Temasek Laboratories and
School of Computer Engineering

Institute for Infocomm Research
A*STAR
Human Language
Technology Department

Keyword search using query expansion for graph-based rescoring of hypothesized detections

Authors: Van Tung Pham[1,2], Haihua Xua[2], Xiong Xiao[2], Nancy F. Chen[3], Eng Siong Chng[1,2], Haizhou Li[1,2,3]

[1]School of computer Engineering, Nanyang Technological University, Singapore

[2]Temasek Laboratories, Nanyang Technological University, Singapore

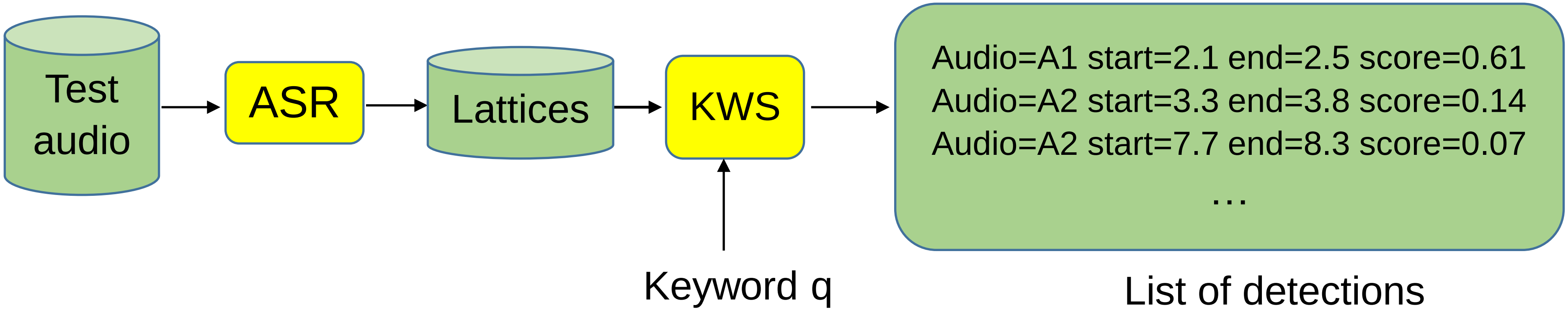
[3]Institute for Infocomm Research, Singapore

Temasek Laboratories and
School of Computer Engineering

Human Language
Technology Department

Introduction

- This work belongs to **Keyword Search (KWS)** - the task of finding all occurrences of a text keyword in a speech corpus



- Detection scores are estimated from a standard model-based, parametric Automatic Speech Recognition (ASR)
- In this work we proposed a novel framework to **rescore** the list of detections using **keyword examples** extracted from **training data**

Introduction (cont.)

- Main idea: if a detection is **acoustically** more similar to the keyword samples, it is more likely to be a correct detection
- The acoustic similarity can be estimated through Dynamic Time Wrapping (DTW)
  - DTW has shown to be successful in the Query-by-example task
  - It is a template-based, non-parametric approach => complementary with ASR scores

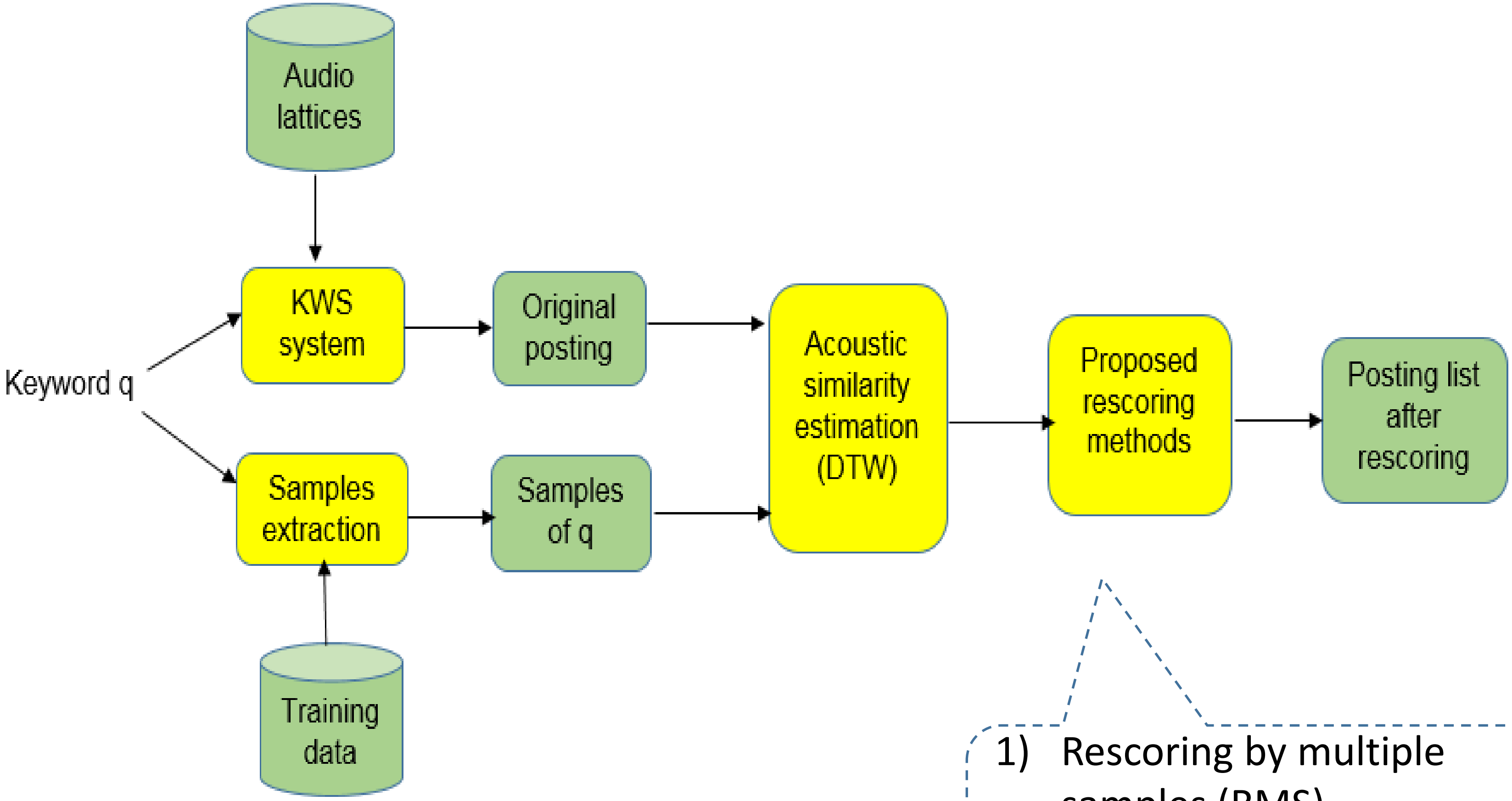
Outline

- Proposed approach
  - The rescoring framework
  - Samples extraction
  - Rescore by multiple samples
  - Rescore by graph-based algorithm
- Experiment
  - Experimental setup
  - Experimental results, analysis and discussion
- Conclusions and future works

Outline

- Proposed approach
  - The rescoring framework
  - Samples extraction
  - Rescore by multiple samples
  - Rescore by graph-based algorithm
- Experiment
  - Experimental setup
  - Experimental results, analysis and discussion
- Conclusions and future works

The rescoring framework



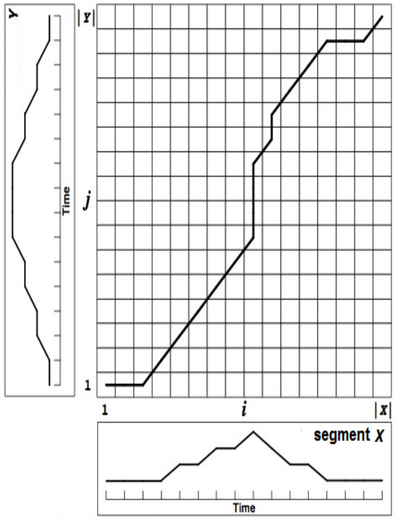
1) Rescoring by multiple samples (RMS)
2) Graph-based rescoring with sample (GBRWS)

Samples extraction

- Estimate the time boundary of each word in training data using forced-alignment
- Consider keyword $q = W_1 W_2 \ldots W_n$
  - If the whole sequence $W_1 W_2 \ldots W_n$ appear in the training data, then we extract the whole speech segment at the found locations as samples
  - Otherwise, find samples of W_i then concatenate them to form sample of q
    - To ensure quality, samples of W_i should belong to same gender
    - Since number of generated samples is large, we randomly select 20 samples.

Acoustic similarity estimation

- First we estimate the dynamic time warping (DTW) between 2 segments



- Then convert the DTW metric to similarity

$$S(X,Y) = 1 - \frac{DTW_{max} - DTW(X,Y)}{DTW_{max} - DTW_{min}}$$

NANYANG TECHNOLOGICAL UNIVERSITY
Institute for Infocomm Research A*STAR
Temasek Laboratories and School of Computer Engineering
Human Language Technology Department

Rescoring by multiple samples (RMS)

- Let d be a detection with raw ASR score C(d)
- Estimate the average similarity between d and all samples

$$\text{AVG\_SIM(d)} = \frac{1}{n}\sum_{i=1}^{n} S(d, x_i)$$

- The final confidence score is

$$C'(d) = C(d)^{\delta}\ \text{AVG\_SIM}(d)^{\,1-\delta}$$

NANYANG TECHNOLOGICAL UNIVERSITY
Temasek Laboratories and School of Computer Engineering
Institute for Infocomm Research A*STAR
Human Language Technology Department

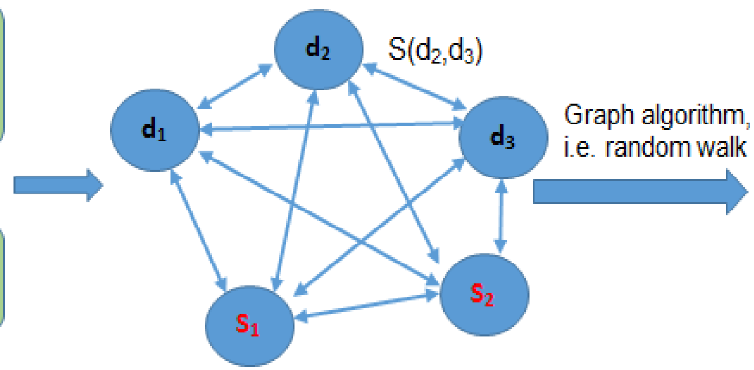
The Graph-based rescoring with sample (GBRWS)

List of detections

Audio=A1 start=2.1 end=2.5 score=0.31 (d_1)
Audio=A2 start=3.3 end=3.8 score=0.14 (d_2)
Audio=A2 start=7.7 end=8.3 score=0.07 (d_3)

Audio=T1 start=1.1 end=1.6 (s_1)
Audio=T2 start=3.6 end=4.2 (s_2)

List of samples



List of rescored detections

Audio=A2 start=3.3 end=3.8 score=0.25 (d_2)
Audio=A1 start=2.1 end=2.5 score=0.20 (d_1)
Audio=A2 start=7.7 end=8.3 score=0.05 (d_3)

$$G(x_i)^t = (1-\alpha-\gamma)C(x_i) + \alpha \sum_{x_j \in D(x_i)} G(x_j)^{t-1} S'(x_i, x_j) + \gamma \sum_{x_j \in E(x_i)} G(x_j)^{t-1} S'(x_i, x_j)$$

Contribution from Initial scores

Contribution from other detections

Contribution from keyword samples

- Previous works [1,2,3] use only detections to build the graph

Outline

- Proposed approach
  - The rescoring framework
  - Samples extraction
  - Rescore by multiple sample
  - Rescore by graph-based algorithm
- Experiment
  - Experimental setup
  - Experimental results
- Conclusion and future work

Experimental setup

- NIST OpenKWS15 data set
  - Language: Swahili – the surprise language of OpenKWS15 Evaluation
  - Training data: FullLP condition 40h.
  - Development data: 10h
  - Evaluation data: 15h evalpart1 released by NIST
  - Keyword list: eval keyword which 1860 keyword appear in evalpart1 data
    - We evaluate the performance of detected keyword

Systems	Detected keywords	Keywords with samples
Word	1711	1509
Subword	1620	1514

Experimental setup (cont.)

- Evaluation metric
  - NIST define the Term-weighted value (TWV) as the metric for KWS

$$TWV(\theta) = 1 - \frac{1}{M}\sum_{k=1}^{M}((P_{miss}(q_k,\theta) + \beta P_{fa}(q_k,\theta))$$

  - We use Maximum TWV (MTWV) as evaluation metric
  - We also report the Detection Error Tradeoff (DET) curves
- Keyword search systems: We build word and subword-based systems using Kaldi toolkit [4]
  - For subword, we use Morfessor toolkit[5] to split both word lexicon and word transcriptions to morpheme-based format.
  - ASR training: fbank feature, 3 gram LM, DNN acoustic model

Experimental results

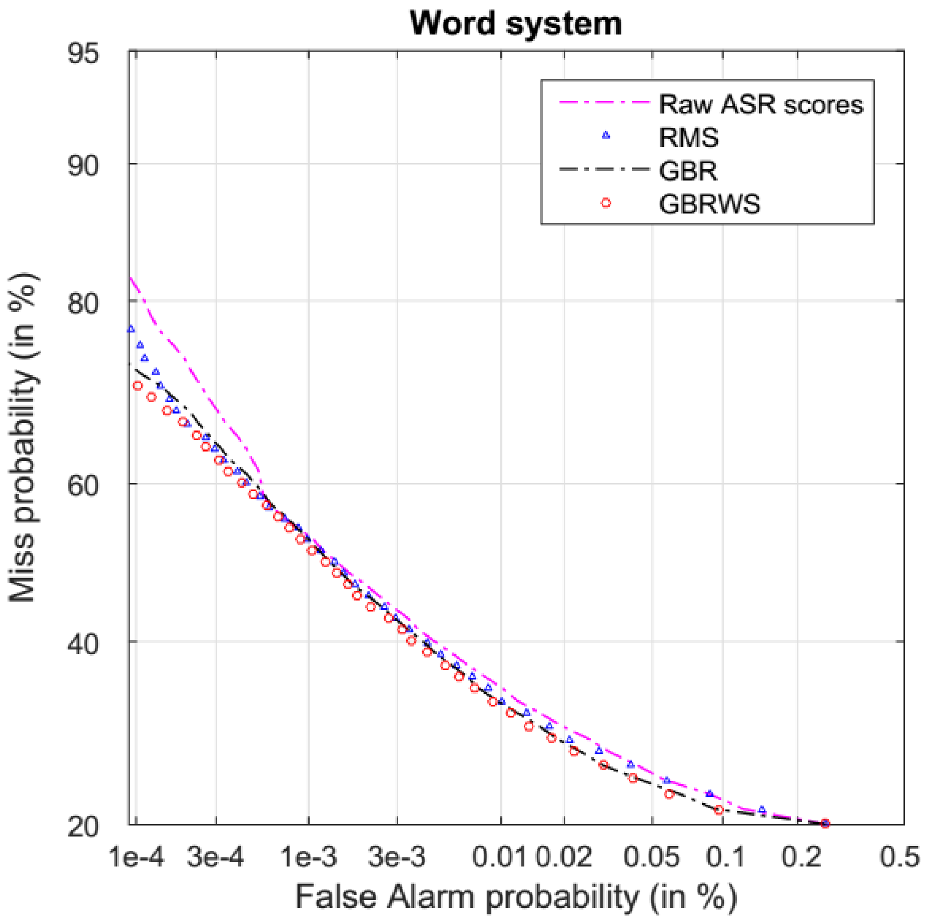
- 2 baselines
  - Raw ASR scores: Original detection scores
  - GBR: Graph based rescoring without training samples [1,2,3]
- MTWV scores

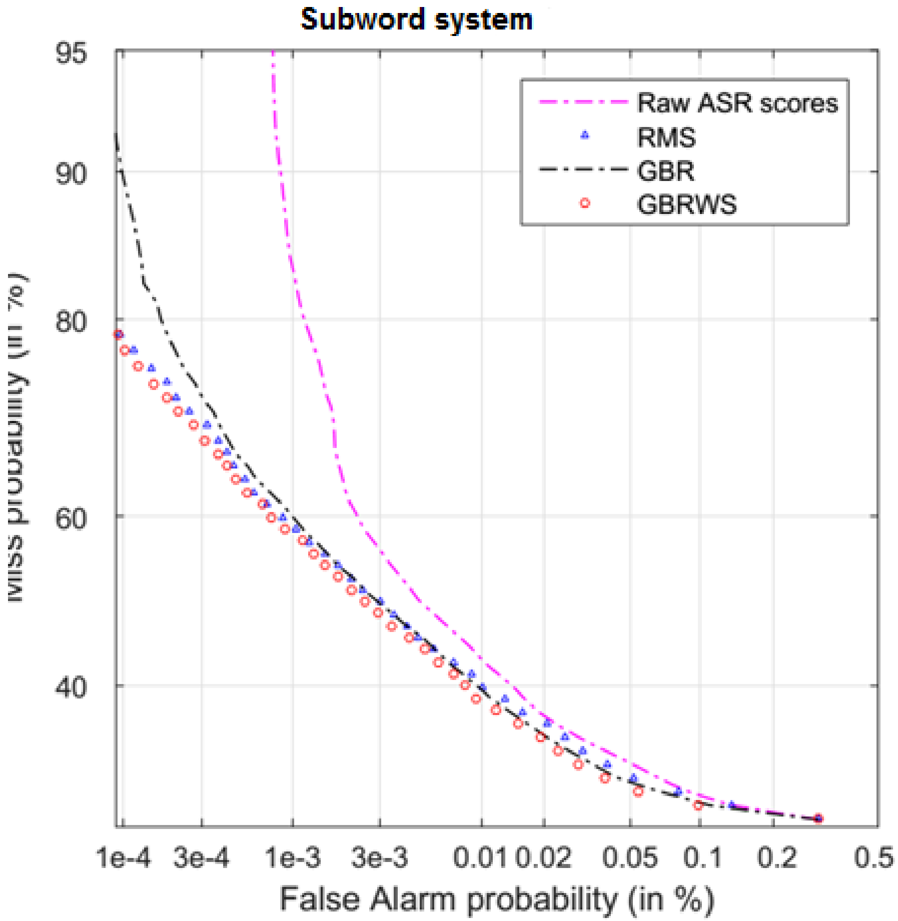
Systems	Raw ASR scores	GBR	RMS	GBRWS
Word	0.5616	0.5797	0.5727	**0.5846**
Subword	0.4716	0.5067	0.5028	**0.5224**

RMS:Rescoring by multiple samples
GBRWS :Graph-based rescoring with sample

NANYANG TECHNOLOGICAL UNIVERSITY
Temasek Laboratories and
School of Computer Engineering
Institute for Infocomm Research
A*STAR
Human Language
Technology Department

Experimental results (cont.)

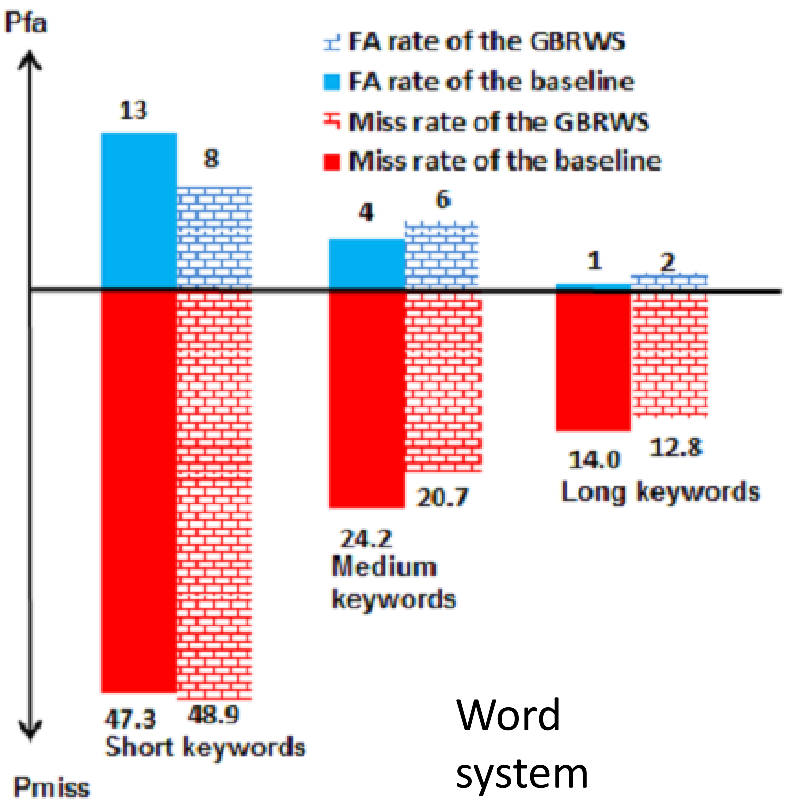




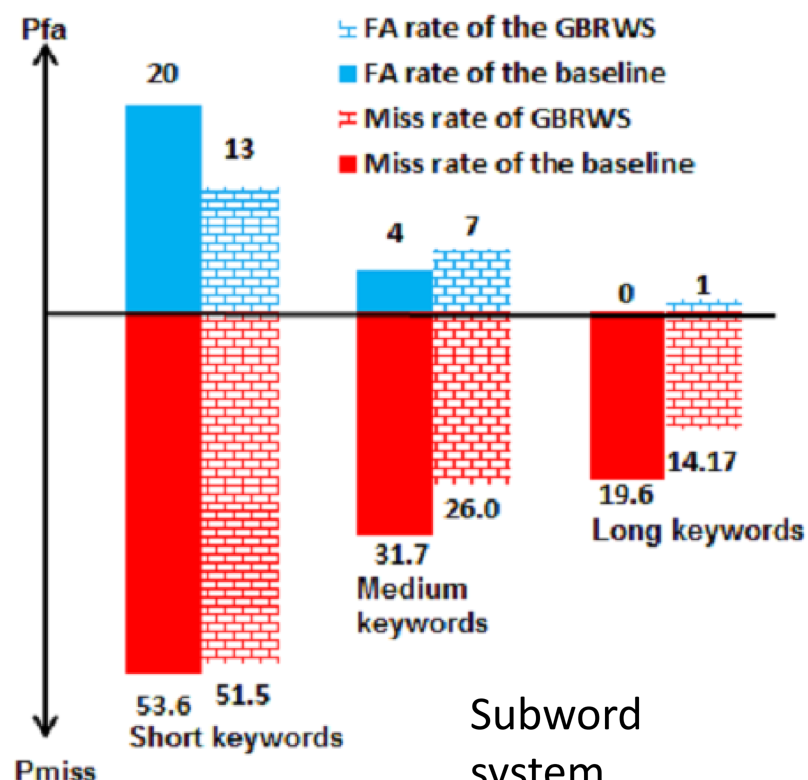
NANYANG TECHNOLOGICAL UNIVERSITY
Temasek Laboratories and School of Computer Engineering
Institute for Infocomm Research A*STAR
Human Language Technology Department

Experimental results (cont.)

- Results for different keyword length



Word system



Subword system

Outline

- Proposed approach
  - The rescoring framework
  - Samples extraction
  - Rescore by multiple sample
  - Rescore by graph-based algorithm
- Experiment
  - Experimental setup
  - Experimental results, analysis and discussion
- Conclusion and future work

Conclusion and future work

- Using keyword samples, together with acoustic similarity, improves the KWS performance
  - The graph based method is more effective than RMS method
  - The proposed approach benefits more for the subword system
  - Much improvement observed on short keywords
- Future work
  - The current method is applicable on seen-word keywords
  - We are investigating way to generate samples for an unseen-word keyword by concatenating samples of its **subwords**

References

[1] H. Y. Lee, Y. Zhang, E. Chuangsuwanich, and J. Glass, "Graph-based re-ranking using acoustic feature similarity between search results for spoken term detection on lowresource," in *Proceedings of ICASSP*, 2013

[2] Y. N. Chen, C. P. Chen, H. Y. Lee, C. Chan, and L. S. Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *Proceedings of ICASSP*, 2011.

[3] A. Norouzian, R. C. Rose, Sina Hamidi Ghalehjegh, and A. Jansen, "Zero resource graph-based confidence estimation for open vocabulary spoken term detection," in *Proceedings of ICASSP*, 2013.

[4] D. Povey et.al, "The kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011

[5] M. Creutz and K. Lagus, "Unsupervised discovery of morphemes," in *In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, 2002

Thank you for listening !

Any question ?