Recent Progress in the CUHK Dysarthric Speech Recognition System

Shansong Liu, Mengzhe Geng, Shoukang Hu, Xurong Xie, Mingyu Cui, Jianwei Yu, Xunying Liu, *Member, IEEE*, and Helen Meng, *Fellow, IEEE*

***Abstract*—Despite the rapid progress of automatic speech recognition (ASR) technologies in the past few decades, recognition of disordered speech remains a highly challenging task to date. Disordered speech presents a wide spectrum of challenges to current data intensive deep neural networks (DNNs) based ASR technologies that predominantly target normal speech. This paper presents recent research efforts at the Chinese University of Hong Kong (CUHK) to improve the performance of disordered speech recognition systems on the largest publicly available UASpeech dysarthric speech corpus. A set of novel modelling techniques including neural architectural search, data augmentation using spectra-temporal perturbation, model based speaker adaptation and cross-domain generation of visual features within an audio-visual speech recognition (AVSR) system framework were employed to address the above challenges. The combination of these techniques produced the lowest published word error rate (WER) of 25.21% on the UASpeech test set 16 dysarthric speakers, and an overall WER reduction of 5.4% absolute (17.6% relative) over the CUHK 2018 dysarthric speech recognition system featuring a 6-way DNN system combination and cross adaptation of out-of-domain normal speech data trained systems. Bayesian model adaptation further allows rapid adaptation to individual dysarthric speakers to be performed using as little as 3.06 seconds of speech. The efficacy of these techniques were further demonstrated on a CUDYS Cantonese dysarthric speech recognition task.**

***Index Terms*—Disordered speech recognition, speaker adaptation, data augmentation, multimodal speech recognition.**

I. Introduction

DESPITE the rapid progress of automatic speech recognition (ASR) technologies in the past few decades, recognition of disordered speech remains a highly challenging task to date. Speech disorders such as dysarthria affect millions of people around the world and introduce a negative impact on their quality of life. Speech disorders are caused by a range of neuro-motor conditions including cerebral palsy [1], amyotrophic lateral sclerosis [2], Parkinson disease [3], stroke or traumatic brain injuries [4]. Common forms of speech disorders such as dysarthria manifest themselves in neuro-motor problems leading to weakness or paralysis of muscles that are used in articulation [5]. This reduces the intelligibility of the resulting speech for human listeners. As the underlying condition deteriorates, people suffering from speech disorders will not only lose their ability to express themselves but also to live independently. Such people often experience co-occurring physical disabilities and other medical conditions at the mean time. Their difficulty in using keyboard, mouse and touch screen based user interfaces makes speech controlled assistive technologies more natural alternatives [6], [7], even though speech quality is degraded. To this end, in recent years there has been increasing research interest in developing ASR technologies that are suitable for disordered speech [8]–[12].

Disordered speech presents a wide spectrum of challenges to current deep neural networks (DNNs) based speech recognition technologies that predominantly target normal speech. First, a large mismatch between disordered and normal speech is often observed. Such difference systematically manifests itself in articulatory imprecision, increased dysfluencies, slower speaking rates and reduced volume and clarity. Furthermore, people suffering from speech impairments tend to use shorter utterances based on isolated words and simple commands, due to the fatigue they encounter when speaking, when communicating with their careers. This limits the long range temporal contexts that current DNN based ASR systems designed for normal speech [13]–[15] can exploit.

A set of publicly available disordered speech corpora are shown in Table I. The Nemours [18] corpus contains less than

TABLE I
Description of Publicly Available Dysarthric Speech Corpora

Corpus	#Hours	#Spk.	Vocab.	Year
Cantonese (CUDYS) [16]	10	16	-	2015
Dutch (EST) [17]	6.3	16	-	2016
English (Nemours) [18]	2.5-3	11	-	1996
English (UASpeech) [19]	102.7	29	455	2008
English (TORGO) [20]	15	15	1573	2012

Manuscript received March 8, 2021; accepted June 13, 2021. Date of publication June 23, 2021; date of current version July 14, 2021. This work was supported in part by Hong Kong Research Grants Council GRF under Grants 14200218, 14200220, in part by Theme based Research Scheme under Grant T45-407/19 N, in part by Innovation and Technology Fund under Grants ITS/254/19, PiH/350/20, InP/275/20, and in part by the Shun Hing Institute of Advanced Engineering Grant MMT-p1-19. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sachin S. Kajarekar. *(Shansong Liu, Mengzhe Geng, Shoukang Hu, and Xurong Xie contributed equally to this work.) (Corresponding author: Xunying Liu.)*

Shansong Liu, Mengzhe Geng, Shoukang Hu, Mingyu Cui, Jianwei Yu, Xunying Liu, and Helen Meng are with the Chinese University of Hong Kong, Hong Kong 999077, china (e-mail: dadinghh2@gmail.com; mzgeng@se.cuhk.edu.hk; skhu@se.cuhk.edu.hk; mycui@se.cuhk.edu.hk; jwyu@se.cuhk.edu.hk; xyliu@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

Xurong Xie is with the Chinese University of Hong Kong, Hong Kong 999077, china, and also with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 100049, China (e-mail: xr.xie@siat.ac.cn).

Digital Object Identifier 10.1109/TASLP.2021.3091805

3 hours of speech from 11 speakers. The similar sized Dutch EST database [17] contains approximately 6 hours of speech. The English TORGO [20] and Cantonese CUDYS [16] corpora are moderately larger. The former contains 15 hours of speech while the latter contains around 10 hours of speech. By far, the largest available and widely used dysarthric speech database, the English UASpeech [19] corpus, contains 102.7 hours of speech[1] recorded from 29 speakers based on single word utterances of digits, computer commands, radio alphabet letters, common and uncommon words, among which 16 are dysarthric speakers while the remaining 13 are healthy control speakers. Compared with more widely available normal speech corpora, such as Switchboard conversational telephone speech [21] or Librispeech [22] containing hundreds of to thousands of audio data, all the existing disordered speech corpora are much smaller in size.

Second, the underlying neuro-motor conditions, often compounded with co-occurring physical disabilities, lead to the difficulty in collecting large quantities of disordered speech required for ASR system development. For data intensive deep learning technologies widely used in current speech recognition systems, large quantities of well-matched, in-domain speech data are essential. Finally, the large variation among speakers with diverse impairment characteristics, severity levels and in different stages of speech disorder progression creates large variation in disordered speech data. This presents a further challenge to the robustness of disordered speech recognition systems. For the above reasons, state-of-the-art ASR systems designed for normal speech often produce very high recognition error rate above 50% when being applied to impaired speech [23], [24].

In order to address these issues, the main part of this paper presents the recent research efforts made at the Chinese University of Hong Kong to significantly improve the performance of current disordered speech recognition systems on the largest available and widely used 102.7-hour UASpeech corpus. A set of purposefully designed modelling techniques were derived to address the aforementioned challenges. Both the description of individual approaches and how they can be integrated together to obtain the best recognition performance are presented.

First, motivated by the large mismatch between normal and disordered speech, a systematic investigation of neural network architecture designs targeting dysarthric speech recognition is conducted. These include state-of-the-art ASR system architectures based on either a hybrid DNN-HMM framework, for example, sequence discriminatively trained time delay neural networks (TDNNs) with phonetic states output targets [25]–[28], or end-to-end approaches represented by connectionist temporal classification (CTC) [29], attention based encoder-decoder models using listen, attend and spell (LAS) [30] and the recent Pychain end-to-end TDNN [31] systems directly modelling grapheme (letter) sequence outputs. A manually designed DNN architecture tailored for the disordered speech data of UASpeech is then proposed. Automatic neural architecture search (NAS) techniques [32]–[37] are further used to refine its structural configurations.

[1]Audio recordings collected from multiple microphone channels were used.

Second, in order to address the data sparsity problem in disordered speech recognition system development, and inspired by the success of data augmentation techniques widely reported in normal speech recognition tasks [15], [38], [39], data augmentation techniques designed to model the spectral-temporal level deviation of disordered speech from normal speech are used. A combined use of speaker independent perturbation of disordered speech and impaired speaker dependent perturbation using normal speech expands the training data quantity by a factor of 4 [40].

Third, in order to model the large variability among disordered speakers in both the original and augmented data, model based DNN adaptation methods represented by, for example, learning hidden unit contributions (LHUC) [41] based speaker adaptive training (SAT) were further applied. Bayesian speaker adaptation approaches were also employed to facilitate rapid, instantaneous adaptation to individual speakers' voice characteristics, using as little as 3.06 seconds of speech per speaker, at the onset of their enrollment to systems.

Lastly, inspired by the bi-modal nature of human speech perception and the success of audio-visual speech recognition (AVSR) technologies when being applied to normal speech [42]–[44], visual information is further incorporated to improve disordered speech recognition performance. In order to address the data sparsity that arises from the difficulty to record large amounts of high quality audio-visual (AV) data, a cross-domain visual feature generation approach [45] was developed. High quality AV parallel data based on normal speech recording of the lip reading sentence (LRS2) dataset [46] was used to build neural AV inversion systems. These were then used to generate visual features for the UASpeech audio data that do not have video recordings available. Cross-domain AV inversion system adaptation was also performed to minimize the mismatch between the LRS2 and UASpeech audio data.

By incorporating all the above techniques, the best recognition system produced an overall word error rate (WER) of 25.21% on the 22.6-hour UASpeech test set containing 16 dysarthric speakers. To the best of our knowledge, this is the lowest WER published so far on the same task reported in the literature [8]–[10], [40], [45], [47]. An overall WER reduction of 5.39% absolute (17.61% relative) was obtained over the CUHK 2018 system featuring a 6-way DNN system combination [10] which defined state-of-the-art performance at the time. A further set of experiments and performance analysis were then conducted on the Cantonese CUDYS [16] corpus which is based on a short sentence recognition task.

The main contributions of this paper are summarized below:

1) To the best of our knowledge, this is the first work to systematically investigate deep neural network architecture design for disordered speech recognition. In contrast, previous research in this area largely focused on using one single type of expert DNN architecture targeting normal speech [23], [48]. Detailed comparison and performance analysis between traditional hybrid DNN-HMM and more recent end-to-end approaches were not conducted in the prior works. In addition, novel auto-configured neural architecture search approaches are proposed in this paper for disordered speech recognition.

2) This paper presents the first work that investigates different data augmentation techniques for disordered speech recognition. Both normal and disordered speech were exploited in the augmentation process and evaluated over a wide range of expert hybrid or end-to-end, manually or automatically designed DNN system architectures. In contrast, previous research focused on using temporal perturbation performed only on normal speech data during augmentation [47], [49].

3) This paper presents the first work on rapid speaker adaptation for disordered speech recognition. The proposed Bayesian DNN adaptation approaches can capture the diverse characteristics among dysarthric speakers using as little as 3.06 seconds of speech. In contrast, previous research focused on batch mode adaptation required significant amounts of speaker level data, for example, over one hour on the UASpeech task [50].

4) This paper presents the first attempt of using cross-domain visual feature generation for audio-visual disordered speech recognition within a state-of-the-art AVSR system. This is contrast to previous AVSR research on disordered speech where the AV data sparsity was largely unaddressed [11], [51].

The rest of this paper is organized as follows. The details of system development on the UASpeech task are presented from Sec. II to V. Among these, a range of hybrid and end-to-end classic ASR system DNN architectures, together with manually designed DNN and neural architecture search (NAS) auto-configured DNN systems as well as their performance across varying speech disorder severity levels are first shown in Sec. II. Disordered speech data augmentation techniques are then presented in Sec. III. Performance of model based dysarthric speaker adaptation methods are shown in Sec. IV. Audio-visual disordered speech recognition systems and their performance are presented in Sec. V. Further performance analysis against recent published state-of-the-art systems constructed on the same UASpeech task is conducted in the same section (Sec. V). A comparable and smaller set of experiments and performance analysis were then conducted on the Cantonese CUDYS corpus to further confirm the trends previously found on the English UASpeech data. The last section draws the conclusions and discusses possible future works. For all results presented in this paper, matched pairs sentence-segment word error (MAPSSWE) based statistical significance test was performed at a significance level $\alpha = 0.05$.

II. ASR System Architecture

In this section, a large set of expert designed hybrid and end-to-end system architectures, together with manually designed DNN and neural architecture search (NAS) auto-configured DNN systems considered in this paper, are extensively evaluated in the experiments of this section on the UASpeech task.

The UASpeech corpus is the largest publicly available disordered speech corpus that is designed as an isolated word recognition task [19]. Approximately 103 hours of speech was recorded from 29 speakers among which 16 are dysarthric speakers while the remaining 13 are healthy control speakers. For speech recognition system development, the entire corpus is further divided into 3 subset blocks per speaker, with each block containing different speech contents based on a mix of common and uncommon words. Among these, the same set of common words contents are used in all three blocks, while the uncommon words in each block are different. The data from Block 1 (B1) and Block 3 (B3) of all the 29 speakers are used as the training set (69.1 hours of audio, 99 195 utterances in total), while the data of Block 2 (B2) collected from all the 16 dysarthric speakers (excluding speech from healthy control speakers) serves as the evaluation data set (22.6 hours of audio, 26 520 utterances in total). After removing excessive silence at the start and end of speech audio segments [10], a combined total of 30.6 hours of audio data from Block 1 and 3 (99 195 utterances) were used as the training set, while 9 hours of speech from Block 2 (26 520 utterances) was used for performance evaluation. Following the configurations specified in [10], [52], recognition was performed using a uniform language model with a word grammar network.

The performance of various expert neural architectures based recognition systems are shown in line 1 to 11 of Table II together with the modelling units, structural configurations, model complexity and error cost functions used in training. These systems include the hybrid frame level cross-entropy (CE) trained DNN model [10], with tied tri-phone state targets (Sys. 1), TDNN system [25]–[28] with tied tri-phone or tri-grapheme state targets (Sys. 2, 3), bi-directional long short-term memory (BLSTM) RNN modelling tied tri-phone state targets [10] (Sys. 4), and a set of end-to-end systems directly modelling phoneme or grapheme (letter) sequence outputs based on either the CTC [29] (Sys. 5, 6), LAS [24], [30] (Sys. 7, 8, 9), or the Pychain TDNN architecture with untied bi-phone or bi-grapheme outputs [31] (Sys. 10, 11). As the UASpeech training data set does not cover all the test data words, direct acoustic to word end-to-end approaches represented by RNN word transducers [53] are impractical for this task. Hence, the scope of the investigation over possible neural network architectures is restricted to those modelling either sub-word phonetic targets or grapheme labels.

The performance of the baseline CE trained hybrid DNN system modelling tied tri-phone state targets is shown in line 12 of Table II. This baseline system architecture was manually designed by applying a series of modifications on top of the first phonetic hybrid DNN system (Sys. 1 in Table II, also served as one of the component branches in our 2018 UASpeech system using system combination [10]). This seed phonetic hybrid DNN (Sys. 1 in Table II), serving as the starting point of our baseline DNN system development, contains 6 hidden layers, each with 2000 neurons using Sigmoid activation functions before the output layer. Acoustic features fed into the network are 80-dimension Mel-scale filter banks (FBKs) and delta features using a context of 9 consecutive frames. Decision tree tied tri-phone states are used and modeled using Softmax function at the output layer.

A set of architecture modifications are then performed on this 6-layer prototype DNN: 1) a 100-dimension bottleneck hidden layer is inserted immediately before the output layer followed with Sigmoid activation functions, in order to constrain the dimensionality while maintaining the necessary information for tri-phone state classification; 2) in order to address the issues of overfitting and vanishing gradient, a group of neural operations

TABLE II
1-BEST AND ORACLE PERFORMANCE AND SYSTEM DESCRIPTION OF EXPERT DESIGNED NEURAL NETWORK ARCHITECTURES (SYS. 1-11), MANUALLY DESIGNED DNN (SYS. 12) AND AUTOMATICALLY SEARCHED NEURAL ARCHITECTURE (SYS. 13). ALL SYSTEMS WERE TRAINED USING 80-DIMENSION INPUT FEATURES BASED ON MEL-SCALE FILTER BANKS (FBKS) AND DELTA FEATURES. "SEEN" AND "UNSEEN" DENOTE TEST SET WORDS OCCURRING IN THE TRAINING DATA OR OTHERWISE. "VERY LOW," "LOW," "MILD" AND "HIGH" DENOTE DIFFERENT INTELLIGIBILITY GROUPS

Sys.	Model	Structure	Obj.	Tgt.	#Param	WER%							Oracle WER%
						Seen	Unseen	Very low	Low	Mild	High	Average	
1	DNN	6-feedforward	CE	phn.	25.50M	21.81	54.24	67.20	36.34	28.63	15.65	35.20	17.35
2	TDNN	6-tdnn	CE	phn.	25.34M	24.17	53.99	69.38	40.28	30.29	14.68	35.80	15.42
3				gph.	25.36M	25.95	62.75	73.84	44.13	35.17	19.47	39.98	21.32
4	BLSTM	4-blstm	CE	phn.	29.52M	21.41	65.62	67.31	40.96	32.69	21.41	38.50	26.87
5	CTC	4-conv2d+ 3-blstm	CTC	phn.	25.99M	33.43	80.06	81.44	56.65	47.72	31.51	51.72	44.03
6				gph.	25.98M	39.37	86.80	87.05	62.42	55.33	37.78	57.98	54.05
7	LAS	4-conv2d+ 3-blstm encoder+ 1-lstm decoder	LAS	gph.	27.28M	27.39	99.24	74.74	56.18	52.55	43.92	55.28	39.48
8	LAS [24] (Librispeech 1000hr)	6-conv2d+ 5-blstm encoder+ 1-lstm decoder			42.78M	73.43	81.00	98.80	90.70	82.90	47.20	76.40	-
9	+domain/speaker adapt					23.03	53.55	68.70	39.00	32.50	12.20	35.00	-
10	Pychain TDNN	6-conv1d	LF-MMI	phn.	25.01M	26.56	39.32	66.50	36.20	24.13	10.27	31.57	12.14
11				gph.	25.84M	27.16	62.35	72.47	46.20	34.25	20.95	40.97	17.23
12	Manual DNN	7-feedforward	CE	phn.	5.86M	21.94	46.18	69.82	32.61	24.53	10.40	**31.45**	15.35
13	NAS DNN				4.73M	22.41	43.85	68.69	33.08	22.86	9.82	**30.83**	12.08

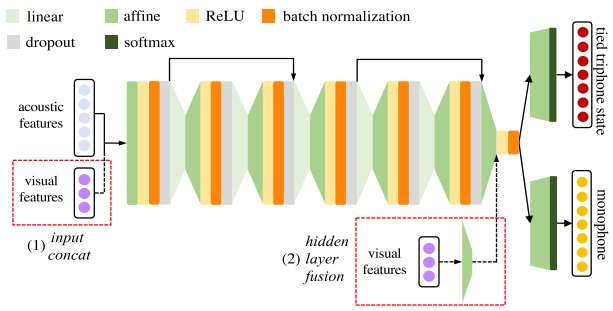


Fig. 1. Baseline system architecture of the manually designed DNN. The parts circled by red dotted boxes will be used later for audio-visual speech recognition system construction in Sec. V.

composed of ReLU activation, batch normalization [54] and dropout [55] are applied to each hidden layer except the newly added seventh bottleneck hidden layer; 3) in order to reduce the overall number of network parameters given the limited dysarthric speech available in the UASpeech corpus, the weight matrices of the first 5 hidden layers positioned before the ReLU activations were further decomposed into 5 pairs of factored 2000×200 linear and 200×2000 affine matrices, similar with those used in the factored TDNN system [25]–[28]; 4) in order to compensate the loss of hidden layer information due to the use of subspace factored weight matrices, skipping connections from the first to third, and from the fourth to sixth hidden layers are also added; 5) in order to reduce the risk of overfitting to unreliable tri-phone state alignments during system training, mono-phone alignments are introduced to form a second auxiliary task, with the multi-task weight set as 0.5. After the aforementioned modifications, the resulting manually designed baseline DNN architecture is shown in Fig. 1 (minus the fusion with video information in dotted parts later used for AVSR systems).

Several trends can be observed from Table II.

1) The hybrid LSTM system (Sys. 4), CTC and LAS based end-to-end systems (Sys. 5, 6, 7) traditionally designed for learning longer temporal contexts are outperformed by the comparable hybrid DNN, TDNN and Pychain TDNN systems (Sys. 1, 2, 3, 10, 11) modelling more restricted contexts.

2) As the training data does not cover all the test data words, the CTC and LAS systems (Sys. 5, 6, 7) produced a large disparity on WER between the seen and unseen words than the other systems in Table II. This may be in part attributed to the poor generalization of CTC and LAS systems when constructed using only the words found in UASpeech training data. Similar performance rank in terms of oracle error rates (last column in Table II) between the CTC and LAS systems against other systems in the table can also be found.

3) The above observations may be attributed to a combination of two factors: the comparatively more limited training data size and the mismatch against normal speech that manifests in, for example, the shorter utterance duration of approximately 3 seconds on average in the UASpeech data. In order to assess the impact from data quantity on the end-to-end systems performance, further experiments are then conducted. A 1000-hour Librispeech [22] (normal speech) data trained larger size (42.78 M parameters) LAS system (Sys. 8) gives a WER of 76.4% [24] when directly used to recognize the UASpeech data. After domain and speaker adaptation, this cross-domain adapted LAS system's WER (Sys. 9) was reduced to 35.0%, on par with the UASpeech data trained hybrid feedforward DNN system (Sys. 1) while still significantly ($\alpha = 0.05$) outperformed by the Pychain TDNN system (Sys. 10) using no out-of-domain speech data by 3.43% absolute in WER.

4) Compared with the starting point DNN system (Sys. 1) in table II, this baseline manually designed hybrid DNN (Sys. 12) produced an overall absolute WER reduction of 3.75%, as well as 77% relative reduction in model size. A similar model size reduction ratio was also obtained over various other hybrid and end-to-end systems (Sys. 2 to 7, 10, 11) in Table II. Based

on its performance and model compactness, the baseline DNN system architecture (Sys. 12) in Table II was used in the following neural architecture search experiments in the rest of this section to automatically learn the optimal subspace projection dimensionality at each hidden layer.

Neural architecture search (NAS) techniques [32] can efficiently automate neural network structure designs that have been largely based on expert knowledge or empirical choice to date. Among existing NAS methods, differentiable neural architecture search (DARTS) [33]–[37] benefits from a distinct advantage of being able to simultaneously compare a very large number of candidate architectures during search time. This is contrast to earlier and more expensive forms of NAS techniques based on, for example, genetic algorithms [56] and Reinforcement learning (RL) [57], [58], where explicit system training and evaluation are required for a large number of candidate structures under consideration.

Architecture search using DARTS is performed over an over-parameterized parent super-network containing paths connecting all candidate DNN structures to be considered. The search is transformed into the estimation of the weights assigned to each candidate neural architecture within the super-network. The optimal architecture is obtained by pruning lower weighted paths. This allows both architecture selection and candidate DNN parameters to be consistently optimized within the same super-network model.

With no loss of generality, we introduce the general form of DARTS architecture selection methods. For example, the l-th layer output $\mathbf{h}^l$ can be computed as follows in the super-network:

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

where l is the layer index, λ_i^l, $\mathbf{z}_i^l$ denote the architecture weight and input vector of the i-th candidate choice in layer l. N^l is the total number of choices of the layer l. The precise forms of neural architectures being considered at this layer is determined by the linear transformation parameter $\mathbf{W}_i^l$ and activation function $\phi_i^l(\cdot)$ used by each candidate system.

In conventional DARTS super-networks, Softmax functions are used to model the architecture selection weight λ_i^l. When the DARTS super-network containing both architecture weights and normal DNN parameters is trained to convergence, the optimal architecture can be obtained by pruning lower weighted architectures that are considered less important. However, when similar architecture weights are obtained using a flattened Softmax function, the confusion over different candidate systems increases and search errors may occur.

In order to address the above issue, Gumbel-Softmax function [59] is used in this paper to sharpen the distribution of architecture weights so that approximately one-hot vectors encoded 1 out of N selection decisions will be obtained. This allows the confusion of choosing different architectures to be minimized. The architecture weights of the Gumbel-Softmax

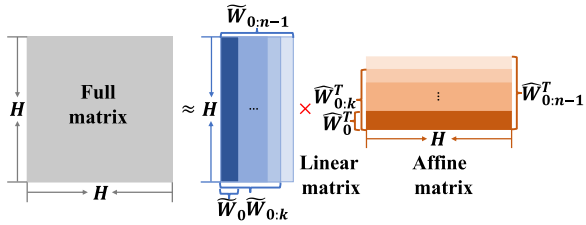


Fig. 2. Example part of a super-network containing different bottleneck projection dimensionality choices in one DNN hidden layer.

DARTS super-network are computed as,

$$\lambda_i^l = \frac{\exp((\log \alpha_i^l + G_i^l)/T)}{\sum_{j=0}^{N^l-1} \exp((\log \alpha_j^l + G_j^l)/T)} \tag{2}$$

where $G_i^l = -\log(-\log(U_i^l))$ is a Gumbel variable, and U_i^l is a uniform random variable. When the temperature parameter T approaches 0, the Gumbel-Softmax distribution is close to a categorical distribution [59]. The temperature parameter T in the Gumbel-Softmax distribution is annealed from 1 to 0.03 throughout our NAS experiments in this paper.

When using the back-propagation algorithm to update the architecture weight parameters, different samples of the uniform random variable U_i^l lead to different values of λ_i^l in Eq. 2. The loss function gradient, in a general form for both CE and LF-MMI criteria, w.r.t $\log \alpha_k^l$ is computed as an average over J samples of the architecture weights,

$$\frac{\partial \mathcal{L}}{\partial \log \alpha_k^l} = \frac{1}{J} \sum_{j=0}^{J} \frac{\partial \mathcal{L}}{\partial \mathbf{h}^{l,j}} \sum_{i=0}^{N^l-1} \frac{1_{i=k} \lambda_i^{l,j} - \lambda_i^{l,j} \lambda_k^{l,j}}{T} \phi_i^l(\mathbf{W}_i^l \mathbf{h}^{l-1,j}) \tag{3}$$

where $\boldsymbol{\lambda}^{l,j}$ is the j-th sample weights vector drawn from the Gumbel-Softmax distrbution in the l-th layer, $\mathbf{h}^{l,j}$ is the output of l-th layer by using the j-th sample $\boldsymbol{\lambda}^{l,j}$. The Gumbel-Softmax variables $\boldsymbol{\lambda}^l$ at different layers are assumed to be mutually independent during the sampling.

In order to find a trade-off between the model performance and complexity, a penalized term is further added to the loss function by incorporating the candidate network sizes,

$$\mathcal{L} = \mathcal{L}_{MTL} + \eta \sum_{l,i} \lambda_i^l C_i^l, \tag{4}$$

where C_i^l is the number of parameters of the i-th candidate considered at the l-th layer, and η is the penalty scaling factor.

In order to facilitate efficient search over a large number of candidate architectures with varying hidden layer specific projection dimensionality settings, parameter sharing among candidate architectures is also used. An example portion of a DARTS super-network containing all the candidate architectures with different projection dimensions is shown in Fig. 2. As this portion of super-network is positioned between the decomposed projection and affine linear layers, the activation function $\phi_i^l(\cdot)$ in Eqn. (1) is set as an identity matrix. Parameter sharing among different candidate architectures' linear matrices $\widetilde{\mathbf{W}}_{0:k}$ (left to right from the first column) and affine matrices $\widehat{\mathbf{W}}_{0:k}$ (bottom

to up from the first row) ($k \in [0, n-1]$) is implemented by the corresponding submatrices extracted from the large matrix $\widetilde{\mathbf{W}}_{0:n-1}$ and $\widehat{\mathbf{W}}_{0:n-1}$. Such sharing allows a large number of projection dimensionality choices at each of the hidden layers, e.g., selected from 8 values {25, 50, 80, 100, 120, 160, 200, 240}, as considered in this paper, to be simultaneously compared for selection during search. This corresponds to a total of $8^5 = 32768$ candidate DNN systems to be selected from. The 1-best auto-configured DNN architecture[2] with 4.73 million parameters produced by the above Gumbel-Softmax DARTS approach is shown in the last line (Sys. 13) in Table II. It not only has 20% fewer parameters than the manually designed baseline DNN system (Sys. 12), but also statistically significantly ($\alpha = 0.05$) reduced the WER by 0.6% absolute WER. The performance of this NAS DNN system, together with the end-to-end systems (Sys. 5 to 7, 10, 11) and the manual DNN baseline (Sys. 12) in Table II will be further evaluated in the following section using data augmentation techniques.

III. Data Augmentation

Current deep learning-based speech recognition systems are data and resource intensive. In order to reduce the risk of overfitting when constructing such systems using limited training data, data augmentation methods have been explored in the context of normal speech recognition tasks. By expanding the limited training data using, for example, tempo, vocal tract length or speed perturbation [60]–[62], spectral distortion and masking [15], [60], stochastic feature mapping [63], cross domain feature adaptation [64], simulation of noisy and reverberated speech to improve environmental robustness [65] and end-to-end back translation in end-to-end systems [66], the coverage of the augmented training data and the resulting speech recognition systems' generalization performance can be improved.

In contrast, so far only limited research on data augmentation targeting disordered speech recognition has been conducted. Motivated by the temporal level differences between disordered speech and normal speech such as slower speaking rates, recent research in this direction has been largely focused on tempo-stretching [47], [49] of normal speech recorded from healthy control speakers. The resulting "disordered like" speech carrying a slower speaking rate is used to augment the limited dysarthric speech training data. Alternative approaches based on cross-domain DNN adaptation [8], [10] and voice conversion [67] have also been investigated.

One issue associated with the above existing approaches is that either only applying a temporal level transformation to the normal speech signals while the spectral envelope remains the same, for example, in tempo-stretching [49], or a spectral level transformation, for example, using cross-domain feature adaptation [8], is applied while the speech tempo remains unaltered. Hence, data augmentation approaches that can exploit the full spectral-temporal differences between normal and disordered speech are preferred, including speaking rate, articulatory imprecision and changes in formant positions and volume. Furthermore, previous researches mainly focused on transforming out-of-domain normal speech to "disordered like" speech [68], [69], while data augmentation directly using existing disordered speech data has been very rarely studied.

TABLE III
Comparison of the Implementation Domain and Effects of VTLP, Tempo Perturbation and Speed Perturbation on Modified Speech Signal. "√" Indicates That Change Occurs After Perturbation

	VTLP	Tempo	Speed
Implement Domain	frequency	time	time
Signal Duration	unchanged	✓	✓
Spectral Envelope	✓	unchanged	✓

In this section, a systematic investigation over data augmentation techniques based on various spectral-temporal transformations is conducted for disordered speech recognition. The resulting augmented speech data is produced from two sources: a) spectral-temporal modification of normal speech of control speakers to "disordered like" speech of a target impaired speaker; and b) spectral-temporal perturbation of existing disordered speech. For each of the two sources, three data augmentation techniques were used. These include i) vocal tract length perturbation (VTLP) designed to only alter the spectral envelope to simulate different vocal tract lengths potentially resulted from imprecise articulators' movements while keeping the speech duration fixed; ii) tempo perturbation modifying the utterance duration to emulate the slower speaking rate in disordered speech while keeping the spectral shape and energy unchanged; and iii) speed perturbation that modifies speech signals in terms of both the duration and shape of the spectral envelope. A summary of these three perturbation methods is presented in Table III.

When performing perturbation of the existing disordered speech training data, a set of global perturbation factors, for example, {0.9,1.1} in case of VTLP and speed perturbation, were used. In contrast, when modifying the normal speech of control speakers to simulate that of a target impaired speaker, speaker-level perturbation factors were calculated as the average phonetic duration ratios between their respective speech obtained using phoneme alignment analysis [47]. Force alignment using a GMM-HMM system constructed using the HTK toolkit [70] was first performed. The resulting frame-level phoneme alignments were then used to compute the disordered speaker specific perturbation factor as $F_{D_j} = \frac{\overline{t_C}}{t_{D_j}}$. Here D_j denotes the j-th dysarthric speaker, $\overline{t_C}$ means the average time duration of all control speakers and t_{D_j} is the time duration of dysarthric speaker D_j.

The data augmentation techniques described in this section were implemented to expand the limited dysarthric speech training data while leaving the test set unchanged. The `HCopy` tool provided by HTK [70] was used to apply VTLP based frequency scaling. The `tempo` command based on the WSOLA algorithm [71] and `speed` command provided in Sox [72] were used for tempo perturbation and speed perturbation respectively.

[2]NAS selected projection dimensions at each layer: {160, 160, 160, 120, 120}. η in Eqn. 4 is set to be 0.21.

TABLE IV
PERFORMANCE OF VARIOUS HYBRID AND END-TO-END SYSTEMS TRAINED USING DIFFERENT DATA AUGMENTATION APPROACHES. "CTL" /"DYS" STANDS FOR NORMAL / DISORDERED SPEECH."TEMPO" /"SPEED" STANDS FOR TEMPO / SPEED PERTURBATION."2x,""4x" AND "6x" REFER TO THE AMOUNT OF AUGMENTED TRAINING DATA. THE #HOURS COLUMN SHOWS THE TOTAL QUANTITY OF SPEECH DERIVED FROM THE ORIGINAL TRAINING DATA SET OR AUGMENTED TRAINING DATA AFTER SILENCE STRIPPING APPLIED AT UTTERANCE BOUNDARIES

Sys.	Model	Tgt.	#Param	Pitch	Data Augmentation			#Hours	WER%						
					Method	CTL	DYS		Seen	Unseen	Very low	Low	Mild	High	Average
1	Manual DNN	phn.	5.86M	✗	✗			30.6	21.94	46.18	69.82	32.61	24.53	10.40	31.45
2				✓					21.33	42.74	67.93	30.35	22.31	9.46	29.73
3				✗	VTLP	1x	-	48.0	21.55	43.99	68.68	31.84	22.71	9.48	30.35
4					Tempo			52.2	23.44	44.68	70.71	32.78	25.12	10.32	31.77
5					Speed			52.2	21.45	43.02	67.52	31.55	21.96	9.57	29.92
6					VTLP	-	2x	65.5	20.85	44.10	69.98	30.08	21.39	9.65	29.97
7					Tempo			65.9	21.78	45.02	69.32	31.75	23.94	10.07	30.90
8					Speed			65.9	20.80	43.71	68.43	29.60	21.37	10.44	29.79
9					Speed	-	4x	100.9	19.95	44.22	67.20	29.86	21.45	10.04	29.47
10							6x	136.7	20.00	44.26	67.15	30.07	21.25	10.17	29.52
11						2x	2x	130.1	19.86	42.46	66.45	28.95	20.37	9.62	28.73
12						4x	4x	207.5	19.46	42.57	66.26	28.60	19.90	9.68	28.53
13				✓	Speed	2x	2x	130.1	20.08	42.10	66.39	29.51	20.56	9.07	**28.72**
14						4x	4x	207.5	19.62	42.50	66.70	28.87	20.49	9.06	28.60
15	NAS DNN		6.17M	✓	Speed	2x	2x	130.1	20.06	41.46	65.63	29.04	20.66	9.08	**28.46**
16	CTC	phn.	25.99M	✓	Speed	2x	2x	130.1	31.02	82.34	79.46	54.26	47.86	32.84	51.15
17		gph.	25.98M						35.42	88.60	84.06	60.29	52.96	37.61	56.27
18	LAS	gph	27.28M						24.18	99.24	72.71	53.13	50.10	43.07	40.67
19	Pychain TDNN	phn.	25.01M						26.37	37.10	66.34	32.98	22.90	10.63	30.58
20		gph	25.84M						31.53	68.17	78.05	51.02	40.84	24.63	45.90

Following [62], three sets of global perturbation factors, {0.9,1.1}, {0.9,0.95,1.05,1.1} and {0.85,0.9,0.95,1.05, 1.1,1.15}, were applied to obtain augmented data based on disordered speech. Speaker dependent perturbation factors discussed above were applied when modifying normal speech to "disordered like" speech.

80-dimension Mel-scale filter bank and delta features were then extracted from the augmented training data. Pitch parameters extracted using the Kaldi toolkit [73] consisting of probability of voicing (POV), normalized pitch, delta-log-pitch and their deltas were also incorporated into the feature front-ends [74] for subsequent system development in all the following experiments of this paper. The CTC, LAS, Pychain systems (Sys. 5 to 7, 10, 11) in Table II, together with the manually designed DNN baseline (Sys. 12) and the NAS auto-configured DNN system[3] (Sys. 13), were retrained using various augmented data sets and their performance contrast is shown in Table IV.

Several trends can be observed from the results of Table IV.

1) Among all the three data augmentation methods (VTLP, tempo or speed perturbation), taking the manually designed baseline hybrid phonetic DNN system (Sys. 3 to 8) for example, with similar amounts of augmented training data being used, speed perturbation (Sys. 5, 8) consistently outperformed the other two methods being perturbation was applied to either the healthy control speakers' data (Sys. 3 to 5) or the original dysarthric speech audio (Sys. 6 to 8).

2) Further experiments conducted on the same manually designed baseline DNN system suggest applying speed perturbation to both healthy control speaker's data and dysarthric speech (Sys. 11) outperformed applying it only to either of the two subsets of training data (Sys. 5 and Sys. 8).

3) Further increasing the amounts of augmented data produced by perturbing both healthy control speaker's data and dysarthric speech from 130.1 hours (Sys. 11, 13) to 207.5 hours (Sys. 12, 14) only led to marginal WER reductions of 0.1%-0.2% absolute, with or without using additional pitch features. Hence, the 130.1-hour augmented data set used by the manual designed DNN (Sys. 13) together with its associated pitch features were used and fixed as the training set for all subsequent UASpeech experiments of this paper.

4) The performance comparison between the manually designed DNN baseline (Sys. 13), the NAS auto-configured DNN,[4] CTC, LAS and Pychain TDNN systems (Sys. 15 to 20) suggests the performance ranking order among all systems constructed using the same 130.1-hour augmented data set in Table IV is consistent with that previously found in Table II where no data augmentation is used.

The manually designed DNN and NAS auto-configured DNN systems (Sys. 13, 15), both having the lowest average word error rates and most compact model sizes among all systems in Table IV, were then selected to conduct the following speaker adaptation and audio-visual recognition experiments in the rest of this paper.

IV. SPEAKER ADAPTATION

A key problem for many speech recognition tasks is to model the systematic and latent variation among diverse speech data. This often creates a large mismatch between the training and evaluation data leading to recognition performance degradation.

[3]Searched over 6 different choices of projection dimensions {80 120 160,200 240,300}.

[4]NAS selected projection dimensions at each layer using the 103.1 h augmented data set: {240 200,200 200,240}. η in Eqn. 4 is set to be 0.

A major source of such variability is attributable to speaker level characteristics representing factors such as accent and idiosyncrasy, or physiological differences that manifests in, for example, age or gender. For disordered speech recognition tasks considered in this paper, in addition to the wide range of variability factors found in normal speech, the underlying causes and severity levels of speech impairment among dysarthric speakers, also compounded by the spectral-temporal perturbations performed during the data augmentation stage discussed in Sec. III, are expected to further increase the diversity among impaired speakers.

To this end, speaker adaptation techniques play a vital role in current ASR systems. These can be characterized into several broad categories: a) auxiliary speaker embedding based approaches that encode speaker-dependent (SD) characteristics in a compact vector representation, for example, using i-vectors [75]; b) feature transformation based methods [76], [77] that are applied to acoustic front-ends and aim to produce canonical, speaker invariant input features using, for example, feature-space maximum likelihood linear regression (f-MLLR) transforms [76]; and c) model-based adaptation techniques [41], [78]–[81] that exploit specially designed SD DNN model parameters to compensate speaker level variability.

Compared with auxiliary feature and feature transformation based adaptation approaches, model-based adaptation methods have two advantages. First, the SD parameters can be jointly estimated together with the speaker independent parameters in the system consistently in speaker adaptive training (SAT) [80], in contrast to the often offline, separate estimation of i-vectors and f-MLLR transforms. Second, the amount of speaker specific adaptation data practically determines the SD parameters' modelling granularity. When a larger amount of speaker level adaptation data is available, the SD parameter's modelling resolution can be accordingly increased. This allows the trade-off between adapting only parts of the whole acoustic model and building complete speaker dependent systems to be flexibly adjusted. In contrast, for auxiliary feature and feature transform based adaptation, SD feature embedding and transforms of fixed sizes are often used.

In this section, several model-based adaptation approaches are first used to construct speaker adaptively trained disordered speech recognition systems based on the manually designed and auto-configured NAS DNN systems (Sys. 13 and 15 of Table IV) in Sec. III using the associated augmented data (using speed perturbation of both control and dysarthric speech). SD transforms that are applied to various parts of the DNN acoustic model, include a) learning hidden unit contributions (LHUC) scaling vectors applied to the hidden layer outputs for each target speaker [78]; b) parameterized activation functions (PAct) with speaker level vector scaling or bias applied to the input feature before fed into the ReLU activations [79]; and c) hidden unit bias vectors (HUB) [81] adding speaker level offset vectors to the hidden unit outputs. The differences between these three model-based adaptation methods when being applied to the hidden layer ReLU activations inside the manually designed and auto-configured NAS DNN systems of line 13 and 15 of Table IV in Sec. III, are illustrated in Fig. 3.

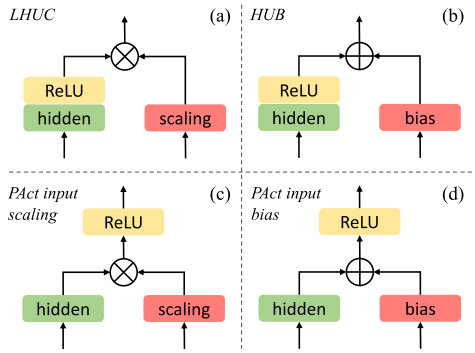


Fig. 3. Schematic representation of model based adaptation methods including learning hidden unit contributions (LHUC), hidden unit bias vectors (HUB) and parameterised activation functions (PAct). Speaker dependent scaling or bias parameters are marked in red.

TABLE V
PERFORMANCE COMPARISON BETWEEN DIFFERENT ADAPTATION METHODS INCLUDING AUXILIARY SPEAKER EMBEDDING (I-VECTOR) AND MODEL BASED ADAPTATION (LHUC, HUB AND PACT). IN MODEL BASED ADAPTATION, THE ADAPTED NUMBER OF HIDDEN LAYERS REMAIN THE SAME DURING SPEAKER ADAPTIVE TRAINING AND TEST TIME ADAPTATION. "IS" AND "IB" STAND FOR INPUT SCALING AND INPUT BIAS, RESPECTIVELY. "†" DENOTES A STATISTICAL SIGNIFICANCE DIFFERENCE IS OBTAINED OVER THE SYSTEM WITH ONE ADAPTED HIDDEN LAYER (SYS. 3 OR 15)

Sys.	Model	Adapt Method		Adapt Pos.	WER%				
					Very low	Low	Mild	High	Avg.
1	Manual DNN	✗		✗	66.4	29.5	20.6	9.1	28.7
2		i-vector			66.4	29.6	19.9	8.1	28.3
3		LHUC		1	63.3	28.2	17.7	7.7	26.7
4				1-2	67.1	27.7	17.8	8.1	27.5
5				1-3	64.2	27.9	18.0	7.6	26.8
6				1-4	63.1	26.5	18.4	7.8	26.4
7				1-5	65.4	26.8	18.0	7.9	26.9
8				1-6	63.8	26.8	17.0	7.8	26.4
9				1-7	62.4	26.8	16.5	7.5	**25.9**†
10		HUB		1-7	63.6	27.8	19.1	8.6	27.2
11		PAct	IS		65.3	27.2	17.8	7.6	26.9
12			IB		64.1	28.9	18.4	8.2	27.4
13	NAS DNN	✗		✗	65.6	29.0	20.7	9.1	28.5
14		i-vector			66.2	28.8	18.3	8.1	27.7
15		LHUC		1	60.7	27.0	18.0	8.0	26.0
16				1-2	62.5	26.7	17.1	7.4	25.9
17				1-3	61.4	26.7	16.2	7.6	**25.6**†
18				1-4	61.7	26.2	16.5	7.7	25.6
19				1-5	61.6	26.1	17.4	7.9	25.8
20				1-6	62.0	26.7	18.2	7.9	26.2
21				1-7	63.8	28.6	18.5	8.4	27.3
22		HUB		1-3	63.1	27.7	18.9	8.6	27.1
23		PAct	IS		61.5	27.1	16.7	7.5	25.8
24			IB		63.2	27.5	17.6	7.6	26.4

The performance of different speaker adaptation techniques when applied to the manually designed DNN and NAS auto-configured DNN systems are shown in Table V. In all the speaker adaptation experiments of this section, the 1-best outputs produced by the un-adapted speaker independent systems (Sys. 1, 13 in Table V, and earlier in Table IV as Sys. 13, 15) served as the supervision for subsequent test time adaptation of speaker adaptively trained (SAT) systems constructed using various adaptation methods introduced in this section. Considering the average amount of speaker specific data used in

these experiments is approximately 34 minutes (after silence stripping) for each dysarthric speaker, considerably larger than that found in other ASR tasks such as the Switchboard corpus, the number of DNN hidden layers on which LHUC, HUB or PAct transforms are applied was also fine-tuned to increase modelling resolution of the SD parameters, and to obtain the best adaptation performance for each technique.

Several trends can be observed from Table V.

1) On both the manually designed DNN and NAS auto-configured DNN systems, all of the model based adaptation methods including LHUC, HUB and PAct (Sys. 3 to 12 and Sys. 15 to 24) consistently outperformed i-vector input feature based adaptation (Sys. 2, 14).

2) Among all three model based adaptation methods, the best performance was obtained using the LHUC adapted manually designed DNN and NAS auto-configured DNN systems (Sys. 9, 17), producing overall statistically significant ($\alpha = 0.05$) WER reductions of 2.8% and 2.9% respectively over the un-adapted baseline systems (Sys. 1, 13).

The above speaker adapted systems' performance were obtained using approximately 34 minutes of adaptation data from each impaired speaker. As discussed in Sec. I, the underlying neuro-motor conditions, when compounded with co-occurring physical disabilities, lead to the difficulty in collecting large quantities of disordered speech from each target speaker. When developing speech based assistive technologies for such people with speech impairment, it is preferable to employ more powerful adaptation approaches to facilitate rapid, instantaneous adaptation to individual speakers' voices. However, when performing adaptation using very little speaker level data, for example, a few seconds of speech, a severe data sparsity issue and the resulting modelling uncertainty need to be addressed.

To this end, the inherent SD parameter uncertainty resulted from limited adaptation data is addressed using Bayesian learning approaches. Rather than learning fixed value estimates of the SD LHUC, HUB or PAct parameters using the standard cross-entropy cost function, the following Bayesian predictive inference (Eqn. 5) incorporating the SD adaptation parameter uncertainty is used instead.

$$P(\tilde{C}^s|\tilde{\mathbf{O}}^s, \mathbf{O}^s, C^s) = \int P(\tilde{C}^s|\tilde{\mathbf{O}}^s, \mathbf{r}^s) p(\mathbf{r}^s|\mathbf{O}^s, C^s) d\mathbf{r}^s \quad (5)$$

where $\mathbf{O}^s$, $\tilde{\mathbf{O}}^s$ denote the adaptation data and test data for speaker s, C^s stands for the corresponding supervision label, $\mathbf{r}^s$ denotes the SD parameters of speaker s, and $\tilde{C}^s$ refers to the output states to be inferred.

The key task of Bayesian adaptation is to learn the underlying SD parameter posterior distribution $p(\mathbf{r}^s|\mathbf{O}^s, C^s)$ used to encode modelling uncertainty. This distribution can be efficiently learned and approximated as a multi-variate Gaussian distribution using a variational inference approach combined with parameter sampling [81], [82]. For efficiency, the expectation of SD parameters can be used to approximate the Bayesian integral in Eqn. 5 for inference during recognition time.

$$P(\tilde{C}^s|\tilde{\mathbf{O}}^s, \mathbf{O}^s, C^s) \approx P(\tilde{C}^s|\tilde{\mathbf{O}}^s, \mathbb{E}[\mathbf{r}^s|\mathbf{O}^s, C^s]) \quad (6)$$

TABLE VI
PERFORMANCE OF BASELINE OR BAYESIAN ADAPTATION ON 130-HOUR AUGMENTED DATA SET TRAINED MANUALLY DESIGNED DNN AND NAS AUTO-CONFIGURED DNN SYSTEMS USING VARYING REDUCING AMOUNTS OF ADAPTATION DATA FROM 80% DOWN TO ONLY 1 UTTERANCE OF SPEECH FROM EACH TARGET DYSARTHRIC SPEAKER (DURATION IN BRACKETS). THE BOLD RESULTS INDICATE THE SMALLEST AMOUNTS OF SPEAKER LEVEL ADAPTATION DATA THAT CAN PRODUCE PERFORMANCE IMPROVEMENTS OVER THE UN-ADAPTED SYSTEMS (SYS. 1 OR 15) AFTER SPEAKER ADAPTATION. "†" DENOTES A STATISTICAL SIGNIFICANCE DIFFERENCE OVER THE UN-ADAPTED SYSTEMS

Sys.	Model	Adapt Method	Apt. Pos.	WER% w.r.t. adapt amounts (utt num)							
				1 Utt (3.06s)	1% (50.8s)	5% (4.2m)	10% (8.5m)	20% (17.0m)	40% (33.9m)	80% (1.1h)	all (1.4h)
1	Manual DNN	✗	✗	-							28.7
2		i-vector									28.3
3		LHUC	1-7	51.6	45.3	37.8	32.2	30.4	**27.7**†	25.9	25.9
4		BLHUC		44.7	36.2	32.4	30.1	29.6	**27.8**†	26.7	26.6
5		HUB		**28.6**	28.4	28.0	28.0	27.6	27.5	27.3	27.2
6		BHUB		**27.5**†	27.6	27.5	27.3	27.5	27.4	27.4	27.4
7		PAct		47.9	43.7	35.7	32.1	30.4	**27.9**†	27.2	26.9
8		BPAct		36.8	32.8	31.6	30.1	**28.7**	27.3	27.0	26.6
9		LHUC	1-3	46.7	41.8	33.5	31.2	29.7	**27.7**†	27.0	26.8
10		BLHUC		40.8	31.8	28.8	**27.9**†	27.5	26.8	26.3	26.2
11		HUB		29.6	29.4	**28.7**	27.8	27.7	27.2	27.7	27.4
12		BHUB		29.1	28.8	**28.3**†	27.7	28.1	27.8	27.9	27.9
13		PAct		47.1	41.7	33.6	30.9	29.0	**27.4**†	26.6	26.6
14		BPAct		35.8	31.1	30.0	**28.7**	27.8	26.9	26.6	26.6
15	NAS DNN	✗	✗	-							28.5
16		i-vector									27.7
17		LHUC	1-3	441.	34.7	29.6	**28.1**†	27.1	26.4	25.7	25.6
18		BLHUC		35.7	31.5	**27.9**†	27.2	26.5	26.2	26.1	26.0
19		HUB		29.9	28.9	**28.2**†	27.7	27.9	27.4	27.1	27.1
20		BHUB		29.6	**26.9**†	26.9	26.7	26.8	26.7	26.8	26.7
21		PAct		35.4	31.2	29.2	**27.9**†	27.5	26.2	25.8	25.8
22		BPAct		33.6	30.9	**28.2**†	27.2	26.5	26.2	26.1	26.0

where $\mathbb{E}[\cdot]$ denotes the expectation. Other symbols in Eqn. 6 are the same with those used in Eqn. 5. An example of applying Bayesian SD estimation to various model-based adaptation approaches of Table V, leading to Bayesian LHUC (BLHUC), Bayesian HUB (BHUB) and Bayesian PAct (BPAct) respectively, are shown in Fig. 4.

A series of Bayesian adaptation experiments were then conducted in order to demonstrate the minimum amounts of dysarthric speaker level data that can produce statistically significant ($\alpha = 0.05$) recognition performance improvements over the un-adapted baseline systems (Sys. 1, 15 in Table VI). This can help improve the resulting ASR system's practical deployment when new dysarthric speakers are freshly enrolled to the system.

During Bayesian adaptation, the SD parameter prior distribution used was empirically set as $\mathcal{N}(0, 0.001)$ for all adaptation methods. The SD adaptation parameters were estimated using either as fixed values in baseline adaptation, or in a Bayesian fashion, while the SI portion of the parameters inherited from the SAT trained systems were kept fixed. Based on the full data set adaptation results in Table V (Sys. 11, 23 vs. Sys. 12, 24), only input scaling based PAct adaptation is considered here together with LHUC and HUB. Performance contrasts between the baseline and Bayesian adaptation methods using varying reduced amounts of speaker level data randomly sampled from 80% down to as little as 1 single utterance (3.06 seconds of speech on average) are shown in Table VI. The following trends can be found.

1) Irrespective of which of the three model adaptation methods being used, Bayesian adaptation consistently outperform the comparable baseline adaptation using fixed value estimation across varying reduced amounts of speaker level data from 40% down to 1 single utterance. For example, on the HUB adapted

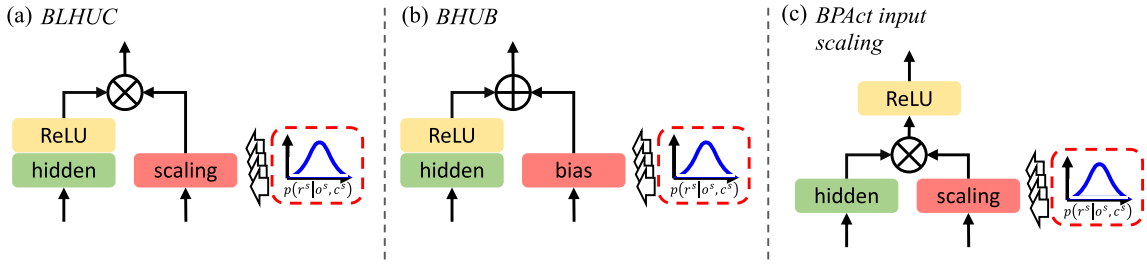


Fig. 4. Schematic representation of different Bayesian adaptation methods including Bayesian LHUC (BLHUC), Bayesian HUB (BHUB) and Bayesian PAct (BPAct). The red dotted box at the bottom right corner of each plot represents the SD parameter uncertainty modelled by Bayesian learning.

NAS auto-configured DNN systems (Sys. 19, 20), when only 1% (50.8 seconds) of speaker level data is used, Bayesian HUB adaptation (Sys. 20) significantly ($\alpha = 0.05$) outperformed the comparable HUB adapted baseline (Sys. 19) by 2.0% and the un-adapted system (Sys. 15) by 1.6% absolute in WER respectively.

2) The bold numbers in Table VI indicate the minimum amounts of speaker level data that can produce statistically significant ($\alpha = 0.05$) recognition performance improvements over the un-adapted baseline systems (Sys. 1, 15) for each adaptation technique. It is clear that for the manually designed DNN system with Bayesian HUB adaptation (Sys. 6), only a single utterance of approximately 3.06 seconds of speech is required to produce a statistically significant ($\alpha = 0.05$) WER reduction of 1.2% absolute over the un-adapted baseline system (Sys. 1).

V. Audio-Visual Speech Recognition

Inspired by the bi-modal nature of human speech perception and the success of audio-visual speech recognition (AVSR) technologies when being applied to normal speech [42]–[44], visual information is further incorporated to improve disordered speech recognition performance. In order to address the data sparsity resulting from the difficulty to collect large amounts of high quality audio-visual (AV) data, a cross-domain visual feature generation approach [45] was developed to generate visual features for the UASpeech original audio data and the augmented audio only data obtained using the speed perturbation method presented in Sec. III. This allows sufficient AV parallel disordered speech data to be used to develop AVSR systems.

High quality AV parallel data based on normal speech recording of the LRS2 dataset [46] was used to construct AV inversion neural network systems. However, the resulting AV inversion system cannot be directly applied to dysarthric speech given its large mismatch against the normal speech data in the LRS2 corpus. This mismatch may render the generated visual features unreliable to use for subsequent AVSR system development [45], [83]. Such mismatch can be compensated using, for example, domain-adversarial neural network (DANN) [84] or multi-level adaptive network (MLAN) [64]. Following the comparative analysis over domain adaptation methods for AV inversion in our previous research [83], the MLAN method was adopted to minimize the domain mismatch between the LRS2 and UASpeech audio data.

An example MLAN network consisting of two DNN components is shown in the left portion of Fig. 5. Each component DNN contains a bottleneck layer positioned immediately

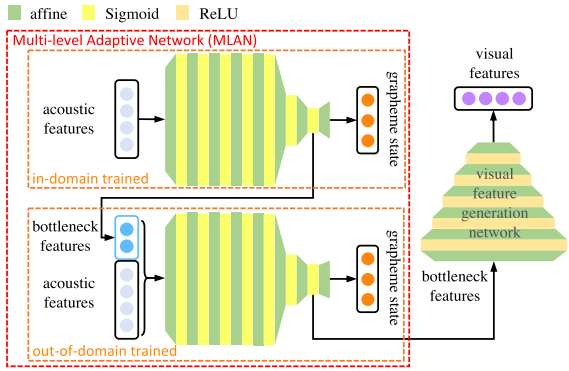


Fig. 5. Cross-domain visual feature generation system. The left part is the MLAN network consisting of two DNN components, while the DNN on the right is the AV inversion model using bottleneck features as cross-domain adapted inputs from the second DNN component of the MLAN network.

before the output layer. The MLAN training process includes the following steps: 1) the first-level DNN was trained with the audio data from the in-domain UASpeech corpus; 2) the resulting in-domain dysarthric speech trained DNN was then used to produce bottleneck features for the out-of-domain data of the LRS2 audio; 3) the second-level DNN was trained using the out-of-domain LRS2 audio data concatenated with the bottleneck features computed from the previous step.

When feedforwarding the UASpeech data into the resulting MLAN network, the combined effect produced by these two cascaded component DNNs is such that the final bottleneck features produced at the second-level DNN component will exhibit smaller mismatch against the bottleneck features obtained by feedforwarding the LSR2 audio into the MLAN network. These cross-domain adapted bottleneck features are used in AV inversion model training and visual feature generation (in the right part of Fig. 5). The dimensionality of these MLAN bottleneck features was set to 80, in line with the settings used in our earlier research [11]. The resulting cross-domain adapted AV inversion system was then applied to 103.1-hour augmented training data previously derived using speed perturbation in Sec. III (used by Sys. 13, 15 in Table IV), as well as the test set, to produce 25-dimension visual features for AVSR systems development.

The performance of various AVSR systems based on either the manually designed DNN or NAS auto-configured DNN architecture and constructed using the above cross-domain generated visual features are shown in lines 2 to 4, and 6 to 8 in Table VII .

TABLE VII
PERFORMANCE OF 103.1-HOUR AUGMENTED TRAINING DATA BASED LHUC SAT ADAPTED AUDIO-ONLY AND AUDIO-VISUAL SYSTEMS USING VARIOUS AV MODALITY FUSION: A) INPUT FEATURE CONCATENATION; B) HIDDEN LAYER FUSION; AND C) SCORE FUSION

Sys.	Model	Vis.	AV Fusion	WER%				
				Very low	Low	Mild	High	Avg.
1	Manual DNN	✗	✗	62.37	26.84	16.47	7.55	25.87
2		✓	input	63.51	27.93	19.84	9.75	27.79
3			7th hidden	63.19	26.60	16.80	8.23	26.28
4			score A+AV	**61.31**	**25.64**	**15.94**	7.68	**25.28**†
5	NAS DNN	✗	✗	61.42	26.72	16.25	7.65	25.63
6		✓	input	63.40	26.97	20.05	9.19	27.37
7			7th hidden	61.81	27.00	15.88	8.40	25.97
8			score A+AV	**60.30**	**26.23**	**15.39**	7.96	**25.21**†

TABLE VIII
PERFORMANCE COMPARISON BETWEEN VARIOUS RECENTLY PUBLISHED SYSTEMS' WERS ON THE UASPEECH TEST SET OF 16 DYSARTHRIC SPEAKERS AND OUR BEST SYSTEM IN THIS PAPER (SYS. 8, TABLE VII)

Systems	WER%
Sheffield-2013 Cross domain augmentation [8]	37.50
CUHK-2021 LAS + CTC + Meta-learning + SAT [24]	35.00
Sheffield-2015 Speaker adaptive training [9]	34.80
Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation [85]	30.76
CUHK-2018 DNN system combination [10]	30.60
CUHK-2021 QuartzNet + CTC + Meta-learning + SAT [24]	30.50
Sheffield-2019 Kaldi TDNN + Data Aug. [47]	27.88
CUHK-2020 Cross-domain AVSR [45]	26.84
CUHK-2020 DNN + Data Aug. + LHUC SAT [40]	26.37
NAS DNN + Data Aug. + LHUC SAT + AV fusion (ours)	**25.21**

In these systems, three forms of audio-visual modalities fusion were used including: a) input audio-visual feature concatenation (Sys. 2, 6 in Table VII, also shown in the left part of Fig. 1); b) hidden layer fusion performed by concatenating the visual features with outputs of the last non-bottleneck hidden layer (Sys. 3, 7 in Table VII, also shown in the right part of Fig. 1); and c) score fusion (Sys. 4, 8 in Table VII) via a linear interpolation over the output layer probability scores of the baseline audio-only ASR system (Sys. 1 or 5), and those of the hidden layer fusion based AVSR systems (Sys. 3 or 7) using equal weights. The results suggest that score fusion between ASR and AVSR systems (Sys. 4, 8) consistently produced statistically significant ($\alpha = 0.05$) WER reductions of 0.4%-0.6% over the comparable audio-only ASR systems (Sys. 1, 5).

The lowest WER of 25.21% was obtained using the NAS auto-configured DNN AVSR system (Sys. 8 in Table VII, again shown in the last line in Table VIII). To the best of our knowledge, this is the lowest WER published so far on the UASpeech test set of 16 dysarthric speakers reported in the literature. Performance contrasts between this system against previously published systems on the same task are shown in Table VIII. In particular, compared with our CUHK 2018 system featuring a 6-way DNN system combination [10] which defined state-of-the-art performance at the time, an overall WER reduction of 5.39% absolute (17.61% relative) was obtained. Furthermore, if excluding the 4 dysarthric speakers of very low intelligibility, the average WER obtained using our final AVSR system (Sys. 8, Table VII) is 15.79%, close to the WERs found on normal speech recognition tasks.

VI. EXPERIMENTS ON THE CANTONESE CUDYS CORPUS

In this section, a comparable set of modelling components and techniques that previously featured in the best performing systems on the English UASpeech task: neural architecture search based DNN auto-configuration of Sec. II, speed perturbation based data augmentation of Sec. III and LHUC speaker adaptive training of Sec. IV, were further evaluated on a Cantonese CUDYS dysarthric speech corpus [16].

The original 10-hour CUDYS corpus was further enlarged with more dysarthric speech collected since its initial release in 2015, and now contains speech from 27 impaired speakers. The development and evaluation sets, which were derived from a subset of 3.6 hours of speech collected from 21 impaired speakers and based on short sentences, were used for performance evaluation. The remaining part of the CUDYS data, after being further supplemented with normal Cantonese speech data from the SpeechOcean collection,[5] formed a baseline training data set of 21.4 hours. After speed perturbation based data augmentation techniques of Sec. III was applied, the training data size was further increased to 33.9 hours. In contrast to the UASpeech task based on single word utterances, each utterance in this task contains an average of more than six characters. Hence, a manually configured lattice-free MMI [26] trained factorized TDNN (f-TDNN) baseline system [25]–[28] with 7 context-splicing layers was used to model longer acoustic contexts. 40-dimension Mel-scale filter banks together with pitch parameters were used as the inputs features for system development. The Gumble-Softmax DARTS based neural architecture search approach of Sec. II was then applied to automatically learn the left and right context offsets[6] and the linear projection layer dimensionality[7] of each factored TDNN hidden layer. Speaker level variability was modelled using LHUC SAT and test time unsupervised adaptation. Due to the poor quality of video recordings in the CUDYS corpus caused by non-frontal face poses, and the difficulty in accessing high quality Cantonese audio-visual normal speech corpora with accurate transcripts required for the MLAN cross-domain visual feature generation approach of Sec. V, experiments were conducted on audio-only ASR systems for the CUDYS task. A 4-gram language model with a 80 K vocabulary was used. The character error rate (CER) metric was used for performance evaluation.

The performance comparison between the baseline manually configured TDNN, NAS auto-configured DNN, before and after data augmentation and LHUC SAT were applied, and further against the comparable graphemic (character) LAS, phonetic CTC and Pychain TDNN systems are shown in Table IX. The same trends as previously observed on the English UASpeech task in Sec. II to Sec. IV can be found. First, data augmentation reduced the CER by a statistically significant ($\alpha = 0.05$) margin of 4.2% absolute when the baseline LF-MMI TDNN system

[5]http://en.speechocean.com/datacenter/recognition.html

[6]Maximum context offset is set to be 6 for both left and right in each layer.

[7]Searched over 6 different choices of projection dimensions {100 120,160 200,240 300}.

TABLE IX
DESCRIPTION AND PERFORMANCE OF MANUALLY DESIGNED, NAS AUTO-CONFIGURED LF-MMI TRAINED FACTORED TDNN SYSTEMS AND VARIOUS END-TO-END SYSTEMS CONSTRUCTED ON THE CUDYS CORPUS. "CTL" MEANS PERTURBING THE HEALTHY SPEECH TOWARDS "DISORDER LIKE" SPEECH AND "DYS" MEANS PERTURBING THE EXISTING DYSARTHRIC SPEECH DATA

Sys.	Model	Tgt.	# Para	LHUC SAT	Data Aug.		CER%				
							dev		eval		Avg.
					CTL	DYS	Low	High	Low	High	
1	TDNN	phn.	9.8M	✗	✗		88.0	13.4	73.6	1.4	19.6
2					✗	✓	85.5	7.1	66.1	1.1	15.8
3					✓		85.3	5.0	70.4	0.7	15.4
4			10.1M	✓			76.9	1.9	63.1	0.6	12.7
5	NAS	phn.	10.2M	✗	✓		81.5	4.1	60.2	0.6	13.5
6	TDNN		10.6M	✓			71.9	3.2	50.2	0.5	11.2
7	CTC	phn.	9.8M	✗	✓		87.3	16.9	78.1	9.1	25.1
8	LAS	char.					81.7	10.4	72.0	8.5	20.9
9	Pychain TDNN	phn.					82.0	6.6	76.2	3.1	17.7

was retrained using the larger augmented data set of 33.9 hours (Sys. 3 vs. Sys. 1). Second, NAS auto-configured TDNN[8] also reduced the CER significantly ($\alpha = 0.05$) by 1.9% absolute prior to speaker adaptation being applied (Sys. 5 vs. Sys. 3). The improvement from the NAS auto-configured TDNN system was retained after LHUC SAT and unsupervised speaker adaptation (Sys. 6 vs. Sys. 4). Large and statistically significant ($\alpha = 0.05$) CER differences were found between the best performing NAS auto-configured TDNN system (Sys. 6) and the end-to-end systems of similar complexity.

VII. CONCLUSION

This paper presented a series of developments associated with the design of state-of-the-art dysarthric speech recognition systems on the largest publicly available UASpeech dysarthric English speech corpus and a Cantonese CUDYS dysarthric speech dataset. Experimental results suggest the following trends.

First, the suitability of current data-intensive deep learning based speech recognition system architectures, for example, end-to-end systems that traditionally benefit from the use of large quantities of data, needs to be re-assessed when being applied to dysarthric speech recognition tasks. This is due to the limited data quantity resulted from the difficulty in impaired speech data collection, and the large mismatch against normal speech. To this end, the auto-configured model structures derived from neural architecture search have been shown to produce better performance than a range of expert designed or manually configured systems of comparable or larger model complexity, before and after data augmentation or domain adaptation is used. Second, data augmentation techniques can effectively expand the limited training data by taking into account the systematic spectral and temporal deviation of dysarthric speech from normal speech. Third, the proposed speaker adaptation techniques can model the large variability among impaired speakers in both the original and augmented data, as well as allow fast adaptation to individual dysarthric speakers to be effectively performed using as little as a few seconds of speech. This user-centric feature is important when practically deploying speech recognition based assistive technologies to serve such people. Lastly, the use of visual features can further improve the recognition performance particularly for impaired speakers of low intelligibility whose voice quality is severely degraded.

The combination of these techniques produced the lowest published word error rate (WER) of 25.21% on the UASpeech test set 16 dysarthric speakers, and an overall WER reduction of 5.39% absolute (17.61% relative) over a very complex CUHK 2018 dysarthric speech recognition system using a 6-way DNN system combination and cross adaptation of out-of-domain normal speech data trained systems. Similar trends of performance improvements obtained using these techniques were also found on the CUDYS Cantonese dysarthric speech recognition task. The average WER over dysarthric speakers on the English UASpeech task obtained by our best AVSR system, if excluding the most difficult speakers of very low intelligibility, is 15.79%. This is considered to be close to the WERs often found on normal speech recognition tasks. Future research will focus on designing neural network architectures and multi-modal speech recognition systems suitable for dysarthric speakers of very low intelligibility.

[8]NAS selected projection dimensions at each layer: {160 160,100 100,120 160,300} and context configurations {-4,6},{-5,4},{-6,6},{-6,6},{-6,6},{-6,6},{-6,6}. η in Eqn. 4 is set to be 0.

ACKNOWLEDGMENT

We thank Disong Wang for sharing their cross-domain LAS system results.

REFERENCES

[1] T. L. Whitehill and V. Ciocca, "Speech errors in cantonese speaking adults with cerebral palsy," *Clin. Linguist. Phonet.*, vol. 14, no. 2, pp. 111–130, 2000.
[2] T. Makkonen, H. Ruottinen, R. Puhto, M. Helminen, and J. Palmio, "Speech deterioration in amyotrophic lateral sclerosis after manifestation of bulbar symptoms," *Int. J. Lang. Commun. Disorders*, vol. 53, no. 2, pp. 385–392, 2018.
[3] S. Scott and F. Caird, "Speech therapy for Parkinson's disease," *J. Mach. Learn. Res.*, vol. 46, no. 2, pp. 140–144, 1983.
[4] P. Jerntorp and G. Berglund, "Stroke registry in malmö, sweden," Stroke, vol. 23, no. 3, pp. 357–361, 1992.
[5] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *J. Commun. Disorders*, vol. 33, no. 5, pp. 391–428, 2000.
[6] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *Augment. Altern. Commun.*, vol. 16, no. 3, pp. 186–196, 2000.
[7] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assist. Technol.*, vol. 22, no. 2, pp. 99–112, 2010.
[8] H. Christensen *et al.*, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 3642–3645.
[9] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proc. SLPAT: 6th Workshop Speech Lang. Process. Assistive Technol.*, 2015, pp. 65–71.
[10] J. Yu *et al.*, "Development of the CUHK dysarthric speech recognition system for the UA speech corpus," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2938–2942.
[11] S. Liu *et al.*, "Exploiting visual features using bayesian gated neural networks for disordered speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4120–4124.
[12] S. Hu *et al.*, "The CUHK dysarthric speech recognition systems for english and cantonese," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3669-3670.
[13] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5934–5938.

[14] C. Lüscher *et al.*, "RWTH asr systems for LibriSpeech: Hybrid vs attention," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 231–235.

[15] Daniel S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.

[16] K. H. Wong, Y. T. Yeung, E. H. Chan, P. C. Wong, G.-A. Levow, and H. Meng, "Development of a cantonese dysarthric speech corpus," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 329–333.

[17] E. Yilmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A dutch dysarthric speech database for individualized speech therapy research," in *LREC*, 2016, pp. 792–795.

[18] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1962–1965.

[19] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 1741–1744.

[20] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.

[21] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1992, pp. 517–520.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[23] E. Hermann and M. M. Doss, "Dysarthric speech recognition with lattice-free mmi," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6109–6113.

[24] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[25] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3214–3218.

[26] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2751–2755.

[27] A. Waibel, H. Sawai, and K. Shikano, "Consonant recognition by modular construction of large phonemic time-delay neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1989, pp. 112–115.

[28] D. Povey *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3743–3747.

[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[30] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[31] Y. Shao, Y. Wang and D. Povey, and S. Khudanpur, "PyChain: A fully parallelized pytorch implementation of LF-MMI for End-to-End ASR," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 561–565.

[32] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no, 1, pp. 1997–2017, 2019.

[33] H. Liu *et al.*, "Darts: Differentiable architecture search," in *Proc. ICLR*, 2019.

[34] S. Xie, H. Zheng, C. Liu, and L. Lin, "Snas: Stochastic neural architecture search," in *Proc. ICLR*, 2019.

[35] H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware," in *Proc. ICLR*, 2019.

[36] S. Hu *et al.*, "DSNAS: Direct neural architecture search without parameter retraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12084–12092.

[37] S. Hu *et al.*, "Neural architecture search for LF-MMI trained time delay neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6758–6762.

[38] T.-S. Nguyen, S. Stüker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7689–7693.

[39] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "SpecSwap: A simple data augmentation method for end-to-end speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 581–585.

[40] M. Geng *et al.*, "Investigation of data augmentation techniques for disordered speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 696–700.

[41] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 171–176.

[42] J. Yu *et al.*, "Audio-visual recognition of overlapped speech for the LRS2 dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6984–6988.

[43] J. Yu *et al.*, "Audio-visual multi-channel recognition of overlapped speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3496–3500.

[44] J. Yu *et al.*, "Audio-visual multi-channel integration and recognition of overlapped speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2067–2082, 2011, doi: 10.1109/TASLP.2021.3078883.

[45] S. Liu *et al.*, "Exploiting cross-domain visual feature generation for disordered speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 711–715.

[46] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2018.2889052.

[47] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5836–5840.

[48] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6395–6399.

[49] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 471–475.

[50] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Comput. Speech Lang.*, vol. 27, no. 6, pp. 1147–1162, 2013.

[51] E. S. Salama, R. A. El-Khoribi, and M. E. Shoman, "Audio-visual speech recognition for people with speech disorders," *Int. J. Comput. Appl.*, vol. 96, no. 2, pp. 51–56, 2014.

[52] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 1159–1163.

[53] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2012.

[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[56] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evol. Comput.*, vol. 10, no. 2, pp. 99–127, 2002.

[57] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. ICLR*, 2017.

[58] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4095–4104.

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

[60] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," *in Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2013, pp. 309–314.

[61] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. Int. Conf. Mach. Learn.*, vol. 117, 2013.

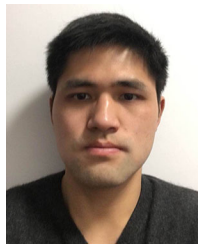
[62] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3586–3589.

[63] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1469–1477, Sep. 2015.

[64] Bell *et al.*, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2012, pp. 324–329.
[65] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5220–5224.
[66] T. Hayashi *et al.*, "Back-translation-style data augmentation for end-to-end ASR," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 426–433.
[67] D. Wang *et al.*, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7744–7748.
[68] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6009–6013.
[69] T. M. Celin, T. Nagarajan, and P. Vijayalakshmi, "Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 346–354, Feb. 2020.
[70] S. Young *et al.*, "The HTK book," *Cambridge Univ. Eng. Dept.*, vol. 3, no. 175, 2006.
[71] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1993, pp. 554–557.
[72] *Sox, audio manipulation tool*. Accessed Feb.10, 2020. [Online]. Available: http://sox.sourceforge.net/.
[73] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automat. Speech Recognit.* Understanding, 2011, pp. 1–4.
[74] S. Liu, S. Hu, X. Liu, and H. Meng, "On the use of pitch features for disordered speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4130–4134.
[75] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2013, pp. 55–59.
[76] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.
[77] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalisation," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 2527–2530.
[78] J. Neto *et al.*, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," *EUROSPEECH*, pp. 2171–2174, 1995.
[79] C. Zhang and P. C. Woodland, "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5300–5304.
[80] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1137–1140.
[81] X. Xie, X. Liu, T. Lee, and L. Wang, "Bayesian learning for deep neural network adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2096–2110, 2021.
[82] X. Xie, X. Liu, T. Lee, S. Hu, and L. Wang, "Blhuc: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5711–5715.
[83] R. Su, X. Liu, L. Wang, and J. Yang, "Cross-domain deep visual feature generation for mandarin audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 185–197, 2020.
[84] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[85] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7424–7428.

Shansong Liu received the B.E. degree in automation from Sichuan University, Chengdu, China, in 2014 and the M.S. degree in control science and engineering from Tsinghua University, Beijing, China, in 2017. He is currently working toward the Ph.D. degree with The Chinese University of Hong Kong, Hong Kong. His current research interests include disordered speech recognition and multimodal speech recognition.

Mengzhe Geng received the B.Sc. degree in 2019 in mathematics and information engineering from The Chinese University of Hong Kong, Hong Kong, where he is currently working toward the Ph.D. degree. His current research interests include data augmentation and speaker adaptation.



Shoukang Hu received the B.E. degree in mechanical and electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently working toward the Ph.D. degree with The Chinese University of Hong Kong, Hong Kong. His current research interests include speech recognition, bayesian modeling, and neural architecture search.

Xurong Xie received the B.S. degree in mathematics and applied mathematics from Sun Yat-sen University, Guangzhou, China, in 2011, the M.S. degree in computational statistics and machine learning from University College London, London, U.K. in 2012, and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong, Hong Kong, in 2020. He is currently a Member of the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Beijing, China, as a Research Assistant Professor. His research interests include adaptation techniques for acoustic modelling, speech processing for disordered speech, and statistical methods applied to speech technologies, including Bayesian learning.

Mingyu Cui reveived the B.E. degree in computer science and software engineering from SouthEast University, Nanjing, China, in 2019. Since 2020, he has been a Research Assistant with The Chinese University of Hong Kong, Hong Kong, and in 2021, will become a Ph.D. Student. His current research interests include language model and neural architecture search.

Jianwei Yu received the B.E. degree from Nanjing University, Nanjing, China, in 2017. He is currently working toward the Ph.D. degree with The Chinese University of Hong Kong, Hong Kong. His current research interests include Language modelling, and multimodel speech enhancement and recognition.

Xunying Liu (Member, IEEE) received the Ph.D. degree in speech recognition and the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K., prior to his undergraduate study with Shanghai Jiao Tong University, Shanghai, China. He was a Senior Research Associate with Machine Intelligence Laboratory, Cambridge University Engineering Department, and since 2016, has been an Associate Professor with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. His current research interests include machine learning, large vocabulary continuous speech recognition, statistical language modelling, noise robust speech recognition, audio-visual speech recognition, computer aided language learning, speech synthesis, and assistive technology. He and his students were the recipient of a number of Best Paper awards and nominations, including the Best Paper Award at ISCA Interspeech2010 for the paper titled Language Model Cross Adaptation for LVCSR System Combination, and the Best Paper Award at IEEE Proceding IEEE International Conference on Acoustics, Speech Signal Process 2019 for their paper titled BLHUC: Bayesian Learning of Hidden Unit Contributions for Deep Neural Network Speaker Adaptation. He is a Member of ISCA.

Helen Meng (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. In 1998, she joined The Chinese University of Hong Kong, Hong Kong, where she is currently the Chair Professor with the Department of Systems Engineering and Engineering Management. She was the former Department Chairman and the Associate Dean of Research with the Faculty of Engineering. Her research interests include human–computer interaction via multimodal and multilingual spoken language systems, spoken dialog systems, computer-aided pronunciation training, speech processing in assistive technologies, health-related applications, and big data decision analytics. Between 2009 and 2011, she was the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. She was the recipient of the IEEE Signal Processing Society Leo L. Beranek Meritorious Service Award in 2019. She was also on the Elected Board Member of the International Speech Communication Association and an International Advisory Board Member. She is a Member ISCA, HKCS, and HKIE.