Keyword Spotting in Audio to Support Video Lecture Indexing

Manish Kanadje

Advisor : Prof Richard Zanibbi

Department of Computer Science, Rochester Institute of Technology

Introduction

- Create a keyword spotting system to locate candidate regions for a query in the lecture audio
  - lectures recorded in a single channel and a single speaker environment
  - query keywords extracted from lectures or recorded on the laptop
- Create indexing tools to facilitate lecture indexing using keyword spotting
  - tools for accessing hits and creating hierarchical annotations

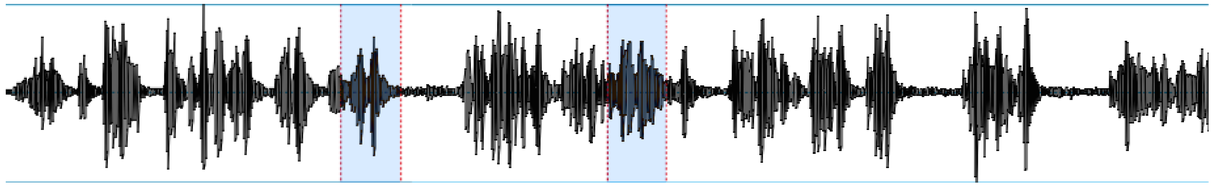

Figure 1: Candidate regions in the lecture audio for a query 'echelon form of a matrix'

Methodology

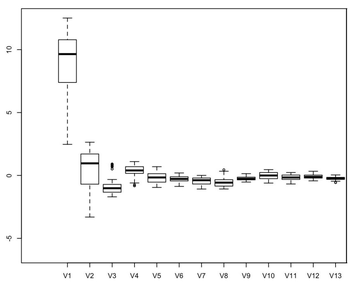

Figure 4: Box plot for raw MFCC features

- Mel Frequency Cepstral Coefficients [2]
- Normalization [1]

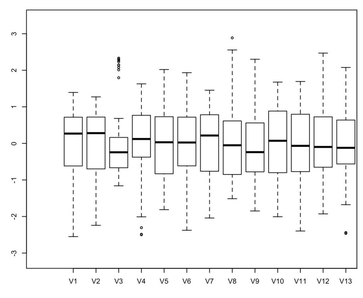

Figure 5: Box plot for 'whitened' MFCC features

- Dynamic Time Warping
- Segmental Dynamic Time Warping [3]

Results

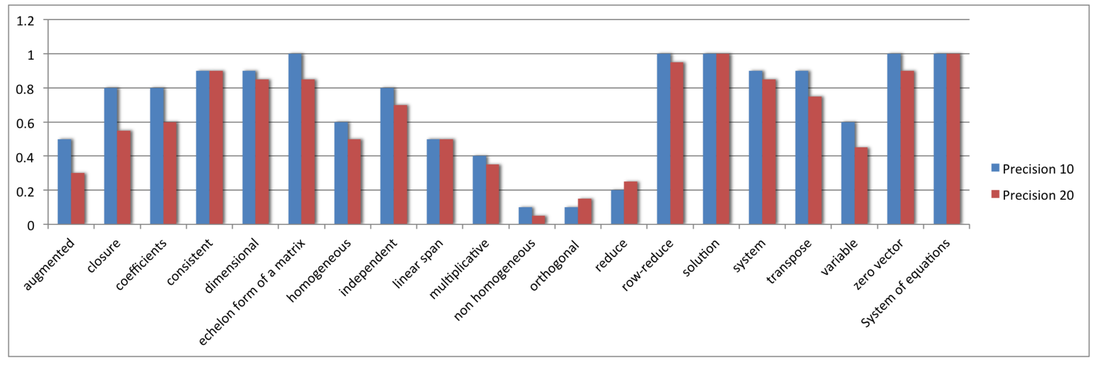

Figure 2: Precision@10 and Precision@20 for 20 queries recorded on the laptop

- Average precision@10 of 70% for 20 queries recorded on the laptop
- Average precision@10 of 76% for 20 queries extracted from lecture audios

Conclusion

- Environmental mismatch affects the performance of the system heavily
- 'Whitening' process reduces the impact of higher values of lower frequencies in laptop recorded queries
- Error reduction of 55.9% over the previous results for laptop recorded queries
- Performance is better when the query has a distinctive pronunciation i.e. 'reduce' vs 'row-reduced'
- Indexing tools facilitate the lecture annotation and searching process

Indexing Tools

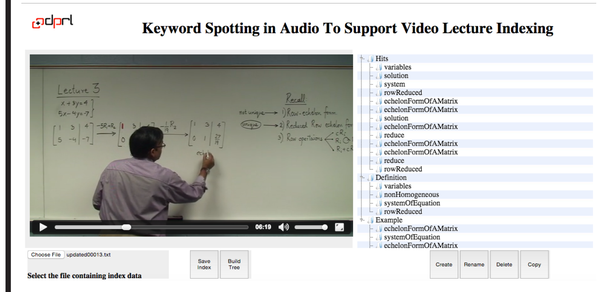

Figure 3: Interface for prototype of indexing tools

- Accessing generated hits
- Hierarchical organization of generated hits

References

[1] M. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy. Comparative evaluation of feature normalization techniques for speaker verification. In *Advances in Nonlinear Speech Processing*, volume 7015 of *Lecture Notes in Computer Science*, pages 246–253. Springer Berlin Heidelberg, 2011.

[2] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transcations on Acoustics, Speech and Signal Processing*, 28(4):357–366, Aug 1980.

[3] A. Park and J. Glass. Towards unsupervised pattern discovery in speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, pages 53–58, Nov 2005.

Future Work

- Include the functionality for performing search using system generated hits
- Link the created index to a database system to create a persistent index
- Create new queries by cropping lecture audio for in-lecture search

Acknowledgement

This work is based upon the work supported by the National Science Foundation (USA) under the Grant No. HCC-1218801

Contact Information

Web www.cs.rit.edu/~dprl

Email mk2852@rit.edu