INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Hyper-parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition

Tianzi Wang[1], *Shoukang Hu*[2], *Jiajun Deng*[1], *Zengrui Jin*[1], *Mengzhe Geng*[1], *Yi Wang*[1], *Helen Meng*[1], *Xunying Liu*[1]

[1]The Chinese University of Hong Kong, China
[2]Nanyang Technological University, Singapore

[1]{twang,jjdeng,zrjin,mzgeng,ywang,hmmeng,xyliu}@se.cuhk.edu.hk,
[2]shoukang.hu@ntu.edu.sg

Abstract

Automatic recognition of disordered and elderly speech remains highly challenging tasks to date due to data scarcity. Parameter fine-tuning is often used to exploit the large quantities of non-aged and healthy speech pre-trained models, while neural architecture hyper-parameters are set using expert knowledge and remain unchanged. This paper investigates hyper-parameter adaptation for Conformer ASR systems that are pre-trained on the Librispeech corpus before being domain adapted to the DementiaBank elderly and UASpeech dysarthric speech datasets. Experimental results suggest that hyper-parameter adaptation produced word error rate (WER) reductions of 0.45% and 0.67% over parameter-only fine-tuning on DBank and UASpeech tasks respectively. An intuitive correlation is found between the performance improvements by hyper-parameter domain adaptation and the relative utterance length ratio between the source and target domain data.

Index Terms: Dysarthric Speech, Elderly Speech, Conformer, Domain Adaptation, Hyper-parameter Adaptation

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) technologies targeting normal speech, accurate recognition of elderly and dysarthric speech remains a challenging task [1–8]. Neurocognitive disorders, e.g. Alzheimer's disease (AD), are often found among older adults [9] experiencing speech impairments [10, 11]. ASR technologies trailed for their needs can improve their quality of life and social inclusion.

Elderly and dysarthric speech exhibit a wide spectrum of challenges for current deep neural networks (DNNs) based ASR technologies that predominantly target normal speech. First, a large mismatch between such data and non-aged, healthy adult voices is often observed. Such difference manifests itself across many fronts including articulatory imprecision, decreased volume and clarity, changes in pitch, increased dysfluencies, and slower speaking rate [12, 13]. Second, the co-occurring disabilities and mobility issues among elderly speakers lead to the difficulty in collecting large quantities of such data that are essential for current data-intensive ASR system development.

A widely adopted solution to address above issues for dysarthric and elderly speech recognition is to use model-based domain adaptation approaches [1, 3, 14–17]. Current practice of domain adaptation of E2E ASR systems mainly perform parameter fine-tuning, while the underlying architecture hyper-parameters remain unchanged during domain adaptation. However, previous studies suggest that the optimal architectural hyper-parameters are heavily domain specific. For example, in [7] demonstrated the optimal settings of hidden layer context offsets in hybrid TDNN acoustic models, which are used to encode hidden layer level temporal context spans, vary significantly between TDNN systems optimized for fluent, normal speech of longer utterances, and those constructed on disfluent, impaired speech utterances of single word commands or short phrases. A similar study that performed architecture adaptation was conducted on CTC-based CNN ASR systems [18] for multilingual speech recognition, revealing that optimal convolutional module hyper-parameters, e.g. the convolution kernal size, vary substantially between languages. In contrast, the hyper-parameters domain adaptation of state-of-the-art end-to-end ASR systems represented by, for example, those based on Conformer models [19–26], remains unvisited for dysarthric and elderly speech recognition.

To this end, this paper presents the study on cross-domain hyper-parameter adaptation from Librispeech [27] pre-trained Conformer ASR systems to two atypical speech recognition tasks: a) the 16-hour DementiaBank Pitt elderly speech corpus (DBank) [28]; and b) the 31-hour UASpeech dysarthric speech [29] dataset. In this work, the hyper-parameter domain adaptation problem is transformed into a cross domain differentiable neural architecture search (DARTS) [30] task. A DARTS super-network is constructed to contain all possible candidate structures associated with varying Conformer encoder and decoder hyper-parameters settings: a) the dimensionality of macron-feedforward and feedforward layers; b) the number and the dimensionality of attention heads; and c) the kernel size of convolution modules. Thus super-network is initially pre-trained on the 960-hour Librispeech corpus with auto-configured hyper-parameter weights, before being domain adapted to the DBank or UASpeech data to extract the optimal target domain specific hyper-parameters. Conformer systems configured using the resulting hyper-parameters setting are then parameter wise pre-trained on the Librispeech data before being cross domain fine-tuned to the DBank or UASpeech datasets respectively.

Experimental results suggest that hyper-parameter domain adaptation produced consistent word error rate (WER) reductions of 0.45% and 0.67% absolute (1.81% and 2.37% relative) for the DBank and UASpeech tasks respectively over the conventional parameter fine-tuning only domain adaptation Conformer systems. Such WER reduction from Conformer hyper-parameter adaptation is found to be intuitively correlated with the average utterance length ratio between the source and target domain data, e.g. Librispeech vs. DBank (12.3s:3.4s) data while Librispeech vs. UASpeech (12.3s:1.1s). This correlation is further validated by evaluating the WER reductions on the subsets of DBank and UASPeech test data of shorter and longer utterance lengths. On the shorter segments based subsets of DBank and UASpeech test data, hyper-parameter adaptation produces larger WER reductions of 1.2% and 0.9% respectively over parameter-only fine-tuning. In contrast, smaller

 10.21437/Interspeech.2023-1263

WER reductions of 0.3%-0.4% were obtained on the longer segments based subsets. Such correlation is also consistent with the corresponding larger or smaller changes of Conformer hyper-parameters: a) the number of attention heads that affect longer range contexts; and b) the kernel sizes of the convolution modules designed to control the span of local contexts.

To the best of our knowledge, this paper presents the first investigation on the hyper-parameter adaptation in Conformer systems for elderly and dysarthric speech recognition. In contrast, the majority of previous research on domain adaptation for the same tasks has been focused on direct parameter fine-tuning [3, 31–33]. Limited previous researches on ASR architecture adaptation were conducted on hybrid TDNN models [7], or CTC-based CNN multilingual ASR [18]. For Conformer ASR systems, only in-domain data based neural architectural search were studied for elderly [34] and Mandarin [35] speech recognition, while the cross domain architecture adaptation problem was not considered.

2. Baseline and Parameter Adaptation

2.1. Conformer Baseline Architecture

This section reviews the Conformer architecture and presents the baseline parameter fine-tuning setup on two atypical ASR tasks: a) the 16-hour DBank elderly speech corpus [28]; and b) the 31-hour UASpeech dysarthric speech dataset [29].

Hybrid CTC/attention based Conformer A novel structure named Conformer was proposed in [19] and achieved state-of-the-art results on LibriSpeech ASR tasks. In this work, we built the sequence trained end-to-end Conformer ASR following the ESPnet [36] hybrid CTC/attention encoder-decoder structure. The encoder contains 2 Convolution blocks to downsample the 80-dimension Mel-scale filter banks and 3-dimension pitch inputs, followed by stacked Conformer blocks. An interpolated CTC+AED cost function (3:7 weighting) was computed over the output vocabulary. The hyper-parameters settings of baseline Conformer system[1] for both DBank and UASpeech are determined following the ESPnet Switchboard recipe.

2.2. Parameter Adaptation

In order to exploit large quantities of out-of-domain, non-aged and healthy adult speech pre-trained Conformer systems, parameter fine-tuning based cross-domain adaptation was considered. A 960-hour Librispeech corpus trained Conformer system was parameter fine-tuned to the DBank or UASpeech data after speed perturbation based data augmentation [15].

3. Hyper-parameter Adaptation

3.1. Differentiable Architecture Search

To automatically learn the suitable Conformer hyper-parameter settings for the target atypical speech domain, DARTS [30] was used to optimize four groups of hyper-parameters inside each Conformer encoder block: a) the feedforward layer dimensionality (FD); b) the number of attention heads (AH); c) the dimensionality of attention head (ADIM); d) the convolution kernel size (CK), as well as three groups of hyper-parameters inside each Transformer decoder block (FD, AH and ADIM). The estimation of standard network parameters inside the super-network model is decoupled from that of the architecture parameters [37, 38]. This leads to a pipelined approach allowing the architectural weights to be learned on separate held-out data. The optimal architecture with the largest weight is selected.

Gumbel-Softmax DARTS: In traditional DARTS methods, when similar architecture weights are obtained using a flattened Softmax function, the confusion over different candidate systems increases and search errors may occur. To this end, a Gumbel-Softmax distribution [37,39,40] is used to sharpen the architecture weights to produce approximately a one-hot vector. The architecture weights are computed as,

$$\lambda_i^l = \frac{\exp(\log(\alpha_i^l + G_i^l)/T)}{\sum_{j=1}^{N^l} \exp(\log(\alpha_j^l + G_j^l)/T)} \quad (1)$$

where $G_i^l = -\log(-\log(U_i^l))$ is the Gumbel variables and U_i^l is a uniform random variable. As the temperature factor T decreases to zero, Eqn. (1) approaches a categorical distribution.

Penalized DARTS: In order to avoid over-parameterization during architecture search, a penalty loss incorporating the number of parameters for each candidate choice was jointly optimized with the original Conformer training loss function:

$$\mathcal{L} = \mathcal{L}_{Conformer} + \eta \sum_{i,l} \alpha_i^l P_i^l \quad (2)$$

where P_i^l is the number of parameters of the i-th candidate architecture at the l-th layer, and η is the penalty scaling factor empirically adjusted for performance vs. complexity trade-off.

3.2. Hyper-parameter Adaptation

Hyper-parameter adaptation of Conformer systems is performed in two stages: **1)** A super-network shown in Fig. 1 (left, purple) that contains all possible candidate structures associated with varying hyper-parameter settings of each encoder block (FD, AH, ADIM, CK) and decoder block (FD, AH, ADIM) using the source domain Librispeech data alone, before being adapted to the target elderly or dysarthric speech domain. In this process, the large number of standard Conformer super-network parameters, often in tens of millions, are inherited and adapted to the limited target domain data, while the comparatively much smaller number of hyper-parameter selection weights, are fine-tuned during domain adaptation. **2)** The differentiable architecture search performed over the resulting domain adapted Conformer super-network will then produce the 1-best hyper-parameter settings with the largest weights. A Conformer system that features the above 1-best hyper-parameter configurations using the source domain training data is then constructed. This is followed by the parameter fine-tuning based domain adaptation as described in Section 2.2. The standard model parameters of this source domain Conformer system are then further adapted via fine-tuning to the target domain speech to produce the full hyper-parameter plus parameter adapted system (Fig. 1, right, yellow).

4. Experiments

The proposed auto-configurable neural architectural and parametric domain adaptation approach was investigated from the Librispeech data with 960-hour audiobooks speech collected from 2338 speakers to two tasks of Conformer systems: a) the 16-hour DementiaBank (DBank) elderly speech corpus; and b) the 31-hour UASpeech dysarthric speech dataset.

DementiaBank elderly speech The DementiaBank (DBank) elderly speech database [28] consists of 16-hours of a training set (29682 utterances) recorded over interviews between the 292 elderly participants and the clinical investigators after silence stripping [5], and is expanded to 59-hours when speed perturbation was performed [15]. The development and evaluation sets contain 2.5-hours and 0.6-hours of audio respectively.

[1] 12 Conformer encoder blocks + 6 Transformer decoder blocks, feed-forward layer dim = 2048, attention heads = 4, dim of attention heads = 256, convolution kernel size=31

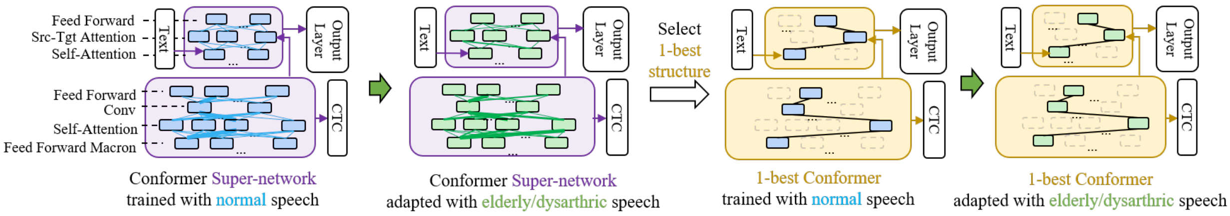

Figure 1: The pipeline of parameter and hyper-parameter domain adaptation of Conformer ASR systems for elderly/dysarthric speech recognition, including adaptation of hyper-parameters inside the super-network model decoupled with standard network parameters (left, blue) from normal speech to elderly/dysarthric speech (green), and parameter adaptation of 1-best domain adapted hyper-parameter based Conformer (right, yellow) with the same source and target speech.

The source-target domain utterance length ratio between Librispeech and DBank is 12.3s:3.4s on average. A 4-gram character language model built with the DBank transcripts was used.

UASpeech dysarthric speech [29, 41] is an isolated word recognition task (including 155 common words and 300 uncommon words) that is collected from 16 dysarthric and 13 control speakers using multiple microphone channels [41]. The training data covers all 29 speakers contains all 155 common words and two third of the uncommon words (31 hours after silence stripping), while the evaluation data was collected from only the 16 dysarthric speakers and cover all 155 common words and the remaining one-third of the uncommon words (9-hours after silence stripping). As E2E ASR systems are sensitive to the training data coverage, B2 data of the 13 control speakers are also used in Conformer training and cross domain adaptation. This produces a 40 hour unaugmented (122392 utt.) and a 190-hour augmented training set (538292 utt.) after applying speaker independent and dependent speed perturbation [15]. The source-target domain utterance length ratio between Librispeech and UASpeech is 12.3s:1.1s on average. A word grammar language model was used in decoding [41].

In the domain adaptation experiments, all the source domain trained Conformer systems used 5000 Byte Pair Encoding (BPE) tokens selected from the 960-hour Librispeech transcripts as the decoder outputs. During domain adaptation, the normal speech pre-trained parameters in Conformer encoder and Transformer decoder were fine-tuned to dysarthric or elderly speech, while the other components of the systems were randomly reinitialized and retrained on the target domain data from scratch. In this process, 100 BPE tokens extracted from the DBank speech transcripts and character level tokens were used in fine-tuning the decoder module for the DBank and UASpeech data respectively.

4.1. Performance on DementiaBank

Baseline and manually designed system The performance of the baseline system as described in Section 2.1 is shown in the first line (Sys. 1) in Tab. 1. We increase the number of decoder Transformer blocks from 6 to 12 and reduce the convolution kernel size from 31 to 7. The resulting manually designed system outperformed the baseline (Sys. 2 vs. Sys. 1, Tab. 1) by WER reduction of 2.75% absolute on average, while the model parameters increased from 43.1M to 52.5M.

Parameter and hyper-parameter adaptation Among the results shown in Tab. 1, several trends can be found. **1)** Using manually crafted hyper-parameters, the parameter domain adaptation of normal non-aged 960-hour Librispeech pre-trained Conformer systems produced WER reductions of 7.17% absolute (22.34% relative) on average (Sys. 6 vs. Sys. 2). **2)** The hyper-parameter adaptation step consistently produced further WER reductions of up to 0.45% on average on top of those obtained by parameter only fine-tuning (Sys. 8 vs Sys. 6) irrespective of the complexity penalty term η setting in Eqn. (2). By further tuning the penalty η during hyper-parameter adaptation, the most compact system achieves 0.24% WER reductions with relatively 11.2% smaller model parameters (Sys. 9 vs Sys. 6). **3)** Compared with hyper-parameter learning using only the in-domain DBank data (Sys. 3-5), the hyper-parameter domain adaptation step can automatically increase the model capacity to better leverage the knowledge distilled from normal non-aged 960-hour Librispeech pre-trained ASR systems (Sys. 7-9).

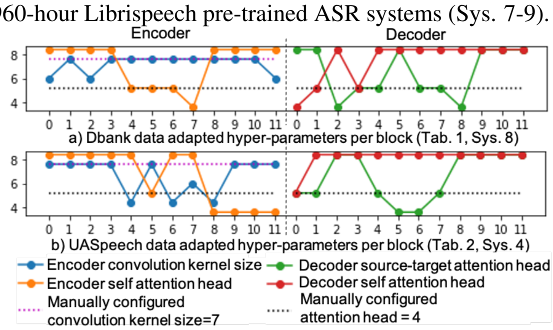

Figure 2: Hyper-parameters after cross-domain hyper-parameter adaptation from Librispeech to DBank (a) and from Librispeech to UASpeech (b). X axis is the encoder and decoder blocks where '0' refers to the very bottom layer close to input and '11' refers to the top layer close to CTC or output layer. Y axis represents two gropus of hyper-parameters including the number of attention heads and the convolution kernel size.

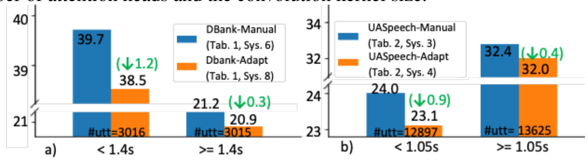

Figure 3: Conformer systems Performance (WER%) on subsets of DBank (a) and UASpeech (b) test data of shorter and longer utterance lengths, with manually designed or cross domain adapted architecture hyper-parameters.

4.2. Performance on UASpeech

Baseline and manually designed system is shown in Sys. 1 and 2 in Tab. 2 with the same hyper-parameter settings as DBank systems. The manually designed system outperformed the baseline (Sys. 2 vs. Sys. 1, Tab. 2) by a WER reduction of 1.65% absolute on average.

Parameter and hyper-parameter adaptation Several trends can be found in Tab. 2. **1)** The parameter adaptation on the UASpeech manually designed system produces a WER reduction of 5.74% absolute (16.85% relative) on average (Sys. 3 vs. Sys. 2). Such relative WER reduction from parameter adaptation is smaller compared to the experiments on the DBank task (22.34% relative, Sys. 6 vs. Sys. 2, Tab. 1). This is due to the larger mismatch in terms of both utterance lengths and speech contents (sentences vs. single words) between the Librispeech and UASpeech data. **2)**, The additional hyper-parameter adaptation step produced further statistically significant consistent WER reductions of up to 0.67% on average of those obtained by parameter only fine-tuning (Sys. 4 vs Sys. 3). The best-performed system achieves 1.34% WER reductions on the most

Table 1: Performance (WER%, #Params) of Conformer systems with baseline, manually designed and automatically learnt hyper-parameter settings trained on in-domain DBank data only (Sys. 1-5), and those cross domain parameter plus hyper-parameter adapted (Sys. 6-10). The manually designed Conformer (Sys. 2) serves as the start point of hyper-parameter in-domain learning and cross-domain adaptation performed over four groups of hyper-parameters inside each Conformer encoder block: a) the macron-feedforward and feedforward layer dimensionality (FD), where the dimensionality indices denote a choice from {512, 1024, 2048, 3072}; b) the number of attention heads (AH); c) the dimensionality of attention head (ADIM), where the indices denote a choice from {16, 32, 64, 96}; d) the convolution kernel size (CK), as well as three groups of hyper-parameters inside each Transformer decoder block (FD, AH and ADIM) with the same corresponding search space as in encoder. η is the penalty factor penalized DARTS of Eqn. (2). † denotes a statistically significant WER difference (MAPSSWE [42], α = 0.05) is obtained over the baseline (Sys. 1).

Sys	Parameter Adaptation	Hyper-parameter Adaptation					Dev		Eval		Avg.	#Params
		Adaptation	SearchMethod	η	Encoder Search Scope	Decoder Search Scope	Inv	Par	Inv	Par		
1	×	×	Baseline	×			21.29	48.90	19.20	37.35	34.84	43.1M
2			Manual				19.53	44.90	19.20	34.66	32.09	52.5M
3			DARTS	0	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}, CK:{3,5,7}	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}	19.32	44.59	17.75	34.95	31.89	53.2M
4				0.003			19.03	44.68	18.64	34.20	31.71	39.3M
5				0.03			19.21	45.55	18.09	35.27	32.29	22.6M
6	Librispeech → Dbank	×	Manual	×			15.27	35.11	17.20	25.61	24.92	52.5M
7		Librispeech → Dbank	DARTS	0.	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}, CK:{3,5,7}	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}	**14.88**†	35.47	15.76	25.74	24.88	69.8M
8				0.003			15.03	**34.41**†	15.42	**25.48**†	**24.47**†	65.7M
9				0.03			15.20	34.71	**14.76**†	25.80	24.68	46.6M

Table 2: Performance (WER%, #Params) of Conformer systems with baseline, manually designed hyper-parameter settings trained on in-domain UASpeech data only (Sys. 1-2), and those cross domain parameter plus hyper-parameter adapted (Sys. 3-6). † and ‡ denote a statistically significant WER difference is obtained over the baseline (Sys. 1) and the parameter-only fine-tuned system (Sys. 3) respectively. Other naming conventions following those of Tab. 1.

Sys	Parameter Adaptation	Hyper-parameter Adaptation					High	Mid	Low	Very Low	Avg.	#Params
		Adaptation	SearchMethod	η	Encoder Search Scope	Decoder Search Scope						
1	×	×	Baseline	×			11.98	35.03	41.85	66.54	35.71	43.1M
2			Manual				9.29	31.90	40.63	67.34	34.06	52.5M
3	Librispeech → UASpeech	×	Manual	×			5.45	23.21	**33.15**†	63.40	28.32	52.5M
4		Librispeech → UASpeech	DARTS	0.	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}, CK:{3,5,7}	FD:{0,1,2,3}, AH:{2,4,8} ADIM:{0,1,2,3}	5.22	**21.35**†‡	33.37	**62.06**†‡	**27.65**†‡	47.8M
5				0.003			5.38	22.49	33.31	63.56	28.23	41.0M
6				0.03			**5.08**†	23.15	34.21	63.10	28.39	29.8M

challenging "Very Low" intelligibility subsets, with relatively 9.0% smaller model parameters (Sys. 4 vs Sys. 3). **3)** Compared with the results on the DBank data of Tab. 1, it is also found that the WER reductions from hyper-parameter domain adaptation are intuitively correlated with the relative utterance length ratio between the source and target domain data. This ratio is smaller between Librispeech and DBank (12.3s:3.4s) while larger between Librispeech and UASpeech (12.3s:1.1s). Such hypothesis is further analysed below in Section 4.3.

4.3. Analysis of the Relation Between Hyper-parameter Adaptation and Source-target Utterance Length Ratio

In this section, the correlation between the performance improvements from Conformer hyper-parameter adaptation and the average utterance length ratio between the source and target domain is further analysed. This is conducted by evaluating the WER reductions on the subsets of DBank and UASPeech test data of shorter and longer utterance lengths. On both tasks the test data are divided using the respective median utterance length into "shorter" and "longer" subsets. As shown in Fig. 3, on the shorter segments based subsets of DBank and UASpeech test data, hyper-parameter adaptation produces larger WER reductions of 1.2% and 0.9% respectively over parameter-only fine-tuning. In contrast, smaller WER reductions of 0.3%-0.4% were obtained on the longer segments based subsets. These trends are also consistent with the corresponding larger changes of Conformer hyper-parameters on UASpeech data observed among more blocks when compared with those on DBank (overlapping with dotted lines of manual hyper-parameter settings indicate no change): a) the number of encoder attention heads that affect longer range contexts (Fig. 2, (a) vs. (b) orange); and b) the kernel sizes of the convolution modules designed to control the span of local contexts (Fig. 2, (a) vs. (b), blue). The performance of the best performing Conformer system using hyper-parameter domain adaptation is further contrasted with recently published results on the UASpeech and DBank task in Table 3.

Table 3: A comparison of average WER between published systems and our hyper-parameter adapted system on DBank and UASpeech.

DBank Systems	Avg. WER.
2021 Kaldi TDNN + Domain Adaptation [5]	32.33
2022 Conformer + Domain Adaptation [34]	25.60
Conformer + Hyper-parameter Adaptation(Sys. 8, Tab. 1)	**24.47**
UASpeech Systems	Avg. WER.
2019 Kaldi TDNN + DA [43]	30.01
2020 DNN + DA [15]	28.73
Conformer + Hyper-parameter Adaptation(Sys. 4, Tab. 2)	**27.65**

5. Conclusions

This paper investigates hyper-parameter adaptation for Conformer ASR systems that are pre-trained on the Librispeech corpus before being domain adapted to the DementiaBank elderly and UASpeech dysarthric speech datasets. Experimental results suggest that conventional parametric domain adaptation from Librispeech data produced consistent WER reductions of up to 7.17% and 5.74% absolute (22.34% and 16.85% relative) for the DBank and UASpeech tasks respectively over the in-domain data trained Conformer models. Hyper-parameter adaptation produced further absolute WER reductions of 0.45% and 0.67% over parameter-only fine-tuning, and can better account for domain mismatch. An intuitive correlation is found between the performance improvements by hyper-parameter domain adaptation and the relative utterance length ratio between the source and target domain data.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200021, 14200218, 14200220, TRS T45-407/19N and Innovation & Technology Fund grant No. ITS/254/19, ITS/218/21.

7. References

[1] H. Christensen, M. B. Aniol, P. Bell, P. D. Green *et al.*, "Combining In-Domain and Out-of-Domain Speech Data for Automatic Recognition of Disordered Speech," in *INTERSPEECH*, 2013.

[2] J. Yu, X. Xie, S. Liu, S. Hu *et al.*, "Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus," in *INTERSPEECH*, 2018.

[3] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*, 2020.

[4] S. Liu, X. Xie, J. Yu, S. Hu *et al.*, "Exploiting Cross-Domain Visual Feature Generation for Disordered Speech Recognition," in *INTERSPEECH*, 2020.

[5] Z. Ye, S. Hu, J. Li, X. Xie *et al.*, "Development of the CUHK Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the Dementiabank Corpus," in *ICASSP*, 2021.

[6] M. Geng, S. Liu, J. Yu, X. Xie *et al.*, "Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition," in *INTERSPEECH*, 2021.

[7] J. Deng, F. R. Gutierrez, S. Hu, M. Geng *et al.*, "Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition," in *INTERSPEECH*, 2021.

[8] M. Geng, X. Xie, Z. Ye, T. Wang *et al.*, "Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition," in *TASLP*, 2022.

[9] A. Association, "2019 Alzheimer's Disease Facts and Figures Report," in *Alzheimer's & dementia*, 2019.

[10] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech," in *Journal of Alzheimer's Disease*, 2016.

[11] A. König, N. Linz, J. Tröger, M. Wolters *et al.*, "Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task," in *Dementia and geriatric cognitive disorders*, 2018.

[12] T. J. Hixon and J. C. Hardy, "Restricted Motility of the Speech Articulators in Cerebral Palsy," in *Journal of Speech and Hearing Disorders*, 1964.

[13] R. D. Kent, J. F. Kent, G. Weismer, and J. R. Duffy, "What dysarthrias can tell us about the neural control of speech," in *Journal of Phonetics*, 2000.

[14] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau *et al.*, "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases," in *INTERSPEECH*, 2021.

[15] M. Geng, X. Xie, S. Liu, J. Yu *et al.*, "Investigation of Data Augmentation Techniques for Disordered Speech Recognition," in *INTERSPEECH*, 2020.

[16] D. Wang, J. Yu, X. Wu, L. Sun *et al.*, "Improved End-to-End Dysarthric Speech Recognition via Meta-learning Based Model Re-initialization," in *ISCSLP*, 2021.

[17] S. Liu, M. Geng, S. Hu, X. Xie *et al.*, "Recent Progress in the CUHK Dysarthric Speech Recognition System," in *TASLP*, 2021.

[18] Y.-C. Chen, J.-Y. Hsu, C.-K. Lee, and H.-y. Lee, "DARTS-ASR: Differentiable Architecture Search for Multilingual Speech Recognition and Adaptation," in *INTERSPEECH*, 2020.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar *et al.*, "Conformer: Convolution-Augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020.

[20] Y. Zhang, J. Qin, D. S. Park, W. Han *et al.*, "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition," in *NeurIPS SAS Workshop*, 2020.

[21] P. Guo, F. Boyer, X. Chang, T. Hayashi *et al.*, "Recent Developments on ESPnet Toolkit Boosted by Conformer," in *ICASSP*, 2021.

[22] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko *et al.*, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *INTERSPEECH*, 2021.

[23] Z. Yao, D. Wu, X. Wang, B. Zhang *et al.*, "WeNet: Production Oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit," in *INTERSPEECH*, 2021.

[24] B. Zhang, D. Wu, Z. Yao, X. Wang *et al.*, "Unified Streaming and Non-streaming Two-pass End-to-end Model for Speech Recognition," in *arXiv preprint arXiv:2012.05481*, 2020.

[25] Y. Zhang, D. S. Park, W. Han, J. Qin *et al.*, "BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition," in *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[26] W. Chan, D. Park, C. Lee, Y. Zhang *et al.*, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," in *arXiv preprint arXiv:2104.02133*, 2021.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[28] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton *et al.*, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," in *Archives of neurology*, 1994.

[29] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson *et al.*, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[30] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable Architecture Search," in *ICLR*, 2018.

[31] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson *et al.*, "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech," in *INTERSPEECH*, 2021.

[32] Y. Lin, L. Wang, S. Li, J. Dang *et al.*, "Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription," in *INTERSPEECH*, 2020.

[33] J. Shor, D. Emanuel, O. Lang, O. Tuval *et al.*, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *INTERSPEECH*, 2019.

[34] T. Wang, J. Deng, M. Geng, Z. Ye *et al.*, "Conformer Based Elderly Speech Recognition System for Alzheimer's Disease Detection," in *INTERSPEECH*, 2022.

[35] X. Shi, P. Zhou, W. Chen, and L. Xie, "Darts-Conformer: Towards Efficient Gradient-Based Neural Architecture Search For End-to-End ASR," in *arXiv preprint arXiv:2104.02868*, 2021.

[36] S. Watanabe, T. Hori, S. Karita, T. Hayashi *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *INTERSPEECH*, 2018.

[37] S. Hu, X. Xie, M. Cui, J. Deng *et al.*, "Neural Architecture Search for LF-MMI Trained Time Delay Neural Networks," in *TASLP*, 2022.

[38] Z. Guo, X. Zhang, H. Mu, W. Heng *et al.*, "Single Path One-Shot Neural Architecture Search with Uniform Sampling," in *ECCV*, 2020.

[39] S. Xie, H. Zheng, C. Liu, and L. Lin, "SNAS: Stochastic Neural Architecture Search," in *ICLR*, 2019.

[40] S. Hu, S. Xie, H. Zheng, C. Liu *et al.*, "DSNAS: Direct Neural Architecture Search without Parameter Retraining," in *CVPR*, 2020.

[41] H. Christensen, S. Cunningham, C. Fox, P. Green *et al.*, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012.

[42] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, 1989.

[43] F. Xiong, J. Barker, and H. Christensen, "Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition," in *ICASSP*, 2019.