IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

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

Jitendra Ajmera [1] Iain McCowan [1]
Hervé Bourlard [1, 2]

IDIAP–RR 01-26

August 2001

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

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

[1] IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P. O. Box 592, CH-1920 Martigny, Switzerland, {jitendra, mccowan, bourlard}@idiap.ch
[2] EPFL, Lausanne

IDIAP Research Report 01-26

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

Jitendra Ajmera Iain McCowan Hervé Bourlard

August 2001

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, the (local) Probability Density Function (PDF) estimators trained on clean, microphone, speech (as used in a standard large vocabulary speech recognition system) 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. Indeed, in the case of entropy, it is clear that, on average, the entropy at the output of the local PDF estimators will be larger for speech signals than non-speech signals presented at their input. In our case, local probabilities will be estimated from an multilayer perceptron (MLP) as used in hybrid HMM/MLP systems, thus guaranteeing the use of "real" probabilities in the estimation of the entropy. The 2-state speech/non-speech HMM will thus take these two dimensional features (entropy and "dynamism") whose distributions will be modeled through (two-dimensional) multi-Gaussian densities or an MLP, whose parameters are trained through a Viterbi algorithm.
Different experiments, including different speech and music styles, as well as different (a priori) distributions of the speech and music signals (real data distribution, mostly speech, or mostly music), will 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.

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.

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 a generalization of 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. In applications like broadcast news transcription, where speech is always mixed with background sounds (music or noise), it is advantageous to find segments which can be recognized properly by an ASR system.

Another issue in the system design is the selection of a classification algorithm. Different classifiers like the Bayesian Information Criterion (BIC) [5], Gaussian likelihood ratio (GLR) [1, 2, 6], quadratic Gaussian classifier (QGC) [7], nearest neighborhood classifier [6, 7] and hidden Markov model (HMM) [3] have been used for this purpose.

Nowadays, BIC [5] is perhaps the most commonly used technique for audio segmentation and assumes that the sequence of acoustic feature vectors is a Gaussian process. BIC then measures the likelihood that two consecutive acoustic frames were generated from two different Gaussian processes than from a single Gaussian process, thus yielding the following definition of the BIC function:

$$BIC(n) = R(n) - \lambda\, P$$

where $R(n)$ is a likelihood ratio calculated as:

$$R(n) = N \ln|\sigma| - N_1 \ln|\sigma_1| - N_2 \ln|\sigma_2|$$

where σ, σ_1 and σ_2 are the sample covariance matrices for all data, the data preceding frame n (N_1 frames), and for the data following frame n (N_2 frames). P is a penalty term and λ is the penalty weight to be tuned for optimal performance. A change point is thus detected where $BIC(n) > 0$.

This BIC scheme makes local decisions at every time instant n, which may be undesirable in applications such as speech/music discrimination, in which it can be reliably assumed that speech or music last for at least a specified minimum duration. In addition, it relies on the setting of a hard decision threshold (via the penalty term) based on experiments, which can be imprecise and undesirable in many situations.

In this work, we use the entropy and dynamism features estimated at the output of the MLP used in a regular hybrid HMM/MLP large vocabulary continuous speech recognition system. Depending on the data presented at the input of the 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 multi-layer perceptron (MLP). This approach thus has two advantages, i.e.:

1. Using more appropriate features, basically "independent" of the acoustic features used for recognition, but mainly based on the quality of the classifier, considered here as a channel model.

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 a multilayer perceptron (MLP) trained to emit posterior probabilities for speech recognition [8], it should therefore be possible to distinguish between speech and non-speech signals by examining the behaviour of these probabilities. In this section, and building upon [6], we define two features, namely *entropy* and *dynamism*, by which we can characterize the distribution of these posterior probabilities.

2.1 Entropy

Entropy is a measure of the uncertainty or disorder in a given distribution [9]. In the case of an 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 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 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:

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

This feature captures the dynamic behaviour of the probability values. As speech is a highly varying signal, it can be expected that the probability values observed at the output of the MLP in the case of speech input will change rapidly from one frame to another. Conversely, music is a harmonic (less varying) signal and hence the dynamism is low, resulting in a higher dynamism for speech than for music.

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 Discrimination

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

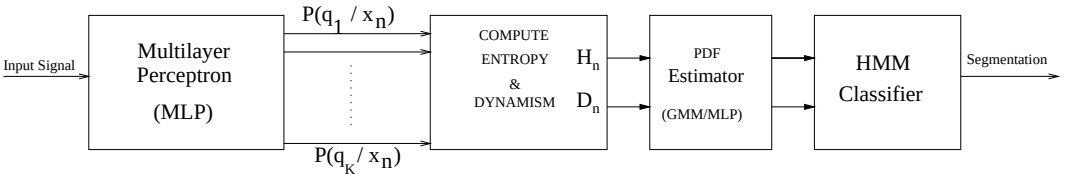

Figure 1: Block diagram of the proposed system where PDF estimator is a GMM or an MLP

3.1 Multilayer Perceptron (MLP)

In the hybrid connectionist-HMM framework for ASR [8], typical practice is to train a neural network to learn the complex temporal structure of tell-tale speech gestures present in phones and their transitions. A signal containing the substantial and balanced collections of these gestures is effectively being recognized as speech, and a signal that rarely passes through these critical patches of feature-space fails to show any resemblance to speech. The phone network classifier thus provides a very specific and significant amplification of the distinction between speech and non-speech segments.

In practice, this neural network 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.

3.2 Feature Computation

The output of the 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 MLP.

3.3.1 Gaussian Mixture Model (GMM)

A GMM is a mixture of several Gaussian distributions and is used to estimate the PDF of a sequence of feature vectors. The likelihood that a particular feature vector y_n was produced from a particular class $C \in \{Speech, Music\}$ is then estimated as:

$$\begin{aligned} p(y_n|C) &= \sum_{i=1}^{M} w_i \mathcal{N}(y_n, \mu_i, \Sigma_i) \\ &= \sum_{i=1}^{M} \frac{w_i}{2\pi\sqrt{|\sum_i|}} \exp\{\frac{-(y_n - \mu_i)^T \sum_i^{-1} (y_n - \mu_i)}{2}\} \end{aligned} \quad (5)$$

where, in our case, y_n is a two-dimensional vector composed of the average entropy H_n and average dynamism D_n, as defined earlier. M is the number of Gaussian distributions in the GMM. The parameters of these distributions, w_i, μ_i and $\sum_i$, are respectively the weight, mean and diagonal covariance matrix of the i^{th} distribution in the GMM. These parameters can be trained by using the standard (supervised or unsupervised) *expectation maximization* (EM) algorithm for both speech and music classes.

The individual PDFs of entropy and dynamism are shown in Figures 2 and 3, respectively. It is clear from these figures that the behaviour of these features for speech and music is quite distinct and that they can be effective discriminatory features. The figures also show both the distributions for the instantaneous (local) values of entropy h_n (1) and dynamism d_n (3), as well as the the average entropy H_n (2) and average dynamism D_n (4). From the figures, the averaged measures exhibit better Gaussian properties (due to the central limit theorem), and are more clearly separated and thus better suited for discrimination purposes.

3.3.2 Multilayer Perceptron (MLP)

An MLP is a special case of feed-forward neural network with one input layer, one or several hidden layers and one output layer. In our case, the 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 MLP, and the output is obtained as the set of

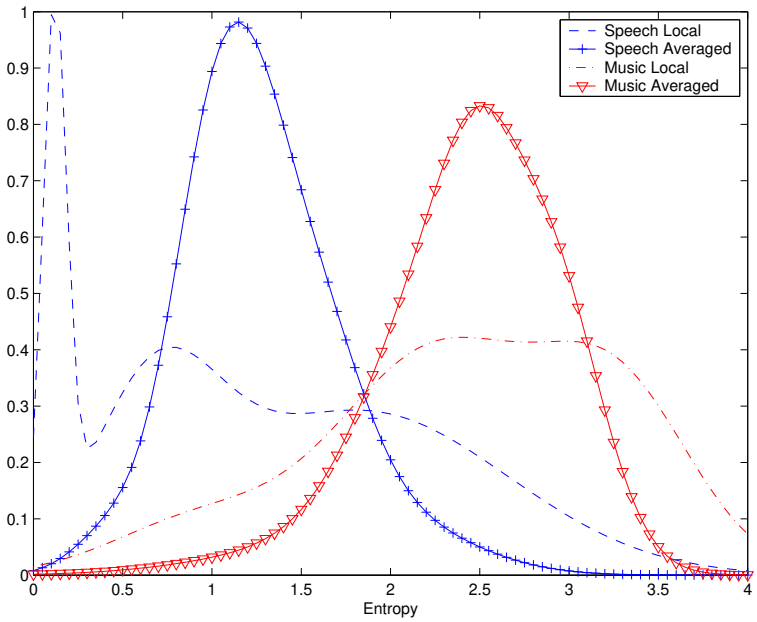

Figure 2: Distribution of local and average entropy for speech and music. As expected, the average entropy is usually higher for music than for speech.

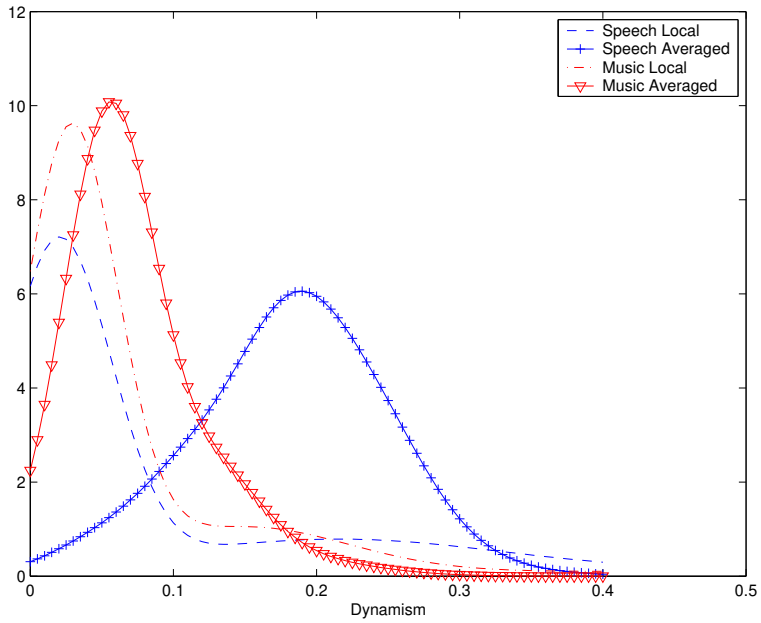

Figure 3: Distribution of local and average dynamism for speech and music. As expected, the speech average dynamism is usually higher than the average music average dynamism.

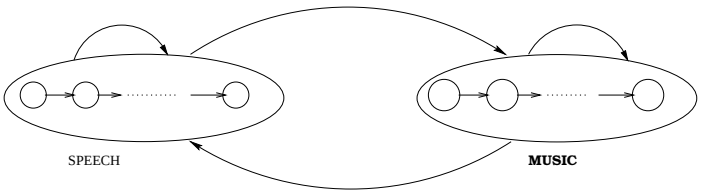

Figure 4: HMM topology for the proposed system

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:

$$\frac{p(y_n|C)}{P(y_n)} = \frac{P(C|y_n)}{P(C)} \tag{6}$$

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.

As there is no *a priori* information available regarding transitions across speech and music segments, transition probabilities are set manually to favour remaining in the current (speech or music) state. Similarly, as there is no information regarding the beginning of the audio data, initial probabilities are set manually to make speech and music segments equally likely. The emission probabilities for the HMM states are estimated by the either a GMM or an MLP expert.

The parameters of MLP are trained via the error back propagation (EBP) algorithm. Equal amounts of labeled clean speech and music data are used for training the MLP. The feature vectors $y_n = (H_n, D_n)$ from the training data are presented at the input layer of the MLP along with several context frames. The parameters of the GMM are trained in a supervised manner using standard EM algorithm. The same data is used to train both the MLP and GMM parameters.

At the time of segmentation, given the observation sequence y_n, the local likelihood of each class is calculated using the GMM (5) or MLP (6) 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, 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. Second, with the HMM it is possible to easily impose the constraint of minimum 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.

4 Evaluation Experiments

4.1 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 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 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. This also makes it easier to use entropy and dynamism individually as discriminatory features if desired. The 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).

4.2 Evaluation

We evaluated the system using 4 labeled data sets, each 10 minutes long. These data sets have a wide variety of speech and music. 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. To observe the effect of durations of sounds, segments of different durations have been mixed together. Three different experiments were performed on each of these data sets using both GMM and MLP experts.

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.

For this data set, the entropy works better than dynamism for both speech and music, although dynamism works much better for music compared to speech. Combining the two features does not significantly improve the performance over entropy. The performances of the GMM and MLP systems are comparable in this case.

4.3.2 Test Set 2

The second test set, which presents a more realistic task, consists of a 10 minute audio stream containing varying lengths of alternate speech and music segments. These segments include very short

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	93.2	98.6	95.6
GMM	Dynamism	73.8	98.3	81.8
GMM	Both	94.2	98.7	96.2
MLP	Entropy	94.1	98.7	96.2
MLP	Dynamism	79.1	93.8	84.8
MLP	Both	94.0	97.3	95.5

Table 1: Classification results for Test Set 1

(2 seconds) as well as long (14 seconds) segment durations. These classification results are shown in Table 2.

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	91.3	97.5	94.2
GMM	Dynamism	81.8	99.5	90.0
GMM	Both	91.6	98.1	94.6
MLP	Entropy	92.3	97.4	94.6
MLP	Dynamism	85.1	92.9	88.6
MLP	Both	94.4	98.6	96.3

Table 2: Classification results for Test Set 2

In the GMM framework, entropy and dynamism work better for speech and music respectively and, once again, the combination does not give a notable improvement over the performance of using only entropy features. However, in the case of the MLP (hybrid HMM/MLP classification), combining the two features does significantly improve the results. One clear observation from Table 2 is that, in the case of the MLP, the error is distributed more evenly between speech and music. This is true even for the individual performance of the two features, especially for dynamism. This observation can be explained by the fact that MLP training exhibits better discriminant capabilities than is the case for GMMs.

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 segment a predominantly speech audio signal for automatic speech recognition. These classification results are shown in Table 3.

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	90.4	94.1	91.5
GMM	Dynamism	89.9	85.8	88.3
GMM	Both	96.2	95.0	95.6
MLP	Entropy	92.5	91.0	91.8
MLP	Dynamism	94.0	76.8	86.2
MLP	Both	97.7	87.3	93.3

Table 3: Classification results for Test Set 3

From these results we see that entropy works better than dynamism for both speech and music segments. In this case, combining the two features gives a noticeable performance improvement over the individual performance of entropy and dynamism. While the overall performances of the GMM and MLP systems are comparable, the GMM and MLP can be seen to work better for music and speech, respectively.

4.3.4 Test Set 4

The final test set contains a 10 minute audio stream mostly consisting 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.

Expert	*Feature*	*Speech*	*Music*	*Total*
GMM	Entropy	86.3	95.8	91.2
GMM	Dynamism	60.9	98.6	81.0
GMM	Both	88.9	98.3	94.3
MLP	Entropy	86.2	95.8	91.7
MLP	Dynamism	77.0	93.8	86.2
MLP	Both	92.0	96.0	94.3

Table 4: Classification results for Test Set 4

Here we again see that entropy generally gives the better discrimination results, although dynamism is quite effective for detecting music segments. Combination of the two features significantly improves the performance over the use of either feature in isolation, especially in the case of the speech segments. This is true for both the GMM and MLP frameworks. Again, while the overall performance of GMM and MLP are comparable, the MLP offers more accurate detection of speech segments, which is highly desirable in the case of speech recognition applications.

4.4 Discussion

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

- Overall, entropy is a better discriminatory feature than dynamism. This is because the entropy for individual frames is generally much higher for music than it is for speech, and this situation does not vary greatly. In contrast, the dynamism for frames within a speech phoneme period can be as low as it would be for a music segment. Even though we compensate for this effect somewhat by averaging these values over a window, the dynamism remains sensitive to the phoneme durations.

- Dynamism is a better discriminatory feature for detecting music segments than it is for detecting speech segments. This is because the behaviour of dynamism does not vary during music segments, and is effectively modelled by statistical approaches such as GMMs and MLPs. In the case of speech segments, however, the dynamism may be affected by factors such as the speaking rate.

- In general, the relative behaviour of entropy and dynamism does not change in the GMM and MLP frameworks. We note, however, that the MLP tends to redistribute the error more evenly among speech and music regions due to its discriminative training.

- As expected, the combination of the two features improves the performance over the individual performance of the two features. This is because the region of overlap (confusion) between speech and music regions in the combined feature space is diminished compared to the regions of

overlap in the individual entropy and dynamism feature spaces. The combination is particularly beneficial for music, due to the more consistent behaviour of the individual features (especially dynamism).

- Overall, while the performances of the GMM and MLP systems are comparable, the MLP gives consistently higher frame accuracy for speech segments. For this reason, in applications such as large vocabulary speech recognition, where identifying true speech regions is more important than ignoring false speech regions, the MLP system would be preferable.

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.

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) [10] and then briefly discuss its use for two different purposes. As the MLP system 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 MLP system, however we note that a similar measure could also be obtained using the GMM system.

5.1 Definition of Confidence Measure

In the context of the 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:

$$R_C(N_1, N_2) = \frac{1}{N_2 - N_1} \sum_{n=N_1}^{N_2} P(C|y_n) \tag{7}$$

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 obseved 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 empirical 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 (MLP system only, using both entropy and dynamism features) of Section 4.3. The results are shown in Table 5.

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

These results demonstrate two important points. First, we can achieve a significant 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 this article 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 application, 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 are 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 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. While the overall performances of the GMM and MLP systems are comparable, we find that the MLP system tends to work consistently better for detecting speech segments, and would thus be preferable when the speech/music segmentation is a pre-processing step for speech recognition.

In addition, 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.

References

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

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

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

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

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

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

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

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

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

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