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

Hiromitsu Nishizaki

Yuto Furuya

Satoshi Natori

Yoshihiro Sekiguchi

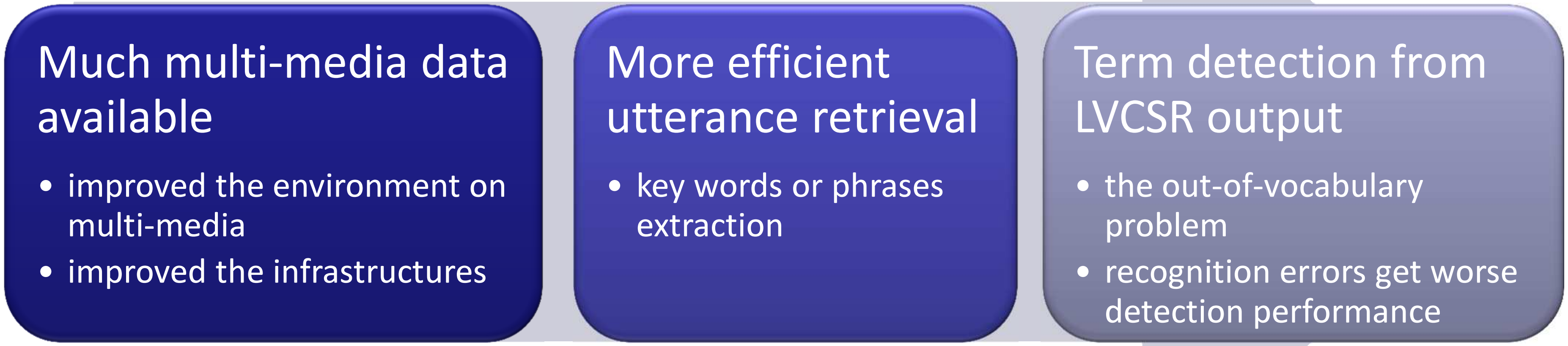
University of Yamanashi, Japan

Outline

- Introduction
- Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result
- False detection control
  - Introducing the control parameters
  - Experimental result
- Conclusion

Introduction

Back ground



Much multi-media data available

- improved the environment on multi-media
- improved the infrastructures

More efficient utterance retrieval

- key words or phrases extraction

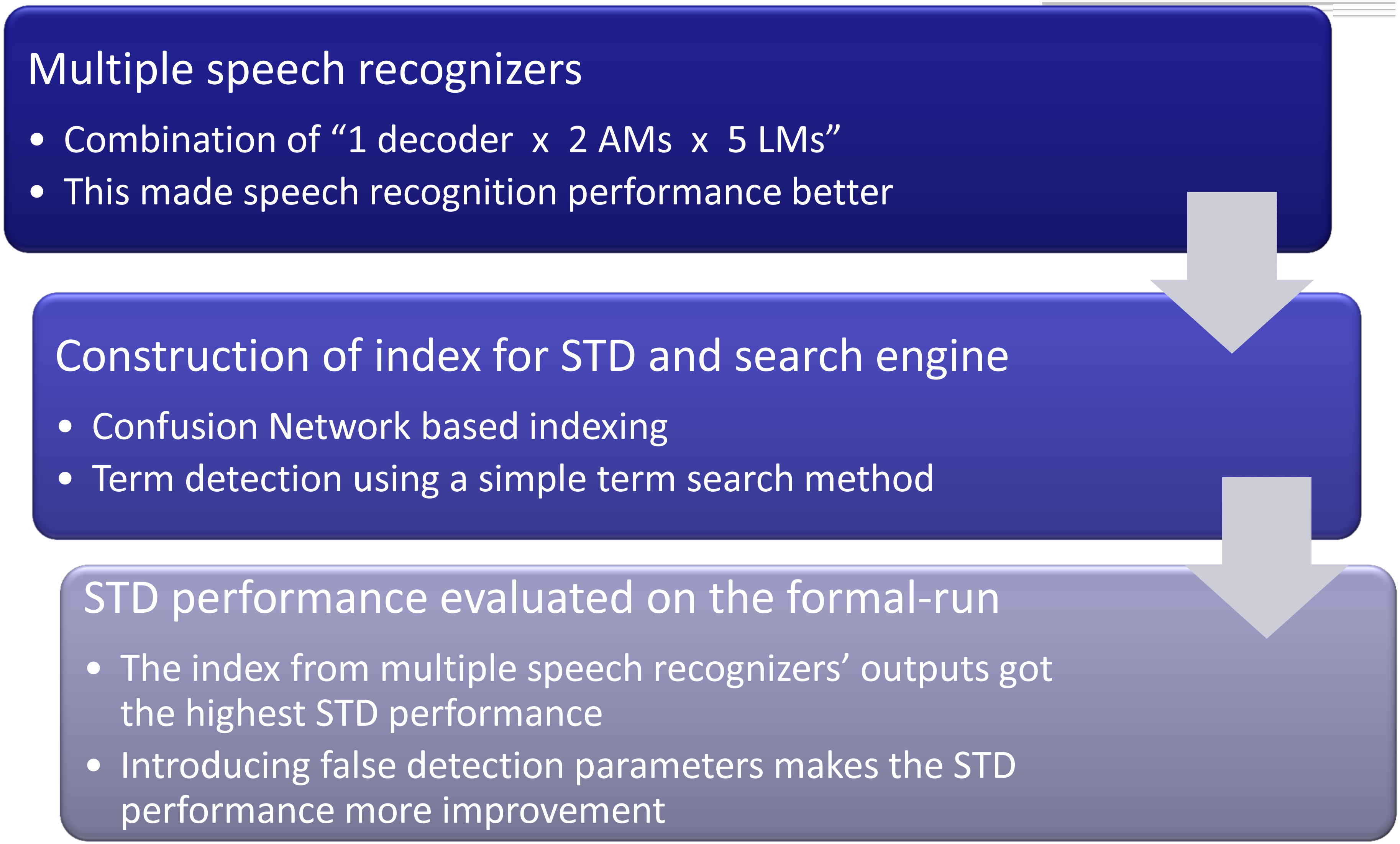
Term detection from LVCSR output

- the out-of-vocabulary problem
- recognition errors get worse detection performance

Our goal

Improving Spoken Term Detection performance

Summary of our research



Multiple speech recognizers

- Combination of “1 decoder x 2 AMs x 5 LMs”
- This made speech recognition performance better

Construction of index for STD and search engine

- Confusion Network based indexing
- Term detection using a simple term search method

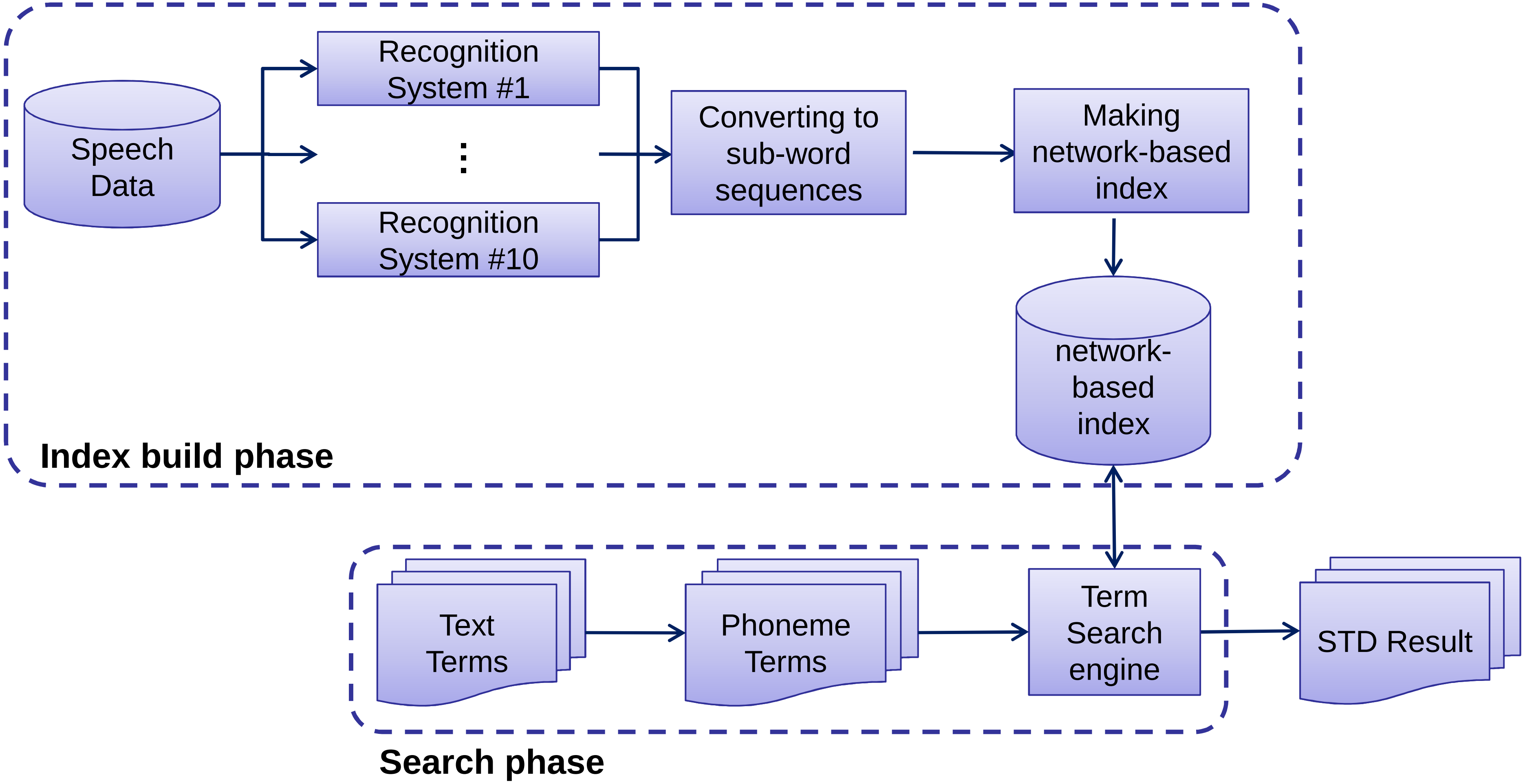
STD performance evaluated on the formal-run

- The index from multiple speech recognizers’ outputs got the highest STD performance
- Introducing false detection parameters makes the STD performance more improvement

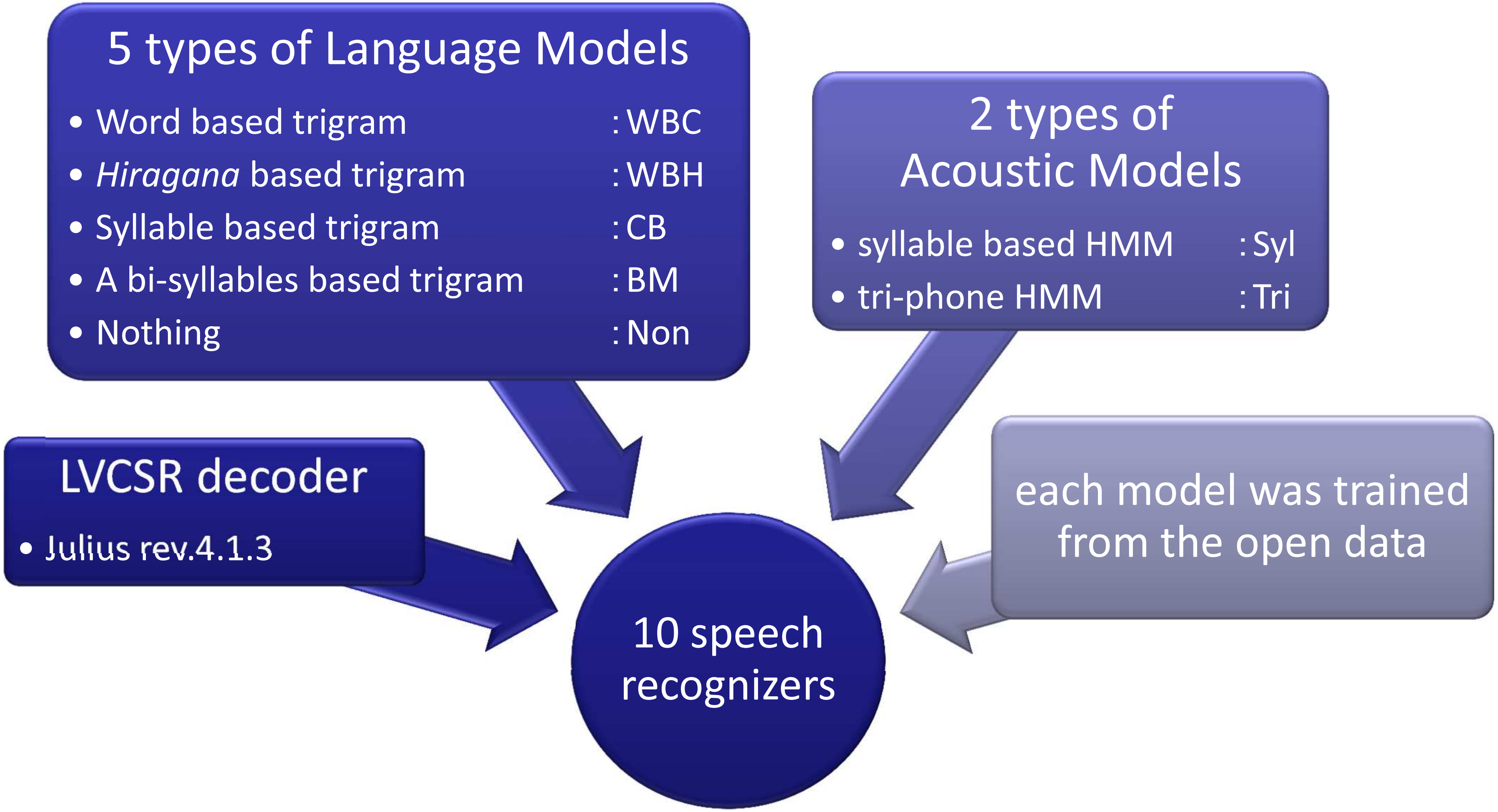
Outline

- ✓ Introduction
- Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result
- False detection control
  - Introducing the control parameters
  - Experimental result
- Conclusion

STD task flow diagram



Multiple speech recognizers



5 types of Language Models

• Word based trigram	: WBC
• *Hiragana* based trigram	: WBH
• Syllable based trigram	: CB
• A bi-syllables based trigram	: BM
• Nothing	: Non

2 types of Acoustic Models

• syllable based HMM	: Syl
• tri-phone HMM	: Tri

LVCSR decoder

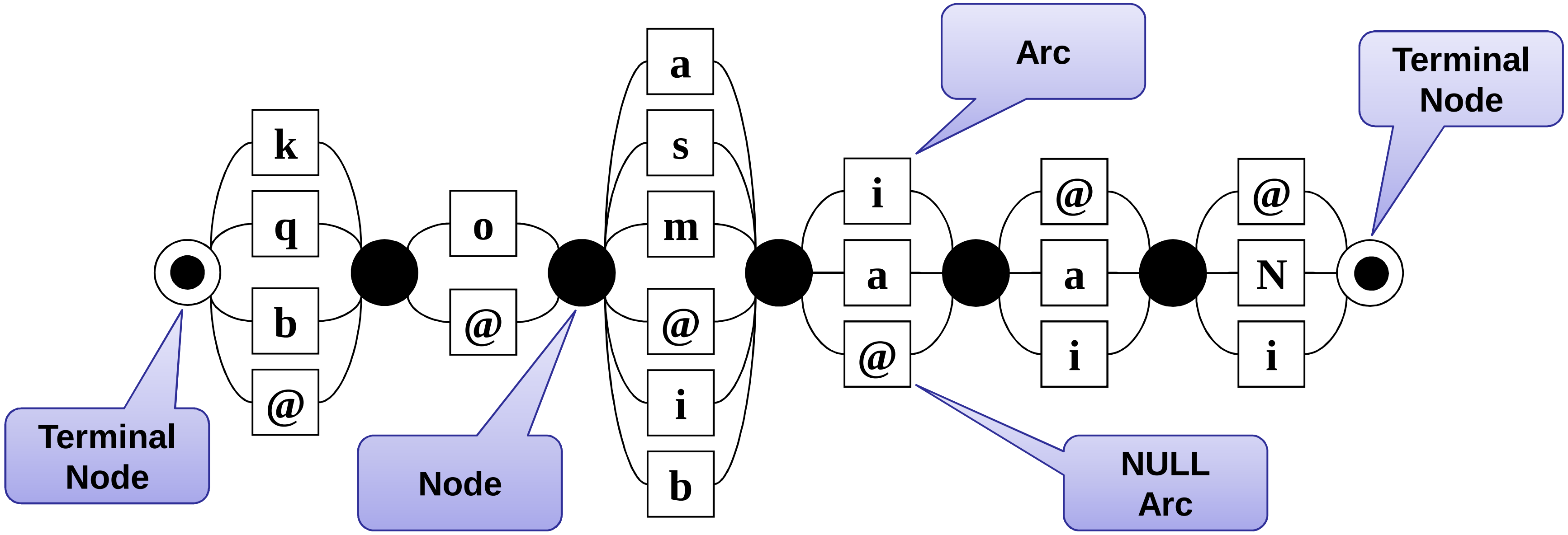
- Julius rev.4.1.3

each model was trained from the open data

10 speech recognizers

Phoneme Transition Network (PTN)

- Phoneme-level Confusion Network based index for STD
  - It called as ``PTN'' (Phoneme Transition Network)
  - PTN is built from multiple speech recognizers' outputs



Example of building PTN-based index

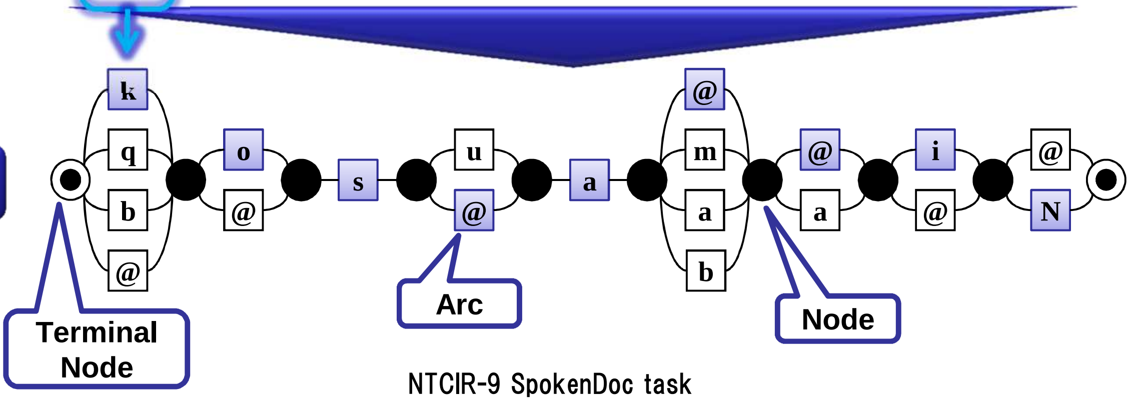
speech utterance “Cosine” (*/k o s a i N/*)

10 systems

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	@
BM/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	@
BM/Syl	@	@	s	@	a	@	@	@	N
Non/Syl	@	@	s	@	a	@	@	@	N

Base output

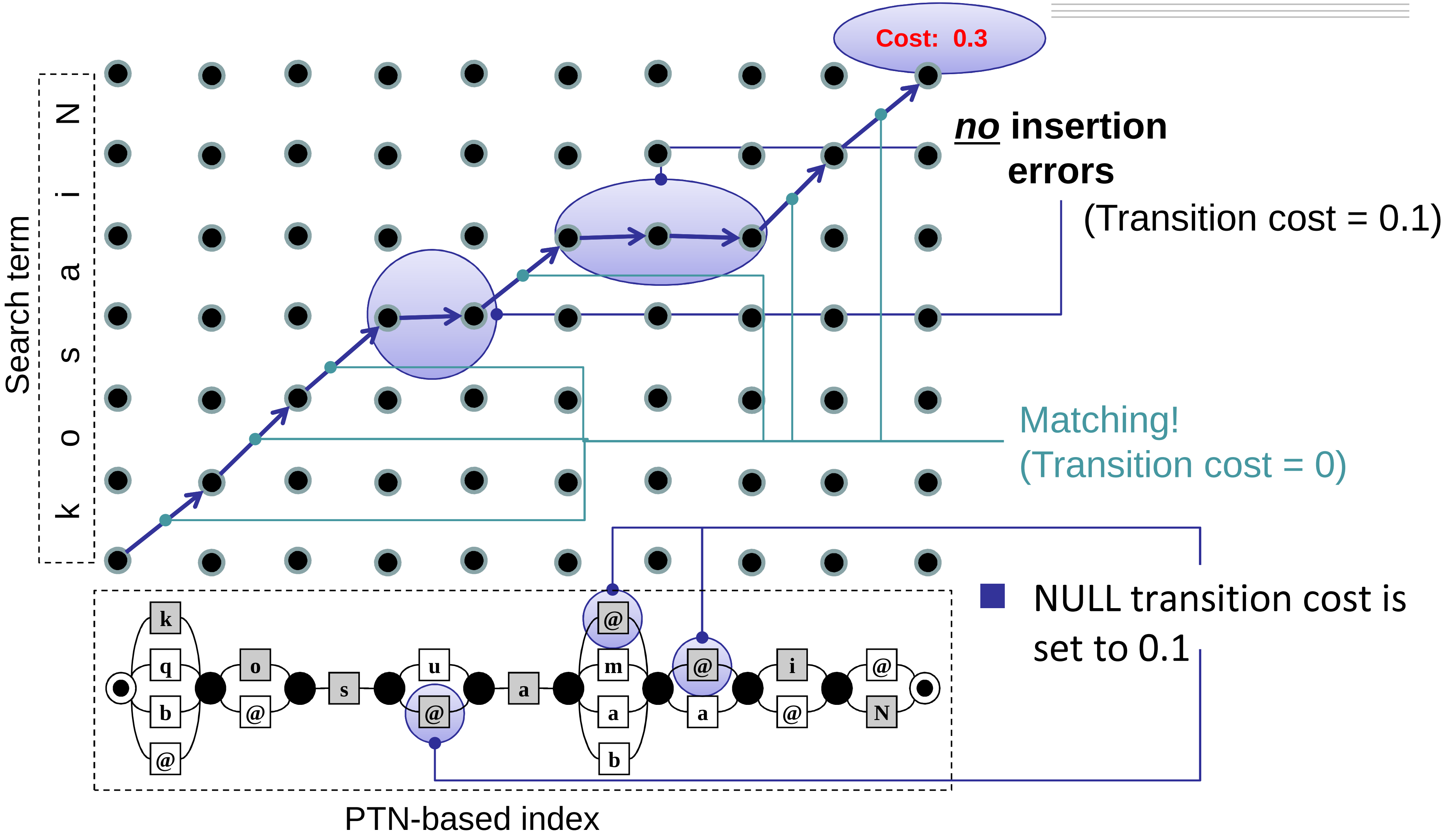
PTN based index



Search engine (no false detection control)

- Simple search engine
  - Dynamic Programming (DP) based engine
  - Both endpoints free
  - Edit distance is used for calculating DP cost between an index and a query term
- We modified the simple DP framework to adapt the PTN-based index

Example of the modified DP framework for PTN-based index (baseline technique)



no **insertion errors**

(Transition cost = 0.1)

Matching!
(Transition cost = 0)

- NULL transition cost is set to 0.1

Experimental setup

Data for STD task

- CORE set of the STD task (about 40 hours, 144 × 10^3 sec.)

Query

- 50 queries for the CORE set
  - Including 31 out-of-vocabulary(OOV) queries

Evaluation measure

- Recall-Precision curve
- F-measure at the maximum point of the curve

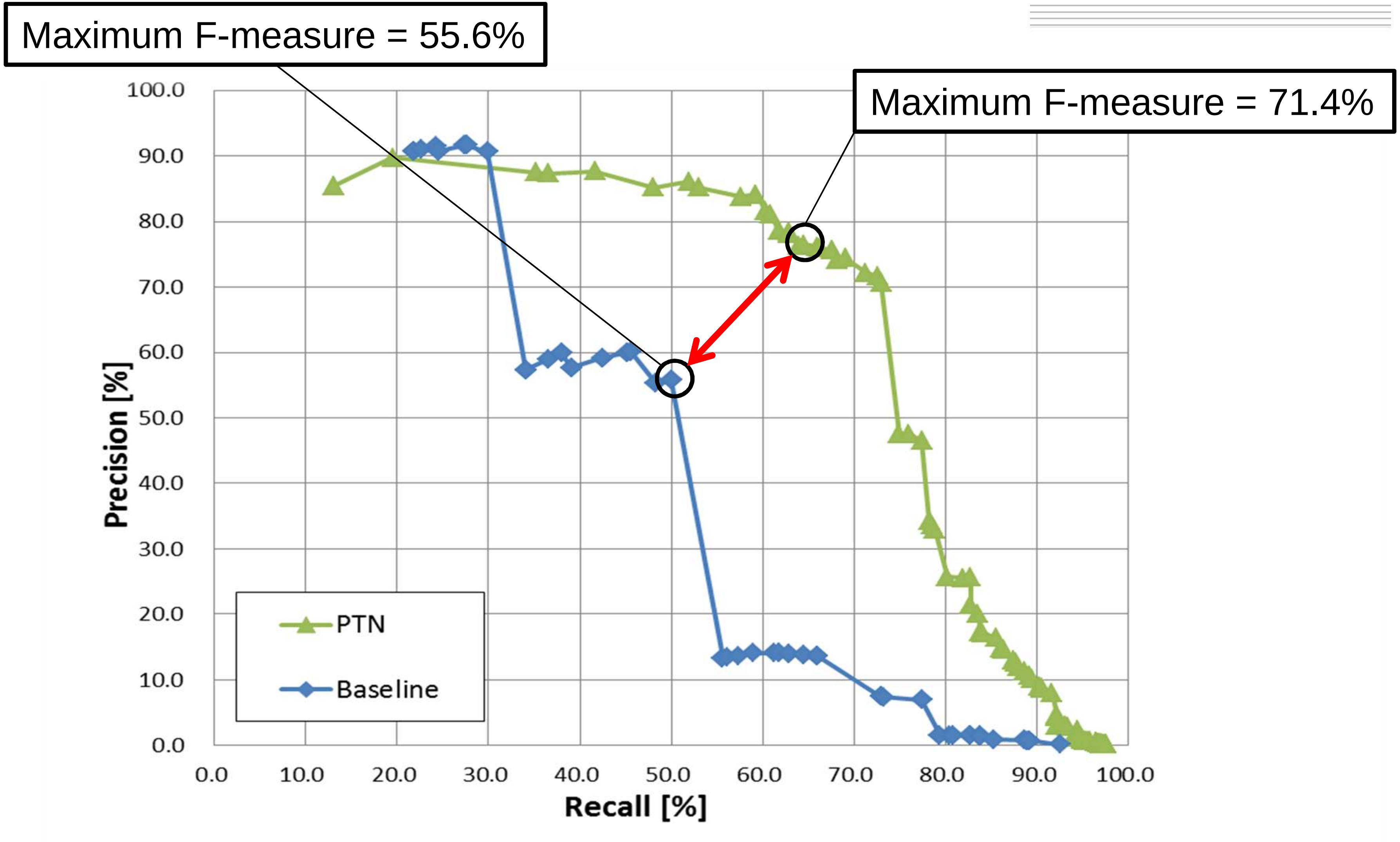
Indices for STD

- Two types of Index

Index	# of hypothesis	type of index	How to make
Baseline	1	Phoneme-base	1-Best output of "CB/Tri"
PTN	10	Phoneme-base	10 types of output

Baseline STD is performed by the simple DP on the transcription of "CB/Tri."

STD results


Maximum F-measure = 55.6%
Maximum F-measure = 71.4%
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Precision [%]
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0
Recall [%]
PTN
Baseline

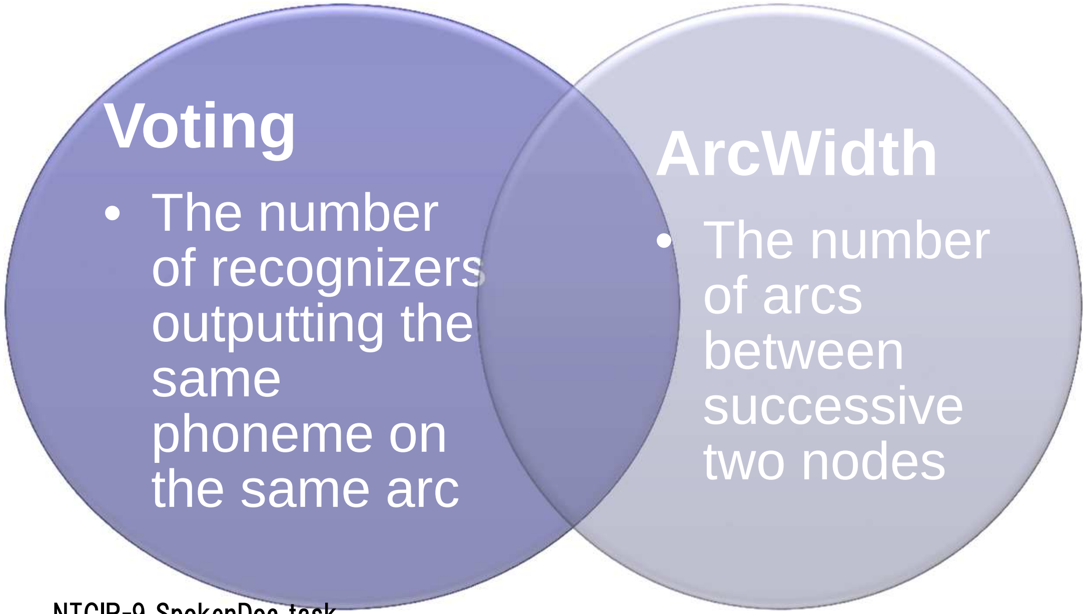
Outline

- ✓ Introduction
- ✓ Spoken Term Detection (STD) using multiple speech recognizers
  - Overview of our STD framework
  - Multiple speech recognizers
  - Phoneme Transition Network (PTN)-based indexing
  - Search engine and experimental result
- False detection control
  - Introducing the control parameters
  - Experimental result
- Conclusion

Robust for false detections

- False detection control for more STD improvement
- Our approach generates many false detections because of :
  - □ using multiple speech recognizers' outputs
  - □ using a network-based index

Two types of control parameters!



Voting

- The number of recognizers outputting the same phoneme on the same arc

ArcWidth

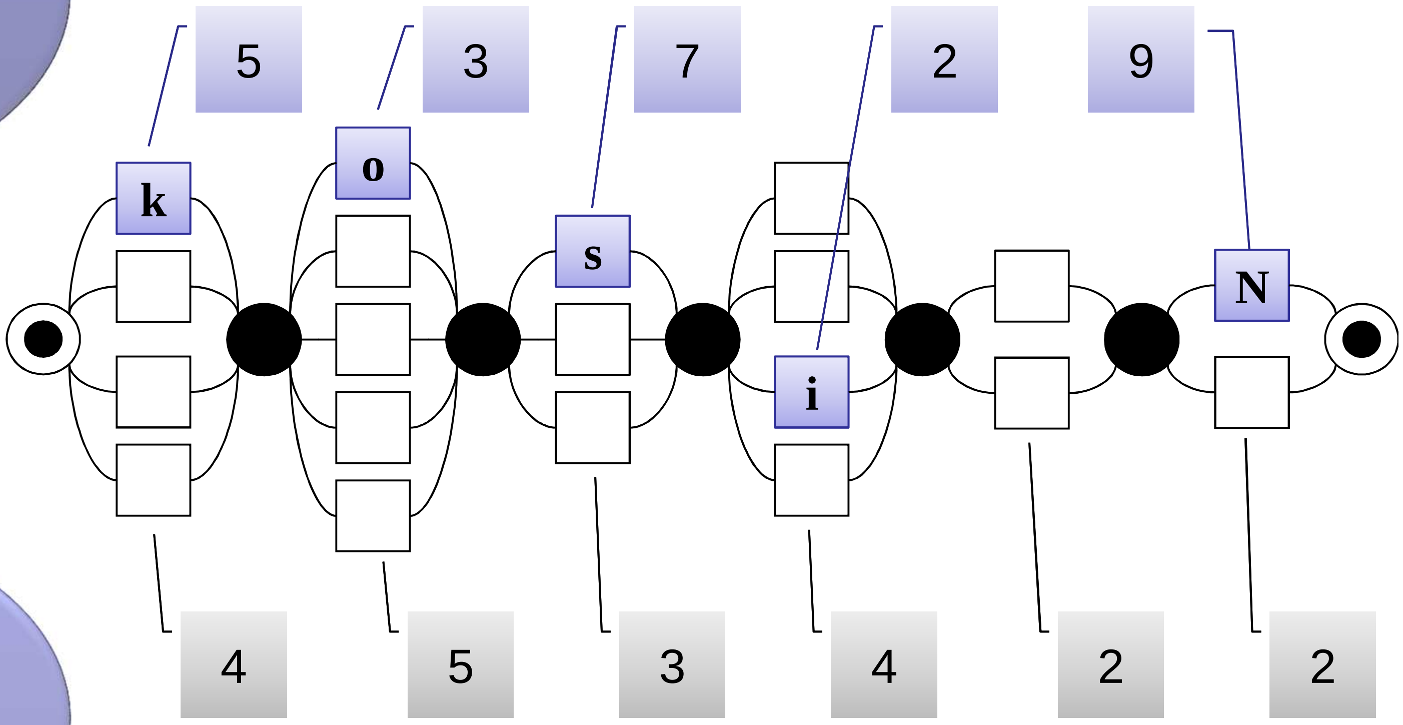
- The number of arcs between successive two nodes

False detection control parameters

Voting

A phoneme from more recognizers may have better confidence

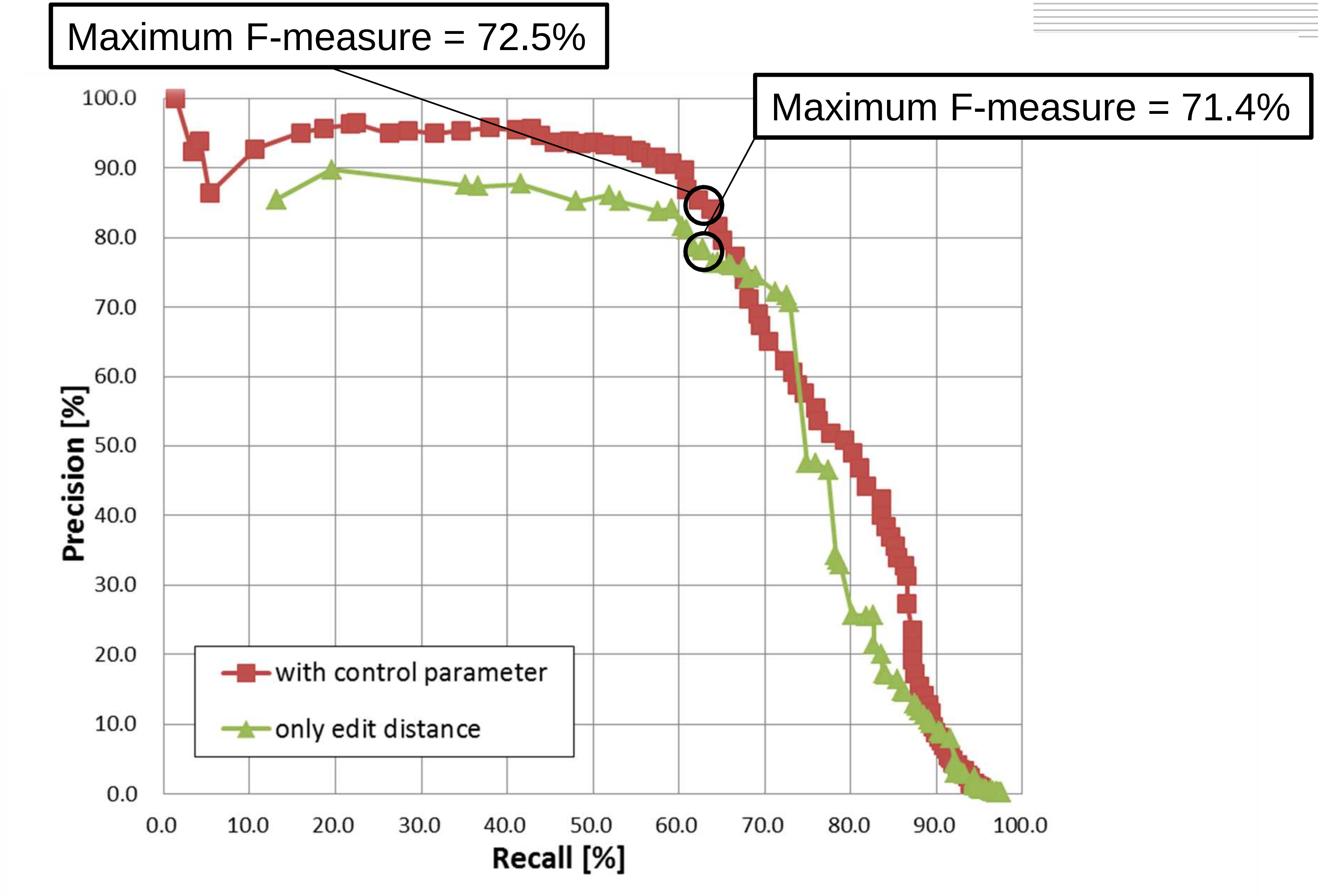
PTN based index



ArcWidth

The less number of arcs may enhance the reliability of the recognized phonemes

Experimental results (with false detection control)



Maximum F-measure = 72.5%

Maximum F-measure = 71.4%

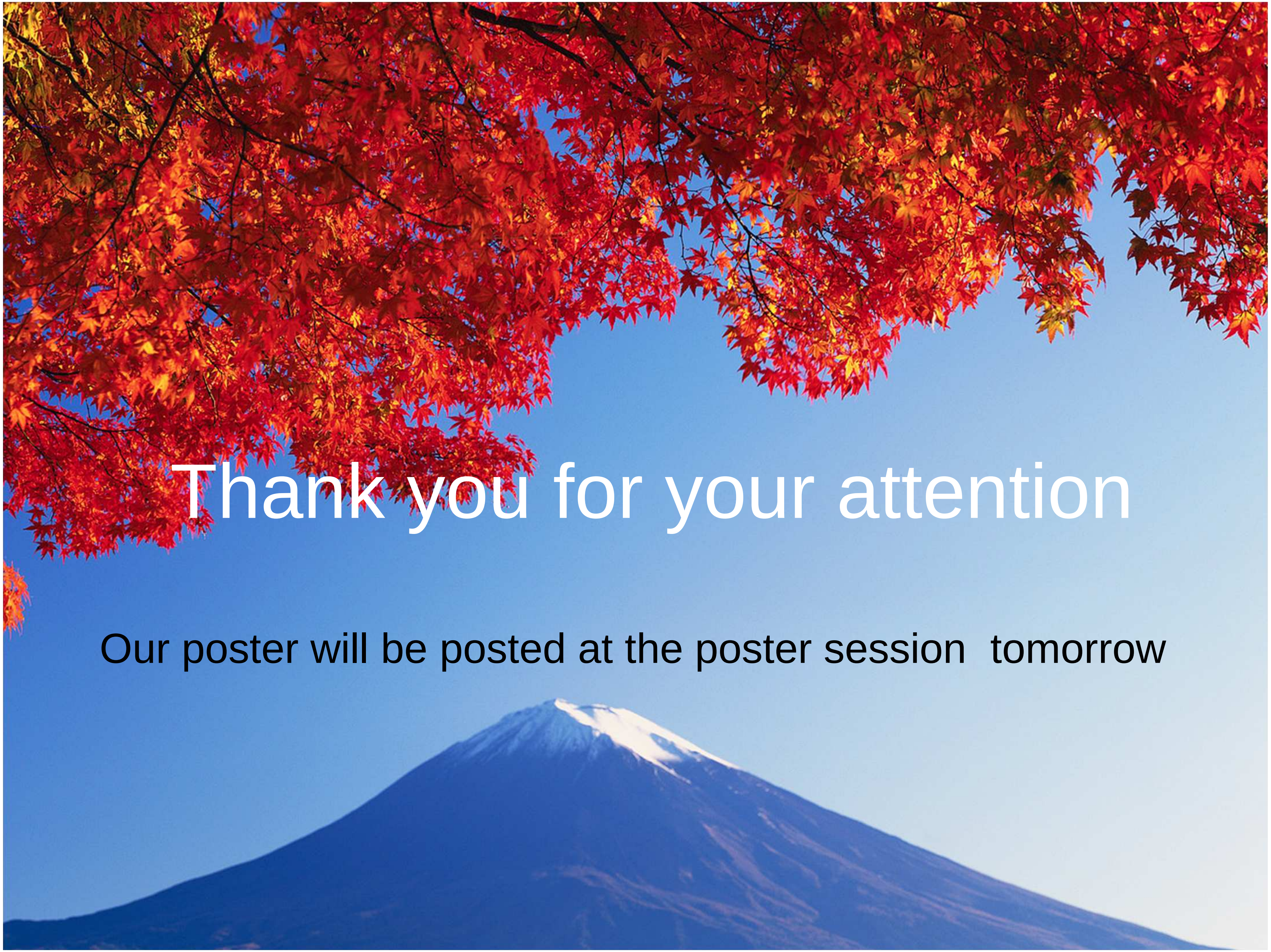
Conclusion

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



Thank you for your attention

Our poster will be posted at the poster session tomorrow