Temporal Features for Broadcast News Segmentation

Michael T. Johnson[1] and Leah H. Jamieson[2]

[1]Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI
mike.johnson@marquette.edu

[2]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN
lhj@ecn.purdue.edu

Abstract

The task of automatically segmenting an acoustic signal into categories (such as speech, speech over background music, or music) is an important step in the transcription process. We are attempting to improve the accuracy of such segmentation systems by incorporating suprasegmental and other temporal information into the frame-based classifiers typically used for this purpose. Two specific approaches are introduced here, one based on using frequency contours to improve the location of segment boundaries and one based on including temporal features directly into the frame-based classifier. Results indicate that improvement in classification accuracy can be achieved through the use of temporal information, particularly for the speech plus music class where methods using traditional features often give poor results.

1. Introduction

The problem of discriminating between speech and music in broadcast news or other real-world domains is one which can be accomplished with a fairly high accuracy (near 98% or better) using traditional spectral-distribution approaches [1-4]. The problem of discriminating simultaneous speech and music from music or speech alone, however, is much more difficult and generally creates many errors during the segmentation process, particularly with regard to exact segment boundaries. Systems for accomplishing this are typically frame-by-frame spectral distribution models such as Gaussian Mixture Models (GMMs). Since prosodic features such as intonation and energy contours have significantly different characteristics within these audio classes, it could be expected that their use, as well as the use of other suprasegmental information, might be able to improve such systems.

The Hub-4 broadcast news project and evaluations have been a primary motivator of research in this area. The task includes both an unpartitioned and a partitioned evaluation, and one of the primary difficulties of the unpartitioned recognition task is that of dividing the 30-minute speech file into separate segments for processing clean speech, noisy speech, telephone speech, and speech over music.

We will focus here on the three-class problem of differentiating speech (S), music (M), and speech over music (SM). Separating this last class from the other two has proven to be fairly difficult, and this ambiguity has resulted in a significant degree of partitioning error in a number of published systems [1, 3, 5], with SM class accuracy ranging from about 53% [5] to 66% [3]. Maximum Likelihood (ML) classification over speech frames is the predominant technique for these and other segmentation systems, often implemented as a Hidden Markov Model (HMM) to include some basic segment duration constraints (in the form of both transition probabilities and hard constraints on the Viterbi search path).

The features used for the segmentation task are frequently the same ones computed for the speech recognition task, such as Mel frequency-warped cepstral coefficients and an energy measure, with velocity/acceleration estimates (in the form of deltas and delta-deltas). For the simpler speech/music segmentation task, short-term features such as pitch and energy have also been examined [6], and have given good results using very few features.

We are investigating the possibility of using the above HMM-based classifier framework to incorporate several types of additional temporal information into the algorithm. Two approaches to this have been developed. The first of these is based on the idea of using a dynamic programming scoring algorithm to estimate how "music-like" the contour of the signal spectrum is and then re-locating the segmentation boundaries in accordance with this score. The second approach is to incorporate temporal characteristics directly into the classifier framework through duration constraints on the recognizer, mean-subtraction, and the addition of pitch and multi-frame statistical features.

Training and testing data is taken from the 1996 Hub-4 Broadcast New corpus, using ABC World News Tonight recordings, with ten half-hour shows used for training and five more reserved for testing.

In Section 2 we describe and evaluate the baseline ML classifier system developed for segmentation. The contour-based method for re-locating segment boundaries is discussed in Section 3, and Section 4 describes the implementation method and results for including duration, pitch, mean subtraction, and multi-frame statistics directly. Section 5 contains conclusions

and a discussion on the impact of utilizing additional temporal information within a standard segmentation framework.

2. Baseline System

The baseline segmentation system is an HMM-based Maximum Likelihood (ML) classifier, using a 39-element vector of 12 Mel-Frequency cepstral coefficients plus sum-squared signal energy, with delta and delta-delta coefficients appended. One HMM per class is trained, with each HMM having a single emitting state. State observation distributions are 16-mixture GMMs.

To incorporate a simple duration model, self-transition probabilities for each class are computed directly using the average number of 10 ms frames (an estimate of the expected duration D) in each segment over the training data set:

$$a_{ij} = 1 - \frac{1}{E\{D\}} \quad (1)$$

In addition, Viterbi constraints are set empirically at a 20 frame minimum, giving a 200 millisecond minimum class length.

The three-class confusion matrix of the baseline system is shown in Table 1. These results are comparable to the broadcast news segmentation results obtained by similar systems [1, 3, 5]. Accuracy is 99.7% for the S class, 95.4% for the M class, and 68.0% for the SM class.

Table 1 Baseline three-class confusion matrix

	S (%)	M (%)	SM (%)
Speech	**99.7**	0.2	0.1
Music	2.2	**95.4**	2.5
Speech over Music	25.8	6.2	**68.0**

3. Boundary re-location using frequency contour scoring

There are distinct differences between the long-term frequency characteristics of speech and music. To illustrate this, Figure 1 shows the fundamental frequency (f_0) contour of a segment of music transitioning into speech, taken from the broadcast news corpus. It is clear where the flat structured pattern of the music contour, consisting of individual notes with fairly clean transitions, changes to the smoother flowing pattern of the speech. The intonation pattern of speech and music together, however, varies depending on the pitch detection algorithm, the distance between the notes of the music and the pitch of the speaker, and in particular the relative amplitudes of the speech and music components of the signal.

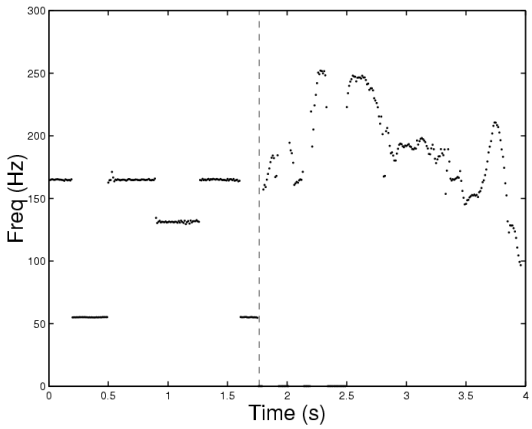

Figure 1 F_0 of music transitioning to speech

3.1. Contour tracking algorithm

In order to identify speech and music together, it is necessary to identify music-generated frequency patterns, even in the presence of louder speech. One approach to do this would be to identify not the fundamental frequency contour across frames, but the contour consisting of the frequency components that together form the most music-like pattern, which is in generally a series of piecewise-constant steps (including the possibility of piecewise-constant pauses, present in both music and speech classes).

The algorithm we are introducing here is based on a dynamic programming approach to the recognition of piecewise-horizontal segments of frequencies in the signal. Its operation is somewhat similar to an approach often used for fundamental frequency identification [7] or for optimized non-linear filtering [8]. However, rather than minimizing on the basis of the relative strength of frequency peaks, it weights all frequency candidates equally (since we are concerned with the presence of the components but not their strength), and finds the sequence of frequency segments with the smallest intra-segment variation and of the longest durations.

A basic block diagram of the system is shown in Figure 2. Spectral analysis and peak picking is accomplished using an autocorrelation measure, then a dynamic tracking algorithm identifies the sequence of frequency candidates that minimizes a scoring function, which incorporates exponentially-increasing penalties for large segment variances or small segment lengths. For full details of the algorithm, see [9].

Figure 3 shows a 250 millisecond differential plot of the algorithm's output on an example segment of speech from broadcast news. This example is taken from a segment with approximately 5 seconds of music, 10 seconds of speech over music, and 5 seconds of speech alone. In this example, the music component is fairly

strong (although below the speech amplitude), then fades and can barely be heard in the background for several seconds. The baseline system correctly classifies the segments but places the second boundary, between the speech over fading music and speech alone, approximately two seconds before the actual final transition.

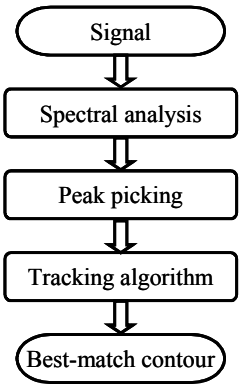

Figure 2 Block diagram of tracking algorithm

Dotted lines within the figure represent the actual boundary points between classes. It can be easily seen that the straight music segment maintains a low score throughout, and that the first part of the speech over music is also identifiable. The transition period, although still somewhat lower than the all-speech section, is more difficult to distinguish.

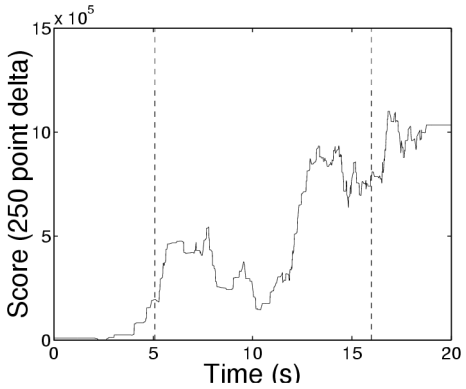

Figure 3 Moving score over M-SM-S segment

3.2. Application to boundary identification

Analysis of the segmentation errors in the baseline system of Section 2 reveals that 85.3% of the total error was due to misplacement of segment boundaries rather than segment misclassification. This suggests that applying the contour scoring algorithm to the problem of boundary relocation could make substantial improvement in system accuracy. To do this, a region of interest is chosen around each segmentation point identified by the recognizer, then a new boundary identified using the score information from our dynamic programming algorithm. This effectively reduces the three-class decision to a two-class decision.

The scoring algorithm creates a cumulative score that increases very slowly for music-like signals and very quickly for speech-like signals, essentially leaving a net score consisting of two approximately linear segments with different slopes. The net change in this slope is indicative of how well the score indicates a specific boundary between the two segments. Thus, by fitting two line segments to the cumulative score over a signal, we can determine both a boundary location from their intersection and an indication of boundary confidence from their angle. An example of this is shown in Figure 4 for a speech to music transition, with dotted lines showing the best-fit lines.

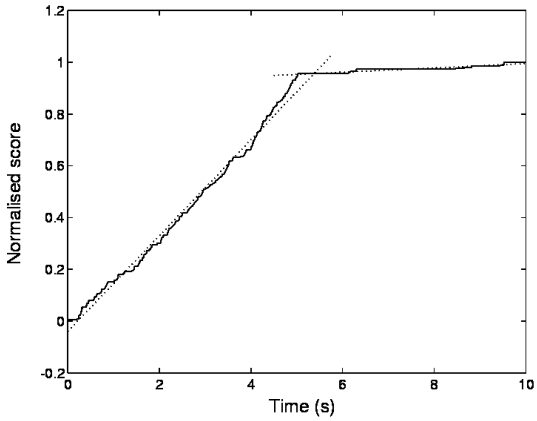

Figure 4 S-M transition cumulative score plot

Computing the best-fit two line segments to fit a given set of data is a straightforward extension of the least-squares linear regression algorithm. Each point k in the data set represents a possible boundary point between a set of points less than or equal to k and a set of points greater than k, partitioning the problem into the sum of two linear regressions. The minimum value of the summed mean squared errors, computed for each possible boundary point k, gives the desired minimum error boundary point, with the corresponding angle between the lines computable from the two slopes.

Since linear regression is an O(N) algorithm and we need to do 2N linear regressions for a segment, the time complexity of the best-fit two line segment problem would be O (N^2) if implemented directly. However, it is possible to keep track of the necessary data (cumulative sums of x, y, x^2, y^2, and xy) in a linear fashion, by adding the appropriate values of the next data point k into the left data set and subtracting them from the right data set, so that the linear regressions do not have to be fully recomputed at each increment. The net complexity is thus still O(N) .

To convert our boundary point and angle information into likelihood numbers that can be

combined with our baseline system, we model the chosen point as the mean of a Gaussian distribution, with a variance determined by a function of the angle between the line segments. This produces frame-level boundary likelihoods which can be combined with the frame-level class likelihoods from the baseline system to produce the best overall segmentation point.

3.3. Results

Results using this approach are shown in Table 2. The results are better than the original baseline system, but still do not represent a large decrease in error, with only a 1.3% net gain in accuracy for the speech over music class compared to the baseline system of Table 1.

Table 2 Results for contour-based boundary relocation

	S (%)	M (%)	SM (%)
Speech	**99.7**	0.2	0.1
Music	2.0	**95.8**	2.2
Speech over Music	24.5	6.2	**69.3**

4. Inclusion of Temporal Features

We now examine the possibility of additional temporally-motivated features for segmentation, under the established HMM framework for ML classification. Temporal concepts we investigate include the following:

- Learned Duration Constraints (both fixed and class-dependent)
- Pitch estimates
- Cepstral mean subtraction
- Multi-frame statistics

Each of these is examined here with respect to its usage and potential impact on the segmentation task. Resulting class accuracies are summarized as a group in Table 3 of Section 6.

4.1. Learned Duration Constraints

One of the most important aspects of the acoustic classification problem is that of duration. Even though we are using a frame-based classifier which indicates the acoustic-class likelihoods at a ten millisecond resolution, we know that because of the basic underlying characteristics of human communication (and news shows in particular), the acoustic class will not be changing rapidly, with average segment lengths of thousands of frames. This duration information can be incorporated into the HMM in two ways: through transition probabilities and through constraints on the minimum segment length allowed during the Viterbi search. The first of these methods is stochastic in nature and assumes a duration distribution which is exponential in nature, while the second is rule-based and allows for fixed constraints to be applied.

The baseline system uses both of these methods to model segment duration. To see how significant duration is, Table 3 also includes accuracy results using the baseline system without either transition probabilities or duration constraints as well as with transition probabilities alone. The results clearly indicate that the frame-based system we are using is already very heavily dependent on temporal characteristics (specifically segment continuity) in many respects.

The transition probabilities used for the baseline are optimal over the training set, chosen using the mean duration of the class segments. The path constraints, however, are chosen fairly arbitrarily at fixed values of 200 milliseconds (20 frames) for all classes. Since these constraints have tremendous impact on accuracy, better methods for selecting limits could potentially lead to significant accuracy improvements.

Such methods can be either class-dependent, i.e. learned from the distribution of segments within each class, or class-independent. One danger here is that although the quantity of training data is large with respect to learning spectral distributions, the number of distinct segments is relatively small. Figure 5 shows a cumulative histogram plot of these distributions.

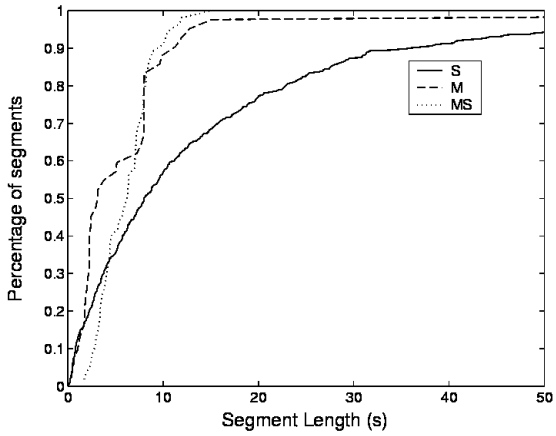

Figure 5 Training data class distribution

The classes have minimum segment durations varying from 80 milliseconds (speech) to 1.73 seconds (speech over music). There are a number of different criteria that could be used to establish well-founded duration constraints for the classes, including using the minimum segment duration itself, the expected minimum segment duration (using the region of the distribution near the origin), or even by running multiple classification passes with different constraints to minimize training set error. For a single constraint parameter, this is straightforward, but for three parameters the number of combined possibilities

requires some implementation care. Because of the small number of segments, care is also needed to ensure robustness to new data.

Noting that the characteristics of the distribution near the origin are fairly linear, we use a regression analysis over that portion of the data to identify a zero-crossing point for each class, to serve as minimum duration estimates. In addition, since the cost of error goes up as the duration constraint is increased (one short segment could create a large number of frame errors), it was arbitrarily decided to limit the duration constraint to no more than 1 second (100 frames) for any class, a condition that affected only the speech over music duration constraint.

As a class-independent approach, we also conducted a sequence of experiments to determine the optimum setting (with respect to training set error) for the minimum duration constraint, keeping the constraints for all three classes equal. Figure 6 shows the resulting training set class accuracies. The accuracy curves show some interesting characteristics, with the music accuracy peaking fairly early and then dropping off. The top overall accuracy is exhibited at a minimum duration of between 600 and 650 milliseconds, after the speech over music accuracy has peaked, but before the music accuracy degrades by very much.

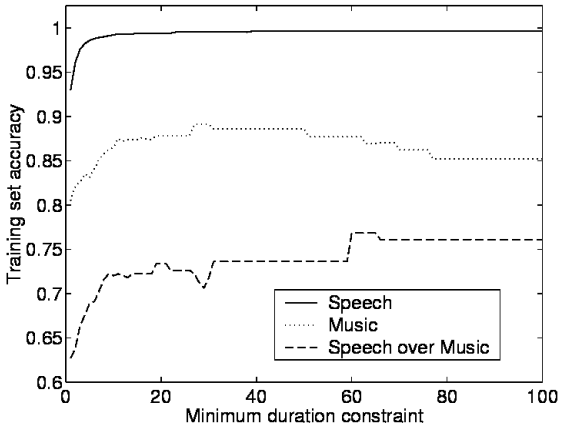

Figure 6 Training accuracy versus duration constraint

Each of these approaches to learned duration constraints had a significant impact on segmentation accuracy, as seen in Table 3. The class-based method yielded no change in the S or M class accuracy but increased the SM class from 68.0% to 80.6%, and a simple change in minimum duration length resulted in a further increase to 87.5% for that class (although also reducing music accuracy to 92.8%).

4.2. Pitch Information

Two methods were used to include pitch (or more precisely fundamental frequency f_0, as an estimate of pitch) as an element of the feature vector. In the first method the output of a f_0 estimation algorithm, as well as its delta and delta-delta, were directly appended to the original vector, for both voiced and unvoiced frames (for which there is no pitch present, designated by a value of zero). As a second method, the Baum-Welch training and Viterbi recognition modules were modified so that frames which had a value of zero for the pitch feature did not use any of the pitch-related features for probability computations or parameter re-estimation.

The second method is clearly more precise with respect to the global mean and variance of class pitch. However, since the occurrence of unvoiced frames varies significantly between classes, in a multiple-mixture system (where we can think of clusters of the feature space as represented by individual Gaussian models), the more direct method presents an opportunity for the HMM to use both voiced/unvoiced information and pitch values.

The direct method resulted in fairly significant improvement to the SM class, to 76.3% accuracy, while the S class and M class exhibited a slight decrease in accuracy, to 99.5% and 92.4% respectively.

4.3. Cepstral Mean Subtraction

Cepstral mean subtraction (CMS) is a very common speech recognition technique for increasing system robustness to noise, channel, and speaker variation. Although the BN task is a multi-speaker environment, it is of interest to see if channel or transmission/recording characteristics (since there is often a 24 hour break or more between each recording) would be significant enough for CMS to have positive impact.

Results of applying this technique to our data are again shown in Table 3. Clearly, the effects are significant here as well, in particular showing an 11.1% increase in the SM classification accuracy.

4.4. Multi-frame statistics

A key feature distinguishing speech and music is that of the length and regularity of pauses [2, 10]. One way to measure this effect quantitatively is through the use of statistical measures over blocks of frames. In [10], several statistical approaches based on the distribution of the amplitude envelope of speech were used to distinguish between spontaneous and non-spontaneous speech with some success. Using a moment expansion to represent the density, the first and third central moments were found to be good representations of this regularity characteristic. We have chosen to use the mean, standard deviation, and skewness (normalized third central moment), over a moving window of feature characteristics, as measures of feature change and structure over time.

Reviewing the work of Spina and Zue [5], whose segmentation work including an analysis of class accuracy versus size of the computation window for

feature averages, we chose to use a 51-frame feature window, over each of the 13 original features. Statistics were appended to the original 39 element feature vector (so that the short-term spectral information was kept as well), giving a feature vector size of 78 elements. The results show relatively little impact due to the added features. Accuracy was unchanged for the S class, down 1.3% for the M class, and up 2.3% for the SM class.

4.5. Results

Class accuracies for each of the above approaches are summarized in Table 3.

Table 3 Individual feature summary

	S (%)	M (%)	SM (%)
Equal likelihoods, no duration constraints	78.6	83.1	53.0
No duration constraints	95.2	91.0	58.7
Baseline	**99.7**	**95.4**	**68.0**
Class-based duration	99.7	92.2	78.5
Independent duration	99.7	92.8	87.5
Appended f_0	99.5	92.4	76.3
Voiced-only f_0	99.7	91.7	65.4
Cepstral mean subtraction	99.8	94.4	79.1
Appended 51-frame statistics	99.7	94.1	70.3

Using all four of these approaches together, a composite segmentation system was built. For duration, the class independent constraints were selected, and for pitch information, the raw f_0 estimates were appended to the feature vector. Results are shown in Table 4. Final class accuracy for the speech class was unchanged at 99.7%, while the music class increased 0.9 to 96.3% and the SM class improved by a net 16.2 to 84.2%, representing a factor of two reduction in class error for combined speech and music.

Table 4 Final cumulative results summary

	S (%)	M (%)	SM (%)
Speech	**99.7**	0.1	0.2
Music	0.3	**96.3**	3.4
Speech over Music	9.3	6.4	**84.2**

5. Conclusions

Improvement in segmentation accuracy due to the application of contour analysis and scoring was insubstantial but did indicate the possibility of using prosodic patterns to directly differentiate classes. Improvement due to applying better duration constraints during frame-based spectral classification was quite significant, and the use of appended features such as fundamental frequency and multi-frame statistics also resulted in a measurable increase in accuracy.

Overall, the results indicate that the incorporation of prosodic or temporally-motivated information can have a significant impact on accurate identification and segmentation of acoustic classes.

6. References

[1] R. Bakis, S. Chen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos, "Transcription of broadcast news - System robustness issues and adaptation techniques," presented at Proc. ICASSP, 1997.

[2] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," presented at Proc. ICASSP, 1999.

[3] J. L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker, "Transcribing broadcast news shows," presented at Proc. ICASSP, 1997.

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

[5] M. S. Spina and V. W. Zue, "Automatic Transcription of General Audio Data: Preliminary Analyses," presented at Proc. ICSLP, 1996.

[6] M. J. Carey, E. S. Paris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," presented at Proc. ICSLP, 1996.

[7] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," presented at Proc. ICASSP, 1983.

[8] H. Ney, "A Dynamic Programming Technique for Nonlinear Smoothing," presented at Proc. ICASSP, 1981.

[9] M. T. Johnson, "Incorporating Prosodic Information and Language Structure into Speech Recognition Systems," in *Electrical and Computer Engineering*. West Lafayette, IN: Purdue University, 2000.

[10] O. P. Kenny, D. J. Nelson, J. S. Bodenschatz, and H. A. McMonagle, "Separation of Non-spontaneous and spontaneous speech," presented at Proc. ICASSP, 1998.