Classification of Music and Speech in Mandarin News Broadcasts

Chuan Liu[1,2], Lei Xie[2,3] and Helen Meng[1,2]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, China
[2]Department of System Engineering and Engineering Management, The Chinese University of Hong Kong
[3]School of Computer Science, NorthWestern Polytechnic University, Xi'an, China

1. Introduction

Motivation: The explosive growth of multimedia content on the Internet presetns a dire need for automated technologies for information processing. Audio scene analysis is an indispensable component in multimedia information processing.
Our work: Automatic classification of streaming audio news broadcasts into silence, speech or music segments as preprocessing for story segmentation.
Other applications: - Filtering of speech segments for speech indexing;
- Speaker diarization;
- Using musical cues to identify landmark regions in the audio repository.

3. Features

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

LPCC: 12 linear predictive cepstrum coefficients with their standard deviation.
LSP: 18 linear spectral pairs together with their standard deviation.
MFCC: 12 Mel-frequency cepstral coefficients and their standard deviation.
STFT: derived from the short-time Fourier transform, including the centroid, rolloff, flux, kurtosis and zero-crossings.

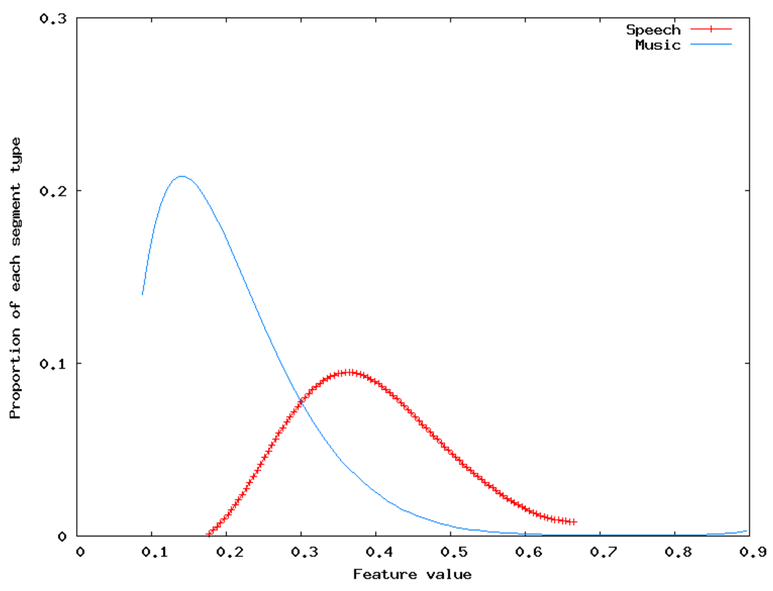


Figure 3.1: The histogram of standard deviation of the second LPCC coefficients.

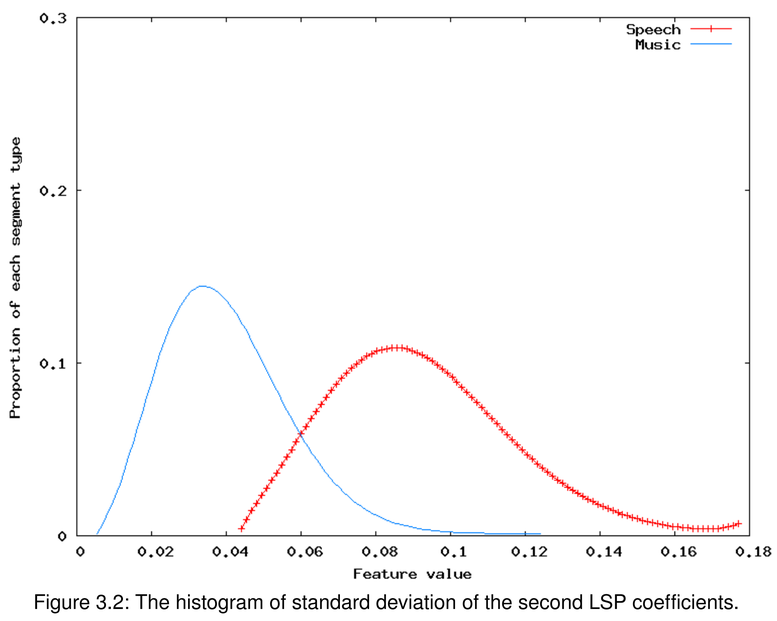


Figure 3.2: The histogram of standard deviation of the second LSP coefficients.

2. Data and Annotation

Data source: Topic Detection and Tracking 2 (TDT2) Mandarin Audio collection.
Annotation: We labeled ten hours of audio semi-automatically into four categories – silence, music, speech and speech with music. We grouped the categories of "speech" and "speech with music" together since our subsequent task of story segmentation with require further processing of segments carrying speech.
Data sets: We randomly partitioned our data into three sets.

Data Sets	Quantity (hours)
Training	4
Development test	2
Evaluation	4

Silence removal: Thresholding on short-time energy.

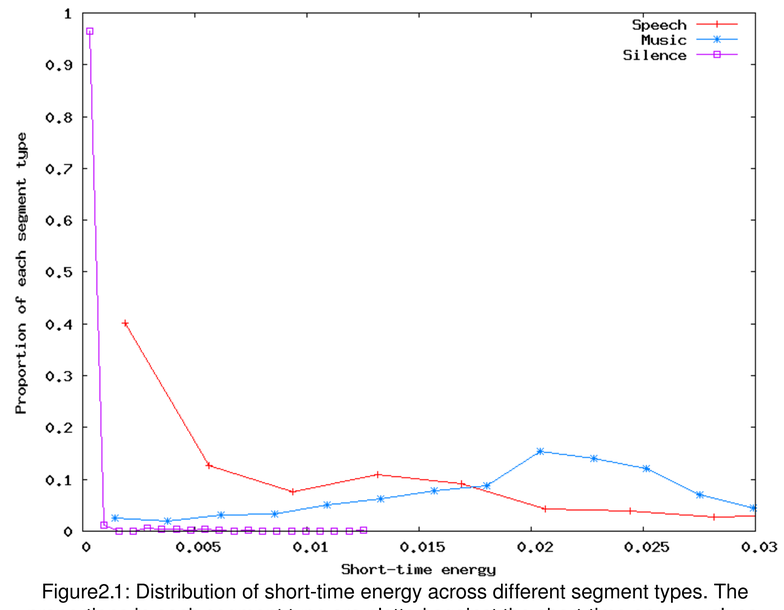
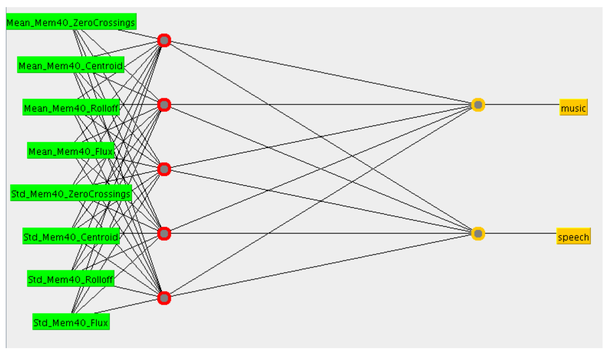


Figure2.1: Distribution of short-time energy across different segment types. The proportions in each segment type are plotted against the short-time energy values.

4. Classifiers

The K-nearest neighbor classifier (KNN): instance-based learning method. Our experimentation involves optimizing the value of k based on the development test set.
The multi-layer perceptron (MLP): artificial neural network. We experimented with the topology of single and double hidden layers. Our MLPs have an input later with 94 nodes and an output layer with 2 nodes.



Support vector machines (SVM): supervised learning methods for classification. In our experiments, we used the linear kernel due to its simplicity.

5. Experiments

Training: carried out with the training set.
Parameter setting: optimized based on the development test set for KNN and MLP.
Criteria: precision, recall and F-measure.
Result:

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 5.1: Experimental results 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	0.980
	Speech	0.996	0.964	0.980
MLP (1 hidden layer, 10 nodes)	Music	0.965	0.998	0.981
	Speech	0.998	0.964	0.981

Table 5.2: Evaluation results based on the test set.

- SVM and MLP achieved perfect scores in the development test set.
- 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

- We have developed an audio classifier that discriminates between speech and music segments in Mandarin broadcasts.
- This serves as a pre-process for our subsequent work on story segmentation.
- 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.
- Future work includes the use of recognition transcripts to index the speech segments to perform story segmentation and speaker diarization.