Audio Segmentation

Presented by Shi Yong

March. 1, 2007

Music Tech @ McGill University

Outline

- Introduction
  - What
  - Why
  - How
- Approaches
- Example

Introduction

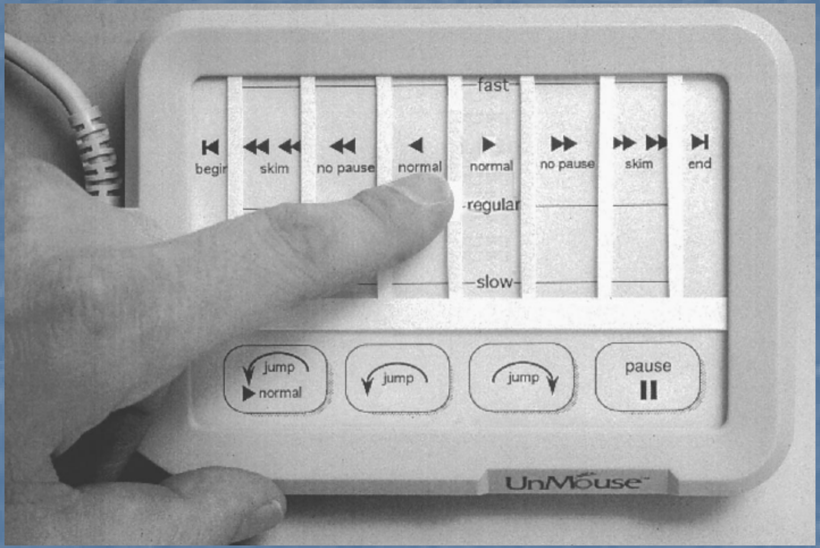
- What is Audio Segmentation?
  - Segmenting the audio stream into homogeneous regions
  - Rule of homogeneity is up to the task, the purpose is to handle regions of different nature differently
    - Music/Noise
    - Speech/Non-speech
    - Male/Female
    - Etc.
  - Often use in conjunction with clustering

Introduction

- Why we need Audio Segmentation?
  - Often used as a pre-processor for further classification of the segments
    - Speaker identification/verification/tracking
    - Automatic speech recognition (ASR)
    - Automatic transcription
    - Segmentation in broadcast news
    - Automatic music analysis, style identification
    - Etc.

Applications

- SpeechSkimmer (Arons97)
  - Allow a user to quickly find what he want to hear
  - Implemented by perceptual segmentation technique and an interactive listener control
- IBM Viavoice (Tritschler99)
  - Real-time broadcast news transcription and speaker identification



SpeechSkimmer (Arons97)

Introduction

- How to do Audio Segmentation?
    - Two steps
        - Features extraction – information need for further processing
            - Temporal domain: ZCR, RMS, etc.
            - Frequency domain: Spectral centroid, Spectral flux, MFCC, LPC, etc.
            - How to find the “best” feature set is an open question.
        - Statistical tools – to find the segment boundaries out
            - GMM, BIC, HMM, etc.
            - What statistical tools shall be chosen? Another open question.
    - Typical methods
        - Energy-based segmentation
        - Model-based segmentation
        - Metric-based segmentation
        - Hybrid methods
        - … maybe more?

Approaches - I

- Energy-based segmentation
  - Detecting silence periods in the audio stream
    - By the location information generated by decoder, such as silences, gender information, etc.
    - By measuring and thresholding the audio energy
  - Segment boundaries are hypothesized in such periods
  - Noise-gate is a very simple example of this approach
- Pros:
  - Easy to implement
    - For commercial products, simple, low-cost, robust are what product developers most concern
- Cons:
  - The boundaries have no direct connection with the acoustic changes
  - E.g., how can we tell a silence period is the pause between the signal of two person or just the pause by one person?
  - E.g., how can we know when a person begin to speak in a continuous music background?

Approaches - II

- Model-based segmentation
  - Modeling: a set of statistical models are defined for each acoustic classes
    - Models: multivariate Gaussian Mixture Model is widely used
    - Classes: speak, music, background noise, silence, telephone speech, etc.
  - Training: model parameters are estimated from the training data
    - For multivariate Gaussian model, the parameters are mean average (mu) and covariance matrix (Sigma).
    - Different solutions have been developed to estimate these parameters: Maximum Likelihood Estimation (MLE), Expectation Maximization (EM), etc.
    - We do not have to dig into all the mathematical details, we can directly use some developed closed-form expression to calculate the parameters
  - Segmentation:
    - Segmentation boundaries are assumed by the boundaries between classes
    - This can be determined by a model selection criterion, such as Bayesian Information Criterion (BIC)
- Pros:
  - Theoretically, acoustic features are connected with the segmentation boundaries
- Cons:
  - Complex (need to use more complex statistical tools)
  - Computational cost (increase the product cost)
  - Due to the statistical nature, the “correct” segmentation is still not guaranteed.

Approaches - III

- Metric-based segmentation
  - Segment boundaries are determined by the contents similarity/distance between two continuing moving adjacent windows
    - We have two neighboring windows (modeled by multivariate Gaussian distributions)
    - Let the two windows move over the audio stream
    - Compute the similarity of the contents of the two windows
    - Segment boundaries are determined by the local maxima and a predefined threshold
  - Algorithms to compute the similarity are called “distance function”
    - Kullback-Leibler Distance
    - Gish Distance
    - Entropy Loss
    - T^2 Distance
    - T^2 - mean Distance
    - Etc.
  - Thing to be considered for designing the metric-based algorithm:
    - Selection of distance function
    - Window size
    - Windows moving speed (time increment)
    - Threshold
    - Etc.
  - Pros and Cons:
    - Like approach II, with a little difference

Approaches - III

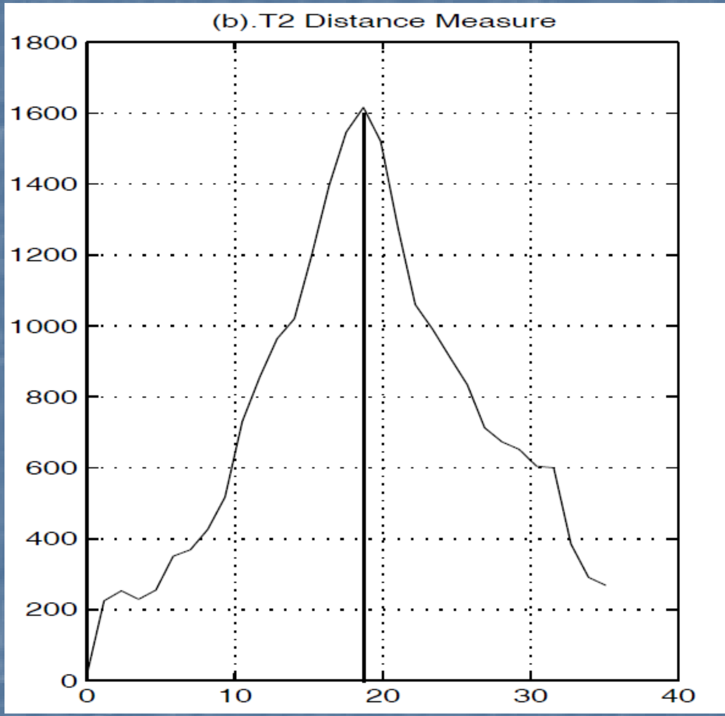
- A glance at T^2 distance
  - Two audio segments modeled by multivariate Gaussian distributions:

$$N(\mu_1, \Sigma_1) \text{ and } N(\mu_2, \Sigma_2)$$

  - T^2 distance is:

$$T^2 = \frac{ab}{a+b}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

a, b are frames numbers within each segments



Huang04

Evaluation Metrics

- How to evaluate the performance of different methods/ models/feature set?
  - Strictly speaking, there is no objective stardard for evaluating the errors in different segmentation methods, because segmentation is very subjective
  - However, by compare the automatic segmentation results with the manual segmentation, we can have some criteria
- Evaluation Criteria (Kemp00)
  - Type I errors (deletion):
    - RCL = number of correctly found boundaries / total number of boundaries
  - Type II erors (false alarm):
    - PRC = number of correctly found boundaries / number of hypothesized boundaries
  - Hybrid measure (combine two number into one)
    - F = (2*PRC*RCL)/(PRC+RCL)
- Now we can have a basic idea of the performance of each method (Kemp00)
  - Energy-based: F = 0.58
  - Model-based: F = 0.62
  - Metric-based (Gish-distance): F = 0.70

Example

Task: detecting the speaker changes in a continuous audio stream (e.g., in a teleconference). Let's try the model-based method.

- First we extract the sequence of feature vectors x (say, ceptral coefficients, $x_i = x_1, x_2, \ldots, x_N$) from the entire audio stream, and assume they are modeled by multivariate Gaussian distribution, denoted as $x_i \sim N(\mu_i, \Sigma_i)$
- Let's begin with the simplest problem: assume only one changing point in the stream, so what is more likely to happen: x as one Gaussian distribution, or x be divided into two part and as two Gaussian distribution?
- Mathematically speaking, we get to testing the two hypothesis:

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

- The changing point is estimated at index i that corresponding to the maximum likelihood ratio R(i)

$$R(i) = N log|\Sigma| - N_1 log|\Sigma_1| - N_2 log|\Sigma_2|$$

Chen98

Using BIC

- Alternately, we can use Bayesian Information Criterion (BIC) value to make our decision: the data is modeled as one Gaussian or two Gaussians?

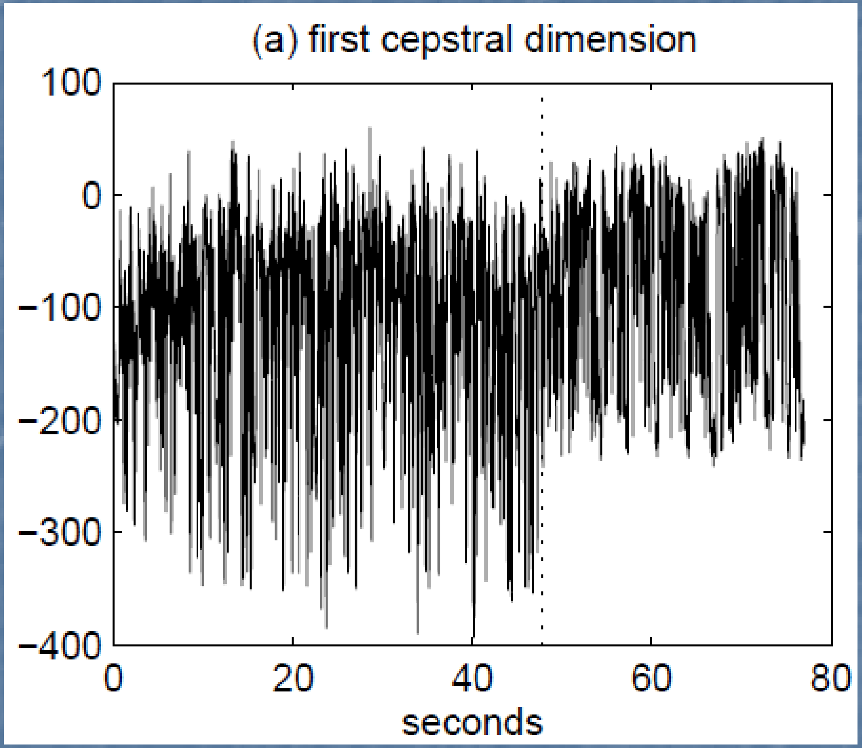
$$BIC(i) = R(i) - \lambda P$$

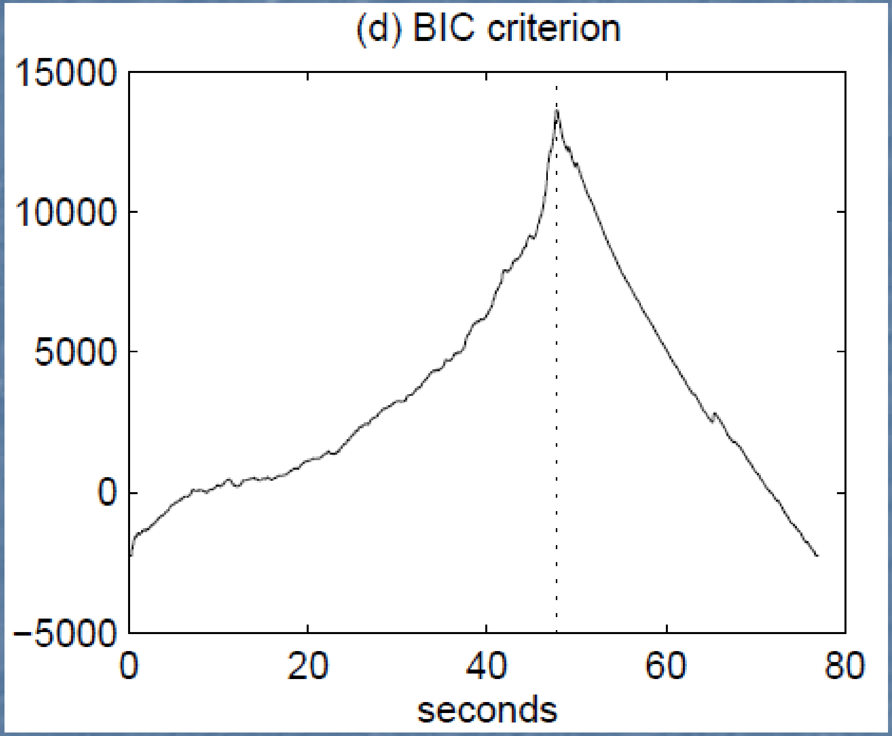
$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N$$

- The segment boundary is decided at the point corresponding to the positive maximum BIC value

Chen98

Depiction





Chen98

Multiple Changing Points

- Multiple changing points detection algorithm is based on the aforementioned method

(1) initialize the interval $[a, b]$: $a = 1; b = 2$.
(2) detect if there is one changing point in $[a, b]$ via BIC.
(3) if (no change in $[a, b]$)
    let $b = b + 1$;
  else
    let $\hat{t}$ be the changing point detected;
    set $a = \hat{t} + 1$; $b = a + 1$;
  end
(4) go to (2).

Reference

- [Arons97] SpeechSkimmer: a system for interactively skimming recorded speech. ACM Transactions on Computer-Human Interaction (TOCHI), ACM Press New York, NY, USA.
- [Chen98] Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, IBM T.J. Watson Research Center: 127-32.
- [Huang04] Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval. IEEE ICASSP-2004: Inter. Conf. on Acoustics, Speech, and Signal Processing.
- [Kemp00] Strategies for automatic segmentation of audio data. IEEE International Conference on Acoustics, Speech, and Signal Processing.
- [Tritschler99] Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion, IBM T.J. Watson Research Center.