

Towards Parameter-Free Classification of Sound Effects in Movies

Selina Chu[†], Shrikanth Narayanan^{†*}, and C.-C. Jay Kuo^{*}

University of Southern California
[†]Department of Computer Science
^{*}Department of Electrical Engineering

August 04, 2005

Outline of Talk

- What is Video Mining?
- Classification of Sound Effects
- Data Representation / Feature Extraction
- Details of Framework
- Experimental Setup and Results
- Conclusions and Future Work

Event Detection/ Video Mining

Pattern Discovery: Detecting intense scenes / highlights in movies

- Preview movies
- Surveillance video

Clustering Multimedia: Group movies together with similar characteristics or styles

Classification: Classify “un-typed” movies based on their characteristics

Classification of Sound Effects

Experiment Performed on:

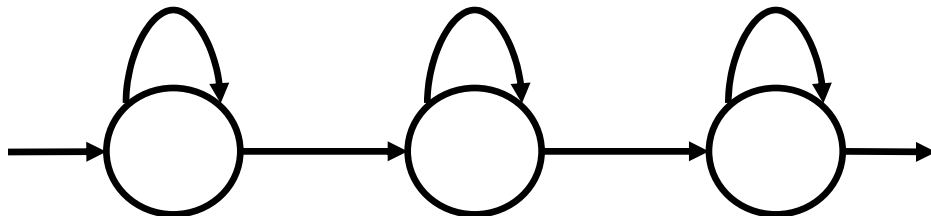
- 4 classes: explosion, glass-shattering, gunshots, and screaming
- Without preprocessing or dimensionality reduction
- Sounds from *action type* movies
- Data consists of 100 examples from each class
- 80% training, 20% testing

Techniques	Accuracy %
Hidden Markov Model (HMM)	87.14
Gaussian Mixture Model (GMM)	70.90
10-Nearest Neighbors (KNN)	83.34
Naïve Bayes	74.52

A Closer Look at HMM

Experiment Performed on:

- Same classes and specifications as before
- Used 4 mixtures per state
- Previously analyzed 2-5 number of states, 2-5, 10 mixtures per state, and different transitions
($4*5*5=100$ different settings!!)



N-State Forward: Each state has transitions to itself and next successor state

N-State Ergodic: Every state can be reached from any other state

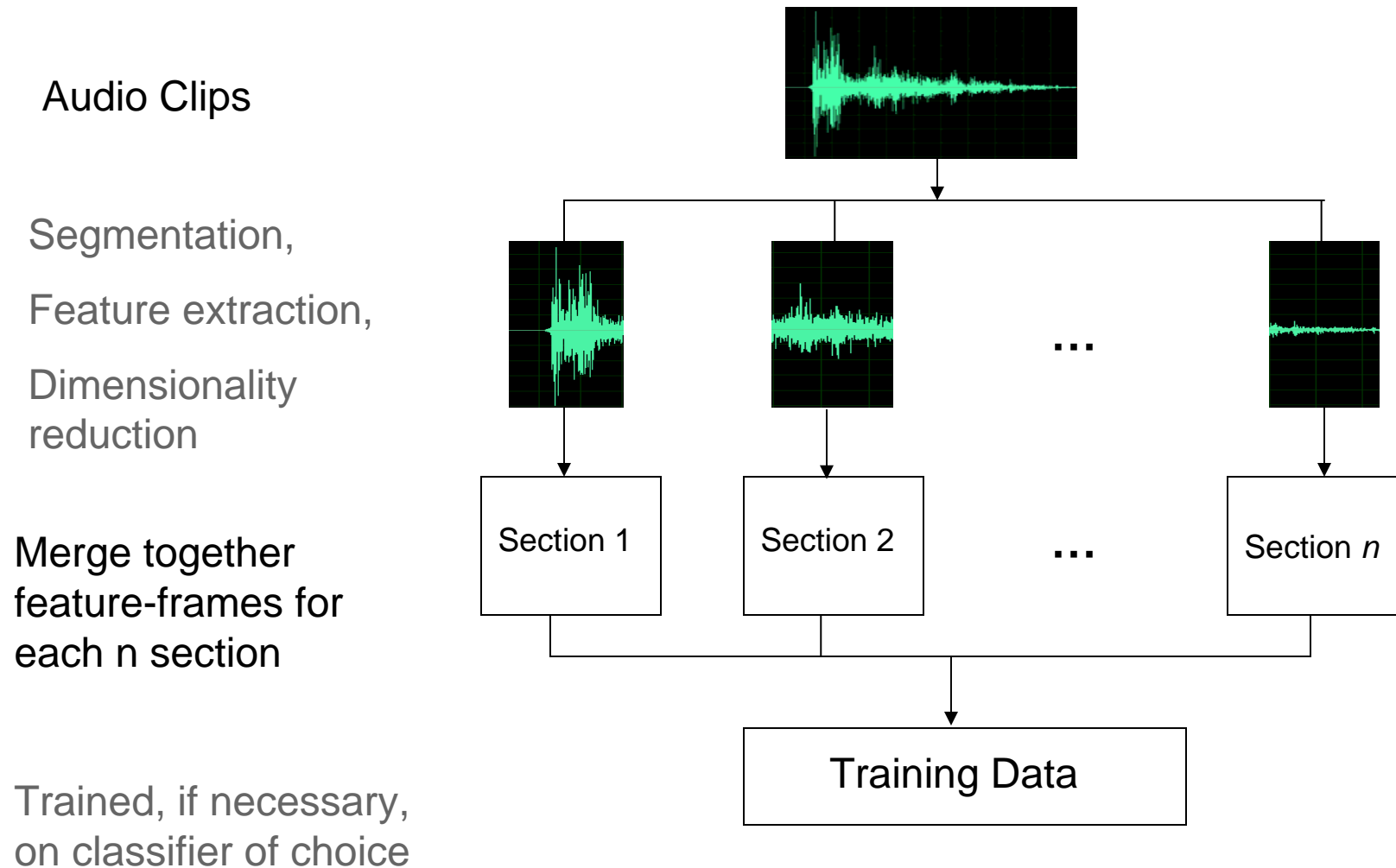
Types of HMM	Accuracy %
3-State Forward	87.14
3-State Ergodic	90.70

- HMM outperforms other classifiers for sound effects.
- HMM challenges
 - fine-tune high number of parameters
 - requires an expert to design and optimize model
 - otherwise, requires many trials



Alternative approach
with less parameter
tweaking??

Framework of Classifier



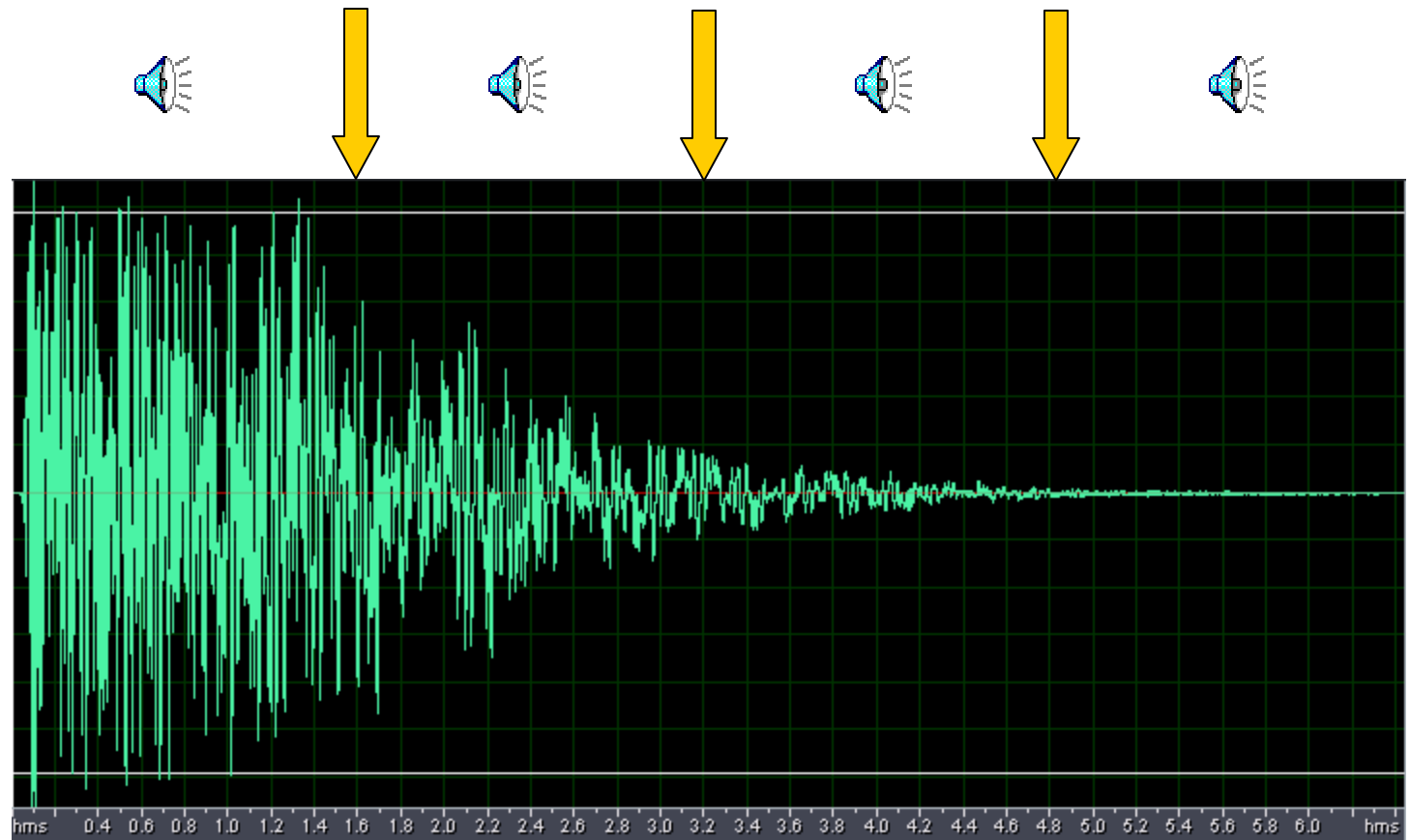
Segmentation

Divide audio clip into n segments based on:

- 1) **Equal-length**: each segment contains same amount of time
- 2) **Equal-energy**: each segment contains same amount of energy (time varies)

Segmentation

Equal-Length
Segmentation



Equal-Energy
Segmentation



Pre-Processing the Data

1. For each audio data sequence,

 Obtain n segments

 For each segment

 Extract MFCC features for each sampling frame (*feature-frame*)

 End

End

2. Combine all *feature-frames* of each segment together

3. Reduce dimensionality of data

This is your training data

Classifying a New Instance

1. Classifying a new instance of sound effect

Obtain n segments

For each segment

Extract MFCC features for each sampling frame (*feature-frame*)

end

2. Combine together all feature-frames of each segment
3. Reduce dimensionality of data
4. Classify each feature-frame individually
5. Use majority voting to determine section classification
6. Each section votes to determine classification of query
(Ties are broken by sections with more frames)

Dataset and Feature Extraction

Dataset: short audio clips of action-type sound effects

- Audio streams are 16 bits, mono-channel, and downsampled to 16 KHz
- Each audio clip is 3-10 seconds
- 100 samples for each class

4 Classes: explosions, glass-shattering, gunshots, and screaming

Feature Extraction: features are analyzed and extracted for every 20ms frame

- Mel-Frequency Cepstral Coefficients (MFCC) to the 40th order per frame

Dimensionality Reduction

- 40 MFCC features are extracted for every 20ms frame
- Each audio clip ranges from 3-10 seconds
 - 150-500 feature-frames per audio sequence
- Use Principal component analysis (PCA)
 - Eliminate redundancy between dimensions based on correlation. Collapse correlated dimensions, leaving uncorrelated ones intact.

	40 dims	12 dims
10-Nearest Neighbors	16.42 min	6.42 min
GMM	10.52 min	4.8 min
Naïve Bayes	0.1 min	0.1 min

Experimental Comparison

- Compared classification rate using our framework along with:
 - GMM, 10-Nearest Neighbors, and Naïve Bayes
- Observe different settings:
 - Variation in dividing into N equal-length segments
 - Equal-length versus Equal-energy as segmentation criteria
 - Effects of PCA on K dimension
- 80% Training, 20% Test
- Measured Classification Accuracy

Classification using Proposed Framework

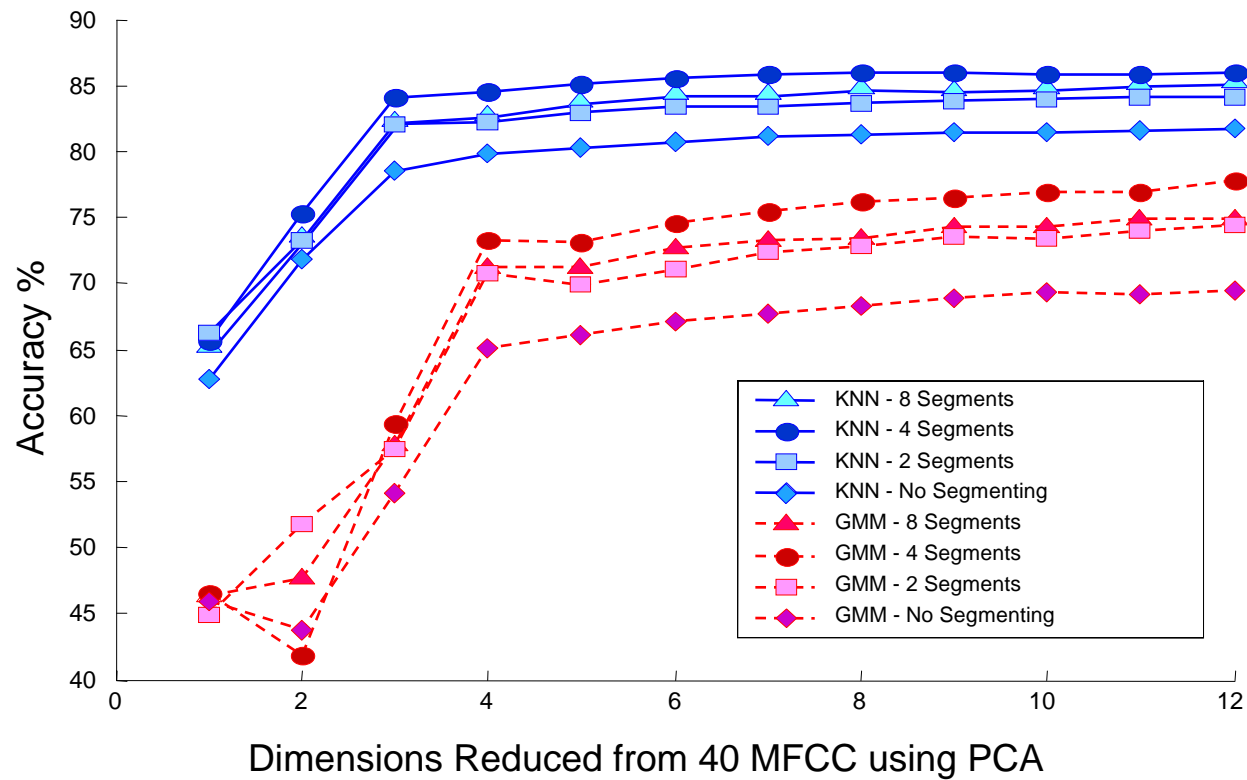
Experiment Performed on:

- 4 classes: explosion, glass-shattering, gunshots, and screaming
- With preprocessing and dimensionality reduction (with exception for HMM)
- Sounds from *action type* movies
- Data consists of 100 examples from each class
- 80% training, 20% testing

	Accuracy % - without preprocessing	Accuracy %
HMM	87.14	N/A
GMM	70.91	84.3
10-Nearest Neighbors	83.34	90.8
Naïve Bayes	74.52	91.3

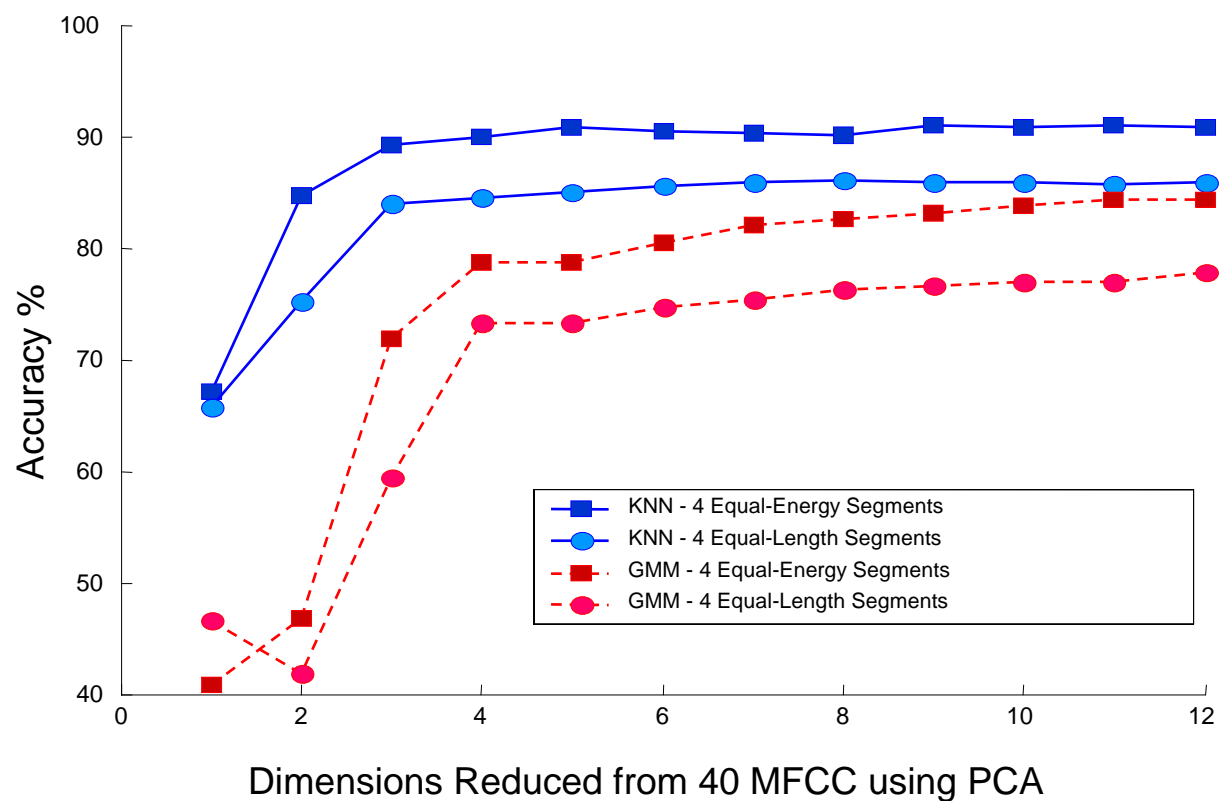
Experimental Results

Accuracy with N Equal-Length Segments



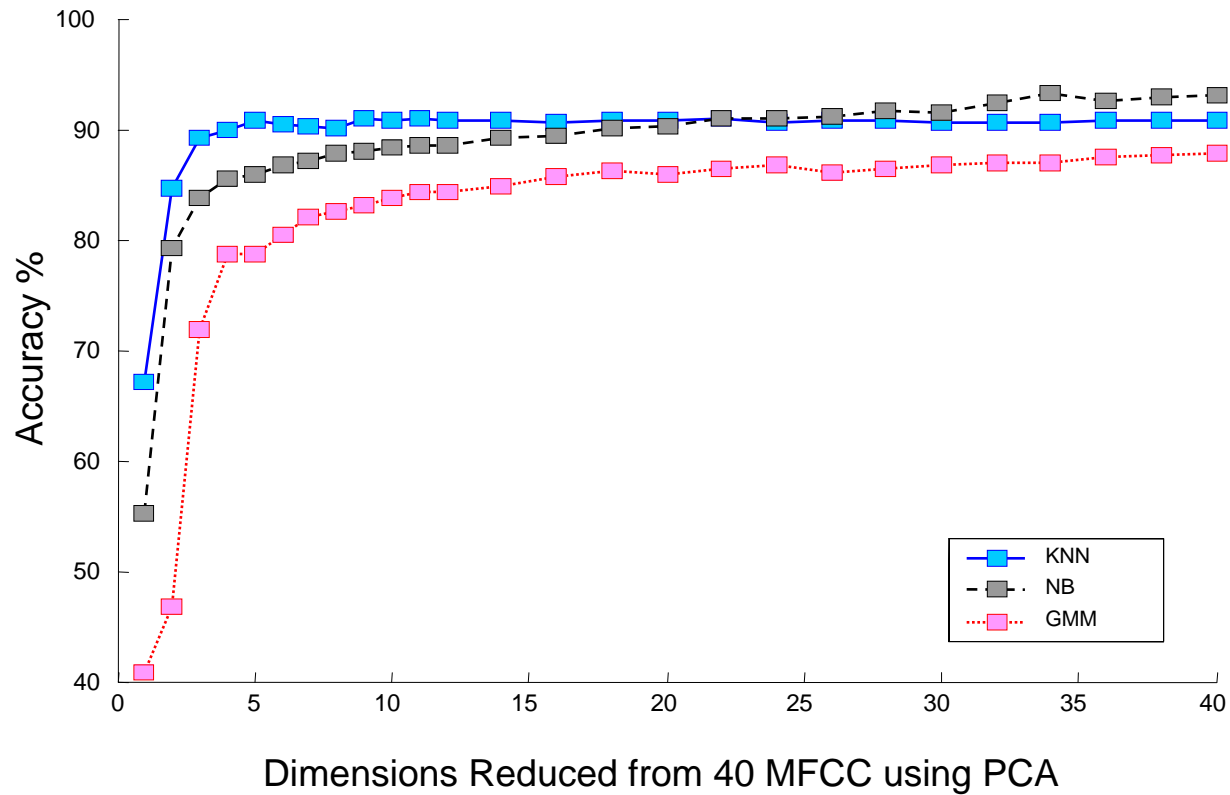
Experimental Results

Accuracy using Different Criteria for Segmentation



Experimental Results

Accuracy with Different Classifiers



Conclusion

- Novel representation of the data for efficient classification of sound effects
 - Segment each audio data sequence by using energy as an indicator
- Introduce a framework using semi- or non-parametric classifiers, along with dimensionality reduction for mining of sound effects
- Perform better than HMM and obviate the need for an expert to design or fine-tune to achieve optimal model

Future work

- Automation of the hierarchical segmentation of data
- Make sections more adaptive using the notion of relevance feedback and weights
- Classify increasing types of sound effects

References

- [1] S. Pfeiffer, S. Fischer and W. Effelsberg, “Automatic audio content analysis,” *Praktische Informatik IV*, Univ. Mannheim, Mannheim, Germany, 1996.
- [2] R. Lienhart, S. Pfeiffer, S. Fischer, “Automatic movie abstracting”, Technical Report TR-97-003, *Praktische Informatik IV*, University of Mannheim, July, 1997
- [3] R. Radhakrishnan, Z. Xiong, A. Divakaran, and Y. Ishikawa, “Generation of sports highlights using a combination of supervised & unsupervised learning in audio domain”, *International Conference on Pacific Rim Conference on Multimedia*, Vol. 2, pp. 935-939, December 2003
- [4] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp.257-287, 1989
- [5] K.-S. Goh, K. Miyahara, R. Radhakrishnan, Z. Xiong, and A. Divakaran, “Audio-visual event detection based on mining of semantic audio-visual labels”, *SPIE Conference on Storage and Retrieval for Multimedia Databases*, Vol. 5307, pp. 292-299, January 2004
- [6] R. Radhakrishnan, A. Divakaran, and Z. Xiong, “A time series clustering based framework for multimedia mining and summarization using audio features”, *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 157-164, October 2004
- [7] T. Zhang and C.-C. Kuo, “Audio content analysis for on-line audiovisual data segmentation”, *IEEE Trans. On Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001
- [8] Divakaran, A., Miyahara, K., Peker, K.A., Radhakrishnan, R., Xiong, Z., “Video mining using combinations of unsupervised and supervised learning techniques,” *SPIE Conference on Storage and Retrieval for Multimedia Databases*, Vol. 5307, pp. 235-243, January 2004
- [9] T. Zhang and C.-C. Kuo. *Content-based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer Publishing Company, 2001
- [10] D. Reynolds and A. Rose. “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. On Speech and Audio Processing*, 3(1):72-82, 1995.
- [11] M. Casey, "Reduced-Rank Spectra and Minimum Entropy Priors for Generalized Sound Recognition", *Proceedings of the Workshop on Consistent and Reliable Cues for Sound Analysis*, EUROSPEECH 2001, Aalborg, Denmark, September 2001.
- [12] E. Scheirer and M. Slaney. “Construction and evaluation of a robust multi-feature speech/music discriminator,” *In Proc. IEEE ICASSP*, Munich, Germany, April 1997.
- [13] M. Rajapakse and L. Wyse. “Generic audio classification using a hybrid model based on GMMs and HMMs,” *IEEE 11th International Conference on Multimedia Modelling*, 2005.
- [14] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 2003
- [15] T. M. Mitchell, *Machine Learning*, Mc Graw-Hill, 1997
- [16] L. I. Smith, “A Tutorial on Principal Components Analysis”, Maintained by Cornell University, 2002.
- [17] A. Moore, “Statistical Data Mining Tutorial on Gaussian Mixture Models”, www.cs.cmu.edu/~awm/tutorials, CMU, 2004.