Classification of Music and Speech in Mandarin News Broadcasts *

Chuan Liu †, Lei Xie‡, Helen Meng§

October 29, 2007

Abstract

Audio scene analysis refers to the problem of classifying segments in a continuous audio stream according to content, e.g. speech versus non-speech, music, ambient noise, etc. Techniques that support such automatic segmentation is indispensable for multimedia information processing. For example, it is a precursor to processes such as indexing of speech segments by automatic speech recognition, automatic story segmentation based on recognition transcripts, speaker diarization, etc. This paper describes our work in the development of a speech/music discriminator for Mandarin broadcast news audio. We formed a high-dimensional feature vector that includes LPCC, LPS and STFT coefficients totaling 94 in all. We also experimented with three classifiers – the KNN, SVM and MLP. Experiments based on the Voice of America Mandarin news broadcasts show high classification performance with F-measure=0.98. The SVM also strikes the best balance in terms of classification performance and computation time (real-time) among the three classifiers.

Keywords: audio scene analysis, speech/music discrimination

1 Introduction

The explosive growth of multimedia content on the Internet presents a dire need for automated technologies for information processing. Audio scene analysis is an indispensable component in multimedia information processing. This paper describes our work in audio scene analysis that involves automatic classification of streaming audio news broadcasts into silence, speech or music segments. Such classification is useful for a diversity of subsequent kinds processing, such as:

- Filtering for speech segments in the broadcast audio for indexing by automatic speech recognition [MSD, 2003];
- Detecting speech segments corresponding to different speakers or segments from the same speaker in different environments (i.e. diarization) for speaker tracking [Tranter and Reynolds, 2006];
- Using speaker-adapted acoustic models to recognize corresponding speech segments for improved processing performance [Reynolds and Torres-Carrasquillo, 2005];
- Using the recognition transcripts to perform automatic story segmentation of the continuous streaming audio [Chan *et al.*, 2007]; and
- Using specific musical cues to identify landmark regions in the audio repository [Reynolds and Torres-Carrasquillo, 2005].

Previous work on audio classification focused on the aspects of both feature extraction as well as classification for the task. In terms of feature extraction, it was reported in [Scheirer and Slaney, 1997] that long-term behavior of audio is important for the development of a robust, multi-feature speech/music discriminator. Additionally, it was also reported in [Carey *et al.*, 1999] that delta features offer good performance for speech/music discrimination. In [Li *et al.*, 2001], 143 features were used for audio classification into 6 categories. It was reported that cepstral-based features such as Mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC)

*Much of this work is conducted as first author's Bachelors thesis research project and during the second author's appointment as postdoctoral fellow at the Chinese University of Hong Kong. The title shows the current affiliations of the authors.

†Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, China

‡School of Computer Science, Northwestern Polytechnical University, Xi'an, China

§Department of System Engineering and Engineering Management, The Chinese University of Hong Kong

provide better performance than temporal and simple spectral features. The work described in [Lu *et al.*, 2002] classified audio into speech, music, environment sounds and silence and involved the use of new features such as band periodicity.

In terms of classification, previous work involved the use of Gaussian mixture models (GMMs), and k-nearest-neighbor (KNN), support vector machines (SVMs), multi-layer perceptrons (MLPs), radial basis functions (RBFs) and Hidden Markov Models (HMMs). In particular, it was reported in [Lu *et al.*, 2003] that SVMs perform better than KNN and GMM in their task. Comparison among MLPs, RBFs and HMMs in [Khan and AI-Khatib, 2006] showed that MLP achieved the best performance in their task.

The objective of this work is to develop a speech/music discriminator for Mandarin news broadcast audio as a pre-process for the subsequent tasks of automatic story segmentation [Xie *et al.*, 2007] [Chan *et al.*, 2007].

2 Data and Annotation

Our experimental data is derived from the Topic Detection and Tracking 2 (TDT2) Mandarin Audio collection [Graff, 2001]. These are recordings of the Voice of America (VOA) Mandarin news broadcasts, collected daily over a period of six months (February to June 1998). The audio files in this corpus are single channel, 16 kHz, 16-bit linear SPHERE files. There are also automatic speech recognition (ASR) transcripts provided by Dragon Systems for the audio. In order to provide a ground-truth reference, we utilized the recognition transcripts and labeled ten hours of audio semi-automatically into four categories – silence, music, speech as well as speech with music. The labeled data included five days of recordings (between 20 to 25 February 2001), with about two hours per day. Table 1 shows the average duration (in seconds) per segment in each category.

Category	Average length (sec)
Silence	0.5
Music	6.0
Speech	2.2
Speech with music	2.5

Table 1: Average segment duration for each of the four categories of audio segments.

In our experimentation, we grouped the categories of "speech" and "speech with music" together since our subsequent task of story segmentation will require further processing of segments carrying speech. In addition, we performed silence removal by thresholding on short-time energy. As shown in Figure 1, the majority of silence segments contain short-time energies with values below 0.0005. This threshold was effective for filtering out the silence segments so that we can focus on speech versus music discrimination.

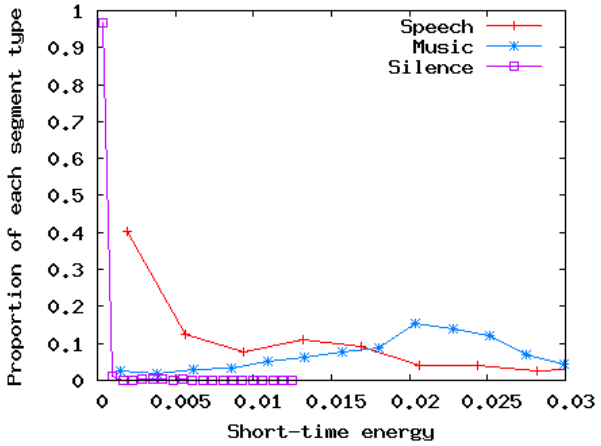

Figure 1: Distribution of short-time energy across different segment types. The proportions in each segment type (y-axis) are plotted against the short-time energy values (x-axis).

We randomly partitioned our data into three sets as illustrated in Table 2.

Data set	Quantity (hours)
Training	4
Development test (for parameter selection)	2
Evaluation	4

Table 2: Data sets for experimentation

We noted that our data include speech from several major speakers (newscasters), speech from interviews with various interviewees as well as different ambient noises, specific music clips that recur to signify the structure of broadcast programs, as well as other segments containing music from a variety of genres ranging from electronic music to rock. Our data do not contain pure vocal music without instrumental accompaniment.

3 Features

With reference to previous work, we included a variety of features in our experimentation. The software Marsyas [Tzanetakis, 2006] is used to extract four types of features and the total count is 94. This feature set includes the "variance" features with reference to [Scheirer and Slaney, 1997] that consist of the standard deviations of the feature vectors calculated within a window of 1.28 seconds (covering 40 frames). Table 3 presents a summary of the feature types followed by a brief description of each type.

Feature type	LPCC	LSP	MFCC	STFT
# dimensions	24	36	24	10

Table 3: The number of features for each of the four feature categories.

LPCC: We used a set of 12 linear predictive cepstral coefficients with their standard deviation.

LSP: This feature set is based on 18 linear spectral pairs together with their standard deviation.

MFCC: This feature set is based on 12 Mel-frequency cepstral coefficients and their standard deviation.

STFT: We used features derived from the short-time Fourier transform, including the centroid, rolloff, flux, kurtosis and zero-crossings:

- The **spectral centroid** is the balancing point of the spectral power distribution.
- The **spectral flux** is the 2-norm of the difference between the magnitudes of the spectrum evaluated at two successive sound frames.
- The **spectral rolloff** point is the t-th percentile of the spectral power distribution, where t is a threshold value. Here we chose $t = 0.90$.
- The **spectral kurtosis** is the 4th order central moment.
- The **zero-crossing rate** is defined as the number of time-domain zero-crossing within the processing window.

In Figures 2 and 3 we plot some feature values based on the training data set to illustrate their utility in discriminating between speech and music. We used the standard deviation of the second coefficient for the LPCC and LSP as examples.

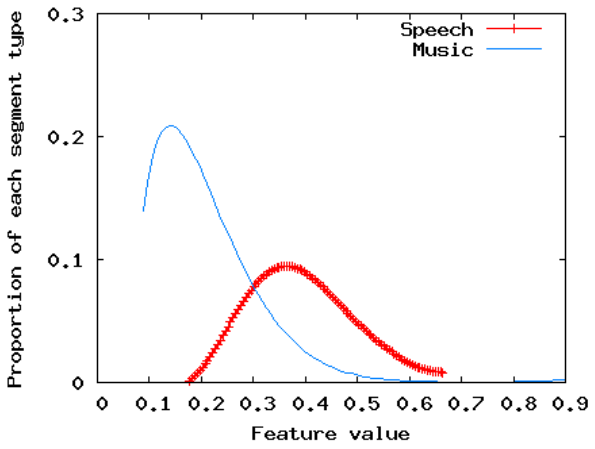

Figure 2: The histogram for standard deviation of the second LPCC coefficients.

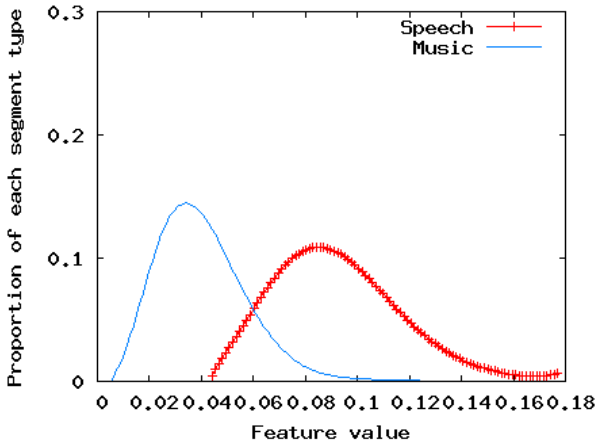

Figure 3: The histogram for standard deviation of the second LSP coefficients.

4 Classifiers

We experimented with three classifiers that have demonstrated favorable performance in previous work, as described in the introductory section. They are:

- k-nearest neighbor (KNN)
- Multi-layer perceptron (MLP)
- Support vector machine (SVM)

KNN adopts an instance-based learning method. It performs classification of a test instance based on the closest instances in the training set. The training phase of the algorithm stores the feature vectors and their corresponding class labels. The testing (or classification) phase measures the distance (often the Euclidean distance) between the test vector and the stored sample vectors. The k closest samples are selected. The new object is classified with the most frequent class label appearing among the k closest instances. Our experimentation involves optimizing the value of k based on the development test set.

MLP is a frequently used artificial neural network. An MLP network has an input layer of source nodes, one or more hidden layers of computation nodes, and an output layer. The MLP solves the classification problem in a supervised manner and is trained with the back-propagation algorithm. With reference to previous work such as [Khan and AI-Khatib, 2006], we experimented with the topology of single and double hidden layers. Our MLPs also have an input later with 94 nodes (corresponding to the input vector) and an output layer with 2 nodes (corresponding to the speech versus music discrimination). We label a test instance based on the output node with the higher value. We also experimented with a different number of nodes in the hidden layer. Figure 4 demonstrates a simplified MLP network structure with an input layer of 94 nodes (a partial set is shown and labeled), one hidden layer with 5 or 10 nodes (the illustration shows 5 nodes) and an output later with two nodes.

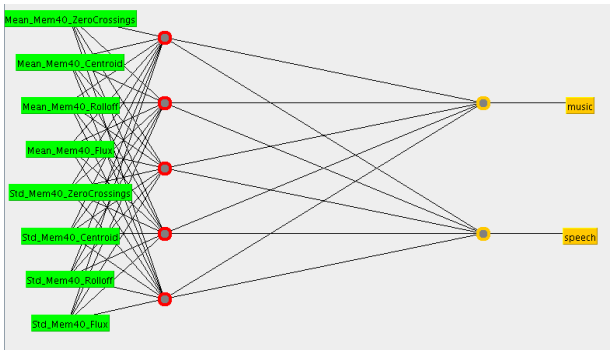

Figure 4: A simplified MLP network structure with 1 hidden layer and 5 neurons.

SVM also use supervised learning methods for classification. SVMs map input vectors to a higher dimensional space. Then a hyperplane is constructed to separate the input vectors. Two parallel hyperplanes are constructed on each side of the hyperplane. The hyperplane that maximize the distance between the two parallel hyperplanes is found to be the solution. In linear non-separable cases, we need a kernel function to transform the original feature space to a higher dimensional space in an implicit way such that the mapped data is linearly separable. Common kernels include polynomial, Gaussian radial basis function, sigmoid, etc. In our experiments, we used the linear kernel due to its simplicity.

5 Experiments

All classifiers were trained with the training set. We also optimized parameter settings for KNN and MLP based on the development test set. Table 5.1 shows the results from the development test set based on the precision (p), recall (r) and F-measure (f) values for retrieval of the music and speech categories respectively. The corresponding equations are:

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$f = \frac{2 \cdot (p \cdot r)}{p + r}$$

where TP is number of truth positive instances, FP is number of false positive instances and FN is number of false negative instances.

All the classifiers performed very well on the development test set. SVM and MLP were able to achieve perfect scores in the development test set. We surmise that this is due to the similarity between the data in the training and development test sets. In terms of training time, the KNN classifier has the lowest requirement as it simply stores the training vectors and their labels. However, KNN requires heavy computation during evaluation and hence longer computation time when compared with the other two classifiers. Computation times for testing with SVM and MLP are comparable and they are sufficiently fast for real-time audio classification performance.

For evaluation with the test set, we selected parameters each classifier based on its performance on the development test set. As shown in the Table 4, the choice of k did not affect the performance of KNN much at all. Therefore, we simply chose $k = 1$ for

evaluation due to its simplicity and fast run-time performance. Table 4 also shows that using a single hidden layer with a greater number of nodes (10 nodes) gave desirable results, so this topology was adopted for the MLP in evaluation. Evaluation results are shown in Table 5. All the classifiers show very close performance. The SVM exhibits the best balance between a high classification performance and low computation time.

6 Conclusions and Future Work

In this work, we have developed an audio classifier that discriminates between speech and music segments in Mandarin broadcast news audio. This serves as a pre-process for our subsequent work on automatic story segmentation of broadcast news. Our experiments are based on a subset of the VOA corpus. We used a high-dimensional vector with 94 features in all, derived from LPCC, LSP, MFCC and STFT respectively. We also experimented with three different kinds of classifiers – KNN, SVM and MLP. Overall, the SVM strikes the best balance between classification performance (F-measure=0.98) and classification speed (i.e. real-time response). Future work includes the use of recognition transcripts to index the speech segments of the audio to perform automatic story segmentation and speaker diarization.

Acknowledgments

This research is partially supported by the CUHK Shun Hing Institute of Advanced Engineering.

References

[Carey *et al.*, 1999] Michael J. Carey, Eluned S. Parris, and Harvey Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proc. ICASSP*, volume 1, pages 149–152, March 1999.

[Chan *et al.*, 2007] Shing-Kai Chan, Lei Xie, and Helen Meng. Modeling the statistical behavior of lexical chains to capture word cohesiveness for automatic story segmentation. In *Proc. Interspeech*, Antwerp, Belgium, August 2007.

[Graff, 2001] David Graff. *TDT2 Mandarin Audio Corpus.* The Linguistic Data Consortium, Philadelphia, 2001. http://projects.ldc.upenn.edu/TDT2/.

[Khan and AI-Khatib, 2006] M. Kashif Saeed Khan and Wasfi G. AI-Khatib. Machine-learning based classification of speech and music. *Multimedia Systems*, 12(1):55–67, August 2006.

[Li *et al.*, 2001] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22:533–544, April 2001.

[Lu *et al.*, 2002] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, October 2002.

[Lu *et al.*, 2003] Lie Lu, Hong-Jiang Zhang, and Stan Z. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8(6):482–492, April 2003.

[MSD, 2003] ISCA workshop on multilingual spoken document retrieval, 2003.

[Reynolds and Torres-Carrasquillo, 2005] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and application of audio diarization. In *Proc. ICASSP*, volume 5, pages 953–956, Philadelphia , PA, March 2005.

[Scheirer and Slaney, 1997] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. ICASSP*, volume 2, pages 1331–1334, April 1997.

[Tranter and Reynolds, 2006] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565, 2006.

[Tzanetakis, 2006] George Tzanetakis. Marsyas: Music analysis, retrieval and synthesis for audio signals, December 2006. Software available at `http://marsyas.sourceforge.net/`.

[Xie *et al.*, 2007] Lei Xie, Chuan Liu, and Helen Meng. Combined use of speaker- and tone-normalized pitch reset with pause duration for automatic story segmentation in Mandarin broadcast news. In *Proc. NAACL/HLT*, pages 193–196, Rochester, New York, April 2007.

Classifier	Parameter	Class	Precision	Recall	F-Measure	Time to train
KNN	$k = 1$	Music	0.996	1	0.998	N/A
		Speech	1	0.996	0.998	
	$k = 2$	Music	0.996	1	0.998	
		Speech	1	0.996	0.998	
	$k = 3$	Music	0.996	1	0.998	
		Speech	1	0.996	0.998	
SVM		Music	1	1	1	0.5 sec
		Speech	1	1	1	
MLP	1 hidden layer; 5 nodes	Music	0.996	1	0.998	41.3 sec
		Speech	1	0.996	0.998	
	1 hidden layer; 10 nodes	Music	1	1	1	98.9 sec
		Speech	1	1	1	
	2 hidden layers; 5 nodes	Music	0.991	0.996	0.993	53.7 sec
		Speech	0.996	0.991	0.993	
	2 hidden layers; 10 nodes	Music	1	1	1	104.3 sec
		Speech	1	1	1	

Table 4: The experiment result based on the development test set.

Classifier	Class	Precision	Recall	F-Measure
KNN ($k = 1$)	Music	0.963	0.996	0.979
	Speech	0.996	0.962	0.979
SVM	Music	0.965	0.996	0.98
	Speech	0.996	0.964	0.98
MLP (1 hidden layer; 10 nodes)	Music	0.965	0.998	0.981
	Speech	0.998	0.964	0.981

Table 5: The classification results based on the test set.