END-TO-END ARTICULATORY MODELING FOR DYSARTHRIC ARTICULATORY ATTRIBUTE DETECTION

Yuqin Lin[1] Longbiao Wang[1] Jianwu Dang[1,3] Sheng Li[2] Chenchen Ding[2]

[1]{linyuqin, longbiao$_{w}ang, jdang\}@tju.edu.cn$ [2]firstname.lastname@nict.go.jp

[1]1Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

1. Introduction

Background and Motivation

- Diseases affect speech articulation leading to unclear, inaccurate and unstable pronunciation.
- Resource of dysarthric speech is limited.
- The articulatory attribute describes the process of human speech production.
- The Automatic speech attribute transcription (ASAT) can assist patients in the treatment of pronunciation disorders.

In this paper, we:

- Present an end-to-end automatic speech attribute transcription (E2E-ASAT) system for dysarthric patients with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS).
  - Directly learns the mapping between acoustic features and articulatory attribute
- Investigate an effective method for dysarthric ASR and ASAT
  - Model refactoring

3. Proposed Method

Refactored Transformer-based Model for Low-resourced Data

- Pre-training of a well-performed ASR model with a large amount of English non-dysarthric speech.
- Refactor the network into fixed-layers and update-layers.
  - Parameters of the fixed-layers are copied from the pre-trained model
  - Only the update-layers are trained
  - Parameters are shared in the update-layers

4. E2E-ASAT for Dysarthric Speech

APL (For comparison)

- Transfers the phone sequences produced by the ASR system into articulatory attributes sequences.

E2E-ASAT

- The E2E-ASAT is based on the method introduced in Section 3.

7.Speech Recognition Evaluation

Table 3. Phone error rate (PER%) of all the methods

Methods	Training data	PER%
S1 (ft-full)	TORGO-trn-DS	66.54
S1 (ft-full, **baseline**)	TORGO-trn-(DS+NS)	**48.35**
S2 (+ DA)	TORGO-trn-(DS+NS) + Libri100	45.57
S3 (ft-decoder)	TORGO-trn-(DS+NS)	**39.53**
S4 (refactor)	TORGO-trn-DS	68.22
	TORGO-trn-DS (+sp)	62.29
	TORGO-trn-(DS+NS)	35.19
	TORGO-trn-(DS+NS) (+sp)	**31.03**
S5 (+ 8-sys. ROVER)	/	**27.13**

- The data augmentation (DA) is not so effective compared to other methods, not to say the large amount of training data causes massive training time.
- the DA is not so effective compared to other methods, not to say the large amount of training data causes massive training time.

2. Articulatory Representations

Articulatory Representations for English Sounds (transcribe phones into the articulatory attributes using the mapping rules)

Table 1. English consonant list with the manner (row) and place (column) attributes

	Labial (L)	Dental (*D*)	Alveolar (*R*)	Post-alveolar (*P*)	Palatal (*T*)	Velar (*V*)	Glottal (*G*)
Plosives (***p***)	p / b		t / d			k / g	
Affricates (***a***)				tʃ/ dʒ			
Nasals (***n***)	- / m		- / n			- / ŋ	
Fricatives (***f***)	f / v	θ/ ð	s / z	ʃ / ʒ			h / -
Approximants (***x***)				- / r	- / j	- / w	
Laterals (***l***)			- / l				

Phones beside / are: voiceless (***s***) / voiced (***v***). Both voiceless and voiced are voicing attributes.

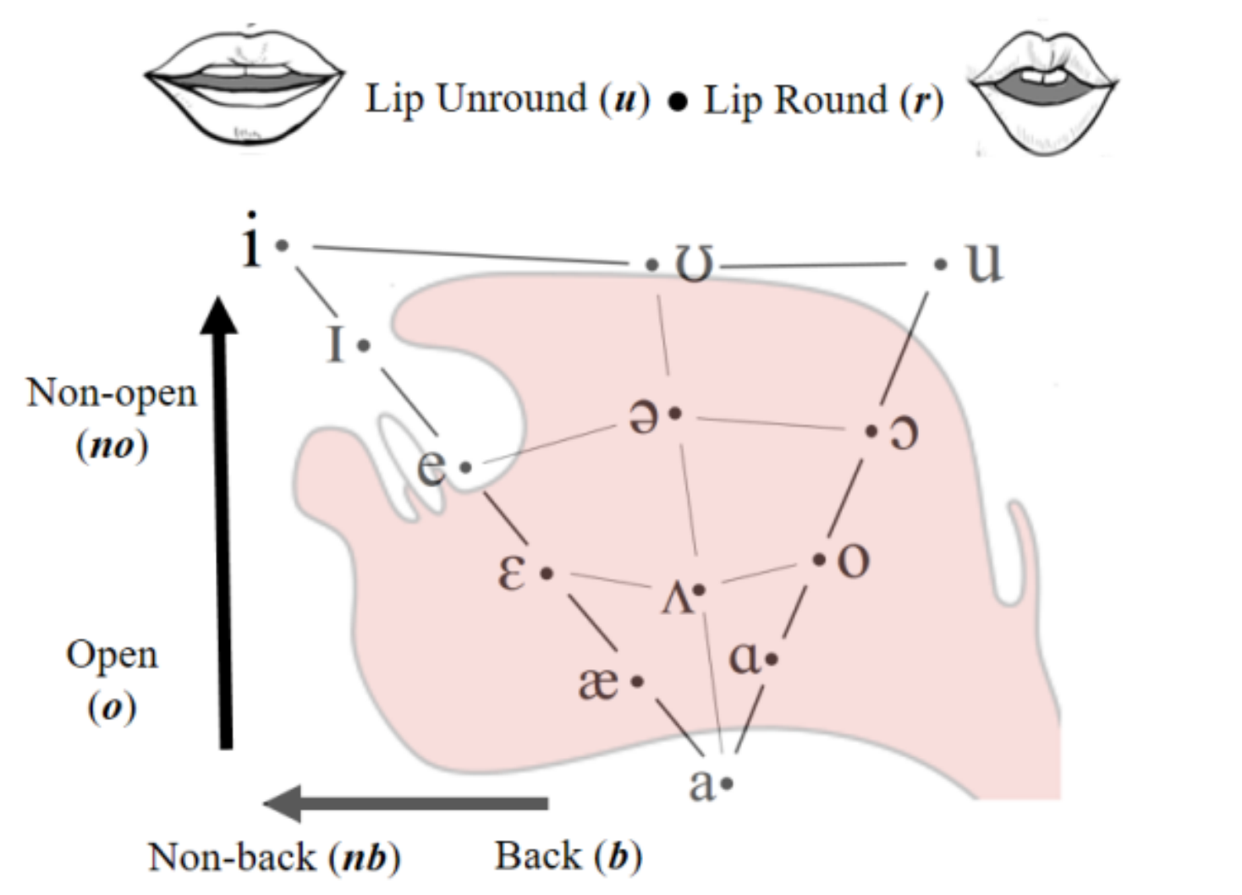


Fig. 1. Schematic diagram of English vowels with attributes

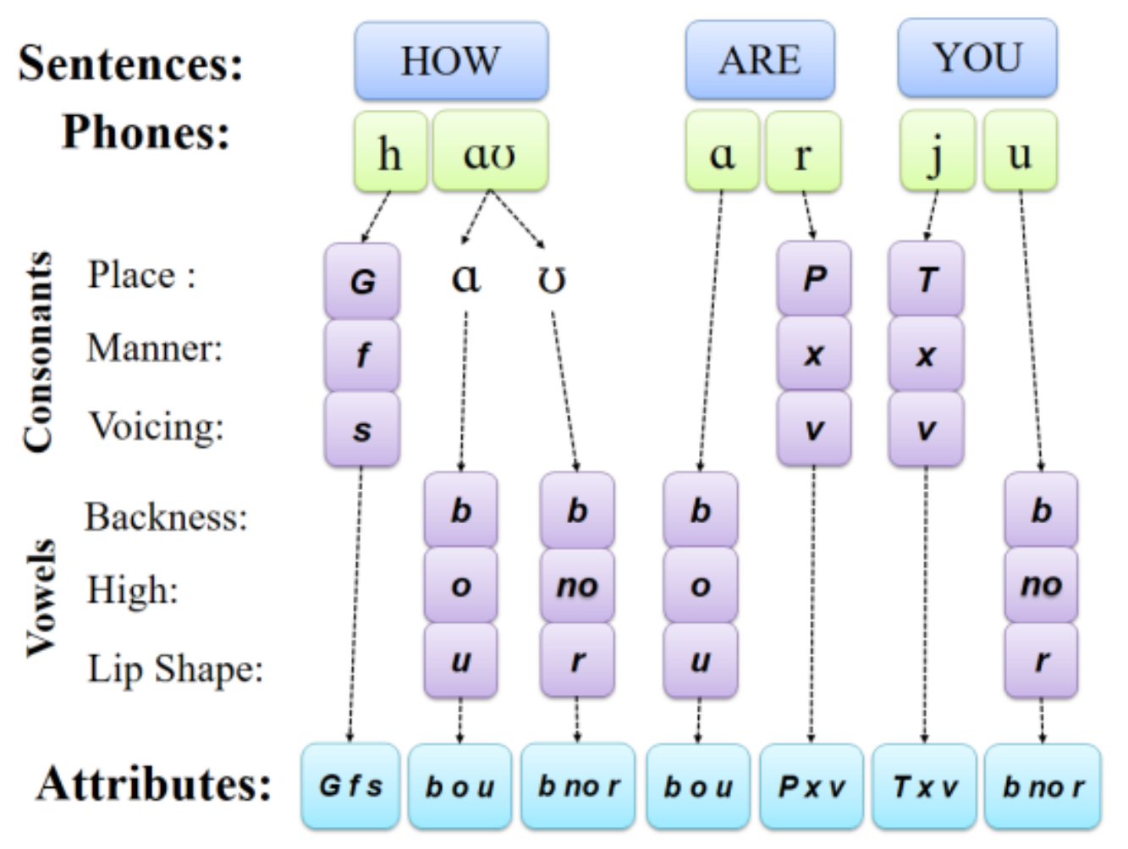


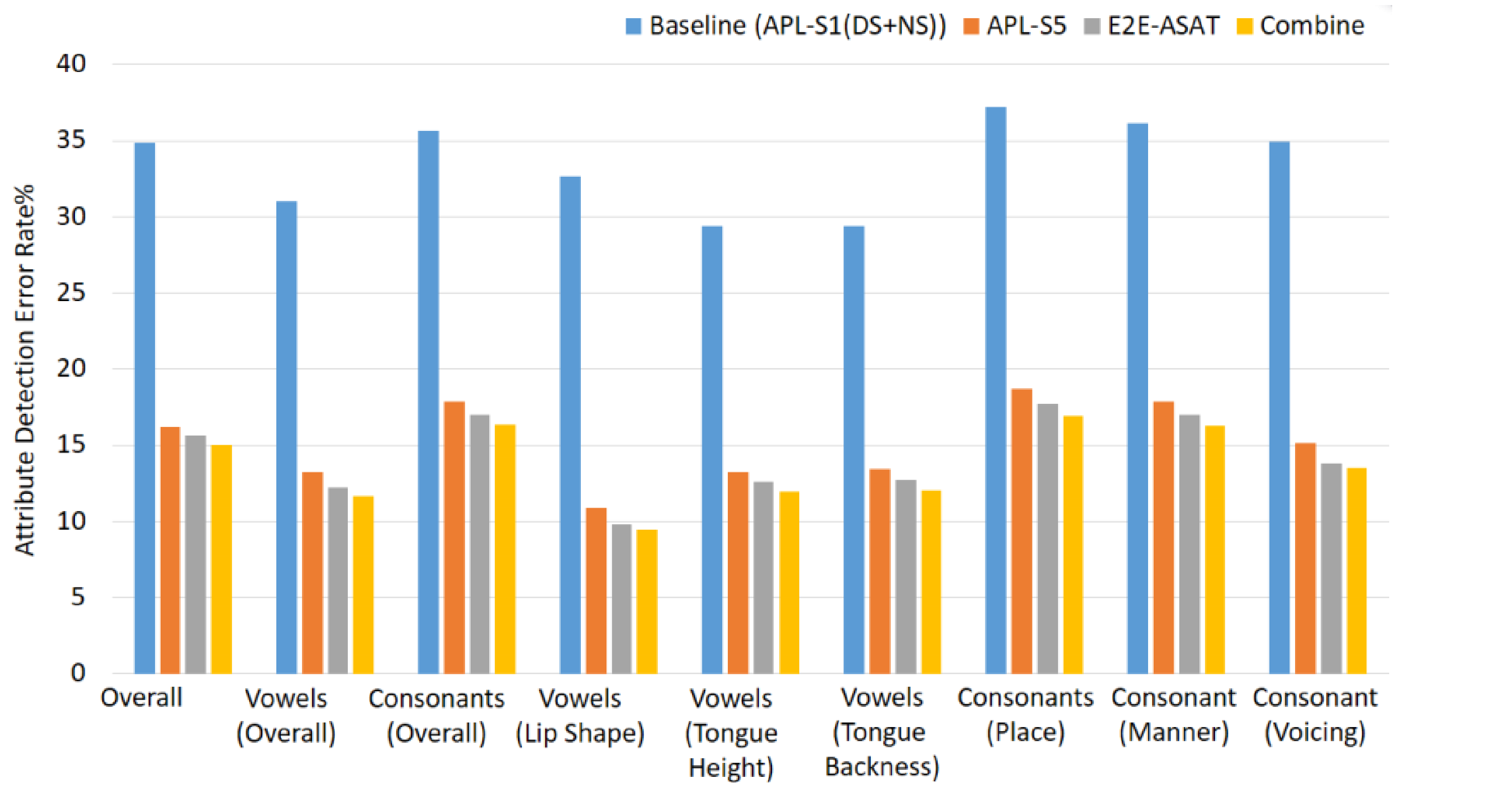
Fig. 2. An example of converting phones to articulatory representations: Glottal (***G***), Post-alveolar (***P***), Palatal (***T***), Fricatives (***f***), Approximants (***x***), Voiceless (***s***), Voiced (***v***), Back (***b***), Open (***o***), Non-open (***no***), Rounded (***r***), Unrounded (***u***)

5. Data Description

Table 2. English data set in dysarthric speech recognition (NS: non-dysarthric speech, DS: dysarthric speech)

	Dataset	Speech Type	Duration (Hours)	Speaker Num.	Utter. Num.
Training	Librispeech	NS	600	1256	63799
	TORGO-trn	NS+DS	6	8	6484
Testing	TORGO-tst	DS	1	3	1207

6. Attributes Detection Evaluation



8. Conclusions and Future Work

Conclusions

- An effective method for training E2E-ASAT system for articulatory attribute detection in patients with dysarthria.
  - Mapping directly within the single network
  - Effective and high precision

Future Work

- Build a concrete E2E-ASAT system for mispronunciation detection
- More dysarthric speech data