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]Tianjin 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

ABSTRACT

In this study, we focus on detecting articulatory attribute errors for dysarthric patients with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). There are two major challenges for this task. The pronunciation of dysarthric patients is unclear and inaccurate, which results in poor performances of traditional automatic speech recognition (ASR) systems and traditional automatic speech attribute transcription (ASAT). In addition, the data is limited because of the difficulty of recording. This study proposes an end-to-end automatic speech attribute transcription (E2E-ASAT) method for detecting articulatory attribute errors more precisely. To use the limited data more effectively, the parameters of the acoustic model are refactored into two layers and only one layer is retrained. Our proposed method showed good performances in both ASR and articulatory attribute detection. Our system has a potential as a rehabilitation tool.

Index Terms— end-to-end model, dysarthric speech recognition, articulatory attribute detection

1. INTRODUCTION

Dysarthria is a clinical category for neurogenic motor speech disorders that associate muscular weakness [1]. Different types of disorders are included in this category, of which cerebral palsy (CP) and amyotrophic lateral sclerosis (ALS) are the two of the most prevalent diseases [2]. CP is caused by cortical lesions, while ALS is due to motor neuron degeneration in the brain stem and spinal cord [3, 4]. These diseases affect speech articulation leading to unclear, inaccurate and unstable pronunciation [5]. Besides, dysarthria often accompanies physical disability, which suggests that speech could be a convenient alternative to remotely control keyboard or PC mouse and other machine interfaces [6]. However, traditional voice systems are not suitable for dysarthric patients because of the distorted voice and the limited dysarthric speech. Therefore, a specialized voice system for dysarthric patients is necessary.

The articulatory attributes describe phonetic properties of human speech according to positions or movements of the tongue, lips, and other organs to produce speech sounds [7]. In many related fields, articulatory information is utilized to assist their research, such as speech comprehension improvement [8], language learning [9, 10] and training of speech perception and production [11, 12]. In speech therapy, rehabilitation training for patients can be completed with the help of articulatory attributes detection [13].

Automatic speech attribute transcription (ASAT) is an automatic speech recognition (ASR) method based on buttomup attribute detection and knowledge integration. In previous studies [9, 10, 14, 15, 16, 17, 18, 19], many detectors are trained to produce a bank of articulatory attributes. These detectors are based on DNN-HMM systems, which requires context-dependent frame-level articulatory labels. To further improve the performance of ASAT, one method is to use end-to-end ASR systems with well performances. For example, automatic phone labeling (APL) is promising, which transfers the phone sequences produced by the ASR system into articulatory attributes sequences.

In this study, we present an end-to-end automatic speech attribute transcription (E2E-ASAT) system for dysarthric patients with CP or ALS. We investigate an effective method of articulatory attribute modeling for articulatory attribute detection, which directly learns the mapping between acoustic features and articulatory attribute based on a recent transformer-based E2E framework. To use the limited data more effectively, the parameters of the acoustic model are refactored into two layers, and only one layer is retrained. Compared with the traditional APL method, the E2E-ASAT has the advantages of high precision and convenience.

The rest of this paper is organized as follows. Section 2 describes our proposed method. Section 3 gives the data description and experiment evaluations. Conclusion and future work are given in Section 4.

*Corresponding author. Sheng Li is the joint first author.

2. PROPOSED METHOD

The proposed method of this paper has three components, and they are described in following subsections.

2.1. Articulatory Representations for English Sounds

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.

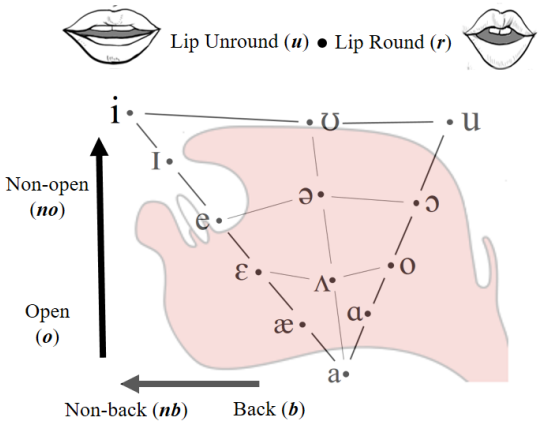

Fig. 1. Schematic diagram of English vowels with attributes

As remarked in Section 1, the articulatory attributes describe the place and manner of articulation with voicing contrasts. The attributes of consonants are determined by the place and manner of articulation, while those of vowels are determined by vertical (high/mid/low) and horizontal (front/central/back) positions of the tongue and the shapes of the lips [18].

In this paper, we transcribe phones into the articulatory attributes using the mapping rules (Table 1 and Fig. 1), which are made according to [7, 18, 20]. In these rules, each consonant has two manner attributes (manner of articulation and voicing), and one place attribute, and each vowel has three place attributes. Diphthongs are regarded as set of two monophthongs. Considering that the poor flexibility of the patient's tongue, we classify tongue-high and tongue-mid as non-open (***no***) attribute and classify the tongue-low as open (***o***) attribute. In addition, the front of the tongue is classified as non-back (***nb***) attribute, and the central/back of the tongue is classified as back (***b***) attribute. Fig. 2 is an example of converting phones to articulatory representations.

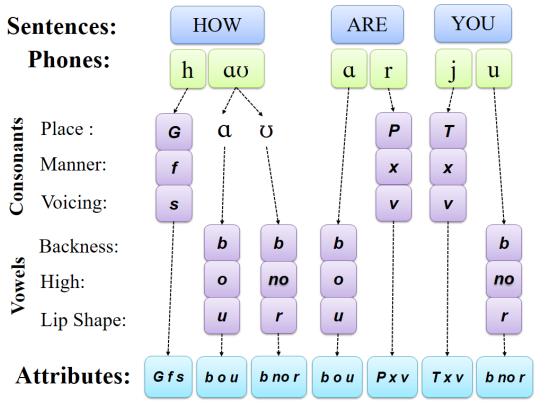

Fig. 2. An example of converting phones to articulatory representations: Glottal (***G***), Post-alveolar (***P***), Palatal (***T***), Fricatives (***f***), Approximants (***x***), Voiceless (***s***), Voiced (***v***), Back (***b***), Open (***o***), Non-open (***no***), Rounded (***r***), Unrounded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

2.2. Refactored Transformer-based E2E Model for Low-resourced Data

The transformer [21, 22] is a sequence-to-sequence attention-based model, which consists of an encoder and decoder. Each encoder and decoder has six blocks. Each block in the encoder contains a multi-head self-attention mechanism (MHA) and fully connected feed-forward network layers. Each block in the decoder has similar structure but with an extra masked MHA layer. The transformer has been demonstrated to have excellent ASR performance [22].

In dysarthric speech recognition, due to the limited training data, directly training a transformer-based model with a large number of parameters is not effective. We adopt the following steps for training. The first step is pre-training of a well-performed ASR model with a large amount of English non-dysarthric speech, which outputs English phone sequences. The next is to 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 during model training and their parameters are shared in the update-layers using the same method in [23]. After the experiments, we found that using the whole encoder as the fixed-layers and the whole decoder as the update-layers can achieve the best result.

2.3. E2E-ASAT for Dysarthric Speech

The E2E-ASAT is based on the method introduced in Section 2.2. As a comparison, several ASR systems for dysarthric

speech based on the same method are trained for APL.

3. EXPERIMENT EVALUATIONS

3.1. Data Description

We use the TORGO database [24] and Librispeech corpus [25] in this experiment. Speech files in the TORGO database are recorded by a microphone array and a head-worn microphone with a 16 kHz sampling rate. This database contains seven patients and seven healthy control speakers (4 males and 3 females). All the patients in this database are with CP or ALS.

Our models are pre-trained with a large amount of English non-dysarthric speech and fine-tuned with a small amount of English non-dysarthric speech and dysarthric speech. For pre-training, we use 500-hour non-dysarthric speech from Librispeech corpus. For fine-tuning, we use 2-hour dysarthric speech (2 males and 2 females) and 4-hour non-dysarthric speech (4 males) in the TORGO database (TORGO-trn). When evaluating the models, we used another 1-hour dysarthric speech (2 males and 1 female) from the TORGO database (TORGO-tst). All of these data sets are listed in Table 2.

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

We use 120-dim log Mel-filterbank energy features (40-dim static, $+\Delta$, and $+\Delta\Delta$), which were mean- and variance-normalized per speaker, and every four frames were spliced (three left, one current and zero right). The lower and higher cut-off frequencies are set to 20 Hz and 8000 Hz. In order to augment the training data, speed-perturbation [26] are used in the fine-tuning stage.

3.2. Model Training

All of our experiments employ implementation of the transformer based machine translation (NMT-Transformer) [21] in tensor2tensor[1]. The training and testing settings are similar to those in [7].

3.3. Speech Recognition Evaluation

As shown in Table 3, we train a series of systems (with method S1 to S5) and evaluate their performance of ASR with the phone error rate (PER%). These systems use 500 hours of non-dysarthric speech from librispeech database for pre-training. The settings are listed as follows:

- S1 (baseline): The full net is fine-tuned (ft-full) using of TORGO-trn dysarthric speech (TORGO-trn-DS) and TORGO-trn non-dysarthric speech (TORGO-trn-NS).
- S2: Based on S1, another 100 hours non-dysarthric speech speech from Librispeech corpus (Libri100) are added for data augmentation (DA).
- S3: Based on S1, only the decoder is fine-tuned (ft-decoder).
- S4: All the TORGO-trn speech is used with speed-perturbation (sp). The network is refactored.
- S5: Based on S4, 8 systems listed from S1 to S4 are combined with ROVER [27].

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**

From the results in Table 3, we observe that the DA is not so effective compared to other methods (especially in ft-decoder), not to say the large amount of training data causes massive training time. The parameter refactoring is a more effective method compared with traditional method (fine-tune full/part net or data augmentation).

3.4. Articulatory Attributes Detection Evaluation

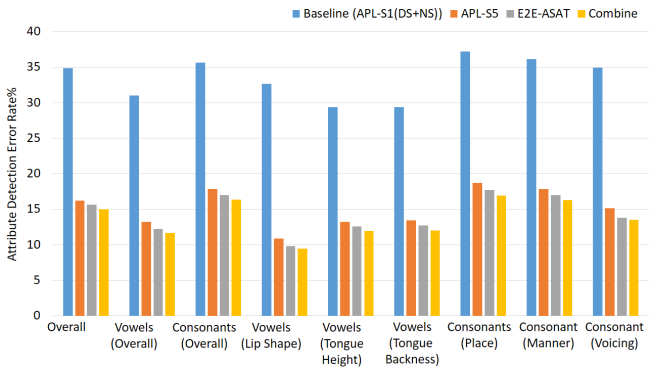

Fig. 3. Detection error rate (DER%) of individual attribute types for the three systems

[1]https://github.com/tensorflow/tensor2tensor

In Fig. 3, we compared the overall and individual attribute detection error rates (DER%) of APL and E2E-ASAT introduced in Section 2.3, as well as the DER% of the system that combines APL-S5 and E2E-ASAT systems with ROVER. APL-S5 means mapping the phones produced by S5-based ASR system to articulatory attributes. As expected, the proposed E2E-ASAT method significantly outperforms APL. As for the rest, the combined system performs better than either individual system in articulatory attribute detection.

Fig. 4 to Fig. 6 are the normalized confusion matrices of vowel and consonant attributes from the combined attribute detection system. The asterisk (*) in the figures indicates the blank. The row with the asterisk indicates insertion errors, and the column with the asterisk indicates deletion errors. From these confusion matrices, we observe that the E2E-ASAT is more accurate than the traditional methods in articulatory attribute detection.

Furthermore, we obtain the following summaries:

1) The consonant detection error rate (DER%) is highest in both affricates (***a***) and laterals (***l***) for the manner of articulation, and it is highest in both dentals (***D***) and glottals (***G***) for the place of articulation. In other words, these attributes are more laborious for dysarthric patients.
2) For voicing of consonants, we observe from the experiments that it is easier for patients to make voiceless consonants (***s***), except for glottal consonants (***G***).
3) Vowels and some consonants are all closely related to the movement of the tongue. As for the place of articulation, the DER% of the consonants is lowest in alveolar (***R***) and post-alveolar (***P***), and the DER% of back (***b***) vowels are higher than that of non-back (***nb***) vowels. It means that dysarthric patients tend to articulate with centralized tongue positions, and it is difficult for them to produce sounds with extreme front or back positions of the tongue (*i.e.*, back vowels (***b***), dental consonants (***D***), palatal consonants (***T***), velar consonants (***V***)).

Above all, these findings can be used for mispronunciation detection in patients with dysarthria.

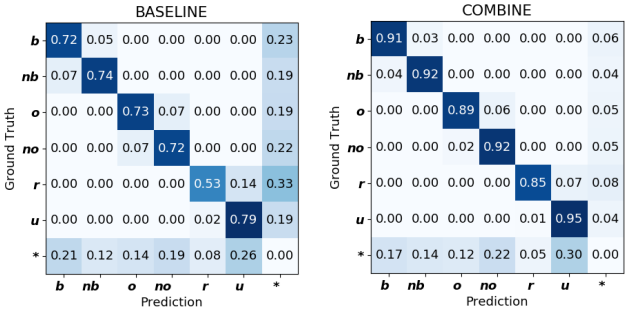

Fig. 4. Confusion matrices of vowels attribute: Open (***o***), Non-open (***no***), Back (***b***), Non-back (***nb***), Rounded (***r***), Unrounded (***u***), Blank (*)

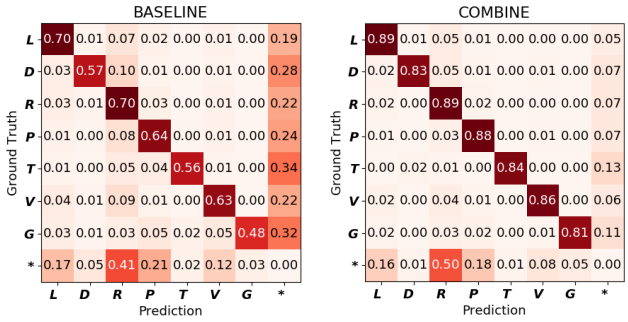

Fig. 5. Confusion matrices of consonants with place attributes: Labial (***L***), Dental (***D***), Alveolar (***R***), Post-alveolar (***P***), Palatal (***T***), Velar (***V***), Glottal (***G***), Blank (*)

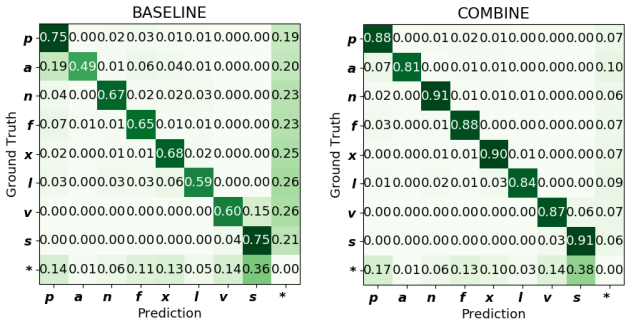

Fig. 6. Confusion matrices of consonants with manner attributes: Plosives (***p***), Affricates (***a***), Fricatives (***f***), Nasal (***n***), Approximants (***x***), Laterals (***l***), Voiced (***v***), Voiceless (***s***), Blank (*)

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed an effective method for training E2E-ASAT system for articulatory attribute detection in patients with dysarthria. Different from traditional APL method, we built acoustic-to-articulatory mapping directly within the transformer based E2E framework. The attribute detection accuracy of our proposed E2E-ASAT model significantly outperformed that of the traditional method. Furthermore, the combined system achieved the highest accuracy and has a potential to assist patients in rehabilitation training. In the future, we will build a concrete E2E-ASAT system for mispronunciation detection with more dysarthric speech data.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61771333, the National Key RD Program of China under Grant 2018YFB1305200, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330. Sheng Li is the joint first author, and he is partially supported by JSPS KAKENHI No. 19K24376 and NICT tenure-track startup fund. Chenchen Ding is partially supported by NICT tenure-track startup fund.

6. REFERENCES

[1] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.

[2] R. Kent and K. Rosen, "Motor control perspectives on motor speech disorders," *Speech motor control in normal and disordered speech*, pp. 285–311, 2004.

[3] K. Pannek, R. N. Boyd, S. Fiori, A. Guzzetta, and S. E. Rose, "Assessment of the structural brain network reveals altered connectivity in children with unilateral cerebral palsy due to periventricular white matter lesions," *NeuroImage: Clinical*, vol. 5, pp. 84–92, 2014.

[4] J. Morris, "Amyotrophic lateral sclerosis (als) and related motor neuron diseases: an overview," *The Neurodiagnostic Journal*, vol. 55, no. 3, pp. 180–194, 2015.

[5] R. Kent, "Research on speech motor control and its disorders: A review and prospective," *Journal of Communication disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[6] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. the Workshop on Machine Learning for Assistive Technologies (in the 24th NIPS)*, 2010, pp. 70–78.

[7] S. Li, C. Ding, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "End-to-End articulatory attribute modeling for low-resource multilingual speech recognition," *Proc. INTERSPEECH*, pp. 2145–2149, 2019.

[8] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, pp. 493–503, 2010.

[9] W. Li, K. Li, S. Siniscalchi, N. Chen, and C. Lee, "Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees." in *Proc. INTERSPEECH*, 2016, pp. 3127–3131.

[10] W. Li, S. Siniscalchi, N. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *Proc. IEEE-ICASSP*, 2016, pp. 6135–6139.

[11] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," in *Proc. the 4th International Conference on Universal Access in Human Computer Interaction*, vol. 4554, 2007, pp. 786–794.

[12] L. Wang, H. Chen, S. Li, and H. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845—856, 2012.

[13] H. Strik, "ASR-based systems for language learning and therapy," in *Proc. IS-ADEPT*, 2012, pp. 9–14.

[14] C.-H. Lee, M. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in *Proc. INTERSPEECH*, 2007, pp. 1825—1828.

[15] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. ICSLP*, vol. 4, 2004.

[16] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Proc. INTERSPEECH*, 2006, pp. 1287—1291.

[17] C.-Y. Lin and H.-C. Wang, "Attribute-based Mandarin speech recognition using conditional random fields," in *Proc. INTERSPEECH*, 2007, pp. 1833—1836.

[18] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning," *IEICE Trans. Information and Systems*, vol. 100, no. 9, pp. 2174–2182, 2017.

[19] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features," in *Proc. INTERSPEECH*, pp. 3885–3889.

[20] S. Li and L. Wang, "Cross linguistic comparison of Mandarin and English EMA articulatory data," in *Proc. INTERSPEECH*, 2012, pp. 903–906.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *CoRR abs/1706.03762*, 2017.

[22] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE-ICASSP*. IEEE, 2018, pp. 5884–5888.

[23] S. Li, R. Dabre, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation," in *Proc. INTERSPEECH*, 2019, pp. 1408–1412.

[24] F. Rudzicz, A. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE-ICASSP*, 2015, pp. 5206–5210.

[26] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, pp. 3586–3589.

[27] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997, pp. 347–354.