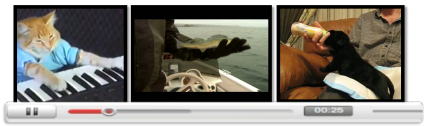
"SEGMENT AND CONQUER"

Benjamin Elizalde[1,2,3], Bhiksha Raj[1], Gerald Friedland[3], Juan Nolazco[2], Leibny Garcia[2]

Problem

Proper **audio** segmentation in consumer-produced aka "wild" videos.

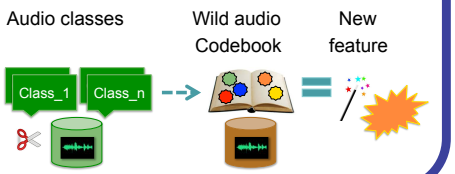


Challenge

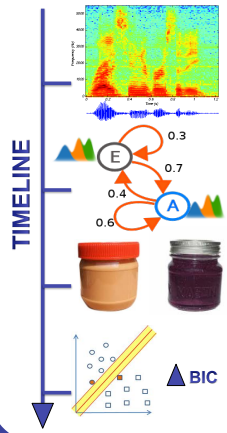
✗ Cannot assume any characteristics in order to draw segment boundaries.

✗ Difficult to pre-train models because of the high variance in the data.

Our approach



Related Work



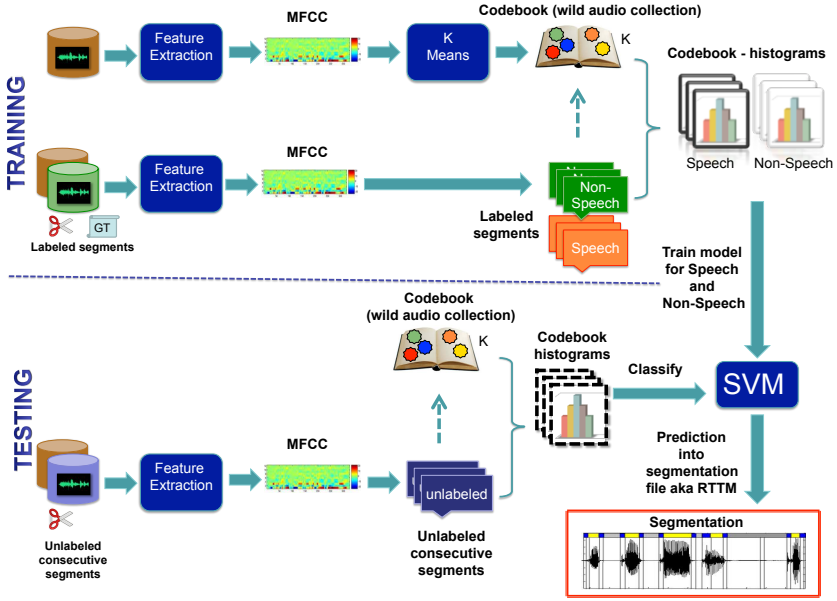
- Thresholds on energy
- Energy based models
- (speech, silence) [1]

- Hidden Markov Models
- Gaussian Mixture Models
- (sounds, speakers)[2]

- Entropy, dynamism
- (music)[3]

- Support Vector Machines[4]
- Bayesian Information Criterion [5]
- (drawing boundaries)

Technical diagram



Results

Speech / Non Speech

Train Test	Clean Clean	Clean Wild	Wild Clean	Wild Wild
SHOUT	15.2%	24.1%	21.6%	28.3%
Our approach	10.3%	36.2%	10.8%	34.6%

Speech Activity Detection Error

Processing the histograms is 5x faster than Multidimensional MFCC's

SHOUT is a "State of the art" speech activity detection system. [6]

Experimental Setup

- TRECVID MED 2011. Codebook: 3 hrs. Segments: Train12 hrs., Test 5 hrs.
- ICSI Meetings Corpus. Train 35 hrs. 1 sec length labeled segments
- ICSI Meetings Corpus. Test 15 hrs. 1 sec length unlabeled consecutive seg.
- MFCC Mel Frequency Cepstral Coefficients 30ms frame, 10ms frame rate, 58 dimensions: 19+D+DD
- K-means output, Dimension:K by 58, 10 iterations
- 2 sets of 1 sec length labeled segments, Meetings: 103k ->Speech, 14k -> N-Speech MED11: 30k->Speech, 30k->N-Speech
- SVM: RBF Radial Basis Function kernel
- Set of histograms. Result of relating each segment's MFCC frame to its closest codebook K-value, Dimension: K by 1 (Occurrences)
- unlabeled: 1 set per test audio file of 1 sec length unlabeled consecutive segments

Conclusions

The codebook approach is promising for segmentation but it needs improvement for wild videos.

The technique is improving the error rate in comparison to the state of the art and is 5x faster.

Future Work

✔ Use GMM models instead of histograms.

✔ Extend algorithm to a multiclass music/speech/non-speech segmentation system.

✔ Smaller size and sliding segments for test.

✔ Try a bigger size codebook.

[1]Carnegie Mellon University
[2]Instituto Tecnológico de Monterrey
[3]The International Computer Science Institute

Literature cited

[1] Marijn Huijbregts, Roeland Ordelman, and Arjan van Hessen. Prosody based boundary detection. 2004.
[2] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and .S.J. Young. Segment generation and clustering in the htk broadcast news transcription system. 1998.
[3] Jitendra Ajmera, Iain McCowan, and Herve Bourlard. Speech and music segmentation using entropy and dynamism features in a hmm classification framework. pages 351–363, 2003.
[4] Mathieu Ramona and Gael Richard. Comparison of different strategies for a svm- based audio segmentation. August 2009.
[5] Steve Cassidy. The macquarie speaker diarisation system for RT04S. proceedings of the NIST RT04s Evaluation Workshop, Montreal, Canada, May 2004.
[6] Marijn Huijbregts. Segmentation Diarization and Speech Transcription: Surprise Data Unraveled. PhD thesis, Universiteit Twente, 2008.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.