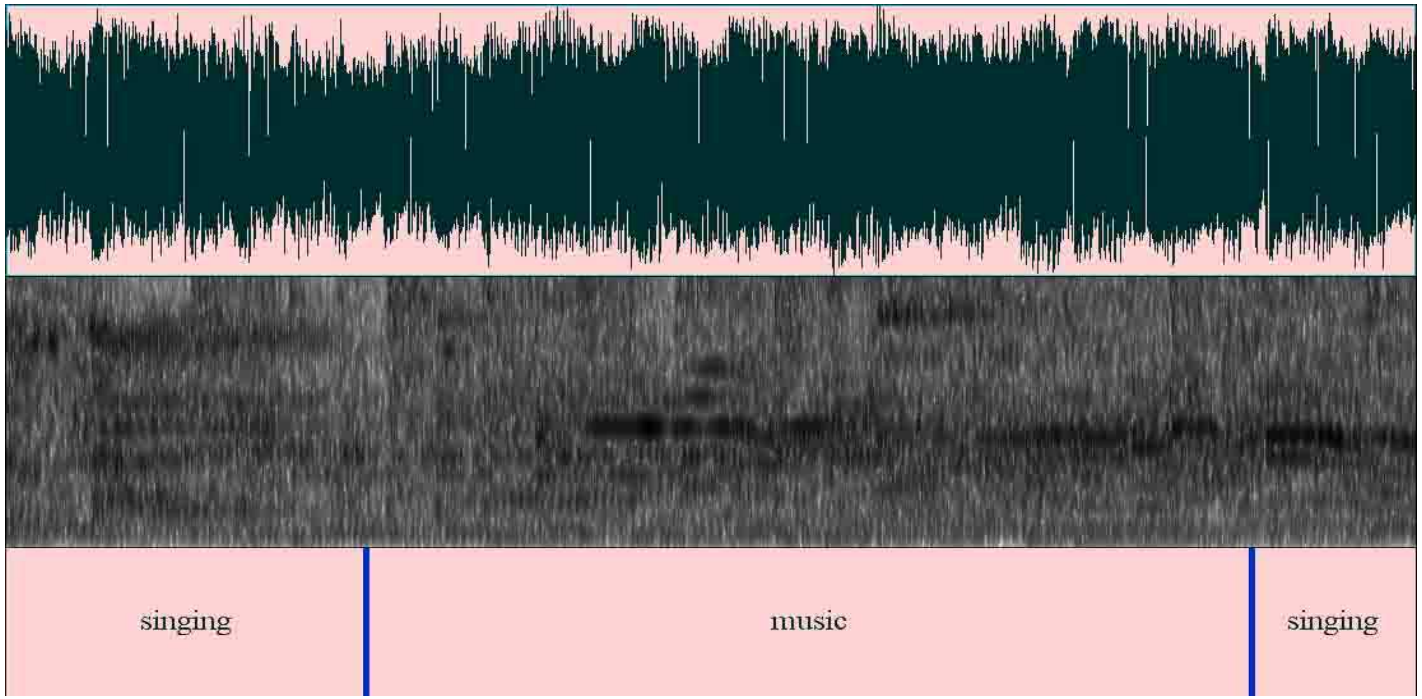
Identifying singing segments in music

Felix Sanchez Garcia

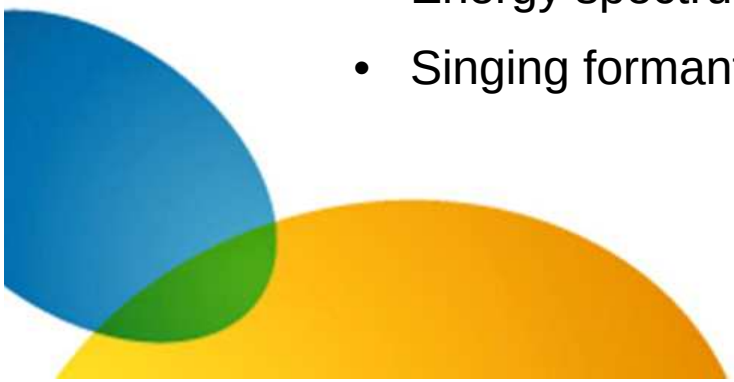
Objective

- Given a music sample, identify singing segments



Difficulties

- It is hard to model singing voice, it differs from normal speech
  - Mixed with aloud, non-stationary background music signal
  - Time-frequency features of singing voice are quite different from those in speaking voice:
    - Voiced
    - Harmonicity
    - Lack of modulation peak at 4Hz syllabic rate
    - Energy spectrum
    - Singing formant



Motivation

- Improve efficiency of artist recognition
  - 57% all segments, 65% voice ,36% no voice (Berenzweig , 2002)
- Song structure and segmentation
- First step for analyzing singing voice
- Ultimate goal : automatic indexing of music data to meet the demand for content-based information

General approach

- Split song into frames
- Extract features for each frame
- Classify each frame
- Smooth using adjacent frames

Features

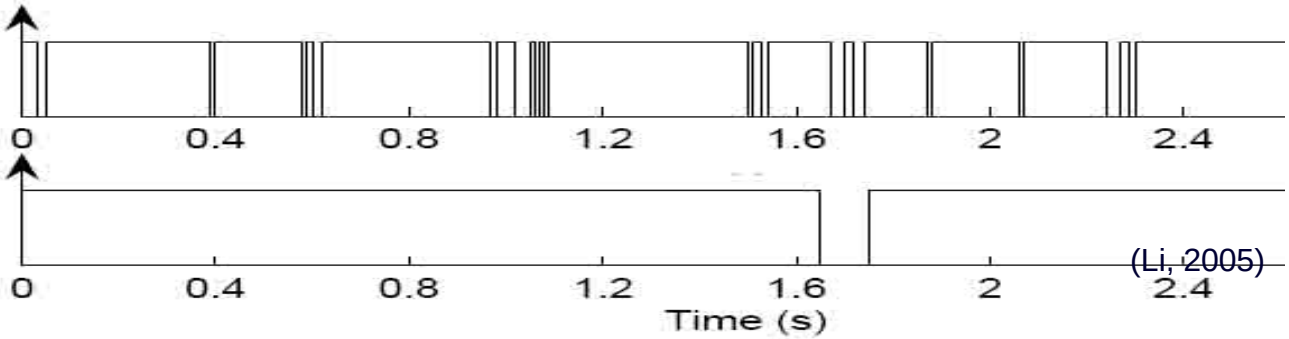
- Harmonicity (Chou, 2001)
- Energy descriptors (Zhang, 2002)
- Timbral descriptors
  - MFCCs (Berenzweig, 2001)
  - LPC (Kim, 2002)
  - LFPC (New, 2004)
  - OSCC (Maddage, 2006)

Classifiers

- Artificial Neural Networks (Berenzweig, 2001)
- Gaussian Mixture Models (Wang, 2006)
- Support Vector Machines (Maddage, 2003)

Smoothing

- Frame-based classification is noisy
- Real data has a important number of singing/non-singing frames in a row.
- Solution: divide in segments and apply class to the whole segment. Several methods to create segments (HMM, clustering, constant sub-segment length,…)



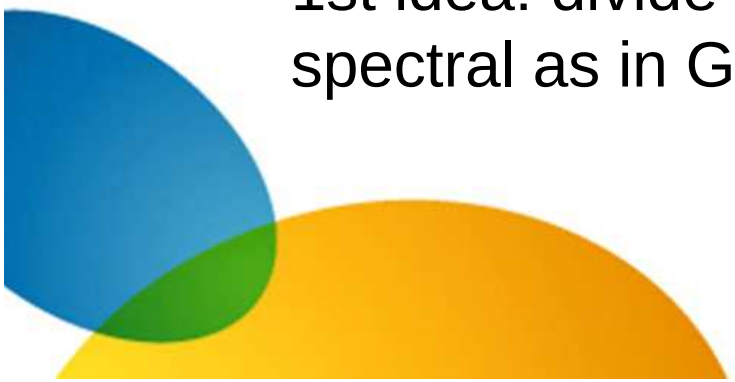
(Li, 2005)

Our system

- Features: MFCCs
- Classifier: GMM
- Smoothing: introduce new contributions

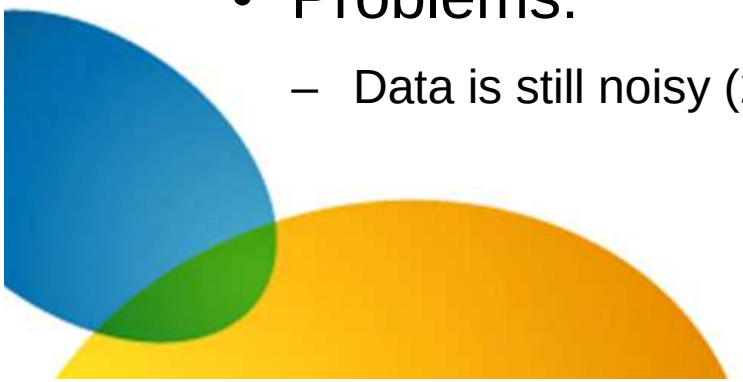
Direction 1

- Smoothing is usually performed without considering correlation between adjacent frames, only their voice/instrumental probabilities.
- Identify features that change on boundaries between segments.
- Candidate segments can use the aggregation of the low level classification to obtain a label
- 1st idea: divide signal in segments with similar spectral as in Gold, 2006



Direction 2

- Smoothing improves classification from 70.3% to 80.4%, that's 15%!
- We might be able to used this new information to update our classifier and perform a 2nd iteration
- Advantages:
  - Model becomes more specific to singer's voice
  - Improves by learning from mislabeled data
- Problems:
  - Data is still noisy (20% misclassified). Maybe use subset?



Dataset

- Music/speech corpus extracted by Eric Scheirer:
  - extracts of 15 second length from the radio
  - Mono, 22Khz
  - 80 extracts in total: 60 from training, 20 for testing
  - Contains time-aligned labels to distinguish between singing and musical accompaniments

Roadmap

- Current state :
  - Base framework :
    - Extraction of MFCCs
    - GMM classifier
- Next steps:
  - Implement smoothing
  - Analyze causes for misclassified frames

Bibliography

- Berenzweig, A. L. and D. Ellis (2001). Locating singing voice segments within music signals. Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the
- Berenzweig, A., D. Ellis, et al. (2002). Using Voice Segments to Improve Artist Classification of Music. {AES} 22nd International Conference. Espoo, Finland.
- Gold, K. and B. Scassellati (2006). Audio Speech Segmentation Without Language-Specific Knowledge. Proceedings of the 28th Annual Meeting of the Cognitive Science Society, Vancouver, Canada.
- Maddage, N. C., C. Xu, et al. (2003). "A SVM-Based Classification Approach to Musical Audio."
- Maddage, N. C., W. Kongwah, et al. (2004). Singing voice detection using twice-iterated composite Fourier transform. Multimedia and Expo, 2004. ICME '04.
- New, T. L., Shenoy, A., and Wang, Y. (2004). Singing voice detection in popular music. Technical report, Department of Computer Science, University of Singapore, Singapore, October 2004.
- Rocamora, M. and P. Herrera (2007). Comparing audio descriptors for singing voice detection in music audio files. Proceedings of 11th Brazilian Symposium on Computer Music, Sao Paulo, Brazil.
- Tin Lay, N., S. Arun, et al. (2004). Singing voice detection in popular music. Proceedings of the 12th annual ACM international conference on Multimedia. New York, NY, USA, ACM.
- Wei-Ho, T. and W. Hsin-Min (2006). "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals." Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on] 14(1): 330-341.
- Zhang, T. (2002). System and method for automatic singer identification. HP Labs Technical Report.