INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition

Jiajun Deng[1*], *Fabian Ritter Gutierrez*[1*], *Shoukang Hu*[1], *Mengzhe Geng*[1], *Xurong Xie*[2], *Zi Ye*[1], *Shansong Liu*[1], *Jianwei Yu*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

{jjdeng,ritter,skhu,mzgeng,zye,ssliu,jwyu,xyliu,hmmeng}@se.cuhk.edu.hk xr.xie@siat.ac.cn

Abstract

Automatic recognition of elderly and disordered speech remains a highly challenging task to date. Such data is not only difficult to collect in large quantities, but also exhibits a significant mismatch against normal speech trained ASR systems. To this end, conventional deep neural network model adaptation approaches only consider parameter fine-tuning on limited target domain data. In this paper, a novel Bayesian parametric and neural architectural domain adaptation approach is proposed. Both the standard model parameters and architectural hyper-parameters (hidden layer L/R context offsets) of two lattice-free MMI (LF-MMI) factored TDNN systems separately trained using large quantities of normal speech from the English LibriSpeech and Cantonese SpeechOcean corpora were domain adapted to two tasks: a) 16-hour DementiaBank elderly speech corpus; and b) 14-hour CUDYS dysarthric speech database. A Bayesian differentiable architectural search (DARTS) supernetwork was designed to allow both efficient search over up to 7^{28} different TDNN structures during domain adaptation, and robust modelling of parameter uncertainty given limited target domain data. Absolute recognition error rate reductions of 1.82% and 2.93% (13.2% and 8.3% relative) were obtained over the baseline systems performing model parameter fine-tuning only. Consistent performance improvements were retained after data augmentation and learning hidden unit contribution (LHUC) based speaker adaptation was performed.
Index Terms: speech recognition, domain adaptation, Bayesian learning, neural architecture search

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) technologies targeting normal speech [1–5] in the past few decades, accurate recognition of atypical speech task domains represented by, for example, elderly and dysarthric speech, remains a highly challenging task to date [6–9].

Ageing presents enormous challenges to health care and current speech technologies. Neurocognitive disorders (NCDs), such as Alzheimer's disease (AD), are often found among older adults [10] and manifest themselves in speech and language impairments including weakened neuro-motor control in speech production and imprecise articulation [11, 12]. Speech disorders such as dysarthria can also be caused by a range of other conditions including cerebral palsy, amyotrophic lateral sclerosis, stroke or traumatic brain injuries [13]. People with speech impairment often experience co-occurring physical disabilities and mobility limitations.

∗ Equal contribution

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 voice 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 [14, 15]. State-of-the-art ASR systems designed for normal speech often produce very high recognition error rate above 40% when being applied to elderly or impaired speech [9, 16, 17]. Second, the co-occurring disabilities, mobility or accessibility limitations often found among elderly and disordered speakers lead to the difficulty in collecting large quantities of such data that are essential for current data intensive deep learning based ASR system development.

To this end, a range of techniques designed to address the above domain mismatch and data sparsity issues have been studied in recent years primarily in the context of dysarthric speech recognition. Motivated by the spectral-temporal level differences of disordered speech from normal speech such as slower speaking rates, recent research in data augmentation has been largely focused on tempo-stretching [18], vocal tract length perturbation (VTLP) [19], and speed perturbation [20] of normal speech recorded from healthy control speakers. The resulting "disordered like" speech carrying a slower speaking rate and modified overall vocal tract spectral shape is then used to augment the limited dysarthric speech training data. Alternative approaches based on cross-domain DNN model or feature adaptation [21–23], domain adversarial training [24], transfer learning [25, 26], knowledge distillation [27], and voice conversion [28, 29] have also been investigated.

Among the above, model based domain adaptation approaches benefit not only from a tight integration of domain dependently estimated parameters with the underlying speech recognition error cost based on, for example, the lattice-free maximum mutual information (LF-MMI) criterion [25], or sequence to sequence learning objective functions, for example, used in recurrent neural network (RNN) transducers [22], but also fine modelling granularity in adapted parameters when sufficient target domain data is available.

However, there are two issues associated with model based domain approaches when being applied to elderly or disordered speech recognition tasks. First, due to the difficulty in collecting large quantities of such data, and the often limited amounts of existing elderly [30] or dysarthric speech datasets [31], direct fine-tuning of large numbers of out of domain, normal speech data estimated DNN model parameters on limited elderly or dysarthric speech data is generally problematic. The severe data sparsity issue and the resulting modelling uncertainty need to be

http://dx.doi.org/10.21437/Interspeech.2021-289

addressed. Second, the underlying neural architecture designs in current ASR systems are often designed using expert knowledge and empirical evaluation within individual task domains, for example, conversational telephone speech [32], or meeting transcription [33]. For example, the left and right splicing context offsets in the hidden layers of state-of-the-art LF-MMI trained time delay neural network (TDNN) systems [3, 34] represent the range of temporal contexts that can be exploited in modelling. The DementiaBank Pitt corpus [30], the largest publicly available elderly speech database, contains 4.8 words per utterance on average, in contrast to the normal speech data from the LibriSpeech corpus [35] of approximately 31 words per utterance. Similar designs based on shorter sentences also feature in current dysarthric speech corpora [31].

In order to address these issues, a novel Bayesian parametric and neural architectural domain adaptation approach is proposed in this paper. Both the standard model parameters and architectural hyper-parameters (hidden layer left and right context offsets[1]) of two LF-MMI factored TDNN systems separately trained using large quantities of normal speech from the English LibriSpeech and Cantonese SpeechOcean corpora were domain adapted to two tasks: a) 16-hour DementiaBank elderly speech corpus; and b) 14-hour CUDYS dysarthric speech database. Bayesian learning of differentiable architectural search (DARTS) [38] super-network was employed to allow both efficient search over up to 7^{28} different TDNN structures during domain adaptation, and robust modelling of parameter uncertainty given limited target domain data. Absolute recognition error rate reductions of 1.82% and 2.93% (13.2% and 8.3% relative) were obtained over the baseline systems performing model parameter fine-tuning only. Consistent performance improvements were retained after data augmentation and learning hidden unit contribution (LHUC) based speaker adaptation was performed. To the best of our knowledge, this is the first work to consider both parametric and architectural cross-domain adaptation for elderly and dysarthric speech recognition. In contrast, the majority of previous researches on domain adaptation for the same tasks have been focused on direct parameter fine-tuning [22, 23, 25, 26] while the data sparsity and architecture mismatch issues remain unsolved.

The rest of this paper is organized as follows. Section 2 presents Bayesian domain adaptation of LF-MMI trained TDNN systems. A novel differentiable architecture search approach automatically learning the L/R context offsets hyper-parameters of Bayesian TDNN systems is proposed in Section 3. Section 4 presents the experiments and results. Finally, the conclusions are drawn in Section 5.

2. Bayesian TDNN Adaptation

In contrast to conventional model adaptation methods performing fixed-value, deterministic parameter fine-tuning given limited target domain data, Bayesian adaptation approaches address the data sparsity issue by modelling parameter uncertainty using the following predictive distribution. Given an adaptation data set $\mathcal{D} = \{\mathbf{O}_r, \mathbf{H}_r\}$, where $\mathbf{O}_r$ and $\mathbf{H}_r$ are the r-th speech utterance and the reference word sequences, respectively. The prediction over the r-th test utterance $\mathbf{O}_r^*$ is given by

$$p(\mathbf{H}_r^*|\mathbf{O}_r^*, \mathcal{D}) = \int p(\mathbf{H}_r^*|\mathbf{O}_r^*, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \quad (1)$$

where $\mathbf{H}_r^*$ denotes the predicted word sequence for the test utterance r, $\mathbf{w}$ is the Bayesian adaptation parameters and $p(\mathbf{w}|\mathcal{D})$ is its posterior distribution learned from the adaptation data.

LF-MMI Trained TDNNs: TDNNs [39] produced state-of-art performance on different tasks [34, 40, 41]. TDNN is an instance of 1-dimension convolutional neural networks with parameters tying over different time steps. The lower TDNN layers are designed to learn narrower, local temporal contexts, while the higher layers learn wider, longer range contexts. The TDNN hidden left and right splicing context offsets are important hyper-parameters controlling its hierarchical temporal modelling ability. This paper adopted the factored TDNN [40].

In contrast to the conventional cross entropy criterion, sequence level error costs more closely related recognition accuracy, for example, the MMI [1] criterion, is widely used in state-of-the-art ASR systems [3, 42, 43].

$$\mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}) = \sum_r \log \frac{p(\mathbf{O}_r|\mathbf{H}_r)^\kappa P(\mathbf{H}_r)}{\sum_{\hat{\mathbf{H}}_r} p(\mathbf{O}_r|\hat{\mathbf{H}}_r)^\kappa P(\hat{\mathbf{H}}_r)} \quad (2)$$

where $\boldsymbol{\Theta}$ contains both hyper-parameters such as hidden layer context offsets and normal TDNN weight parameters, κ is the acoustic scaling factor and $\hat{\mathbf{H}}_r$ is the possible word sequence in the decoded speech lattice for utterance r. The efficient lattice-free MMI training [3] that alleviates the explicit denominator lattice generation is considered in this paper.

Bayesian TDNN Model Adaptation: During domain adaptation, the parameter posterior distribution $p(\mathbf{w}|\mathcal{D})$ required in the form of Bayesian prediction in Eqn. (1) can be learned by maximising the following MMI criterion marginalisation over all parameter estimates.

$$\mathcal{F} = \log \int \exp\{\mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta})\} P_r(\mathbf{w}) d\mathbf{w} \quad (3)$$

where $\mathbf{w} \in \boldsymbol{\Theta}$ and $P_r(\mathbf{w})$ is the prior distribution of adaptation parameters. Direct optimisation of the above integral is nontrivial. An alternative more efficient variational inference is utilized to learn the adaptation parameter posterior distribution by optimising the following lower bound,

$$\begin{aligned} \mathcal{F} &\geq \int q(\mathbf{w}) \mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}) d\mathbf{w} - KL(q(\mathbf{w})||P_r(\mathbf{w})) \\ &= \mathcal{L}_1^{MMI} - \mathcal{L}_2^{MMI} = \mathcal{L}^{MMI} \end{aligned} \quad (4)$$

where $q(\mathbf{w})$ is the variational approximation of the posterior distribution $p(\mathbf{w}|\mathcal{D})$ and $KL(q(\mathbf{w})||P_r(\mathbf{w}))$ is the Kullback-Leibler (KL) divergence between $q(\mathbf{w})$ and $P_r(\mathbf{w})$. For efficiency, and based on the previous research findings [41], both $q(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and $P_r(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu_r}, \boldsymbol{\sigma_r^2})$ are assumed to be Gaussian distributions. The first term $\mathcal{L}_1^{MMI}$ is approximated with Monte Carlo sampling method, which is given by

$$\mathcal{L}_1^{MMI} \approx \frac{1}{N} \sum_{k=1}^{N} \mathcal{F}_{MMI}(\mathcal{D}; \boldsymbol{\Theta}, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon_k}) \quad (5)$$

where ϵ_k is the k-th Monte Carlo sampling value drawn from the standard normal distribution $\mathcal{N}(0, 1)$ and $\odot$ is the Hadamard product. The KL divergence based second term $\mathcal{L}_2^{MMI}$ in Eqn. (4) can be explicitly calculated as

$$\mathcal{L}_2^{MMI} = \frac{1}{2} \sum_i \left(\frac{\sigma_i^2 + (\mu_i - \mu_{r,i})^2}{\sigma_{r,i}^2} + 2 \log \frac{\sigma_{r,i}}{\sigma_i} - 1 \right) \quad (6)$$

[1] Prior researchers suggested [36, 37] that TDNN context offset settings significantly affect the resulting system's temporal modelling resolution and recognition performance, while other hyper-parameters, e.g. the hidden layer dimensionality, were used to control the overall system complexity, thus not considered here.

where $\{\mu_{r,i}, \sigma_{r,i}\}$ and $\{\mu_i, \sigma_i\}$ are the i-th hyper-parameters of the prior distribution $\{\boldsymbol{\mu_r}, \boldsymbol{\sigma_r}\}$ and variational distribution $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$, respectively. The variational distribution parameters are updated during back propagation.

Implementation Details over several crucial settings are: 1) The first TDNN layer parameters practically exhibit more uncertainty due to the larger input data variability than those observed at higher layers designed to produce more invariant features. Based on our previous findings [41, 44], Bayesian domain adaptation was applied to the first layer of all TDNN systems in this paper, while the other higher layers parameters were fine-tuned to the target domain data. 2) The prior for all Bayesian adapted TDNN systems is based on the comparable fully converged standard fixed-parameter fine-tuning adapted TDNN systems. Other parameters in the Bayesian adapted TDNN systems are initialized using those of the halfway fine-tuned TDNN systems during adaptation. 3) The variational distribution variance is shared among all nodes of the first layer, which allows the number of parameters in Bayesian adapted TDNN system to be comparable to that of the standard fixed-parameter adapted system. 4) For efficiency, only one parameter sample is drawn in Eqn. (5) to ensure the computational cost in Bayesian adaptation to be comparable to that of the standard fine-tuning adapted TDNN system. During recognition time, the predictive inference integral in Eqn. (1) is efficiently approximated by the expectation of Bayesian adapted TDNN model parameters.

3. TDNN Architecture Adaptation

The general problem of TDNN hyper-parameter domain adaptation is transformed into a domain adaptive neural architecture search [36, 45] task within the DARTS [38] framework that allows both the architecture hyper-parameters and TDNN, or Bayesian TDNN parameters to be optimized consistently during adaptation to elderly or disordered speech data. An over-parameterized super-network containing paths connecting all neural architecture candidates is trained first, before the selection weights over each neural architecture candidate within the super-network are learned in the search stage. On convergence of the super-network model, the optimal architecture is obtained by pruning lower weighted paths. For example, the output $\mathbf{h}^l$ of l-th layer in the DARTS super-network is given by

$$\mathbf{h}^l = \sum_{i=0}^{N^l-1} \lambda_i^l \phi_i^l(\mathbf{W}_i^l \mathbf{h}^{l-1}) \tag{7}$$

where N^l denotes the number of architecture candidate selections in the l-th layer and λ_i^l is the weight of the i-th architecture candidate in the l-th layer. $\mathbf{W}_i^l$ and ϕ_i^l are the linear transformation parameter matrix and activation function for the i-th candidate system in the l-th layer, respectively.

Pipelined Gumbel-Softmax DARTS: In order to minimize the confusion between different architectures found in conventional Softmax based DARTS [38], a Gumbel-Softmax distribution [46] is used to produce approximately a one-hot vector, categorical architecture weights as the following

$$\lambda_i^l = \frac{\exp((\log \alpha_i^l + G_i^l)/T)}{\sum_{j=0}^{N^l-1} \exp((\log \alpha_j^l + G_j^l)/T)} \tag{8}$$

where α_i^l is the parameter in the Gumbel-Softmax distribution, $G_i^l = -\log(-\log(U_i^l)))$ is the Gumbel variable, U_i^l is a random variable sampled from a uniform distribution. T is the temperature hyper-parameter annealed from 1 to 0.03 in this paper.

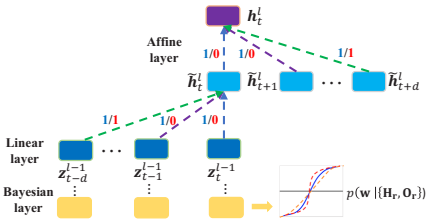

Figure 1 *Example DARTS super-network for Bayesian TDNNs (Bayesian layer in yellow square). Dashed lines in different colors are different Left/Right context offsets. The blue integers denote the super-network system using all context offsets, while the red integers represent a candidate offset choice of* $\pm$ *2.*

Following [36, 46], the update of TDNN parameters and architecture weights were performed in two stages, in a pipelined fashion, to avoid sub-optimal selection of architectures. In order to prevent overfitting to the training data, a separate held-out data set taken out of the original training data is used. In the first stage, the TDNN parameters are updated to convergence using the training data first, while randomly sampled one-hot architecture weights drawn from a uniform distribution are used in back-propagation. In the second stage, the TDNN parameters estimated in the first stage in the super-network are fixed and the architecture weights are updated using the held-out data.

TDNN-F Context Offset Search Space: The context offsets of TDNNs are crucial for modeling long temporal information in speech. Manually setting these hyper-parameters by evaluating a large number of possible system configurations is impractical. To this end, parameter sharing within the super-network can be used [47]. As shown in Fig. 1, all possible choices of context offsets to the left $(\{\text{-}d, 0\}, \cdots, \{1, 0\}, \{0, 0\})$ and right $(\{0, 0\}, \{0, 1\}, \cdots, \{0, d\})$ at each layer are incorporated into the TDNN-F super-network. The super-network designed for a L hidden layers TDNN contains $(d{+}1)^{2L}$ possible candidate models, each of which is indicated by setting the corresponding connecting weights as 1, while others as 0.

Architectural and Parametric Adaptation of Bayesian TDNNs is performed in three stages: a) **Architecture adaptation** is performed by first constructing a Bayesian TDNN super-network shown in Fig. 1 that contains all possible hidden layer context offset settings using the source domain data alone, before being adapted to the target domain. In this process, the very large number of standard TDNN parameters, often in tens of millions, are Bayesian adapted to ensure robustness on limited target domain data as described in Section 2, while the comparatively much smaller number of architecture selection weights, $2L(d{+}1)$ in total, linearly related to the number of hidden layers L and maximum context offset d, are fine-tuned during adaptation. b) **Architecture search** performed over the resulting domain adapted Bayesian TDNN super-network will then be searched over to produce the 1-best TDNN context offset settings. c) **Bayesian model adaptation** is finally performed by first constructing a TDNN system that features the above adapted architecture configurations but uses the source domain training data. The standard model parameters of this prior TDNN system are then further adapted in a Bayesian fashion to the target domain speech to produce the full architecture plus parameter adapted system.

4. Experiments

The proposed Bayesian parametric and architectural domain adaptation approach was investigated on two tasks for LF-

MMI factored TDNN systems: 1) from the English LibriSpeech speech corpus to DementiaBank elderly database; 2) from the Cantonese SpeechOcean corpus to the CUDYS dysarthric data.

Experimental Setup: The data sets and the baseline systems used in the two adaptation tasks are described below.

English Elderly Adaptation Task: A 1000 Hour LibriSpeech data set [35] is adopted as the source domain data. The DementiaBank database [30] is the target domain data, which includes 15.74-hour training set (9.72-hour elderly participant and 6.03-hour investigator data) and 3.14-hour test set (1.93-hour elderly participant and 1.21-hour investigator data) after silence stripping [8]. The word and duration per utterance on average in LibriSpeech (DementiaBank) corpus are 31 (4.8) and 11.3 (1.9) second, respectively. The training set was expanded to 59 hours when speed perturbation was performed. A 4-gram language model described in [8] was used.

Cantonese Dysarthric Adaptation Task: A Cantonese CUDYS dysarthric speech corpus [31] containing a 14.09-hour training set and 3.61-hour testing set with low and high intelligibility groups after silence stripping and speed perturbation [19] is utilized as the target domain data. 19.77-hour external data extracted from 163-hour Cantonese SpeechOcean normal speech corpus was mixed with 14.09-hour CUDYS training set for source-domain acoustic model training. The word and duration per utterance on average in the SpeechOcean (CUDYS) data set are 9.3 (1.6) and 4.0 (1.6) second, respectively. A 80k word 4-gram language model in [19] was used in recognition.

Baseline TDNN Systems: For the two domain adaptation tasks, LF-MMI factored TDNNs of 14 (English) and 7 (Cantonese) hidden layers were used[2], with the GMM-HMM system configuration the same as [8]. 40-dim filter-bank input features were used in both tasks. 100-dim i-vector features were appended for Librispeech and DementiaBank systems, while 3-dim pitch features were used for Cantonese SpeechOcean and CUDYS systems. For both tasks, the Bayesian architecture and parameter adaptation procedure in Section 3 was performed[3].

Performance of English Elderly DementiaBank: Table 1 demonstrates the performance of the DementiaBank corpus. Several trends are observed. First, the systems considering both architectural and parametric adaptation (Sys. 5, 6) outperform the corresponding systems only considering parameter adaptation (Sys. 3, 4) by up to **2.55%** absolute word error rate (WER) reductions. Second, further improvement by **0.38%** absolute WER reduction was obtained in the Bayesian architectural and parametric adapted systems (Sys. 6) over the corresponding architectural and parametric adapted systems without Bayesian estimation (Sys. 5). Finally, consistent performance improvements were retained after data augmentation and LHUC based speaker adaptation. In the cross-domain adapted systems, the largest absolute WER reduction up to **2.93%** was achieved by the Bayesian parametric and architectural adapted system (Sys. 6) over the parameter fine-tuning system (Sys. 3).

Table 1 *WERs (%) of TDNN systems trained using LibriSpeech or DementiaBank data alone, before domain adaptation of model parameters and optionally architecture (context offsets) w/o Bayesian estimation. (a,b) in the "context offsets" column denotes the context offsets {-a, 0} to left and {0, b} to right. † denotes a statistically significant difference obtained over the parametric fine-tuning baseline system (Sys. 3, 8, 13).*

Sys.	Data sets	Domain adaptation Arch.	Domain adaptation Para.	Context offsets 1-th to 14-th layer	Data aug.	LHUC SAT	DEV. PAR	DEV. INV	Eval. PAR	Eval. INV	ALL
1	LIBRI	✗	✗	1,1 1,1 0,0 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3 3,3	✗	✗	-	-	-	-	99.59
2	DEMEN	✗	✗	same as Sys. (1)	✗	✗	51.16	22.01	38.78	21.53	36.34
3	LIBRI → DEMEN	✗	FineTune				49.70	20.77	38.97	21.31	35.28
4		✗	Bayes [8, 41]				47.48	20.01	36.72	19.09	33.65†
5		DARTS	FineTune	5,0 0,4 0,2 0,4 5,4 2,6 0,6 0,6 0,5 6,5 6,5 6,6 6,5 6,6			46.45	**19.16**	35.78	**18.87**	32.73†
6		Bayes	Bayes	1,1 4,3 5,3 5,2 4,3 6,2 4,4 4,5 4,5 4,6 5,6 5,6 6,6 6,6			**45.31**	19.86	**34.35**	19.53	**32.35**†
7	DEMEN	✗	✗	same as Sys. (1)	✓	✗	46.94	20.06	36.97	19.98	33.53
8	LIBRI → DEMEN	✗	FineTune				46.91	19.29	36.64	20.09	33.15
9		✗	Bayes[8, 41]				45.90	19.84	35.15	19.53	32.71
10		DARTS	FineTune	0,6 0,5 5,5 4,5 6,6 6,5 0,6 0,6 0,6 0,6 0,5 0,6 0,6 6,5			45.25	**18.94**	35.46	21.09	32.19†
11		Bayes	Bayes	1,1 4,6 4,5 4,4 4,4 3,3 3,6 4,3 4,5 6,6 6,6 6,6 6,5 6,6			**44.56**	19.66	**33.68**	**17.87**	**31.81**†
12	DEMEN	✗	✗	same as Sys. (1)	✓	✓	44.95	18.52	35.33	**17.54**	31.77
13	LIBRI → DEMEN	✗	FineTune				44.16	19.12	34.16	19.42	31.56
14		✗	Bayes[8, 41]				44.08	19.11	34.22	18.87	31.52
15		DARTS	FineTune	same as Sys. (10)			43.75	**18.37**	33.84	19.53	31.04†
16		Bayes	Bayes	same as Sys. (11)			**43.36**	19.07	**32.08**	17.98	**30.83**†

[2]The DARTS systems perform the search over 7^{28} (English) and 7^{14} (Cantonese) TDNN-F choices with the maximum contexts of ± 6.

[3]A matched pairs sentence-segment word error based statistical significance test was performed at a significance level α=0.05.

Performance of Cantonese Dysarthric CUDYS: Results conducted on the CUDYS corpus are presented in Table 2 with similar trend to the DementiaBank task, absolute character error rate (CER) reductions of up to **1.61%** were obtained in the systems considering both architectural and parametric adaptation (Sys. 6, 7) over the corresponding parameter fine-tuning adapted systems (Sys. 4, 5). Second, the Bayesian architectural and parametric adapted systems (Sys. 7, 12) perform the best among other systems before and after speaker adaptation. In the cross-domain systems, the greatest absolute CER reduction up to **1.82%** was obtained by the Bayesian parametric and architectural adapted TDNN system (Sys. 7) over the parameter fine-tuning TDNN system (Sys. 4).

Table 2 *CERs (%) of TDNN systems trained using SpeechOcean or CUDYS data alone, before domain adaptation of model parameters and optionally architecture (context offsets) w/o Bayesian estimation. † denotes a statis. sig. diff. obtained over the parametric fine-tuning baseline system (Sys. 4, 9).*

Sys.	Data sets	Domain adaptation Arch.	Domain adaptation Para.	Context offsets 1-th to 7-th layer	LHUC SAT	DEV. High	DEV. Low	Eval. High	Eval. Low	ALL
1	SPOC CUDY	✗	✗	1,1 1,1 0,0 3,3 3,3 6,6 6,6	✗	32.92	98.51	14.24	94.97	36.12
2						9.94	88.09	1.30	85.89	19.80
3	SP. & CU. → CUDY	✗	✗	same as Sys. (1)	✗	4.97	85.32	**0.72**	70.36	15.37
4		✗	FineTune			5.67	79.47	1.02	57.14	13.74
5		✗	Bayes			**4.13**	74.68	0.90	52.30	12.22†
6		DARTS	FineTune	5,6 6,6 6,6 5,6 6,6 6,5 6,6		5.14	71.17	1.17	**48.88**	12.13†
7		Bayes	Bayes	6,4 5,6 6,6 6,6 6,6 6,6 6,6		4.77	**66.06**	1.49	49.05	**11.92**†
8	SP. & CU. → CUDY	✗	✗	same as Sys. (1)	✓	1.85	76.91	0.60	63.12	12.74
9		✗	FineTune			1.91	75.74	0.44	52.07	11.15
10		✗	Bayes			**1.15**	71.81	0.44	50.92	10.51†
11		DARTS	FineTune	same as Sys. (6)		2.01	**67.87**	**0.40**	45.49	9.96†
12		Bayes	Bayes	same as Sys. (7)		1.51	69.47	0.69	**45.51**	**9.41**†

5. Conclusions

The paper proposed a Bayesian parametric and neural architectural domain adaptation approach to rapidly port LF-MMI trained TDNNs based state-of-the-art ASR systems developed using large amounts of normal speech data to elderly and disordered speech task domains of more limited quantities. Experimental results suggest Bayesian adaptation can effectively mitigate the risk of overfitting when directly cross domain fine-tuning systems containing a large number of parameters. Architecture adaptation can further improve the generalization of systems using parameter adaptation only. Future research will focus on the adaptation of more advanced neural architectures.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200218, 14200220, TRS T45-407/19N, ITF grant No. ITS/254/19, and SHIAE grant No. MMT-p1-19.

7. References

[1] L. Bahl, P. Brown, P. De Souza *et al.*, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP*, 1986.

[2] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[3] D. Povey, V. Peddinti, D. Galvez *et al.*, "Purely sequence-trained neural networks for asr based on lattice-free MMI," in *INTERSPEECH*, 2016.

[4] W. Chan, N. Jaitly, Q. Le *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[5] Y. Wang, A. Mohamed, D. Le, C. Liu *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP*, 2020.

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

[7] J. Yu, X. Xie, S. Liu *et al.*, "Development of the CUHK dysarthric speech recognition system for the ua speech corpus," in *INTERSPEECH*, 2018.

[8] Z. Ye, S. Hu, J. Li *et al.*, "Development of the CUHK elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *ICASSP*, 2021.

[9] D. Wang, J. Yu, X. Wu *et al.*, "Improved End-to-End dysarthric speech recognition via meta-learning based model re-initialization," in *ISCSLP*, 2021.

[10] A. Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[11] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *J. Alzheimer's Dis.*, vol. 49, no. 2, pp. 407–422, 2016.

[12] A. König, N. Linz, J. Tröger *et al.*, "Fully automatic speech-based analysis of the semantic verbal fluency task," *Dementia and Geriatric Cognitive Disorders*, vol. 45, no. 3-4, pp. 198–209, 2018.

[13] W. Lanier, "Speech disorders." in *Greenhaven Publishing LLC*, 2010.

[14] T. Hixon and J. C. Hardy, "Restricted motility of the speech articulators in cerebral palsy." *J SPEECH HEAR DISORD*, vol. 29, pp. 293–306, 1964.

[15] R. D. Kent, J. F. Kent, G. Weismer *et al.*, "What dysarthrias can tell us about the neural control of speech," *J. Phonetics*, vol. 28, pp. 273–302, 2000.

[16] H. Albaqshi and A. Sagheer, "Dysarthric speech recognition using convolutional recurrent neural networks," *INT J INTELL SYST*, vol. 13, pp. 384–392, 2020.

[17] E. Hermann and M. M. Doss, "Dysarthric speech recognition with lattice-free MMI," in *ICASSP*, 2020.

[18] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*, 2019.

[19] M. Geng, X. Xie, S. Liu *et al.*, "Investigation of data augmentation techniques for disordered speech recognition," in *INTERSPEECH*, 2020.

[20] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *INTERSPEECH*, 2018.

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

[22] J. Shor, D. Emanuel, O. Lang *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," in *INTERSPEECH*, 2019.

[23] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *INTERSPEECH*, 2020.

[24] D. Woszczyk, S. Petridis, and D. Millard, "Domain adversarial neural networks for dysarthric speech recognition," in *INTERSPEECH*, 2020.

[25] F. Xiong, J. Barker, Z. Yue *et al.*, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*, 2020.

[26] Y. Takashima, R. Takashima, T. Takiguchi *et al.*, "Dysarthric speech recognition based on deep metric learning," in *INTERSPEECH*, 2020.

[27] Y. Lin, L. Wang, S. Li *et al.*, "Staged knowledge distillation for End-to-End dysarthric speech recognition and speech attribute transcription," in *INTERSPEECH*, 2020.

[28] D. Wang, J. Yu, X. Wu *et al.*, "End-to-End voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *ICASSP*, 2020.

[29] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang *et al.*, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," in *INTERSPEECH*, 2020.

[30] J. T. Becker, F. Boller, O. L. Lopez *et al.*, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Arch. Neurol.*, vol. 51, no. 6, pp. 585–594, 1994.

[31] K. H. Wong, Y. T. Yeung, E. H. Chan *et al.*, "Development of a Cantonese dysarthric speech corpus," in *ISCA*, 2015.

[32] W. Xiong, J. Droppo, X. Huang *et al.*, "Toward human parity in conversational speech recognition," *IEEE TASLP*, vol. 25, pp. 2410–2423, 2017.

[33] T. Hain, L. Burget, J. Dines *et al.*, "Transcribing meetings with the amida systems," *IEEE TASLP*, vol. 20, pp. 486–498, 2012.

[34] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[35] V. Panayotov, G. Chen, D. Povey *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[36] S. Hu, X. Xie, S. Liu *et al.*, "Neural architecture search for LF-MMI trained time delay neural networks," in *ICASSP*, 2021.

[37] T. Moriya, T. Tanaka, T. Shinozaki *et al.*, "Evolution-strategy-based automation of system development for high-performance speech recognition," *IEEE TASLP*, vol. 27, no. 1, pp. 77–88, 2018.

[38] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *ICLR*, 2018.

[39] A. Waibel, "Consonant recognition by modular construction of large phonemic time-delay neural networks," in *ICASSP*, 1989.

[40] D. Povey, G. Cheng, Y. Wang *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018.

[41] S. Hu, X. Xie, S. Liu *et al.*, "Bayesian learning of LF-MMI trained time delay neural networks for speech recognition." *IEEE TASLP*, vol. 29, pp. 1514–1529, 2021.

[42] H. Hadian, H. Sameti, D. Povey *et al.*, "End-to-end speech recognition using lattice-free mmi," in *INTERSPEECH*, 2018.

[43] W. Michel, R. Schlüter, and H. Ney, "Comparison of lattice-free and lattice-based sequence discriminative training criteria for LVCSR," in *INTERSPEECH*, 2019.

[44] S. Hu, X. Xie, S. Liu *et al.*, "LF-MMI training of Bayesian and Gaussian process time delay neural networks for speech recognition." in *INTERSPEECH*, 2019.

[45] S. Xie, H. Zheng, C. Liu *et al.*, "SNAS: Stochastic neural architecture search," in *ICLR*, 2019.

[46] C. J. Maddison, A. Mnih, and Y. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR*, 2017.

[47] H. Pham, M. Y. Guan, B. Zoph *et al.*, "Efficient neural architecture search via parameter sharing," in *ICML*, 2018.