# Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription

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

longbiao_wang@tju.edu.cn, jdang@jaist.ac.jp

## Abstract

This study proposes a staged knowledge distillation method to build End-to-End (E2E) automatic speech recognition (ASR) and automatic speech attribute transcription (ASAT) systems for patients with dysarthria caused by either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). Compared with traditional methods, our proposed method can use limited dysarthric speech more effectively. And the dysarthric E2E-ASR and ASAT systems enhanced by the proposed method can achieve 38.28% relative phone error rate (PER%) reduction and 48.33% relative attribute detection error rate (DER%) reduction over their baselines respectively on the TORGO dataset. The experiments show that our system offers potential as a rehabilitation tool and medical diagnostic aid.

**Index Terms**: knowledge distillation, dysarthric speech recognition, articulatory attribute detection, End-to-End

## 1. Introduction

The development of artificial intelligence contributes to improving the quality of our lives. In recent years, automatic speech recognition (ASR) systems have become popular and widely used for services such as personal assistance on smartphones, home control via smart speakers, etc. People free their hands and eyes and make their lives more convenient by replacing them with voice. Obviously, these services are icing on the cake for common people. However, dysarthria patients cannot benefit from any of today's voice assistance services. Dysarthria is a category of motor speech disorder associated with muscular weakness [1, 2], including many different types of disorders such as cerebral palsy (CP) [3]. This condition is caused by disruptions of the motor-speech system and characterized by poor pronunciation [4, 5, 6]. Moreover, the diseases causing potential physical disabilities make patients' lives inconvenient. The need for an ASR system for faster communication and easier access to services is suggested [7]. Therefore, a specialized dysarthria intelligent voice system must provide timely assistance for patients.

ASR is one of the most prevalent intelligent voice systems. There are two challenges to the study of automatic dysarthric speech recognition (dysarthric ASR). First, the patients' pronunciation differs from that of healthy people due to their speech disorders, and these differences lead to personalized manners of pronunciation [5, 8]. Hence, the ASR systems for healthy people are not suitable for dysarthria patients. Furthermore, muscular weakness causes the patients to overburden their speech muscles during pronunciation [5, 9]. For this reason, the limited resources of dysarthric speech make it difficult to train an effective dysarthric ASR system.

Automatic speech attribute transcription (ASAT) is another useful intelligent voice system. It can be used for patient rehabilitation training [10]. Articulatory attributes describe speech production features. In previous studies, although many detectors have been studied to generate attributes, they still face the challenge of low resources, as with dysarthric ASR. Lin et al. [11] reports that using well-performing End-to-End (E2E) systems is one effective way to further improve the performance of ASAT.

In recent years, two main approaches have been utilized to overcome the limited resources challenge. One is voice conversion (VC), which transforms dysarthric speech into normal speech for data augmentation [12, 13], and the other is to improve the training strategy for limited ASR resources, such as by using multiple databases [5], by joint articulatory and acoustic features [14], by using finetuning techniques [15], and so on. However, disadvantages still exist for these state-of-the-art approaches: VC requires sufficient dysarthric speech data, and the improved training strategies are usually effective for voice systems with limited resources, but not for dysarthric voice systems. In the latest study, staged training has been shown to be effective for dysarthric speech [16].

This paper focus on patients with CP and amyotrophic lateral sclerosis (ALS), which are two of the most prevalent causes of dysarthria [3]. The acoustic model is based on a recent competitive End-to-End ASR framework called Speech Transformer [17]. The present work investigates the knowledge distillation method [18] and proposes staged conditional teacher-student learning to overcome the challenge of limited resources. The proposed knowledge distillation-based E2E Modeling is applied to enhance the dysarthric ASR and ASAT tasks.

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

## 2. Knowledge Distillation-based E2E Modeling for Dysarthric ASR and ASAT

### 2.1. Knowledge Distillation-based Dysarthric E2E-ASR

Speech Transformer [17, 19] is a sequence-to-sequence attention-based model and it has been demonstrated to perform effectively in ASR [19]. The staged teacher-student learning method is proposed to overcome the limited resource challenge

*Corresponding author.

mentioned in Section 1.

Teacher-student learning/knowledge distillation is a compression framework [20]. It trains a compact student network using the output of a high-performance teacher network as soft labels for knowledge transfers. The student network can explore not only the information provided by the ground truth but also the knowledge learned by the teacher network.

Given input speech features $\mathrm{X} = \{\mathrm{x}_1, ...\mathrm{x}_L\}$ with $L$ length, and ground truth $\mathrm{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_N\}$ with $N$ length, the teacher network is trained by optimizing the loss between the ground truth Y and the output softmax of the teacher $\mathrm{O}^t = \{\mathrm{o}_1^t, ..., \mathrm{o}_N^t\} \in \mathbb{R}^{N\times D}$. $D$ is the number of target classes.

In conditional teacher-student learning [18], the student network is trained to learn from selected labels, that is made up of the ground truth (hard labels) Y, and the outputs softmax of the teacher network (soft labels) $\mathrm{O}^t$. The loss function is defined between the selected labels and the outputs softmax of the student network (predicted label) $\mathrm{O}^s = \{\mathrm{o}_1^s, ..., \mathrm{o}_N^s\}$ $\mathrm{o}_i^s \in \mathbb{R}^{N\times D}$. However, this approach is ineffective when the resources of data are limited as in dysarthric speech. To make a full use of limited data resources, a staged training strategy is adopted. The latest study in [16] shows staged training (first adapted to multiple dysarthric speakers, and then to the target speaker) is effective for dysarthric speech. Different from [16], the student model in our method is first adapted to the mixture of multiple dysarthric and common speakers, and then the adapted model is further adapted for the target dysarthric speakers.

The selected labels are defined differently at different stages. In the first stage, the selected labels are made up of the hard labels Y and soft labels $\mathrm{O}^t$. In the second stage, the selected labels are made up of the hard labels Y and predicted labels $\mathrm{O}^s$. The specific definition is as follows:

$$\tilde{\mathrm{y}}_i(\mathrm{o}_i) = \begin{cases} \mathrm{o}_i, & \underset{j\in\{1,2,..,D\}}{\arg\max}\ \mathrm{o}_{i,j} = \underset{k\in\{1,2,..,D\}}{\arg\max}\ \mathrm{y}_{i,k} \\ \mathrm{y}_i, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathrm{o}_i$ represents the $i$-th output softmax of the teacher network or student network. That is to say, the student at first learns knowledge from both the teacher and ground truth and then focuses on the more difficult aspects when it has learned most of the knowledge.

A boundary value $\lambda$ is introduced to divide the two stages. The training process enters the second stage when the prediction accuracy of the student network is higher than $\lambda$. In brief, the student network is trained to optimize the following loss function:

$$\mathcal{L}_{TS} = \begin{cases} -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D}\tilde{\mathrm{y}}_{i,j}(\mathrm{o}_i^t)\log \mathrm{o}_i^s, & \mathrm{acc} \le \lambda \\ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D}\tilde{\mathrm{y}}_{i,j}(\mathrm{o}_i^s)\log \mathrm{o}_i^s, & \text{otherwise} \end{cases} \tag{2}$$

$$\mathrm{acc}_i = \begin{cases} 1, & \underset{j\in\{1,2,..,D\}}{\arg\max}\ \mathrm{o}_{i,j}^s = \underset{k\in\{1,2,..,D\}}{\arg\max}\ \mathrm{y}_{i,k} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\mathrm{acc} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{acc}_i \tag{4}$$

where $\lambda$ is the tunable tradeoff parameter.

### 2.2. Knowledge Distillation-based Dysarthric E2E-ASAT

Table 1: *English consonant list with 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 | | | | |

Phonemes beside/are: -Voiced (***s***) / +Voiced (***v***). Both -Voiced and +Voiced are voicing attributes.

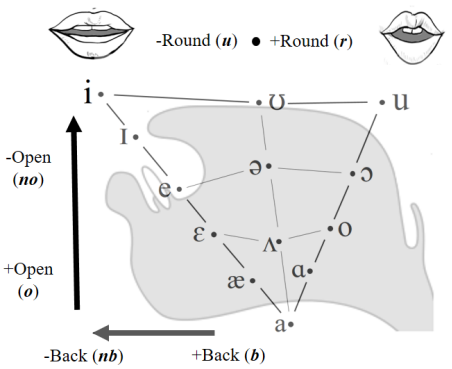


Figure 1: *Schematic diagram of English vowels with attributes*

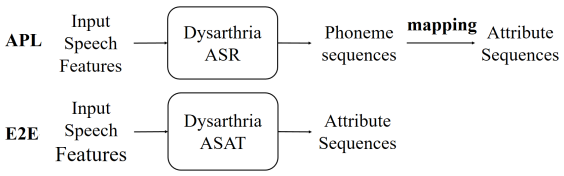


Figure 2: *Flowcharts of two E2E-ASAT methods (APL and E2E)*

In this paper, the proposed knowledge distillation-based modeling in Subsection 2.1 is also applied to enhance the dysarthric E2E-ASAT methods proposed in [11].

As a brief review, phonemes are transcribed into the articulatory attributes by using the mapping rules according to [11, 21, 22, 23]. Table 1 and Figure 1 show the mapping rules for consonants and vowels, respectively. In these rules, each consonant has two manner attributes (manner of articulation and voicing), along with one place attribute, and each vowel has three place attributes. Diphthongs are regarded as sets of two monophthongs. Considering the poor mobility of the patient's tongue, we classify tongue-high and tongue-mid as -Open (***no***) attribute and classify tongue-low as +Open (***o***) attribute. In addition, the front positioning of the tongue is classified as -Back

(***nb***) attribute, and the central or back positioning is classified as +Back (***b***) attribute.

In order to detect the articulatory attributes of patients' speech, two E2E-ASAT methods are proposed in [11], as shown in Figure 2.

1. APL approach: Reuse a well-performed phone-level ASR system. The phone-level ASR result is mapped to articulatory attributes according to the phoneme-attribute mapping rules in Table 1 and Figure 1.
2. E2E approach: Directly map the acoustic features to articulatory attributes.

In this paper, the proposed knowledge distillation-based modeling is also applied to enhance these two dysarthric E2E-ASATs.

# 3. Experiments

## 3.1. Experimental Setups

The TORGO database [24] and Librispeech corpus [25] are used in this experiment. All the data sets are for the English speech recognition task. Speech files in the TORGO database are recorded by a microphone array and a head-worn microphone with a 16kHz sampling rate. This database contains the speech data obtained from seven patients and seven healthy control speakers (4 males and 3 females). All the patients in this database were diagnosed with varying degrees of CP or ALS. The Librispeech corpus we use in this paper contains approximately 600 hours of English speech read by healthy speakers and sampled at a rate of 16kHz.

All the models are pre-trained with 500-hour normal speech from the Librispeech corpus. For retraining, due to the limited resources of dysarthric speech, we use all speech samples in the TORGO database, which contains 2 hours of dysarthric speech (2 males and 1 female) and 4 hours of normal speech (4 males). When evaluating the models, we used another 1 hour of dysarthric speech (2 males and 2 females) from the TORGO database. The other 100-hour normal speech from the Librispeech corpus (100h-Libri) is used for data augmentation.

Table 2: *Data set for dysarthric speech recognition*

| Dataset | Speech | Hours | Utterances |
|---|---|---|---|
| Training | Librispeech (Libri) | 600 | 63,799 |
| | TORGO (T-train) | 6 | 6,484 |
| Testing | TORGO (T-test) | 1 | 1,207 |

The input features are 120-dimensional log Mel-filterbank energy features (40-dim static, $+\Delta$, and $+\Delta\Delta$) Each feature was mean- and variance-normalized, and every four frames were spliced (three left, one current and zero right). The lower and higher cutoff frequencies were set to 20 Hz and 8,000 Hz, respectively. To augment the training data, the standard 3-way speed-perturbation with factors of 0.9, 1.0 and 1.1 [26] was used in the fine-tuning stage.

All of the experiments are used for the implementation of the Transformer-based machine translation (NMT-Transformer) [27] in tensor2tensor [1]. The training and testing settings are similar to those in [21].

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

## 3.2. Speech Recognition Evaluation

In Table 3, the effectiveness of the proposed staged conditional teacher-student method (TS2) for limited data resources are shown, together with a series of the systems (S1 to S4) and their ASR performance based on the phoneme error rate (PER%) for comparison. Additionally, the conditional teacher-student learning proposed by [18] (TS1) is used for comparison.

S1: The full net is fine-tuned (ft-full).

S2: The data augmentation with 100h-Libri (+DA) are adopted, and the full net is fine-tuned.

S3: Only the decoder is fine-tuned (ft-decoder).

S4: The model is refactored (refactor) [11]. The layers in the decoder are shared [28], and the parameters of the shared layers are fine-tuned by using all the speech data from the TORGO database with speech perturbation (sp).

TS1: The conditional teacher-student learning proposed by [18] (TS1). The teacher model is a data-augmentation model (S2) or refactoring model (S4), and the student model is a refactored model trained by using all the speech from the TORGO database with speed perturbation (sp).

TS2: The proposed staged teacher-student learning (TS2). The same teacher, student model, and training data are used as TS1. The tunable tradeoff parameter $\lambda$ is set to 0.95.

Table 3: *Phone error rate (PER%) of all methods for limited data resources*

| Methods | Training data | PER% |
|---|---|---|
| S1 (ft-full, **baseline**) [11] | T-train | 48.35 |
| S2 (ft-full + DA) [11] | T-train + 100h-Libri | 45.57 |
| S3 (ft-decoder) [11] | T-train | 39.53 |
| S4 (Refactor) [11] | T-train | 35.19 |
| | T-train (+sp) | **31.03** |
| TS1-DA (TS1 + Teacher:S2) | T-train (+sp) | 30.76 |
| TS1-R (TS1 + Teacher:S4) | T-train (+sp) | 32.40 |
| TS2-DA (TS2 + Teacher:S2) | T-train (+sp) | **29.84** |
| TS2-R (TS2 + Teacher:S4) | T-train (+sp) | 31.42 |

From the results (S1 to S4) in Table 3, the DA is not as effective when compared with other methods (especially the ft-decoder) because the large quantity of training data requires time. The model-refactoring method (S4) is a more effective method than the traditional method (S1 to S3). And the proposed staged conditional teacher-student learning method (TS2-DA) is relatively improved by 38.28% compared with the baseline (S1) and by 2.99% compared with traditional conditional teacher-student learning methods (TS1-DA).

Experimental results in Table 4 prove the effectiveness of the proposed method by evaluating the phone error rate (PER%) for the four patients (F01, F02 are female; M01, M02 are male.) included in the test set. Further more, the significant difference between the baseline method (S1) and the proposed method (TS2-DA) at the level of 0.01 are shown.

Table 4: *Phone error rate (PER%) of both baseline (S1) and the proposed method (TS2-DA) on four patients.* ***Bold*** *means there is no significant difference between the performance of the two methods, otherwise there is a significant difference between that of these two methods.*

| | Errors type | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patients | Overall | | Substitution | | Insertion | | Deletion | |
| F01 | 50.17 | 26.89 | **21.07** | **10.31** | 9.49 | 9.76 | 19.60 | 6.80 |
| F02 | 33.59 | 15.62 | 14.81 | 4.80 | 7.04 | 4.92 | 11.73 | 5.88 |
| M01 | 56.06 | 35.98 | 24.17 | 18.32 | 6.51 | 8.73 | 25.36 | 8.92 |
| M02 | 50.92 | 28.32 | **23.83** | **12.48** | 10.37 | 9.10 | 16.71 | 6.72 |

PERs beside | are: baseline (S1) | proposed (TS2-DA).

Table 5: *Detection error rate (DER%) of individual attribute types for the five systems*

| Method | Overall | Vowels | Consonants |
|---|---|---|---|
| APL-baseline | 34.86 | 31.00 | 35.64 |
| APL-S4 | 19.35 | 15.06 | 20.49 |
| APL-TS | 18.01 | 14.24 | 19.60 |
| E2E-S4 | 19.25 | 14.93 | 20.30 |
| E2E-TS | 19.70 | 15.45 | 20.78 |
| Combined | 14.35 | 10.88 | 15.77 |

### 3.3. Evaluation of Articulatory Attribute Detection

Table 5 compares the overall attribute detection error rates (DER%) of two E2E-ASATs (APL and E2E introduced in Subsection 2.2) as well as the DER% of the system that combines APL and E2E systems with ROVER [29]. All the models are trained with the model-refactoring approach for low-resourced data and proposed staged teacher-student (TS2) learning methods. The APL-baseline and APL-S4 map phonemes produced by S1 and S4 with the whole TORGO database (+sp), respectively. The proposed TS2 method (APL-TS) improved the APL-baseline model by 48.33% of relative DER%. The combined system performs better than any individual system, and is used for the study the patterns of patient pronunciation in what follows.

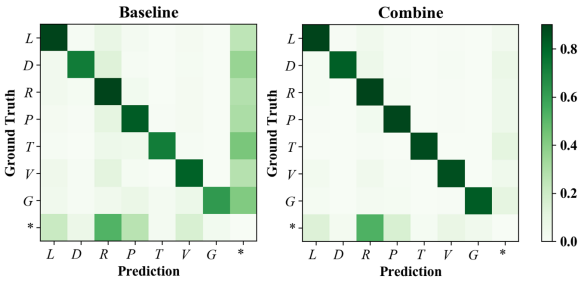


Figure 3: *Confusion matrices of consonants with place attributes: Labial (***L***), Dental (***D***), Alveolar (***R***), Post-alveolar (***P***), Palatal (***T***), Velar (***V***), Glottal (***G***), Blank (*)*

Figure 3 and Figure 4 are the normalized confusion matrices of consonant attributes from the combined attribute detection system. The asterisk (*) in the figures indicates the blank.

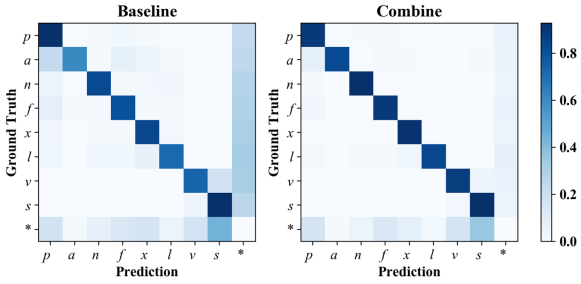


Figure 4: *Confusion matrices of consonants with manner attributes: Plosives (***p***), Affricates (***a***), Fricatives (***f***), Nasal (***n***), Approximants (***x***), Laterals (***l***), Voiced (***v***), Voiceless (***s***), Blank (*)*

The row with the asterisk indicates insertion errors, and the column with the asterisk indicates deletion errors. From these confusion matrices, the proposed TS2 enhanced E2E-ASAT (combined in Table 5) is more accurate than the traditional methods in articulatory attribute detection (APL-baseline in Table 5).

Furthermore, we obtain the following summaries:

1) The sounds with dentals (***D***) and glottals (***G***) are more laborious for dysarthric patients.

2) Affricates (***a***) and laterals (***l***) are more laborious than other consonants for dysarthric patients. Voiceless consonants are easier for them.

3) Pronunciations are closely related to articulatory movement of the tongue. The tongue moves from high to low positions in the sounds with dentals (***D***), alveolars (***R***), post-alveolars (***P***), palatal (***T***), and velar (***V***). From the matrices, dysarthric patients tend to articulate with centralized tongue positions, and it is difficult for them to produce sounds with extreme positions of the tongue.

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

## 4. Conclusion and Future Work

This paper proposed an effective staged teacher-student learning to tackle the low resource challenge in training End-to-End (E2E) voice systems (ASR and ASAT) for patients with dysarthria. The accuracy of our proposed models (knowledge distillation-based dysarthric E2E ASR and ASAT) significantly outperforms the traditional methods. Furthermore, the higher precision of E2E-ASAT offers potential functions as a rehabilitation tool and medical diagnostic aid. In the future, we will improve our dysarthric E2E ASR and ASAT systems with larger data sets and use them as comprehensive tools for effective analysis and evaluation of dysarthric speech.

## 5. ACKNOWLEDGEMENTS

This study is supported by JSPS KAKENHI Grant (20K11883), and partially by the National Natural Science Foundation of China under Grant 61771333, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330, JSPS KAKENHI No. 19K24376, NICT tenure-track startup fund, NICT tenure-track startup fund.

# 6. References


[1] P. C. Doyle, H. A. Leeper, A.-L. Kotler, N. Thomas-Stonell, C. O'Neill, M.-C. Dylke, and K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility," *Journal of rehabilitation research and development*, vol. 34, pp. 309–316, 1997.

[2] 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.

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

[4] S. B. O'Sullivan, T. J. Schmitz, and G. Fulk, *Physical rehabilitation*. FA Davis, 2019.

[5] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-End dysarthric speech recognition using multiple databases," in *Proc. ICASSP*, 2019, pp. 6395–6399.

[6] 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.

[7] 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.

[8] J. R. Duffy, *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[9] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences." in *Proc. INTERSPEECH*, 2018, pp. 3403–3407.

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

[11] Y. Lin, L. Wang, J. Dang, S. Li, and C. Ding, "End-to-End articulatory modeling for dysarthric articulatory attribute detection," in *Proc. ICASSP*, 2020, pp. 7349–7353.

[12] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-discriminative features for dysarthric speech conversion." in *Proc. INTERSPEECH*, 2017, pp. 3374–3378.

[13] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. ICASSP*, 2018, pp. 6009–6013.

[14] E. Ylmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for asr of pathological speech," in *proc. INTERSPEECH*, 2018, pp. 2958–2962.

[15] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," in *Proc. INTERSPEECH*, 2019, pp. 784–788.

[16] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *Proc. IEEE-ICASSP*, 2020, pp. 6104–6108.

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

[18] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional teacher-student learning," in *Proc. IEEE-ICASSP*, 2019, pp. 6445–6449.

[19] S.Zhou, L.Dong, S.Xu, and B.Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in Mandarin Chinese," in *Proc. INTERSPEECH*, 2018, pp. 791–795.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS*, 2015.

[21] 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.

[22] 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.

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

[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. ICASSP*, 2015, pp. 5206–5210.

[26] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[28] 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. 4400–4404.

[29] 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.