INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Use of Speech Impairment Severity for Dysarthric Speech Recognition

Mengzhe Geng[1,3,*], *Zengrui Jin*[1], *Tianzi Wang*[1], *Shujie Hu*[1], *Jiajun Deng*[1], *Mingyu Cui*[1], *Guinan Li*[1], *Jianwei Yu*[3],*Xurong Xie*[2], *Xunying Liu*[1]

[1]The Chinese University of Hong Kong
[2]Institute of Software, Chinese Academy of Sciences
[3]Tencent AI Lab

{mzgeng,zrjin,twang,sjhu,jjdeng,mycui,gnli,xyliu}@se.cuhk.edu.hk, tomasyu@tecent.com, xurong@iscas.ac.cn

Abstract

A key challenge in dysarthric speech recognition is the speaker-level diversity attributed to both speaker-identity related factors (e.g. gender) and speech impairment severity. Most prior researches on addressing this issue focused on using speaker-identity only. To this end, this paper proposes a novel set of techniques to use both severity and speaker-identity in dysarthric speech recognition: a) multitask training incorporating severity prediction error; b) speaker-severity aware auxiliary feature adaptation; and c) structured LHUC transforms separately conditioned on speaker-identity and severity. Experiments conducted on UASpeech suggest incorporating additional speech impairment severity into state-of-the-art hybrid DNN, E2E Conformer and pre-trained Wav2vec 2.0 ASR systems produced statistically significant WER reductions up to 4.78% (14.03% relative). Using the best system the lowest published WER of 17.82% (51.25% on very low intelligibility) was obtained on UASpeech.

Index Terms: Disordered Speech, Speech Recognition, Dysarthric Speech, Speech Impairment Severity

1. Introduction

In spite of the rapid development of automatic speech recognition (ASR) techniques targeting normal speech in recent decades, accurate recognition of dysarthric speech remains a highly challenging task to date [1–28]. Dysarthric speech imposes several challenges to current deep learning based ASR techniques prevailingly targeting normal speech. These include: a) data scarcity due to the difficulty in collecting a large amount of such speech from dysarthric speakers who often suffer from physical disabilities and mobility issues; b) mismatch against healthy speech; and c) large diversity among dysarthric speakers, when sources of variability commonly found in normal speech including accent or gender are aggregated with those over speech impairment severity. For example, dysarthric speakers of very low speech intelligibility exhibit clearer patterns of articulatory imprecision, decreased volume and clarity, increased dysfluencies, slower speaking rate and changes in pitch [29], while those diagonalized with mid or high speech intelligibility are closer to normal speakers. Such heterogeneity further increases the mismatch against normal speech and the difficulty in both speaker-independent (SI) ASR system development using limited impaired speech data and fine-grained personalization to individual users' data [3, 25, 30].

So far the majority of prior researches to address the dysarthric speaker level diversity have been focused on using speaker-identity only either in speaker-dependent (SD) data augmentation [7, 9, 13, 14, 18, 27], or in speaker adapted or dependent ASR system development [1, 3, 4, 7, 11–13, 19, 22, 25, 31–33]. In contrast, very limited prior researches have used speech impairment severity information for dysarthric speech recognition. Dysarthria severity-dependent GMM-HMM state distributions were proposed in [2]. Severity-dependent MLLR and CMLLR adaptations were studied in the context of GMM-HMM models in [34]. Severity-dependent tempo perturbation of dysarthric speech to match healthy speech was investigated in [35]. Speech impairment severity was also used in multi-speaker text-to-speech (TTS) systems to control pitch, energy and duration when synthesizing additional dysarthric training speech data for ASR system development [36]. However, there is a notable lack of solutions that can account for the dysarthric speech data heterogeneity that is attributed to both speaker-identity and speech impairment severity. In particular, the use of speech impairment severity in current end-to-end (E2E) and pre-trained ASR systems has been rarely visited.

To this end, this paper investigates a novel set of techniques to incorporate speech impairment severity into state-of-the-art hybrid DNN [13], end-to-end Conformer [37] and self-supervised learning (SSL) based pre-trained Wav2vec 2.0 [38] ASR systems. These include the use of: a) multi-task [39] training cost interpolation between the ASR loss and speech impairment severity prediction error; b) spectral basis embedding (SBE) [11, 25] based speaker-severity aware adaptation features that are trained to simultaneously predict both speaker-identity and impairment severity; and c) structured learning hidden units contribution (LHUC) [40] transforms that are separately conditioned on speaker-identity and impairment severity. These are used to facilitate both speaker-severity adaptive training of ASR systems and their test-time unsupervised adaptation to both factors of variability. Learning both speech impairment severity and speaker-identity serves as a dual-purpose solution. First, it allows speaker and impairment severity invariant "canonical" ASR systems to be constructed. Second, these two sources of variability can be flexibly factored in and combined for fine-grained adaptation to diverse dysarthric speakers.

Experiments were conducted on the largest available and most widely used UASpeech [41] dysarthric speech dataset. Experimental results suggest the incorporation of speech impairment severity produced statistically significant [42] word error rate (WER) reductions up to 3.95%, 4.78% and 4.37% absolute (12.56%, 14.03% and 16.53% relative) for hybrid DNN, E2E Conformer and cross-domain fine-tuned Wav2vec 2.0 models. Modeling both severity and speaker-identity produced further improvements. The lowest published WER of 17.82% (51.25% and 17.41% on very low and low intelligibility) was obtained on the UASpeech test set of 16 dysarthric speakers by combining the best-performing hybrid DNN, E2E Conformer and

* Part of this work was done while the author was an intern at Tencent AI Lab.

10.21437/Interspeech.2023-322

fine-tuned Wav2vec 2.0 systems via two pass rescoring [43].

The main contributions of the paper are summarized below:

1) To the best of our knowledge, this paper presents the first work of systematically incorporating speech impairment severity into hybrid DNN, E2E Conformer and cross-domain fine-tuning of Wav2vec 2.0 pre-trained ASR models for dysarthric speech recognition. A set of novel techniques and recipe configurations were proposed to learn both speech impairment severity and speaker-identity when constructing and personalizing these systems. In contrast, prior researches mainly focused on using speaker-identity only in speaker-dependent data augmentation [7,9,13,14,18,27] and speaker adapted or dependent ASR system development [1,3,4,7,11,13,19,22,23,25,31–33]. Very limited prior researches utilized speech impairment severity information [2,11,25,34–36]. None of these was conducted in the context of fine-tuning state-of-the-art pre-trained ASR models for dysarthric speech recognition, as considered in this paper.

2) The final system combining the best-performing hybrid DNN, E2E Conformer and fine-tuned Wav2vec 2.0 systems via two pass rescoring gave an overall WER of 17.82% on the UASpeech test set of 16 dysarthric speakers. This is the best performance reported so far on UASpeech as far as we know.

2. Incorporation of Speech Impairment Severity into Hybrid ASR Systems

In this section, we propose a novel set of techniques to incorporate speech impairment severity into the hybrid DNN [13] systems (Fig. 1). These include the use of: 1) speaker-severity aware auxiliary features serving as front-ends; 2) structured learning hidden unit contributions (LHUC) transforms separately conditioned on speaker-identity and severity; and 3) multitask learning incorporating severity prediction error.

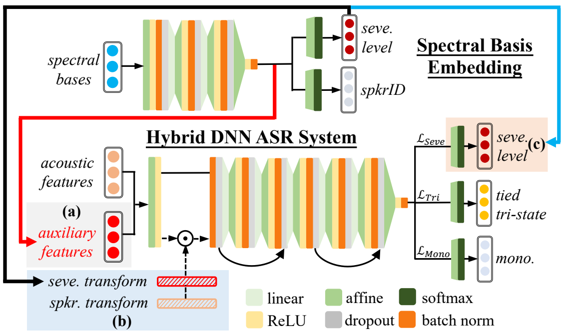

Figure 1: *Incorporating speech impairment severity (seve.) into the hybrid DNN system via **(a)** model input using speaker-severity aware auxiliary features, **(b)** structured speaker-severity LHUC transforms, and **(c)** model output using multi-task training with severity prediction error.*

2.1. Speaker-Severity Aware Auxiliary Features

The underlying variability of dysarthric speech manifesting in changes of spectral envelope, volume reduction, and imprecise articulation, can be modeled via disentangling the speech spectrum into time-invariant and time-variant subspaces [11,30] learned in a supervised manner [11, 25]. The resulting spectral basis deep embedding (SBE) features are more effective in encoding latent attributes of impaired speech [25] than classical speaker embeddings such as iVectors [44] and xVectors [45]. Hence, we adopt SBE features as speaker-severity aware auxiliary inputs to the hybrid DNN systems to incorporate both speaker-identity and impairment severity into the front-ends.

Following [11], singular value decomposition (SVD) is first conducted on the speech spectrum to generate decomposed spectral and temporal subspaces. Top-2 bases from the spectral subspace are then fed into a 3-layer DNN embedding network (Fig. 1 upper) with severity levels and speaker IDs as targets [25]. During training, the 25-dim compact features extracted from the last bottleneck layer serve as the auxiliary features (Fig. 1 red bold line). During test time evaluation, spectral bases of impaired speech utterances are fed into the trained embedding network to generate auxiliary features and also predict the speech impairment severity level for each test speaker.

2.2. Structured Speaker-Severity LHUC Adaptation

LHUC [40] adaptation can model the large variability among dysarthric speakers [9, 13], with speaker-dependent LHUC transforms applied to DNN hidden layers. Inspired by the factorized LHUC representations respectively modeling the speaker and the acoustic environment [40], we propose structured speaker-severity LHUC adaptation separately conditioned on speaker-identity ($\boldsymbol{A}^s_{spkr}$) and speech impairment severity ($\boldsymbol{A}^s_{seve}$). Such LHUC transforms are further restricted as diagonal matrices to reduce the number of parameters [40], equivalent to using scaling vectors to modify the amplitudes of the ReLU activation in the first layer of the DNN (Fig. 1 (b)).

Let $\boldsymbol{r}^s_{spkr}$ and $\boldsymbol{r}^s_{seve}$ denote the speaker-dependent and severity-dependent scaling vectors for speaker s. The hidden layer output adapted to speaker s is given as [change]:

$$\boldsymbol{h}^s = \xi(\boldsymbol{r}^s_{spkr}) \odot \xi(\boldsymbol{r}^s_{seve}) \odot \boldsymbol{h} \tag{1}$$

where $\odot$ is the Hadamard product and $\xi(\cdot)$ is the element-wise $2 \times sigmoid(\cdot)$. During unsupervised test time adaptation, severity levels of test speakers are automatically assessed using the spectral basis embedding network (Fig. 1 black bold line).

2.3. Multitask Learning

Multitask learning (MTL) [39] is proven to be helpful in improving the generalization ability of each task [46]. To this end, incorporating severity prediction error in the training criteria of the DNN ASR systems can help produce a neutral, canonical model that is invariant to speech impairment severity. An interpolation between the cross entropy (CE) loss on the frame-level tied triphone states (tri-states) ($\mathcal{L}_{Tri}$), monophone alignments ($\mathcal{L}_{Mono}$) and speech impairment severity levels ($\mathcal{L}_{Seve}$) is utilized as the multitask loss function, given as:

$$\mathcal{L}_{MTL_{DNN}} = \omega_1 \cdot \mathcal{L}_{Tri} + \omega_2 \cdot \mathcal{L}_{Mono} + \omega_3 \cdot \mathcal{L}_{Seve} \text{ [1]} \tag{2}$$

Incorporating $\mathcal{L}_{Seve}$ increases the generalization ability of the model to test speakers with diverse severity levels. During test time adaptation, the unsupervised severity label of the test speaker is derived from automatic assessment via the spectral basis embedding network (Fig. 1 blue bold line).

2.4. Severity-Dependent Systems with KLD Regularization

Motivated by [2] where separate GMM-HMM acoustic models are trained for each severity group, we develop severity-dependent DNN systems as an ablation study. To prevent overfitting to the limited amount of severity-dependent data, we further introduce Kullback-Leibler divergence (KLD) based regularized adaptation [47] by adding the KLD between the out-

[1] Weights are empirically set as $\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$.

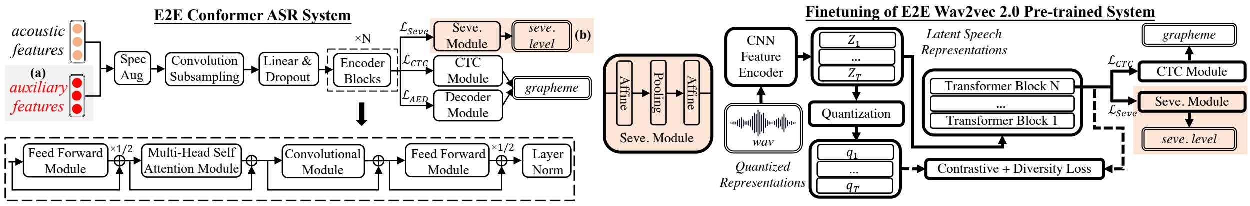

Figure 2: *Incorporating speech impairment severity into the E2E Conformer system (left) using **(a)** speaker-severity aware auxiliary features and **(b)** multitask training with severity prediction error. Speech impairment severity is incorporated during the fine-tuning of the E2E Wav2vec 2.0 pre-trained system (right) via multitask learning. Components of the severity module are shown in the middle.*

put distributions of the unadapted SI and the severity-dependent models into the training cost. This is given as:

$$\mathcal{L}_{KLD} = (1-\lambda)\mathcal{L} + \frac{\lambda}{N}\sum_{t=1}^{N} p^{SI}(y_t|x_t)\log p(y_t|x_t) \quad (3)$$

where $\mathcal{L}$ is the standard CE loss to train a DNN model and N is the number of frames. $p^{SI}(\cdot)$ is the output distribution of the SI model. λ is a tunable regularization weight empirically set to 0.5 in the experiments.

3. Incorporation of Speech Impairment Severity into E2E Systems

3.1. E2E Conformer Systems

We incorporate speech impairment severity into E2E Conformer systems through two approaches: 1) using speaker-severity aware auxiliary features (Fig. 2 left (a)) and 2) adding a CE-based severity prediction error L_{Seve} into the training criteria (Fig. 2 left (b)). The MTL training cost is given as:

$$\mathcal{L}_{MTL_{CONF}} = \alpha_1 \cdot \mathcal{L}_{CTC} + \alpha_2 \cdot \mathcal{L}_{AED} + \alpha_3 \cdot \mathcal{L}_{Seve}{}^{2} \quad (4)$$

where L_{CTC} and L_{AED} denote the Connectionist Temporal Classifcation loss [48] and the attention-based loss [49].

3.2. Fine-tuning of SSL Wav2vec 2.0 pre-trained Systems

CE-based severity prediction error in the training cost during the cross-domain fine-tuning of SSL-based Wav2vec 2.0 pre-trained systems (Fig. 2 right). The MTL criterion is given as:

$$\mathcal{L}_{MTL_{W2V}} = \beta_1 \cdot \mathcal{L}_{CTC} + \beta_2 \cdot \mathcal{L}_{Seve}{}^{3} \quad (5)$$

4. Experiments and Results

4.1. Task Description

The UASpeech [41] corpus is the largest publicly available and widely used dysarthric speech dataset. It is an isolated word recognition task containing 103h speech from 16 dysarthric and 13 control speakers. The dysarthric speakers are further divided into speech intelligibility subgroups "very low", "low", "mid" and "high". For each speaker, the data is split into three blocks B1, B2 and B3, each with the same 155 common words and a different set of 100 uncommon words. The training set includes the data from B1 and B3 of all 29 speakers (69.1h), while the test set includes the data from B2 of all 16 dysarthric speakers (22.6h, excluding speech from control speakers). Silence stripping using an HTK [50] trained GMM-HMM system [13] produces a 30.6h training set (99195 utt.) and a 9h test set (26520 utt.). Data augmentation via speed perturbation [9] produces a 130.1h augmented training set (399110 utt.). As E2E systems are sensitive to the training data coverage (Sys.2 vs. Sys.1 in Table 3-4), B2 data of the 13 control speakers are further folded in during the training of Conformer and fine-tuning of Wav2vec 2.0 systems following [24]. This produces a 40h unaugmented (122392 utt.) and a 190h augmented training set (538292 utt.).

4.2. Experiment Setup

Hybrid DNN systems: The 7-layer hybrid DNN systems were implemented using Kaldi [51] following [13]. The inputs were 80-dim filter-bank (FBK) + Δ features, optionally plus 25-dim speaker-severity aware auxiliary features.
E2E Conformer systems: The graphemic Conformer systems were implemented via ESPnet [52][4]. The inputs were 80-dim FBK + Δ features, optionally plus 25-dim auxiliary features.
Wav2vec 2.0 pre-trained systems: The Wav2vec 2.0 pre-trained systems (Fig. 2 right) contained three components: 1) a speech feature encoder with CNN convolution blocks, 2) a contextual transformer network and 3) a quantization module. The inputs were raw speech waves. During pre-training, an interpolation between the contrastive and the diversity loss was used to train the model. During fine-tuning, the CTC loss, optionally with an interpolation of the severity prediction error, was used. The large Wav2Vec 2.0 model[5] pre-trained using 60k hours of Libri-light data and fine-tuned using 960h Librispeech data was used as the base for cross-domain fine-tuning on UASpeech.

4.3. Result Analysis

Table 1 shows the performance comparison[6] on the 30.6h training set between separately trained severity-dependent DNN models, speaker-independent (SI) models, and models with severity incorporated via auxiliary features, multitask learning or structured LHUC transform. Several trends can be observed: **1)** The SI system (Sys.3) outperforms the severity-dependent systems (Sys.1-2) and thus is chosen as the baseline for further experiments. **2)** Incorporating severity via auxiliary features, multitask learning, or structured LHUC transforms into the systems (Sys.4-6) all produce statistically significant improvements over the SI baseline (Sys.3). Combining two approaches (Sys.7-9) produce WER reductions up to **3.95% absolute (12.56% relative)** over the baseline (Sys.7 vs. 3). Combining all three approaches (Sys.10) leads to no further improvement. **3)** When further combined with LHUC-SAT speaker adaptation, similar trends are observed (Sys.12-14,15-17,18 vs. Sys.11). A similar set of experiments on the 130h augmented training set serve to further evaluate the above techniques pro-

[2] Weights are empirically set as $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$.
[3] Weights are empirically set as $\beta_1 = \beta_2 = \frac{1}{2}$.
[4] 12 encoder layers + 12 decoder layers, feed-forward dim = 2048, 4 attention heads of 256 dimensions.
[5] https://huggingface.co/facebook/wav2vec2-large-960h-lv60
[6] A matched pairs sentence-segment word error (MAPSSWE) based statistical significance test [42] was done at significance level $\alpha = 0.05$.

posed for modeling severity, as shown in Table 2.

Table 1: *Performance of incorporating severity into DNN models on the 16 UASpeech dysarthric speakers using the 30.6h training set. "Aux. Feat." and "Trn. Tar." denote auxiliary feature and training target. "Seve." and "Intel." stand for severity and intelligibility. † denotes a statistically significant improvement ($\alpha = 0.05$) is obtained over the comparable baseline systems without using severity information (Sys.3,11).*

Sys.	Model	KLD	#Hrs	Impair. Seve. Aux. Feat.	Impair. Seve. Trn. Tar.	Impair. Seve. LHUC (Seve.)	LHUC SAT (Spkr)	WER% Intel. Subgroup VL	L	M	H	Unseen	Seen	All
1	Seve.	✗	30.6	✗	✗	✗	✗	70.96	35.29	26.35	11.05	47.47	23.60	32.96
2	DNN	✓						70.14	34.84	25.76	10.52	46.67	23.15	32.37
3	DNN	/		✗	✗	✗	✗	69.82	32.61	24.53	10.40	46.18	21.94	31.45
4	DNN	/	30.6	✓	✗	✗	✗	64.43†	29.71†	19.84†	8.57†	42.82†	18.51†	**28.05**†
5				✗	✓	✗	✗	69.09†	29.35†	21.01†	9.78†	43.43†	20.62†	29.57†
6				✗	✗	✓	✗	67.21†	29.67†	20.82†	9.07†	43.16†	19.83†	28.98†
7				✓	✓	✗	✗	**63.83**†	**28.00**†	**19.80**†	**8.68**†	**41.71**†	**18.33**†	**27.50**†
8				✗	✓	✓	✗	67.18†	30.18†	21.27†	9.00†	43.57†	19.87†	29.17†
9				✓	✗	✓	✗	64.56†	28.97†	19.01†	8.53†	42.49†	18.17†	27.71†
10				✓	✓	✓	✗	64.52†	28.77†	19.90†	8.31†	42.33†	18.33†	27.75†
11	DNN	/	30.6	✗	✗	✗	✓	64.39	29.88	20.27	8.95	42.32	19.23	28.29
12				✓	✗	✗	✓	63.40†	28.90†	18.64†	8.13†	41.79†	17.84†	**27.24**†
13				✗	✓	✗	✓	65.70	28.40†	19.43†	9.06	42.24	18.91†	28.06
14				✗	✗	✓	✓	65.09	28.83†	19.58†	8.56†	42.85	18.25†	27.90†
15				✓	✓	✗	✓	64.52	28.10†	18.56†	8.01†	41.35†	18.08†	27.21†
16				✗	✓	✓	✓	64.49	29.29†	18.15†	8.52†	42.06	18.28†	27.61†
17				✓	✗	✓	✓	**63.65**†	**28.11**†	**18.31**†	**8.41**†	**41.99**†	**17.52**†	**27.12**†
18				✓	✓	✓	✓	64.86	28.16†	19.29†	8.23†	42.13	18.08†	27.52†

Table 2: *Performance of incorporating severity into DNN models on UASpeech using the 130.1h or 190h augmented training set. † denotes a statistically significant improvement ($\alpha = 0.05$) over the baseline systems (Sys.3,7,11,13).*

Sys.	Model	KLD	#Hrs	Impair. Seve. Aux. Feat.	Impair. Seve. Trn. Tar.	Impair. Seve. LHUC (Seve.)	LHUC SAT (Spkr)	WER% Intel. Subgroup VL	L	M	H	Unseen	Seen	All
1	Seve.	✗	130.1	✗	✗	✗	✗	66.61	29.52	22.70	9.90	43.55	20.35	29.45
2	DNN	✓						66.54	29.38	22.00	9.38	43.22	19.98	29.09
3	DNN	/		✗	✗	✗	✗	66.45	28.95	20.37	9.62	42.46	19.86	28.73
4	DNN	/	130.1	✓	✗	✗	✗	**61.55**†	**27.52**†	**17.31**†	**8.22**†	**40.18**†	**17.28**†	**26.26**†
5				✓	✓	✗	✗	63.83†	28.00†	19.80†	8.68†	40.63†	17.54†	26.60†
6				✓	✗	✓	✗	62.24†	27.81†	17.25†	8.15†	40.89†	17.13†	26.45†
7	DNN	/	130.1	✗	✗	✗	✓	62.50	27.26	18.41	8.04	40.01	17.85	26.55
8				✓	✗	✗	✓	**59.83**†	**27.16**	**16.80**†	**7.91**	**39.55**†	**16.60**†	**25.60**†
9				✓	✓	✗	✓	61.49†	27.05	17.19†	7.84	40.01	16.92†	25.98†
10				✓	✗	✓	✓	61.21†	27.62	17.60†	7.80	40.47	16.88†	26.13†
11	DNN	/	190	✗	✗	✗	✗	67.41	29.17	18.90	7.48	39.50	20.56	27.99
12				✓	✗	✗	✗	61.76†	27.11†	15.96†	6.54†	36.85†	17.97†	**25.37**†
13				✗	✗	✗	✓	62.60†	26.65†	15.80†	5.64†	36.98	17.43	25.10†
14				✓	✗	✗	✓	62.31	25.15†	13.00†	4.66†	33.43†	17.55	**23.78**†

From Table 2, the SI system (Sys.3) outperforms the separately trained severity-dependent systems (Sys.1-2) on the 130.1h augmented training set. The best-performing 130.1h systems with or without standard LHUC SAT are obtained by incorporating severity through auxiliary features (Sys.4,8). Their improvements on the 190h augmented training set are also consistently observed (Sys. 12 vs. 11, Sys. 14 vs. 13).

Tables 3 and 4 show the performance of incorporating severity into Conformer and Wav2vec 2.0 systems[7]. Using severity in Conformer produces statistically significant WER reductions up to **4.78% absolute (14.03% relative)** (Table 3, Sys. 3 vs. 2), while on Wav2vec 2.0 producing WER reduction of **4.37% absolute (16.53% relative)** (Table 4, Sys. 3 vs. 2).

4.4. System Combination

The best-performing hybrid DNN, E2E Conformer and finetuned Wav2vec 2.0 systems are further combined using two-pass rescoring [43], where the DNN system produces N-best outputs ($N = 50$) in the 1^{st} decoding pass while the E2E systems perform 2^{nd} pass rescoring using score interpolation. The final system (Table 5 Sys.3) gives an overall WER of 17.82% (51.25% on very low intelligibility) on the UASpeech test set of 16 dysarthric speakers. To the best of our knowledge, this is the best performance reported so far on UASpeech.

Table 3: *Performance of incorporating severity into E2E Conformer (CONF.) models on UASpeech. † denotes a significant improvement ($\alpha = 0.05$) over the baseline (Sys.2).*

Sys.	Model	#Hrs	Impair. Seve. Aux. Feat.	Impair. Seve. Trn. Tar.	WER% Intel. Subgroup VL	L	M	H	Unseen	Seen	All
1	CONF.	130.1	✗	✗	73.88	53.12	49.92	42.03	99.14	23.51	53.17
2		190	✗	✗	65.70	40.63	33.39	9.53	57.40	19.03	34.07
3			✓	✗	**65.18**†	**34.90**†	**24.21**†	**5.00**†	**47.77**†	**17.20**†	**29.19**†
4			✗	✓	66.95	39.08†	29.78†	8.03†	56.01†	17.72†	32.74†
5			✓	✓	68.03	36.36†	23.29†	4.76†	47.16†	18.79	29.91†

Table 4: *Performance of finetuing Wav2vec 2.0 (W2V.) pretrained model on UASpeech. † denotes a statistically significant improvement ($\alpha = 0.05$) over the baseline (Sys.2).*

Sys.	Model	#Hrs	Seve. Task	WER% Intel. Subgroup VL	L	M	H	Unseen	Seen	All
1	W2V.	30.6	✗	68.25	37.81	24.00	8.13	52.25	18.29	31.61
2		40	✗	69.04	29.86	14.75	3.68	36.52	19.92	26.44
3			✓	**59.38**†	**23.91**†	**12.10**†	**2.91**†	**33.27**†	**14.85**†	**22.07**†

Table 5: *Performance of combining the best DNN (Table 2 Sys.14), E2E Conformer (Table 3 Sys.3) and finetuned Wav2vec 2.0 (Table 4 Sys.3) systems via two-pass rescoring [43].*

Sys.	Model	#Hrs	WER% Intel. Subgroup VL	L	M	H	Unseen	Seen	All
1	DNN $\rightarrow$ CONF.	190	60.66	22.82	11.37	3.80	31.62	16.15	22.22
2	DNN $\rightarrow$ W2V.		52.00	18.26	8.35	2.58	26.41	12.92	18.21
3	DNN $\rightarrow$ CONF. + W2V.		**51.25**	**17.41**	**8.16**	**2.66**	**26.45**	**12.26**	**17.82**

Table 6: *Performance comparison against recently published systems on **UASpeech**. "DA" denotes data augmentation.*

Sys.	VL	All
Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation (15spk) [53]	68.24	30.76
CUHK-2020 DNN + DA + LHUC SAT [9]	62.44	26.37
CUHK-2021 QuartzNet + CTC + Meta-learning + SAT [54]	69.30	30.50
Sheffield-2022 DA + Source Filter Features + iVector adapt [26]	-	30.30
BUT-2022 Wav2vec 2.0 + fMLLR + xVectors (15spk) [23]	57.72	22.48
CUHK-2022 DA + TDNN + Wav2vec 2.0 feat. + system combination [28]	52.53	22.83
DA + severity incorporation + system combination (Table 5 Sys.3, ours)	**51.25**	**17.82**

5. Conclusions

This paper investigates a novel set of techniques to incorporate speech impairment severity into state-of-the-art hybrid DNN, E2E Conformer and pre-trained Wav2vec 2.0 systems for dysarthric speech recognition. These techniques include the use of multitask learning, auxiliary features and structured LHUC transforms. Experiments conducted on the UASpeech dataset suggest that incorporating severity and combining the best-performing systems produces the lowest published WER of 17.82%. Future research will focus on more advanced techniques of incorporating severity into the E2E systems.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200220, 14200021, Innovation Technology Fund grant No. ITS/218/21, National Natural Science Foundation of China (NSFC) Grant 62106255, and Youth Innovation Promotion Association CAS Grant 2023119.

[7] Using the 30.6h and the 130.1h training set on UASpeech produce comparable performance when fine-tuning Wav2vec 2.0 models [28].

7. References

[1] H. Christensen *et al.*, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012.

[2] M. J. Kim *et al.*, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models," in *INTERSPEECH*, 2013.

[3] S. Sehgal *et al.*, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *SLPAT*, 2015.

[4] N. M. Joy *et al.*, "Improving acoustic models in TORGO dysarthric speech database," *IEEE T NEUR SYS REH*, 2018.

[5] J. Yu *et al.*, "Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus," in *INTERSPEECH*, 2018.

[6] S. Hu *et al.*, "The CUHK Dysarthric Speech Recognition Systems for English and Cantonese," in *INTERSPEECH*, 2019.

[7] F. Xiong *et al.*, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*, 2019.

[8] S. Liu *et al.*, "Exploiting cross-domain visual feature generation for disordered speech recognition," in *INTERSPEECH*, 2020.

[9] M. Geng *et al.*, "Investigation of Data Augmentation Techniques for Disordered Speech Recognition," in *INTERSPEECH*, 2020.

[10] J. Harvill *et al.*, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *ICASSP*, 2021.

[11] M. Geng *et al.*, "Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition," in *INTERSPEECH*, 2021.

[12] X. Xie *et al.*, "Variational Auto-Encoder Based Variability Encoding for Dysarthric Speech Recognition," in *INTERSPEECH*, 2021.

[13] S. Liu *et al.*, "Recent Progress in the CUHK Dysarthric Speech Recognition System," *IEEE T AUDIO SPEECH*, 2021.

[14] Z. Jin *et al.*, "Adversarial Data Augmentation for Disordered Speech Recognition," in *INTERSPEECH*, 2021.

[15] J. Deng *et al.*, "Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition," in *INTERSPEECH*, 2021.

[16] S. Hu *et al.*, "Neural Architecture Search For LF-MMI Trained Time Delay Neural Networks," *IEEE T AUDIO SPEECH*, 2022.

[17] M. Geng *et al.*, "On-the-fly feature based speaker adaptation for dysarthric and elderly speech recognition," *arXiv preprint arXiv:2203.14593*, 2022.

[18] Z. Jin *et al.*, "Personalized adversarial data augmentation for dysarthric and elderly speech recognition," *arXiv preprint arXiv:2205.06445*, 2022.

[19] D. Mulfari *et al.*, "Exploring ai-based speaker dependent methods in dysarthric speech recognition," in *CCGrid*, 2022.

[20] S. Hu *et al.*, "Exploiting cross-domain and cross-lingual ultrasound tongue imaging features for elderly and dysarthric speech recognition," *arXiv preprint arXiv:2206.07327*, 2022.

[21] S. Hu *et al.*, "Exploiting Cross Domain Acoustic-to-articulatory Inverted Features for Disordered Speech Recognition," in *ICASSP*, 2022.

[22] J. Tobin *et al.*, "Personalized automatic speech recognition trained on small disordered speech datasets," in *ICASSP*, 2022.

[23] M. K. Baskar *et al.*, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *INTERSPEECH*, 2022.

[24] A. Hernandez *et al.*, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *INTERSPEECH*, 2022.

[25] M. Geng *et al.*, "Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition," *IEEE T AUDIO SPEECH*, 2022.

[26] Z. Yue *et al.*, "Acoustic modelling from raw source and filter components for dysarthric speech recognition," *IEEE T AUDIO SPEECH*, 2022.

[27] Z. Jin *et al.*, "Adversarial data augmentation using vae-gan for disordered speech recognition," in *ICASSP*, 2023.

[28] S. Hu *et al.*, "Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition," in *ICASSP*, 2023.

[29] R. D. Kent *et al.*, "What dysarthrias can tell us about the neural control of speech," *J PHONETICS*, 2000.

[30] I. Kodrasi *et al.*, "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection," *IEEE T AUDIO SPEECH*, 2020.

[31] C. Bhat *et al.*, "Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation," in *INTERSPEECH*, 2016.

[32] M. Kim *et al.*, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE T NEUR SYS REH*, 2017.

[33] R. Takashima *et al.*, "Two-step acoustic model adaptation for dysarthric speech recognition," in *ICASSP*, 2020.

[34] M. B. Mustafa *et al.*, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PloS one*, 2014.

[35] C. Bhat *et al.*, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *SPECOM*, 2016.

[36] M. Soleymanpour *et al.*, "Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition," in *ICASSP*, 2022.

[37] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020.

[38] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ADV NEUR IN*, 2020.

[39] R. Caruana, "Multitask learning," *MACH LEARN*, 1997.

[40] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE T AUDIO SPEECH*, 2016.

[41] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[42] D. S. Pallet *et al.*, "Tools for the analysis of benchmark speech recognition tests," in *ICASSP*, 1990.

[43] M. Cui *et al.*, "Two-pass decoding and cross-adaptation based system combination of end-to-end conformer and hybrid tdnn asr systems," in *INTERSPEECH*, 2022.

[44] A. Senior *et al.*, "Improving DNN speaker independence with i-vector inputs," in *ICASSP*, 2014.

[45] D. Snyder *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[46] G. Pironkov *et al.*, "Multi-task learning for speech recognition: an overview." in *ESANN*, 2016.

[47] D. Yu *et al.*, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, 2013.

[48] A. Graves *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[49] A. Vaswani *et al.*, "Attention is all you need," *ADV NEUR IN*, 2017.

[50] S. Young *et al.*, "The HTK book," *Cambridge university engineering department*, 2002.

[51] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[52] S. Watanabe *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *INTERPSEECH*, 2018.

[53] F. Xiong *et al.*, "Source Domain Data Selection for Improved Transfer Learning Targeting Dysarthric Speech Recognition," in *ICASSP*, 2020.

[54] D. Wang *et al.*, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *ISCSLP*, 2021.