Audio Segmentation

Shi Yong
yong.shi2@mail.mcgill.ca

1. Introduction

This summary attempts to introduce the basic idea of the audio segmentation technique. In many applications, we are interested in segmenting the audio stream into homogeneous regions. For different cases, we may have different ideas about how to define the rules of homogeneity, but the key question is the same -- to find the changing points and mark them, so that different regions can be handled differently.

Generally, the task of audio segmentation can be divided into two sub-tasks. The first step is to extract a suitable set of features from the audio stream data. After that, different segmentation methods can be used to estimate the changing points. There are three typical segmentation methods, namely energy-based segmentation, model-based segmentation, and metric-based segmentation.

2. Features Extraction

The design of a set of good features is very important for building an audio segmentation system. In (Tzanetakis and Cook 2002), three different categories of features, that based on Timbral Texture, Rhythmic Content, and Pitch Content respectively, are proposed for automatic musical genre classification purpose. In the audio segmentation context, the most useful features are those based on timbral texture.

Timbral textural features may be either extracted from time domain or from frequency domain. The most often used time domain features might be Zero Crossing Rate (ZCR) and RMS. On the other hand, there are a number of candidates from the frequency domain a few examples are listed below (Tzanetakis and Cook 2002):

- Spectral Centroid, defined as the weighted mid-point of the spectrum: $C_t = \frac{\sum_{n=1}^{N} M_t[n]*n}{\sum_{n=1}^{N} M_t[n]}$, where the $M_t[n]$ is the magnitude of the spectrum of frequency bin n.
- Spectral Rolloff, defined as $\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n]$, this is another representation of spectrum shape.

- Spectral Flux: $F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2$, where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at frame t and t-1, respectively.
- Mel-Frequency Cepstral Coefficients (MFCC): the frequency bins are wrapped to the Mel-frequency scale, and then MFCC coefficients are computed via a DCT. For speech representation, typically 13 coefficients are used. For music instrument representation, 18 cepstral coefficients are appropriate (Brown 1999). MFCC are reported to be successfully used in speech recognition, they can also be used in audio segmentation.

3. Segmentation Approaches

In (Chen and Gopalakrishnan 1998; Kemp et al. 2000) , typical segmentation methods were categorized into three groups, namely energy-based, metric-based, and model-based.

The energy-based algorithm only makes use of the running power in time domain. On the other hand, both the metric-based and the model-based method are based on statistical models, say, multivariate Gaussian distributions. That means, rather than using the feature values directly, the running means and variances of them are modeled by a multidimensional Gaussian distribution.

3.1. Energy-based algorithm

The energy-based algorithm can be very easily implemented. Silence periods, that measured by the energy value and a predefined threshold, are assumed to be the segment boundaries. However, since there is no direct connection between the segment boundaries and the acoustic changes, this method can be problematic for many applications, such as gender detection, and speaker identification, etc.

3.2. Model-based algorithm

In the model-based algorithm, statistical distribution models are used for each acoustic class (e.g., speak, music background, noise background, etc.) The boundaries between classes are used as the segment boundaries. Typically, Bayesian Information Criterion (BIC) is used to make the decision if the changing point turns out, which is essentially a hypothesis testing problem.

As introduced in (Chen and Gopalakrishnan 1998), the model-based method can be examined in a simplified problem, that assumes only one changing point at time i in the audio stream. The maximum likelihood ratio R(i) is estimated to test the two hypotheses:

$$H_0 : x_1 ... x_N \sim N(\mu, \Sigma) \text{ versus } H_1 : x_1 ... x_i \sim N(\mu_1, \Sigma_1); x_{i+1} ... x_N \sim N(\mu_2, \Sigma_2)$$

where $x_1,\ldots,x_N$ is the sequence of cepstral vectors extracted from the entire audio stream. R(i) is

computed by: $R(i) = N \log|\Sigma| - N_1 \log|\Sigma_1| - N_2 \log|\Sigma_2|$, where $\Sigma, \Sigma_1, \Sigma_2$ are covariance matrices. The changing point i corresponds to the maximum value of R(i). The BIC value can be interpreted as the weighted version of R(i), defined as $BIC(i) = R(i) - [\frac{1}{2}(d + \frac{1}{2}d(d+1))\log N]$, where d is the space dimension. The changing point is estimated at the time i corresponding to the local positive maximum BIC(i). The algorithm of detecting multiple changing points can be developed from this simplified method by an iteration algorithm, as proposed in (Chen and Gopalakrishnan 1998). Furthermore, two improvements aims for realtime application were proposed in (Tritschler and Gopinath 1999) , that the accuracy can be improved by a variable window scheme, and some computations of BIC testing can be reduced in some cases.

3.3. Metric-based algorithm

In the metric-based algorithm, statistical distribution models are also used for modeling the feature space. Gaussian model is a typical choice, but some other distributions can also be used. For example, Chi-squared distribution are found to be appropriate and with less computational cost in (Omar 2005). The sound transition is measured by the distance between the distributions of two adjacent windows. The local maximum of distance value suggests a changing point. Different distance functions can be used here, for example, Mahalonobis distance is used in (Tzanetakis and Cook 1999), which is defined as $D(x,y) = (x-y)^T \Sigma^{-1}(x-y)$, where Σ is the feature covariance matrix. Other distance functions are also reported and tested (Kemp et al. 2000; Huang and Hansen 2004), such as Kullback-Leibler Distance, Gish Distance, Entropy Loss, T^2 Distance, T^2 - mean Distance, etc.

4. Evaluation Metrics

It is not easy to evaluate the performances of different segmentation methods. First, different methods are designed for dealing with different kind of audio signal. Second, the segmentation can be very subjective. People may have different ideas about what is a good segmentation result. However, it is still possible to compare different automatic segmentation results with a given manual segmentation result. Two errors are usually found in the automatic segmentation results, thus they can be used as the criteria to evaluate the segmentation performance. Error type I is missing correct boundaries, error type II is false alarm. in (Kemp et al. 2000), A hybrid measure that combines both error ratios were used in the comparison of three kinds of segmentation methods. The result showed that: 1, model-based and metric-based segmentation algorithms outperform the energy-based segmentation algorithm. 2, Model-based algorithm achieves higher precision while metric-based algorithm achieves higher recall.

5. Conclusion

Hopefully, the basic idea of audio segmentation has been presented in this quick summary. While for a practical segmentation problem, the optimization of features design and the adaptation of segmentation algorithm can be challenging.

References

Brown, J. C. 1999. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features, J. Acoust. Soc. AM.

Chen, S.S., and P.S. Gopalakrishnan. 1998. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion: IBM T.J. Watson Research Center.

Huang, R., and J.H.L. Hansen. 2004. Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval. Paper read at IEEE ICASSP-2004: Inter. Conf. on Acoustics, Speech, and Signal Processing.

Kemp, T., M. Schmidt, M. Westphal, and A. Waibel. 2000. Strategies for automatic segmentation of audio data. Paper read at IEEE International Conference on Acoustics, Speech, and Signal Processing.

Omar, A.H. 2005. Audio Segmentation and Classification, Technical University of Denmark.

Tritschler, A., and R. Gopinath. 1999. Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion: IBM T.J. Watson Research Center.

Tzanetakis, G., and P. Cook. 1999. Multifeature Audio Segmentation for Browsing and Annotation. Paper read at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 17-20, 1999, at New Paltz, New York.

Tzanetakis, G., and P. Cook. 2002. Musical Genre Classification of Audio Signals. Paper read at IEEE Transactions on Speech and Audio Processing July 2002.