	81.0
GMM	**Both**	97.2	92.7	**94.3**
MLP	Entropy	93.3	91.0	91.7
MLP	Dynamism	90.7	83.7	86.2
MLP	**Both**	93.2	95.1	**94.3**
GMM	**MFCC**	95.9	92.2	**93.5**

Table 4: Classification results for Test Set 4

$Test$	$Total$	
	$Before$	$After$
1	95.5	96.1
2	96.3	96.1
3	93.3	95.6
4	94.3	94.9
Avg	94.8	95.7

Table 5: Comparison of results before and after using confidence measures