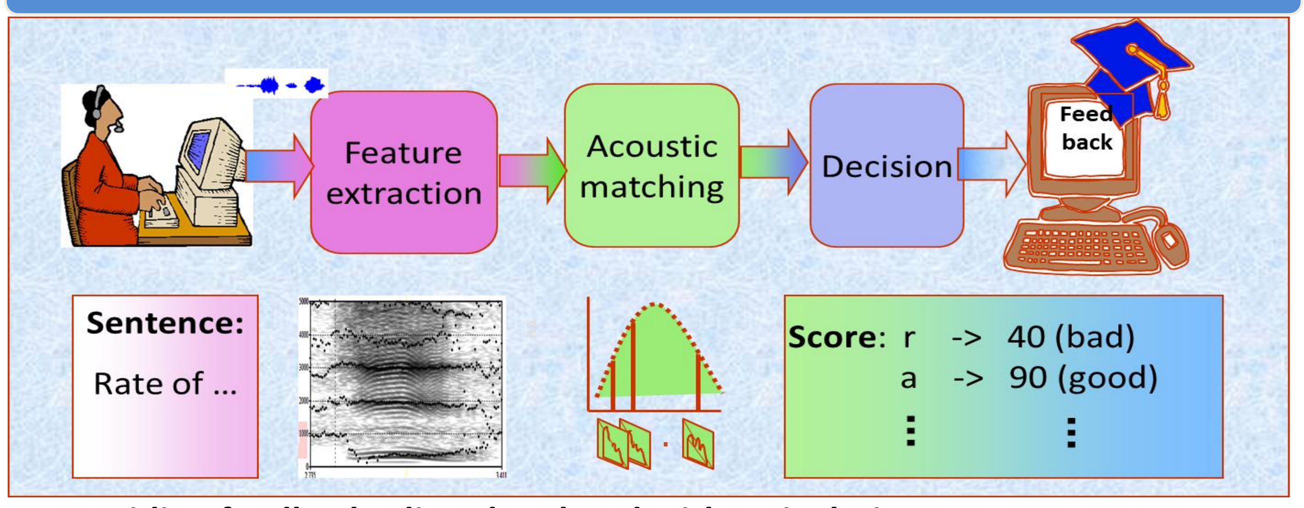
Pronunciation Error Detection using DNN Articulatory Model based on Multi-lingual and Multi-task Learning

Richeng Duan[1] Tatsuya Kawahara[1] Masatake Dantsuji[2] Jinsong Zhang[3]
[1]School of Informatics, Kyoto University
[2]Academic Center for Computing and Media Studies, Kyoto University
[3]School of Information Science, Beijing Language and Culture University

Introduction



- Providing feedbacks directly related with articulation.

Challenge & Proposed method

- **Challenge**
  - Non-native corpus collection in a large scale is not easy .
  - Precisely annotating non-native speech is difficult.
- **Proposed method**
  - Modeling articulatory attributes **without** non-native training data.
  - Enhancing articulatory models with multi-task learning.
  - Learning better feature representation using Multi-lingual learning .

Context-dependent attribute modeling

Definition of articulatory attribute

CH (Chinese) learning by JP (Japanese) students

Manner	Phone set
Aspirated-stop	CH: p t k
Unaspirated-stop	CH: b d g
nasal	CH: m n JP: m n N
Unvoiced-fricative	CH: f s sh JP: s sh h
Unvoiced-stop	JP: p t k
Voiced-stop	JP: b d g

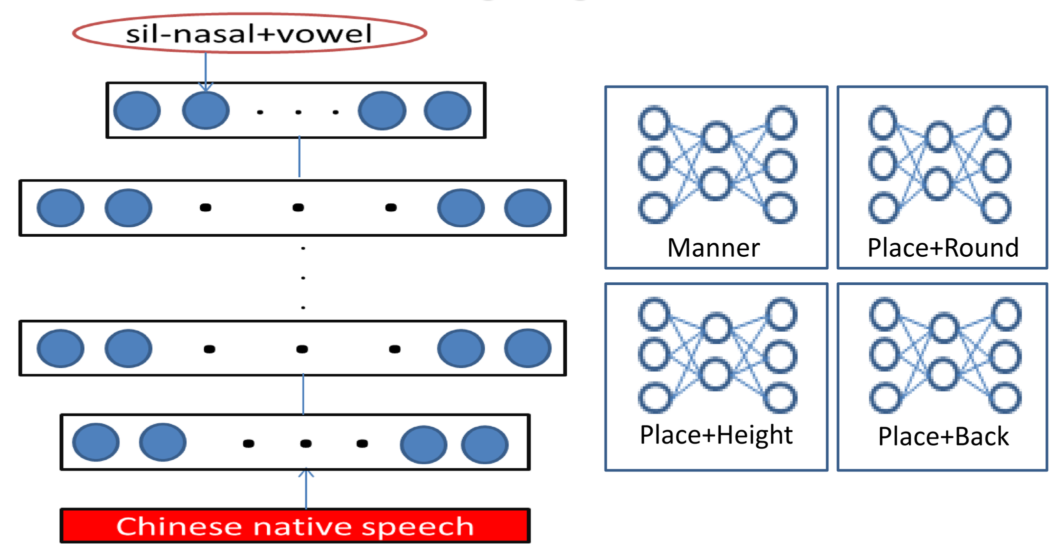
Place	Phone set
Retroflex	CH: zh ch sh r
Bilabial	CH: b p m JP: b p m
Velar	CH: g k h JP: g k
glottal	JP: h

Attribute		Phone set
Backness	Anterior	CH: i ü JP: i e
	Central	CH: a JP: a u
	Back	CH: e u o JP: o
Height	High	CH: i u ü JP: i u
	Mid	CH: o e JP: e o
	Low	CH: a JP: a
Roundedness	Unroundedness	CH: a i e JP: a i e
	Roundedness	CH: o u ü JP: o

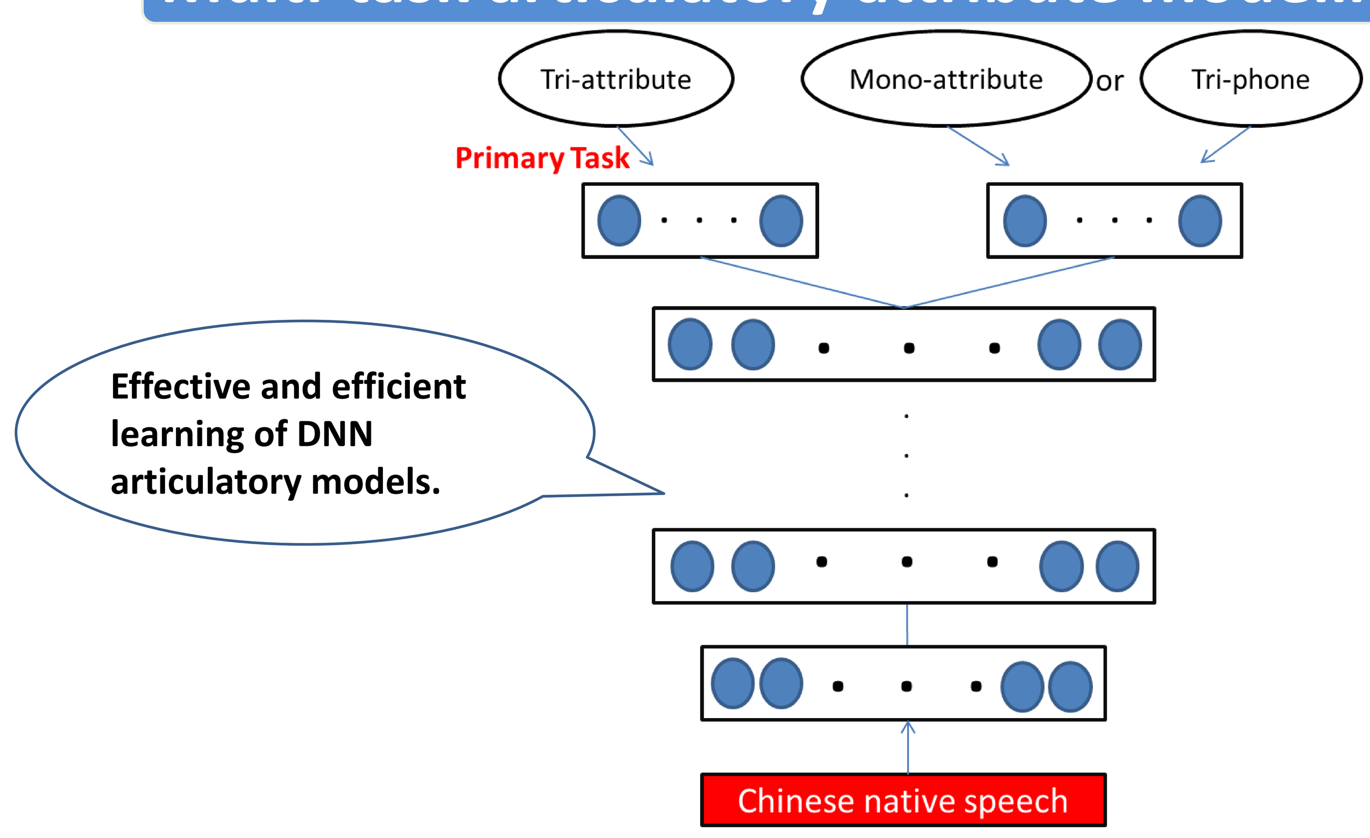
Articulatory attribute transcription

Sentence	你好（HELLO）					
Phone	sil	n	i	h	ao	sil
Manner	sil	nasal	vowel	unvoiced-fricative	vowel	sil
Place & Backness	sil	alveolar	anterior	velar	central back	sil
Place & Height	sil	alveolar	high	velar	low middle	sil
Place & Roundedness	sil	alveolar	unroundedness	velar	unroundedness roundedness	sil

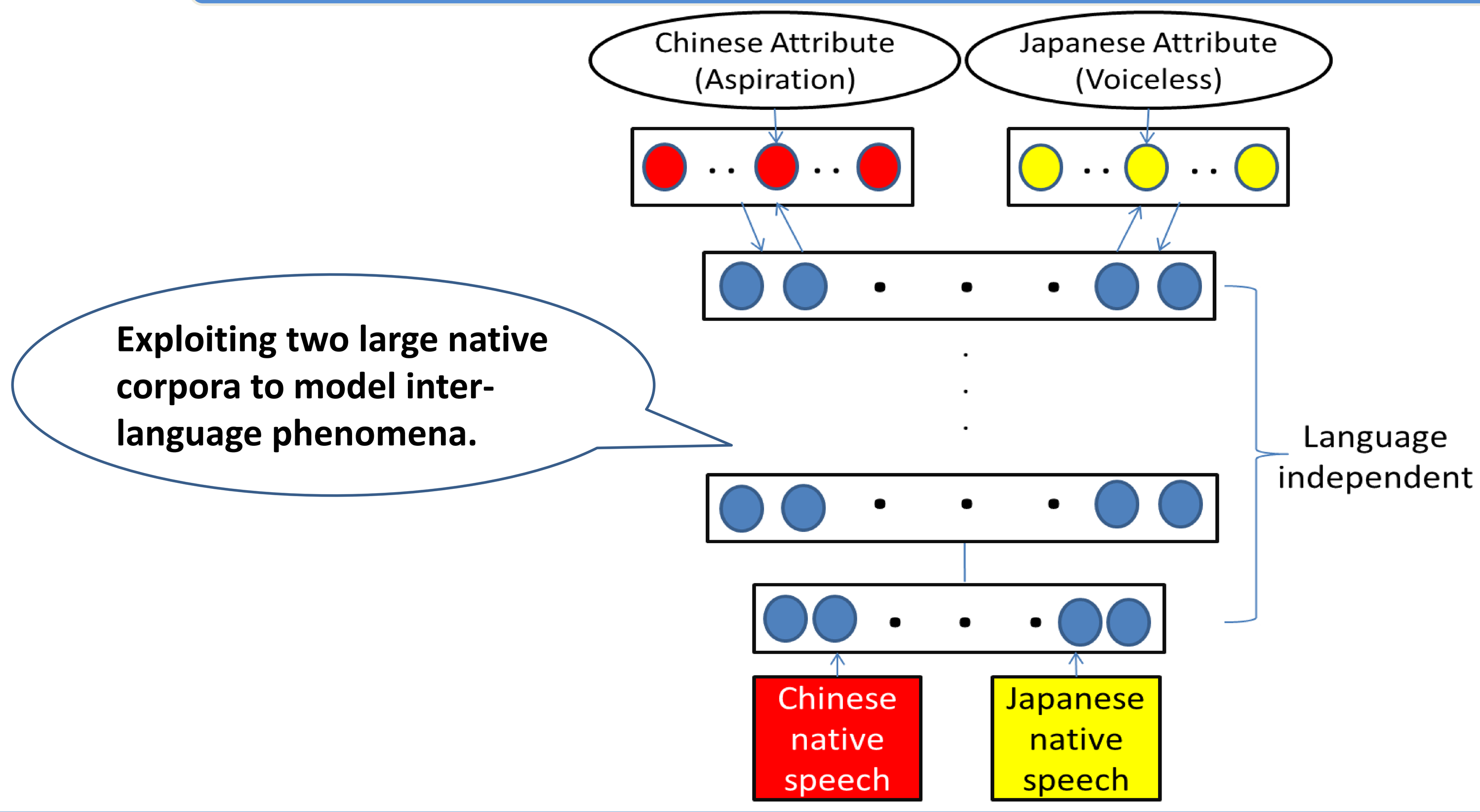
Modeling the golden attribute



Multi-task articulatory attribute modeling

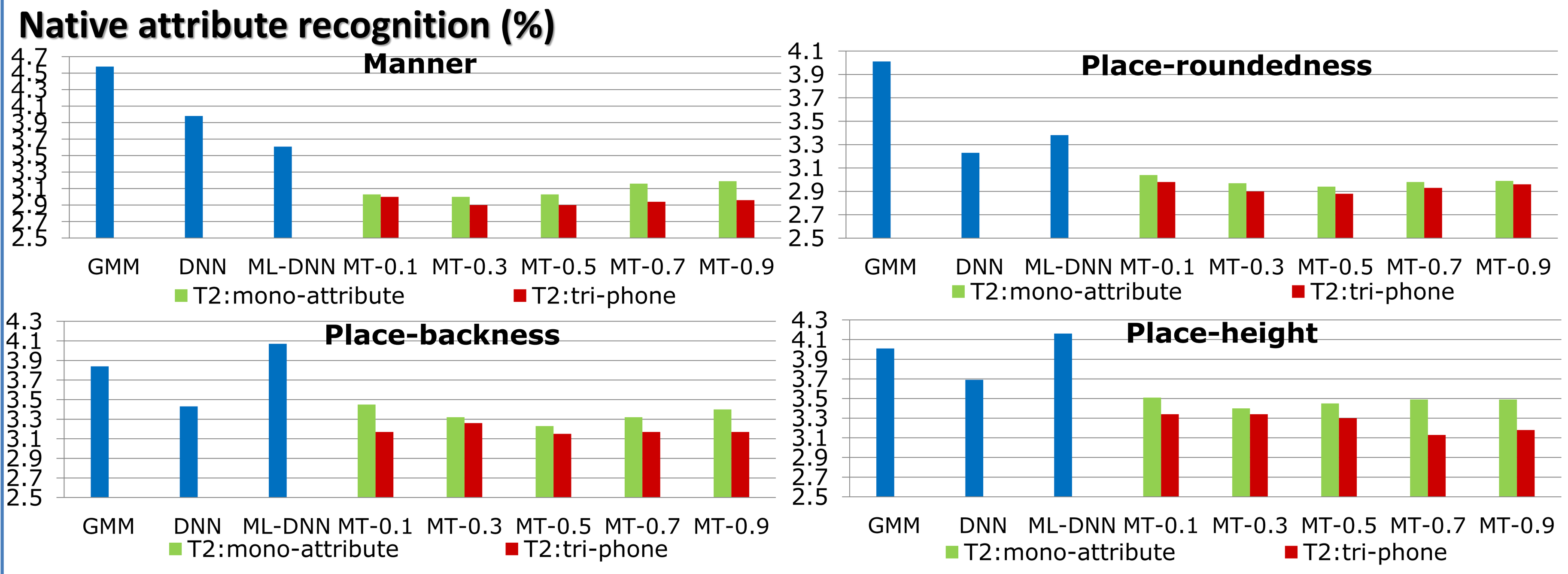


Multi-lingual articulatory attribute modeling

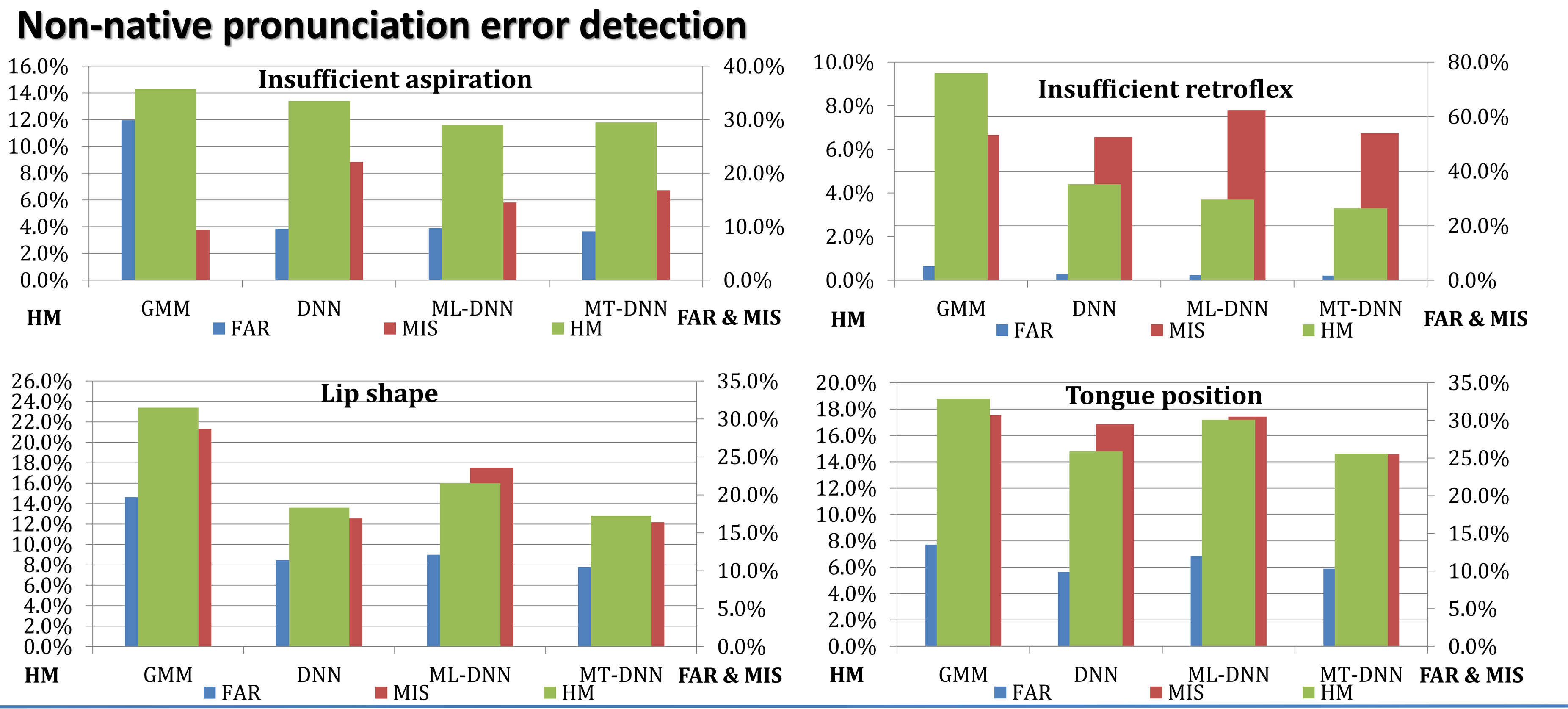


Experiments

Native attribute recognition (%)



Non-native pronunciation error detection



Experimental data

	Data set	Amount of data
Train	Chinese native training data	42h (28males, 36females)
	Japanese native training data	42h(80males, 73females)
Test	Chinese native testing data	5.3h (5males,3females)
	Chinese non-native testing data	1896 utterances (7 female Japanese students)

Error type focused

Insufficient aspiration: Insufficient aspiration when producing aspirated constants (e.g. p)

Insufficient retroflex: Insufficient retroflex when producing retroflex constants (e.g. r)

Lip roundedness: Vowels with spread lips have problems of rounded sound (e.g. ü)

Backness: Inappropriate tongue position with a little back (e.g. an)

Evaluation metrics

- *False Alarm Rate (FAR)* : rate of correct pronunciation that are detected as pronunciation pronunciation errors by the system.
- *Miss Rate (MIS)* : rate of true pronunciation errors that are missed detected by system.
- *Harmonic Mean (HM)*: harmonic mean of FAR and MIS.

Conclusions

We address effective articulatory models without non-native training data.

Multi-task learning method can enhance DNN articulatory modeling.

Multi-lingual learning method is effective for modeling non-native speech.