Spoken Term Detection Using Multiple Speech Recognizers' Outputs at NTCIR-9 SpokenDoc STD subtask

Hiromitsu Nishizaki, Yuto Furuya, Satoshi Natori and Yoshihiro Sekiguchi,
University of Yamanashi, Japan

1. Introduction

Much multi-media data available
- improved the environment on multi-media
- improved the infrastructures

More efficient utterance retrieval
- key words or phrases extraction

Term detection from LVCSR output
- the out-of-vocabulary problem
- recognition errors get worse detection performance

Goal ⇨ improving Spoken Term Detection performance

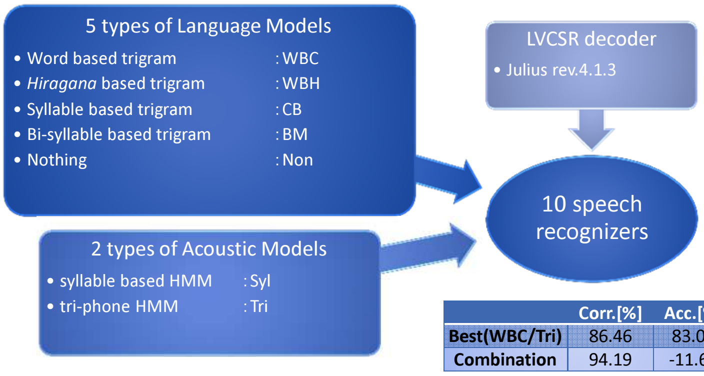
Ideas

✓ **Using Network-formed index from multiple speech recognizers' outputs**
✓ **Introduction of false detection parameters**

2. Proposed STD technique

Multiple speech recognizers

■Using multiple speech recognizers' outputs is very effective in improving syllable-based speech recognition performance



5 types of Language Models
- Word based trigram : WBC
- *Hiragana* based trigram : WBH
- Syllable based trigram : CB
- Bi-syllable based trigram : BM
- Nothing : Non

2 types of Acoustic Models
- syllable based HMM : Syl
- tri-phone HMM : Tri

LVCSR decoder
- Julius rev.4.1.3

10 speech recognizers

	Corr.[%]	Acc.[%]
Best(WBC/Tri)	86.46	83.01
Combination	94.19	-11.67

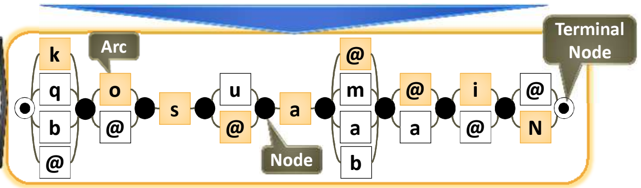
Input voice data
Cosine (*/k o s a i N/*)

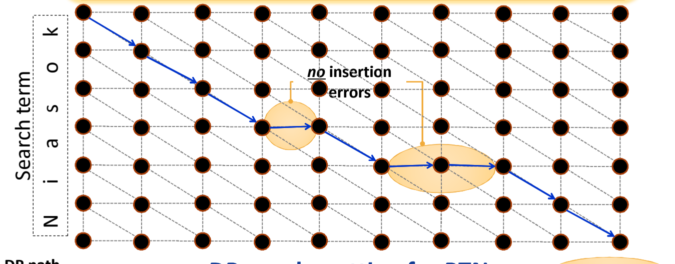
Construction Network-formed Index

LM/AM	Outputs of 10 recognition systems (all outputs are converted into phoneme sequence)								
WBC/Tri	k	o	s	@	a	@	@	i	@
WBH/Tri	q	o	s	u	a	@	a	@	N
CB/Tri	k	o	s	@	a	m	a	i	@
CSB/Tri	k	o	s	@	a	@	@	@	N
Non/Tri	k	o	s	@	a	@	@	@	N
WBC/Syl	@	@	s	@	a	@	@	@	N
WBH/Syl	b	o	s	@	a	a	a	@	@
CB/Syl	@	@	s	@	a	b	@	i	@
CSB/Syl	@	@	s	@	a	@	@	@	N
Non/Syl	@	@	s	@	a	@	@	@	N

NULL

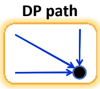
Phoneme Transition Network (PTN)





DP word-spotting for PTN

Distance: 0.3



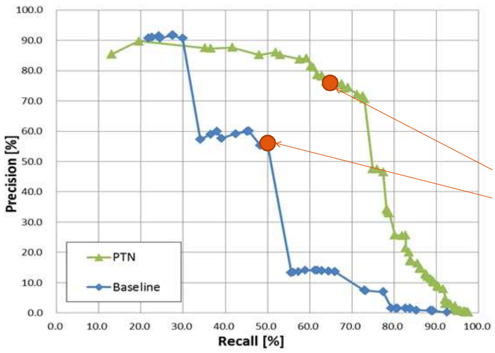
■DP path cost : Edit Distance(error cost only 1.0)
■NULL transition cost : 0.1(heuristic cost)

3. Experiment setup and result

Data for STD task
- CORE set of the STD task (about 40 hours, 144×10³ sec.)

Query
- 50 queries for the CORE set
  - Including 31 out-of-vocabulary queries

Evaluation measure
- Recall-Precision Curve
- F-measure (maximum point on the curve)
- Mean Average Precision (MAP)



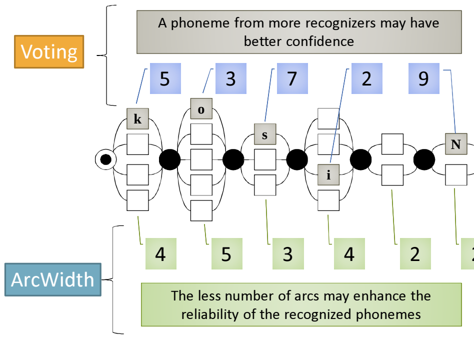
✓ Baseline STD is performed by the DP without NULL transition on the transcription of "CB/Tri."
✓ Baseline STD is performed by the simple DP on the transcription of "CB/Tri."

✓ The maximum F-measure of "PTN" is 71.4%
✓ The maximum F-measure of "Baseline" is 55.6%

■MAP:
✓ PTN 0.757
✓ Baseline 0.595

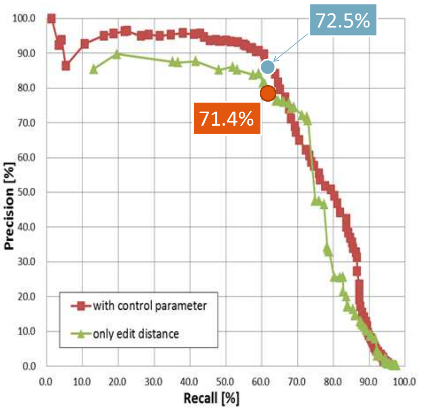
4. False detection control in DP framework

- Introduction of the false detection control parameters
  - ✓ "Voting": the number of recognizers outputting the same phoneme on the same arc
  - ✓ "ArcWitdh": the number of arcs between successive two nodes


Voting
A phoneme from more recognizers may have better confidence
5
3
7
2
9
k
o
s
i
N
4
5
3
4
2
2
ArcWidth
The less number of arcs may enhance the reliability of the recognized phonemes

- The parameters are installed to the calculation of DP cost

STD performance with the control parameters



MAP
✓ With control parameters 0.837
✓ Only edit distance 0.757

5. Conclusion

Summary
- Using multiple speech recognizers for STD
  - Multiple recognizers make STD performance better
  - Integrating multiple recognizers' output in to PTN was very powerful to improve the performance

Future works
- Improving index
  - Reduction of unnecessary information
- Improving search engine
  - Developing new control parameters in the STD engine
  - Customizing the engine depending on an inputted query