A Semi-manual Annotation Approach for Large CAPT Speech Corpus

Yanlu Xie, Xin Wei, Wei Wang, Jinsong Zhang

Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University, Beijing 100083, China

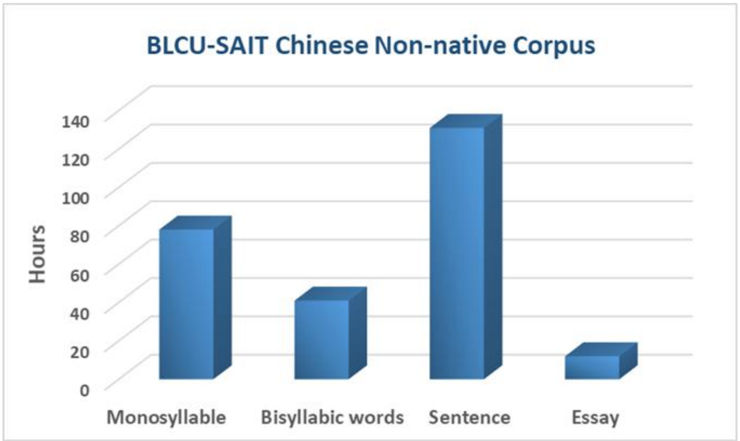
语言资源高精尖创新中心
Beijing Advanced Innovation Center for Language Resources

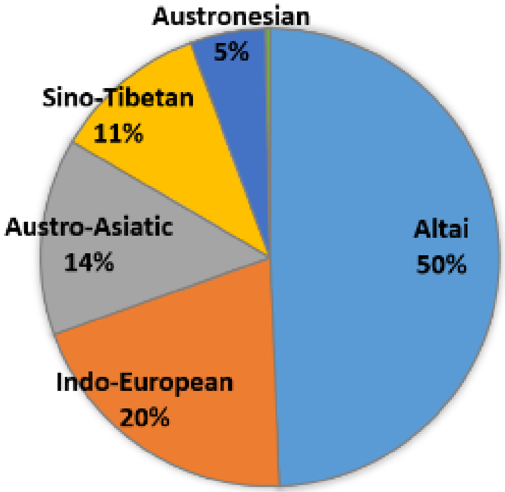
Outline

- □ BLCU-SAIT CAPT Speech Corpus
- □ Semi-manual Annotation Methods
- □ Annotation Evaluation Methods
- □ Annotation Results
- □ Conclusion

BLCU-SAIT CAPT Speech Corpus

- Aiming at Computer Assistant Pronunciation Teaching
- 243 hours’ nonnative data from 618 Speakers
- 21 kinds of native language backgrounds





BLCU-SAIT CAPT Speech Corpus

Sentence Set:

- 103 declarative sentences + 35 question/exclamatory sentences
- cover 97% tri-tone types bounded by prosodic boundary
- cover 96% syllable types

Word Set:

- 284 bi-syllable words
- cover 97% Chinese segmental phonemes
- cover 20 kinds of bi-tone types

Monosyllable Set:

- 1520 tonal syllables
- 98% base syllables

A Discourse:

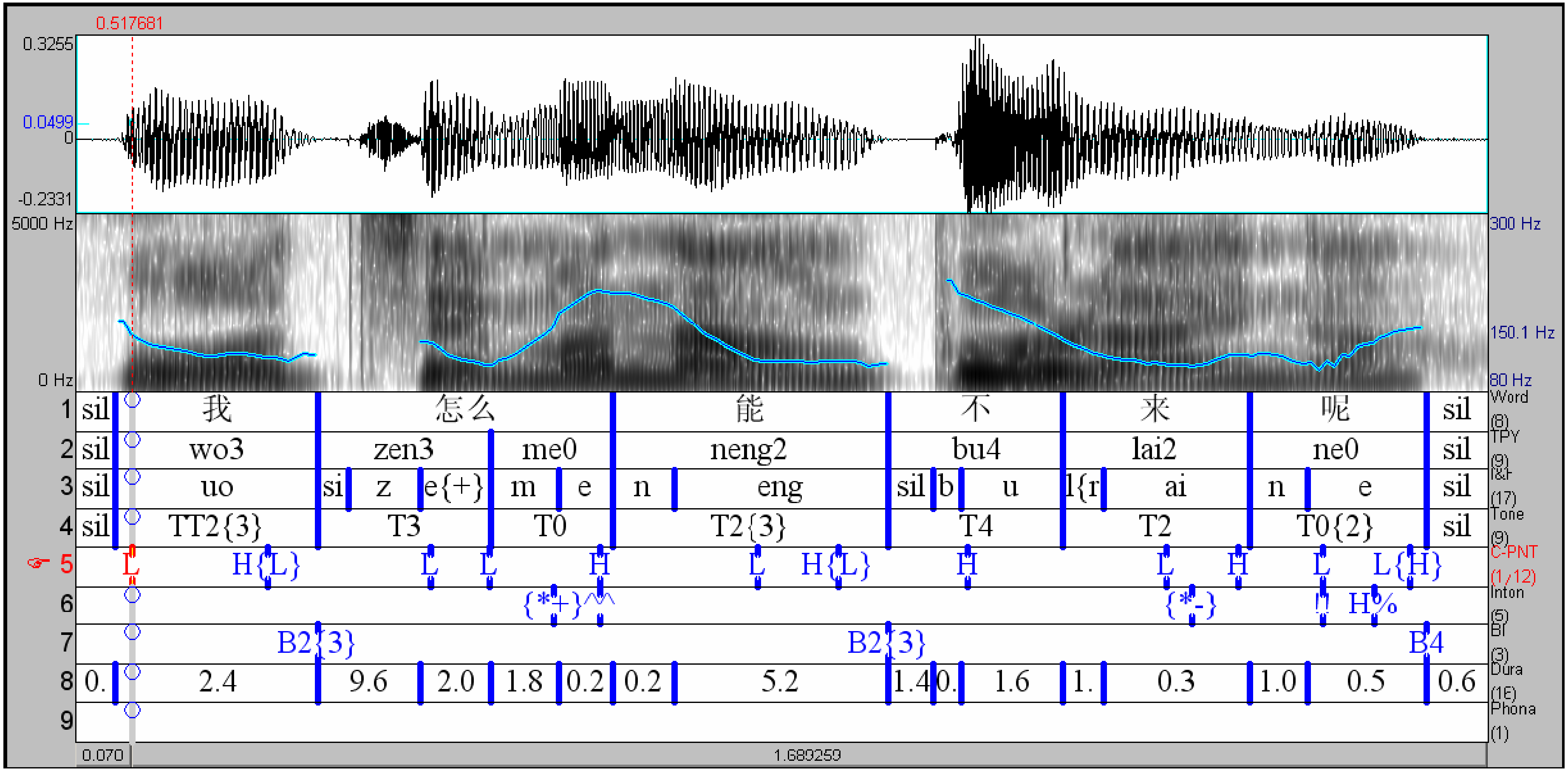
- The North Wind and the Sun
- 7 sentences, 143 Chinese characters

The Challenge in speech annotation

- □ Annotation plays an important roles in speech database.

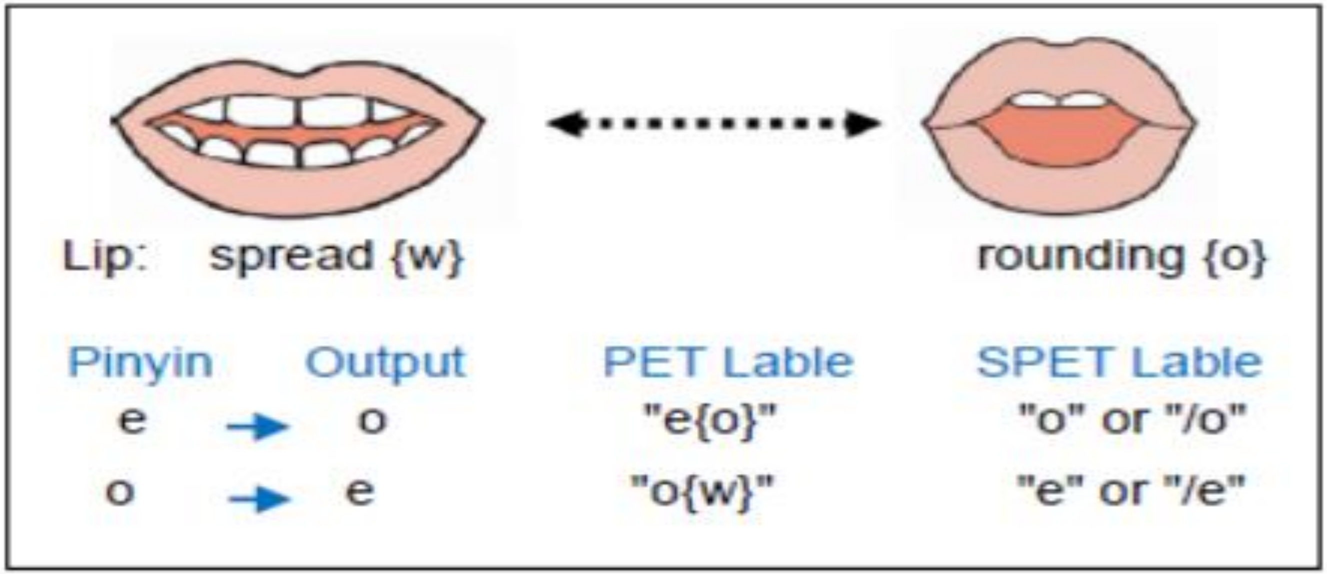
- □ Annotation is time and annotators consuming.
  - SLAM and Speech Analyzer POSCAT (Kim, B., 2000)(Godwinjones, R. 2009)
  - CHAT (Codes for the Human Analysis of Transcripts) (MacWhinney, 2000).
  - DARCLE Annotation Scheme (DAS) (Marisa Casillas, 2017).

Phonetic Labels



Pronunciation Erroneous Tendencies

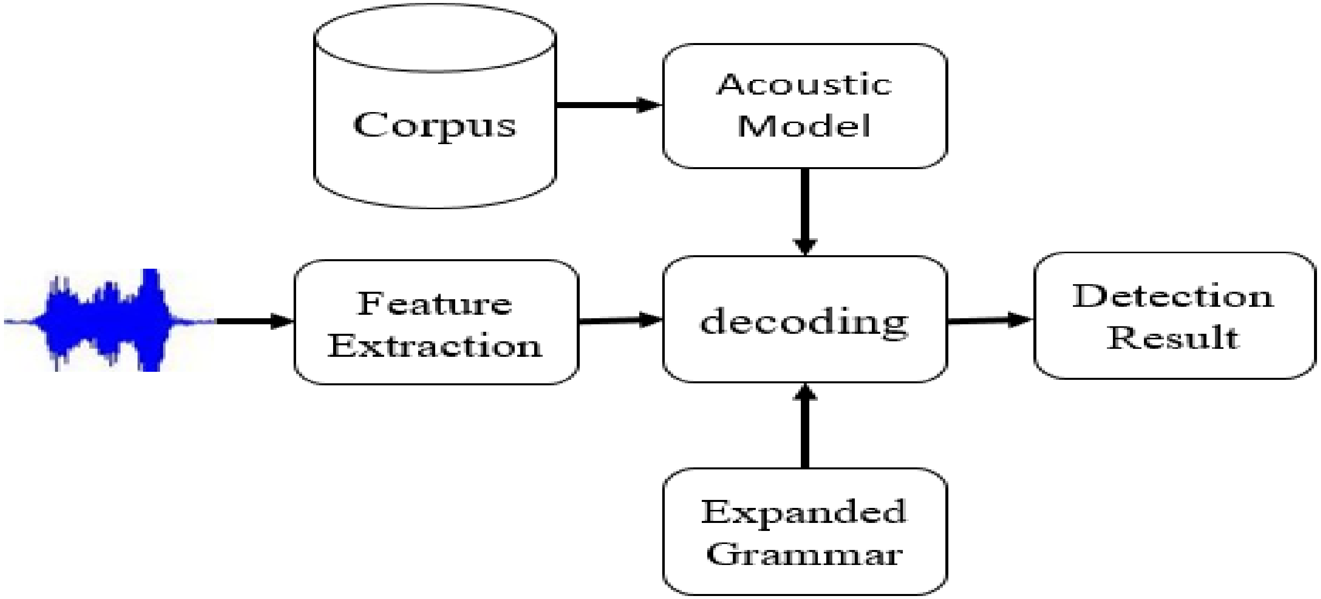
- □ Pronunciation Teaching
- □ PET/SPET



Pinyin		Output	PET Lable	SPET Lable
e	→	o	"e{o}"	"o" or "/o"
o	→	e	"o{w}"	"e" or "/e"

Semi-manual Annotation Automatic Label(1)

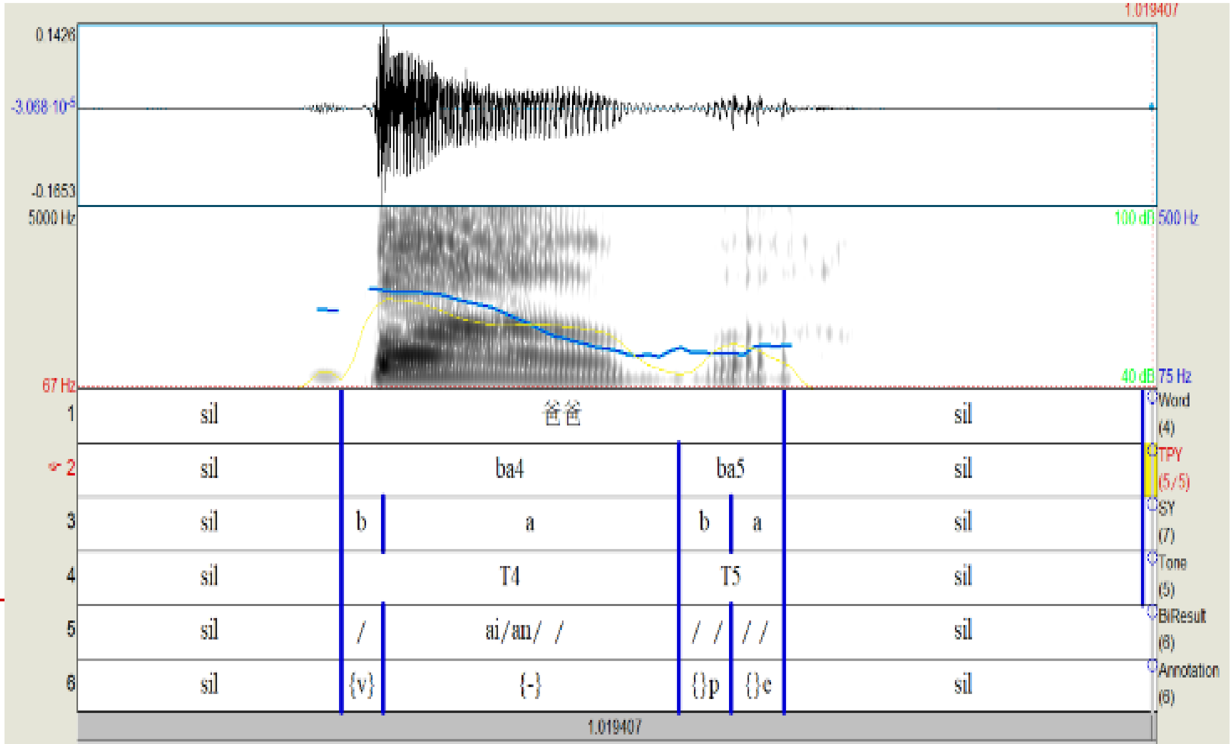
- □ state-of-the-art ASR: LSTM/Chain model
- □ an expanded grammar according to the length of input speech



Semi-manual Annotation Manual Label(2)

open-ended questions (choose one from all the initials and finals)

multiple-choice questions (choose one from four candidates)



Annotation Evaluation

Mean consistency rate (MCR):

In an extreme case, if the erroneous is very little and one annotator is lazy and labels zero erroneous. The consistency rate will also be high.

Posterior Probability Annotation Evaluation

$$F_1 = \frac{2\,\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

the ground truth?

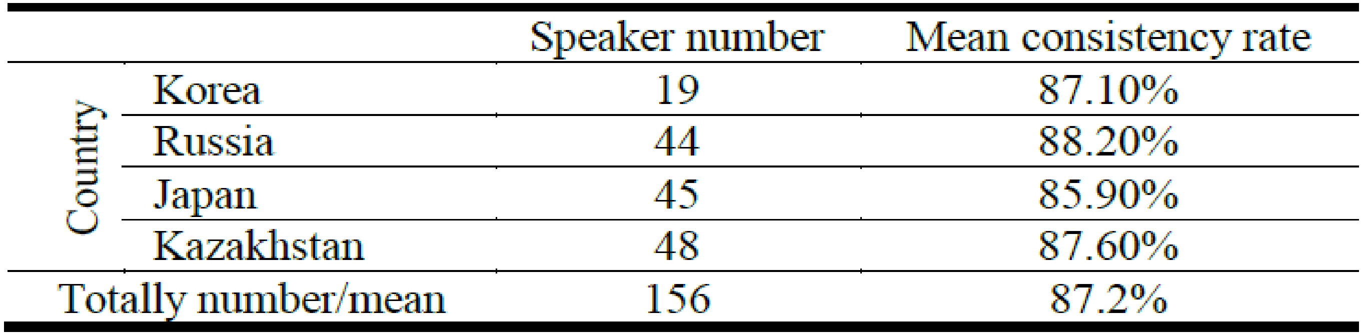
$$F_1p = \frac{2\,\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} * MCR$$

Posterior F1(F1p):

Annotation Results

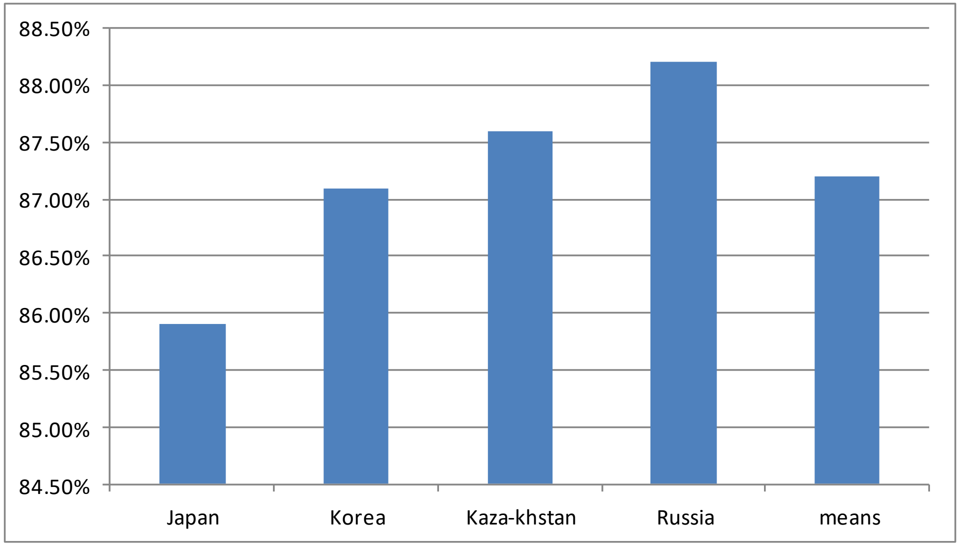
- □ 156 Speakers’ Word Set
- □ 284*156=44304 bi-syllable words
- □ Three annotators for each words

Speaker numbers & Mean consistency rate of phoneme annotations



		Speaker number	Mean consistency rate
Country	Korea	19	87.10%
	Russia	44	88.20%
	Japan	45	85.90%
	Kazakhstan	48	87.60%
Totally number/mean		156	87.2%

Annotation Results

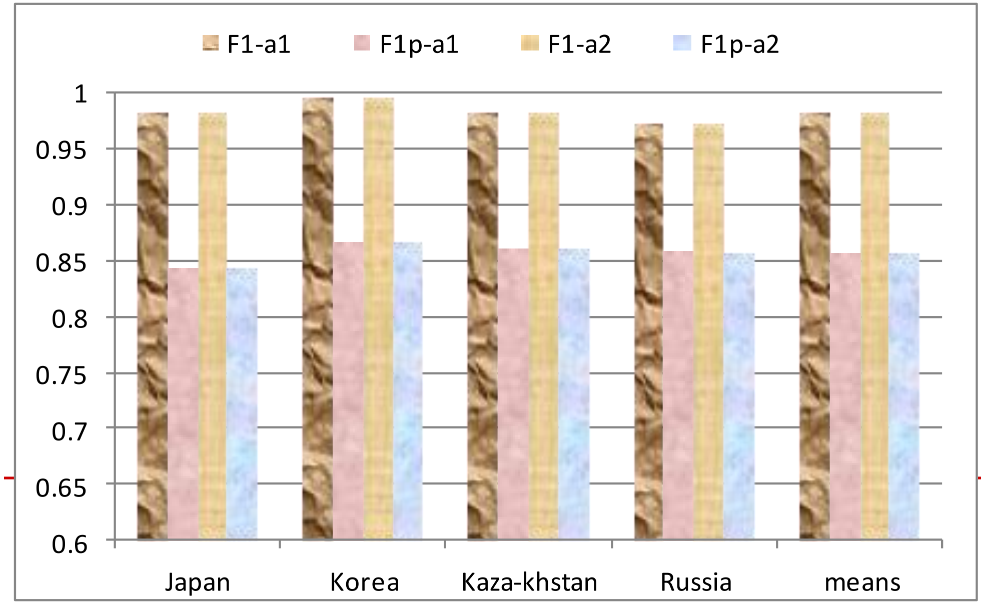


Mean consistency rate of each annotators

the consistency rate of the two annotators in this study raised from 80.7% to 87.2%,

Annotation Results

- Granted that the third annotator's label result is the ground truth.
- F1-a1 and F1-a2 are the F1 score of the first annotator and the second annotator



The F1 is extremely high.

The posterior F1 score is 0.857.

Conclusion

- □ Semi-manual annotation is a promising method in labelling speech data.
- □ The posterior F1 could measure the annotation result more reasonable.
- □ Annotation is still a challenge task.

Thanks