Keyword Spotting in Audio for AccessMath

{ Zachary Miller, Roger Gaborski, Richard Zanibbi }
Computer Science Department, Rochester Institute of Technology

Introduction

- Implement a Keyword Spotting System.
- Purpose: To allow a Math Professor to easily search through his or her lecture recordings.
- Goal: To implement a number of state-of-the-art Dynamic Time Warping based techniques

Figure 1: Visualization of Keyword Spotting Problem:
Green Boxes are returned hits from the spotter, Red Boxes are where the keyword occurs.

Materials & Methods

- Mel Frequency Cepstral Coefficients [1]
- Gaussian Mixture Models
- Dynamic Time Warping
  - Non-Linearly matches 2 sequences
- Segmental Dynamic Time Warping [2]
  - Allows small queries to be matched to portions of the longer sequence
- Lower Bound Estimate IP [3]
  - Speeds up Segmental DTW through an estimation technique

Experiments

Experiments were conducted using the following framework:

- 2 Lectures for training
- 6 Lectures for testing
- 4 Keywords
  - *Solution*
  - *System*
  - *System of Equations*
  - *Variable*
- Used Precision@N to evaluate results

Results

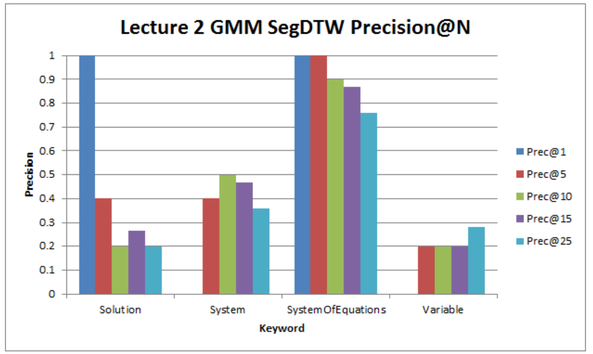

Figure 2: Precision at N of Segmental DTW searching for 4 keywords using Posteriorgram Features

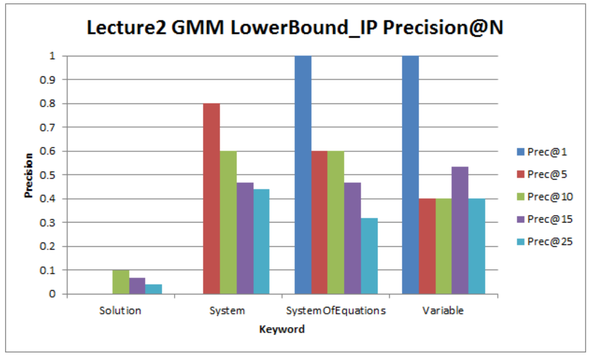

Figure 3: Precision at N of Lower Bound IP searching for 4 keywords using Posteriorgram Features

Conclusion

- Both techniques perform well
- Segmental DTW generally has better performance
- Lower Bound IP generally is faster
- Longer queries tend to perform better than shorter ones as both *System Of Equation* and *Variable* returned better overall results.
- Common false hits include
  - Words having similar structure
  - Words containing similar sounding syllables
  - Empty noise due to no preprocessing of the audio
- Performance is heavily based on the Gaussian Mixture Model Trained

References

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

[2] Yaodong Zhang and J.R. Glass. Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams. In *Automatic Speech Recognition Understanding, IEEE Workshop on*, pages 398–403, Nov 2009.

[3] Yaodong Zhang and J.R. Glass. An inner-product lower-bound estimate for dynamic time warping. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5660–5663, May 2011.

Future Research

- Noise Reduction on input audio.
- Try More Keywords.
- Try More Lectures.

Acknowledgments

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

Contact Information

Web cs.rit.edu/~dprl
Email zrm6085@rit.edu