INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

On-the-Fly Feature Based Rapid Speaker Adaptation for Dysarthric and Elderly Speech Recognition

Mengzhe Geng[1,4,], Xurong Xie[2], Rongfeng Su[3], Jianwei Yu[4], Zengrui Jin[1], Tianzi Wang[1], Shujie Hu[1],Zi Ye[1], Helen Meng[1], Xunying Liu[1]*

[1]The Chinese University of Hong Kong, [2]Institute of Software, Chinese Academy of Sciences
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, [4]Tencent AI Lab

{mzgeng,zrjin,twang,zye,hmmeng,xyliu}@se.cuhk.edu.hk, xurong@iscas.ac.cn, rf.su@siat.ac.cn, tomasyu@tecent.com

Abstract

Accurate recognition of dysarthric and elderly speech remain challenging tasks to date. Speaker-level heterogeneity attributed to accent or gender, when aggregated with age and speech impairment, create large diversity among these speakers. Scarcity of speaker-level data limits the practical use of data-intensive model based speaker adaptation methods. To this end, this paper proposes two novel forms of data-efficient, feature-based on-the-fly speaker adaptation methods: variance-regularized spectral basis embedding (SVR) and spectral feature driven f-LHUC transforms. Experiments conducted on UASpeech dysarthric and DementiaBank Pitt elderly speech corpora suggest the proposed on-the-fly speaker adaptation approaches consistently outperform baseline iVector adapted hybrid DNN/TDNN and E2E Conformer systems by statistically significant WER reduction of 2.48%-2.85% absolute (7.92%-8.06% relative), and offline model based LHUC adaptation by 1.82% absolute (5.63% relative) respectively.

Index Terms: Speaker Adaptation, Rapid Adaptation, Dysarthric Speech, Elderly Speech, Speech Recognition

1. Introduction

Despite the breakthroughs in automatic speech recognition (ASR) technologies targeting normal speech, accurate recognition of dysarthric and elderly speech remains highly challenging tasks to date [1–18]. Speech impairments are commonly found among dysarthric speakers and the elderly experiencing natural aging and neurocognitive disorders [19]. ASR technologies tailored to their needs can improve their quality of life.

Dysarthric and elderly speech presents a prominent challenge to current ASR technologies primarily targeting normal speech in many aspects. Heterogeneity commonly found in normal speech sourcing from accent or gender, when further combined with that over age and speech impairment severity, create large diversity among dysarthric and elderly speakers [20, 21]. Such diversity is further aggregated when spectral or temporal perturbation based data augmentation techniques [6,22,23] are used. To this end, speaker adaptation techniques play a crucial role in the personalization of ASR systems for such users.

Speaker adaptation techniques for normal speech have been widely studied in three broad categories: 1) speaker-dependent (SD) auxiliary speaker embedding features [24–26]; 2) feature transformations generating canonical features at acoustic front-ends [27, 28]; 3) model based approaches using specially designed SD DNN parameters [29, 30].

In contrast, there are limited prior researches on dysarthric and elderly speaker adaptation. Earlier works were mainly conducted on HMM based ASR systems, including MLLR and MAP adaptation [1, 31–33] and their combination with speaker adaptive training (SAT) [2], feature-space MLLR (f-MLLR) based SAT [34] and regularized speaker adaptation via Kullback-Leibler (KL) divergence [35]. More recent researches applied model based adaptation to current DNN based ASR systems, including direct parameter fine-tuning based adaptation in both hybrid TDNN [4, 36] and end-to-end RNN-T [37, 38] systems, LHUC [8, 11] and Bayesian speaker adaptation [39]. Spectro-temporal basis embedding features (SBE) based offline adaptation via speaker-level averaging was studied in [10, 16].

One major issue associated with the prior researches above is the lack of suitable rapid, on-the-fly adaptation techniques targeting dysarthric and elderly speech. Such methods serve as dual-purpose solutions to handle not only the difficulty in collecting large quantities of data from such speakers with mobility issues that are essential for model based adaptation but also their latency issues due to multi-pass decoding and SD parameter estimation. The Bayesian model based adaptation using very limited speaker data [11, 39, 40] only addressed the aforementioned data scarcity issue, but the latency problem remains unvisited. Similarly, the spectro-temporal deep embedding features [10] computed and averaged over all speaker-level data incurred latency and precluded the use of on-the-fly adaptation.

In order to address this issue, two novel forms of feature-based on-the-fly rapid speaker adaptation approaches are proposed in this paper. The first is based on speaker-level variance-regularized spectral basis embedding (SVR) features. An additional variance regularization term is included when training spectral basis embedding DNNs [10, 16] to ensure speaker-level homogeneity of the resulting embedding features and thus allow them to be applied on the fly during test time adaptation. The second approach uses on-the-fly feature-based LHUC (f-LHUC) transforms conditioned on spectral features. Specially designed regression TDNN [41] predicting speaker-level LHUC transforms are used to directly generate and apply such parameters during test time adaptation and thus resolve the latency due to multi-pass decoding. Experiments were conducted on the largest available and most widely used UASpeech [42] dysarthric and DementiaBank Pitt [43] elderly speech datasets. Consistent performance improvements were obtained by our proposed on-the-fly speaker adaptation approaches using both hybrid DNN/TDNN and E2E Conformer [44] systems.

The main contributions of the paper are summarized below:

1) This paper presents the first work of on-the-fly feature-based fast speaker adaptation targeting dysarthric and elderly speech. In contrast, previous works on feature-based dysarthric and elderly speaker adaptation utilize all speaker-level data and

* Part of this work was done while the author was an intern at Tencent AI Lab. Codes are available at https://github.com/timspeech/on_the_fly_adapt

 10.21437/Interspeech.2023-301

operate in batch mode [10, 16] while those using model based adaptation not only require the usage of all speaker-level data, but also additional multiple decoding passes and explicit parameter estimation in test time [6, 8, 11, 39]. These prior works incur significant latency and are not the on-the-fly, rapid speaker adaptation approaches considered in this paper.

2) Our proposed SVR features can instantaneously extract homogeneous dysarthric and elderly speaker characteristics on the fly. Experiments conducted on benchmark UASpeech dysarthric and DementiaBank Pitt elderly speech datasets suggest the proposed on-the-fly speaker adaptation approaches consistently outperform the baseline iVector adapted hybrid DNN/TDNN and E2E Conformer systems by statistically significant word error rate (WER) reduction of 2.48%-2.85% absolute (7.92%-8.06% relative), and offline model based LHUC adaptation by 1.82% absolute (5.63% relative) respectively.

2. Variance-Regularized Spectral Features

To model the latent diversity in dysarthric and elderly speech, singular value decomposition (SVD) is performed on Mel-filterbank log amplitude spectrum $\mathbf{S}$ [45], given as:

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{1}$$

where the top-d principal spectral bases are retrieved from the column vectors of $\mathbf{U}$. Following [10, 16], further supervised learning is performed via constructing DNN speech intelligibility or age classifier (the upper part of Fig.1). The inputs are the selected spectral bases, and the targets are speech intelligibility groups + speaker IDs for the UASpeech corpus and binary aged vs. non-aged annotations for the DementiaBank Pitt corpus.

To ensure speaker-level homogeneity of the embedding features, a pair of such DNN classifiers (Fig.1) are constructed. The 25-dim embedding features taken from the bottleneck of the upper classifier are averaged by speaker and serve as the regression targets of the lower DNN classifier (Fig.1 blue bold line) for variance regularization. A multitask learning (MTL) cost function is used to train the lower classifier, using interpolation between 1) the cross-entropy (CE) error computed over speech intelligibility or age labels, optionally plus that over speaker IDs and 2) the mean squared error (MSE) computed between the lower DNN bottleneck features and the corresponding embedding produced by the upper DNN, which is given as:

$$\mathcal{L}_{MTL} = \omega_1 \cdot \mathcal{L}_{MSE} + \omega_2 \cdot \mathcal{L}_{CE_{group}} + \omega_3 \cdot \mathcal{L}_{CE_{ID}} {}^{1} \tag{2}$$

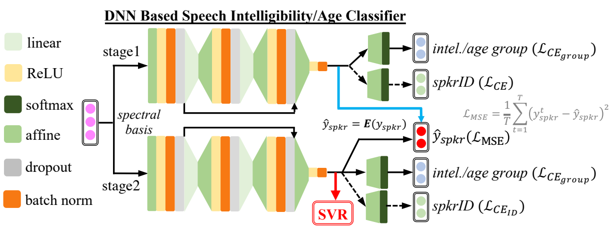

Figure 1: *Our proposed DNN speech intelligibility or age classifier with a bottleneck layer to extract variance-regularized spectral basis embedding (SVR) features for speaker adaptation. Here "intel." denotes speech intelligibility.*

The 25-dim variance-regularized spectral basis embedding (SVR) features are then taken from the bottleneck layer of the lower classifier (Fig.1 red bold line), and appended to the acoustic features at the front-end of hybrid DNN/TDNN (Fig.4) and E2E Conformer systems (Fig.5), to facilitate on-the-fly test time adaptation. The aforementioned procedure is summarized in Fig. 2. Directly using the top d spectral bases or intermediate embedding features are unsuitable for speaker adaptation [16].

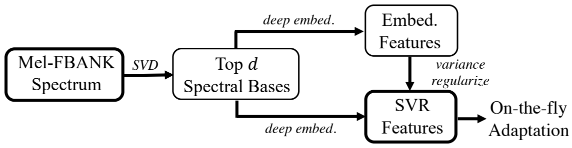

Figure 2: *Procedure of generating the SVR features. "deep embed." refer to the DNN embedding process shown in Figure 1.*

3. On-the-Fly F-LHUC Transforms

In feature-based learning hidden unit contributions (f-LHUC) based adaptation approaches [41], LHUC transforms are predicted from the acoustic features on the fly. Supervised estimation of LHUC transforms on the training data is first conducted via standard speaker adaptive training (SAT). Principal component analysis (PCA) is further applied to produce compressed LHUC vectors encoding the most distinctive speaker-level features. These serve as the output targets for the TDNN based LHUC regression network (Fig.3), with a specially designed online averaging layer [41] given as:

$$\boldsymbol{m}_{\boldsymbol{s}}^{\boldsymbol{i}} = \frac{\sum_{t=1}^{T_i} \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{i}} + \alpha \cdot \boldsymbol{G}_{\boldsymbol{s}}^{\boldsymbol{i-1}}}{T_i + \alpha \cdot N_s^{i-1}} \tag{3}$$

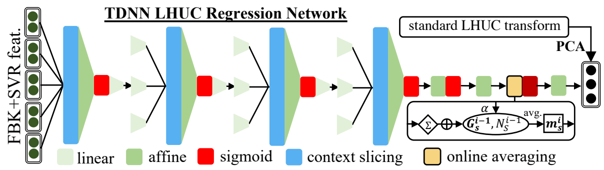

Figure 3: *TDNN based LHUC regression network.*

where $\boldsymbol{G_s}$, N_s and $\boldsymbol{m_s}$ denote the accumulated hidden vector, frame count, and averaged hidden vector till the i^{th} segment of speaker s. The i^{th} utterance contains T_i frames, and the hidden vector of the t^{th} frame is $\boldsymbol{h_t^i}$. $\alpha \in [0, 1]$ is the history interpolation weight. Different from [41], FBK + SVR features are used to train the LHUC regression network (Fig 3). An additional affine transformation is further trained to map the predicted low-dimensional LHUC features for training speakers to corresponding LHUC transforms. During test time on-the-fly adaptation, the regression network (Fig. 3) and the affine transformation (Fig.4 circled in green) are applied in turn to generate speaker-level LHUC transforms using FBK + SVR features.

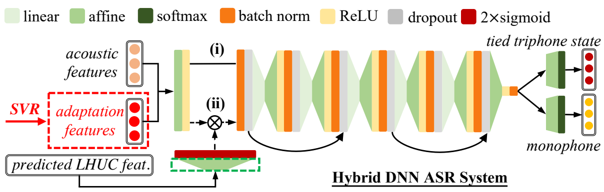

Figure 4: *Incorporation of variance-regularized spectral basis embedding (SVR) features at front-end of hybrid DNN ASR systems [11]. Selecting path (i) leads to systems with auxiliary feature based adaptation only, while selecting (ii) leads to systems with additional feature-based LHUC (f-LHUC) adaptation.*

[1]Empirically set for the UASpeech corpus as $\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$, while for the DementiaBank Pitt corpus $\omega_1 = \omega_2 = \frac{1}{2}, \omega_3 = 0$.

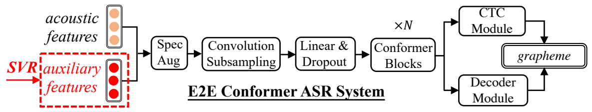

Figure 5: *Incorporating SVR features into E2E Conformer.*

4. Experiments and Results

4.1. Experiments on the UASpeech Dataset

Task Description: UASpeech is the largest publicly available and widely used dysarthric speech dataset [42], which is an isolated word recognition task containing 103h speech from 16 dysarthric and 13 control speakers. It is split into blocks B1, B2 and B3, each with the same 155 common words and different 100 uncommon words. The training set includes B1 and B3 data of all 29 speakers (69.1h), while the test set includes B2 data of 16 dysarthric speakers (22.6h, excluding speech from control speakers). Silence stripping using an HTK trained GMM-HMM system [11] produces a 30.6h training set (99195 utt.) and a 9h test set (26520 utt.). Data augmentation [6] produces a 130.1h augmented training set (399110 utt.). The average utterance length is 1.2 seconds. As E2E systems are sensitive to the training data coverage, B2 data of the 13 control speakers and their speed perturbed versions are also used for Conformer system training. This creates a 190h training set (538292 utt.).

Experiment Setup: The 7-layer hybrid DNN and E2E graphemic Conformer systems were implemented using Kaldi following [11] and ESPnet[2]. The inputs were 80-dim filter-bank (FBK) + Δ features plus 25-dim variance-regularized spectral basis embedding (SVR) features or 100-dim iVectors[3] (Fig. 4-5). Top 2 spectral bases [10, 16] were used to train the DNN intelligibility classifier (Fig.1). The history interpolation weight α for LHUC regression was set to 0.9 with the context slicing indices as $\{-2,-1,0,1,2\}$,$\{-2,0,2\}$,$\{-3,0,3\}$ and $\{-4,0,4\}$. A uniform language model (LM) was used in decoding [1]. As an ablation study, we used iVectors as the inputs to the DNN classifier (Fig.1) for variance regularization and generate variance-regularized iVectors (iVRs).

Result Analysis: Table 1 shows the performance comparison[4] between the proposed SVR feature adaptation, spectral feature driven f-LHUC adaptation, iVector adaptation and offline LHUC adaptation on the UASpeech corpus. Several trends can be observed: **1)** On-the-fly SVR adaptation (Sys.5,15,26) consistently and statistically significantly outperform iVector adaptation (Sys.2,12,23) with various amounts of training data by up to **2.48% absolute (7.92% relative)** overall WER reduction for hybrid DNN (Sys.5 vs. Sys.2), and **2.85% absolute (8.06% relative)** reduction (Sys.26 vs. Sys.23) for Conformer systems, respectively. **2)** The improvements from offline LHUC adaptation (Sys.6,16) over the SI systems (Sys.1,11) are largely retained (by 82%) and comparable to those obtained using on-the-fly SVR adaptation (Sys.5,15). **3)** Compared with improvements over SI systems (Sys.1,11) obtained by offline SBE adaptation [10] that requires expensive speaker-level averaging (Sys.3,13), on-the-fly SVR adaptation (Sys.5,15) produces comparable performance. **4)** The spectral feature (FBK+SVR) driven f-LHUC adapted systems (Sys.10,20) outperform both the FBK or FBK+iVector driven f-LHUC adapted systems (Sys.8,9,18,19) and the SVR adaptation alone (Sys.5,15). **5)** Frame-level log-likelihood score combination between on-the-fly SVR adaptation and FBK+SVR driven f-LHUC adaptation leads to further improvements (Sys.5+10, Sys.15+20). **6)** The SVR on-the-fly adapted systems (Sys.5,15,26) consistently outperform comparable variance-regularized iVector (iVR) adaptation (Sys.4,14,25). **7)** A comparison between published systems on UASpeech and ours is shown in Table 4. Our best-performing system (Table 1, Sys.15+20) produces the lowest WERs among all systems using online speaker adaptation.

[2] 12 encoder layers + 12 decoder layers, feed-forward dim = 2048, 4 attention heads of 256 dimensions, interpolated CTC+AED cost.

[3] Kaldi: egs/wsj/s5/local/nnet3/run_ivector_common.sh. Changing the dimensionality of iVectors produces margin effect [16].

[4] A matched pairs sentence-segment word error (MAPSSWE) based statistical significance test [46] was done at significance level $\alpha = 0.05$.

Table 1: *Performance of the proposed variance-regularized spectral basis embedding (SVR) feature adaptation, iVector adaptation and LHUC adaptation on the* ***UASpeech*** *test set of 16 dysarthric speakers. "SBE" denotes spectral basis embedding features. "VL/L/M/H" refer to intelligibility very low, low, mid and high. "On Fly" indicates using on-the-fly adaptation.* † *denotes a statistically significant improvement ($\alpha = 0.05$) obtained over iVector adapted systems (Sys.2,12,23).*

Sys.	Model (#Para.)	Data Aug.	#Hrs	Adapt. Feat.	LHUC SAT	f-LHUC	On Fly	WER% VL	L	M	H	All
1				✗			-	69.82	32.61	24.53	10.40	31.45
2				iVector			✓	69.46	33.78	22.58	10.45	31.33
3				SBE [10]	✗	✗	✗	64.43	29.71	19.84	8.57	28.05
4				iVR			✓	68.66	33.72	22.84	9.83	30.99
5	Hybrid			SVR			✓	65.04†	30.90†	20.70†	10.15†	**28.85**†
6	DNN	✗	30.6	✗	✓	✗	✗	64.39	29.88	20.27	8.95	28.29
7	(6M)			SBE [10]			✗	63.40	28.90	18.64	8.13	27.24
8				✗		(FBK)	✓	66.47†	29.55†	21.00†	8.99†	28.80†
9				iVector	✗	(+iVector)	✓	64.86	36.44	21.17	9.03	30.29
10				SVR		(+SVR)	✓	65.75†	29.80†	19.07†	8.99†	**28.31**†
5+10				-			✓	**64.36**†	**29.68**†	**18.96**†	**8.89**†	**27.96**†
11				✗			-	66.45	28.95	20.37	9.62	28.73
12				iVector			✓	65.73	30.10	20.21	9.03	28.65
13				SBE [10]	✗	✗	✗	61.55	27.52	17.31	8.22	26.26
14				iVR			✓	66.02	29.52	19.56	9.32	28.53
15	Hybrid			SVR			✓	62.54†	30.22	18.54†	8.59†	**27.54**†
16	DNN	✓	130.1	✗	✓	✗	✗	62.50	27.26	18.41	8.04	26.55
17	(6M)			SBE [10]			✗	59.83	27.16	16.80	7.91	25.60
18				✗		(FBK)	✓	65.06	27.94†	18.76†	8.39†	27.45†
19				iVector	✗	(+iVector)	✓	63.63	32.56	18.52	8.31	28.28
20				SVR		(+SVR)	✓	61.56†	28.81	18.39†	8.50†	**26.90**†
15+20				-			✓	**60.80**†	**28.19**†	**17.72**†	**8.23**†	**26.36**†
21			130.1	✗			-	73.88	53.12	49.92	42.03	53.17
22				✗			-	65.70	40.63	33.39	9.53	34.07
23	Conformer	✓		iVector	✗	✗	✓	69.05	42.45	33.60	9.74	35.37
24	(52M)		190	SBE [10]			✗	65.18	34.90	24.21	5.00	29.19
25				iVR			✓	68.94	42.00	32.19	8.52	34.55
26				SVR			✓	67.52†	**38.85**†	**28.60**†	**7.88**†	**32.52**†

4.2. Experiments on the DementiaBank Pitt Dataset

Task Description: The DementiaBank Pitt [43] dataset contains 33h speech recorded over interviews between 292 elderly participants and clinical investigators. After split of the data and silence stripping [8], the training set contains 15.7h speech from 244 elderly and 444 investigators (29682 utt.) while the development and evaluation sets contain 2.5h (5103 utt.) and 0.6h (928 utt.) speech from 43 elderly and 76 investigators[5]. Data augmentation [8] produced an 58.9h augmented training set (112830 utt.). The average utterance length is 1.9 seconds.

Experiment Setup: The inputs to the hybrid TDNN systems[6] and E2E graphemic Conformer systems[7] were 40-dim FBK + 25-dim SVR features or 100-dim iVectors. Top 3 spectral bases [16] served as the inputs to the DNN age classifier (Fig.1). A word 4-gram LM [8] and a 3.8k recognition vocabulary covering all words in DementiaBank Pitt corpus was used.

[5] The evaluation set is based on exactly the same 48 speakers' Cookie task recordings following [47] while the development set contains the recordings of these speakers in other tasks if available.

[6] 14 context slicing layers with a 3-frame context.

[7] 12 encoder layers + 12 decoder layers, feed-forward dim = 2048, 4 attention heads of 256 dimensions, interpolated CTC+AED cost.

Table 2: *Performance of the proposed variance-regularized spectral basis embedding (SVR) feature adaptation, iVector adaptation and LHUC adaptation on augmented* ***DementiaBank Pitt*** *corpus. "INV" and "PAR" denote investigator and elderly.* † *denotes a statistically significant improvement (*$\alpha = 0.05$*) over both the iVector adaptation (Sys.2) and offline LHUC adaptation (Sys.6), while* ‡ *denotes a statistically significant improvement (*$\alpha = 0.05$*) over the iVector adaptation (Sys.12).*

Sys.	Model (#Para.)	#Hrs	Adapt. Feat.	LHUC SAT	f-LHUC	On Fly	WER% Dev INV	WER% Dev PAR	WER% Eval INV	WER% Eval PAR	WER% All
1	Hybrid TDNN (18M)	58.9	✗	✗	✗	-	19.91	47.93	19.76	36.66	33.80
2			iVector			✓	19.97	46.76	18.20	37.01	33.37
3			SBE [16]			✗	18.61	43.84	17.98	33.82	31.12
4			iVR			✓	19.19	47.64	18.65	35.80	33.26
5			SVR			✓	18.72†	44.67†	18.65	34.03†	**31.55**†
6			✗	✓	✗	✗	19.26	45.49	18.42	35.44	32.33
7			SBE [16]			✗	17.41	40.94	17.98	31.89	29.16
8			✗	✗	(FBK)	✓	19.61	45.40	18.87	34.77	32.33
9			iVector		(+iVector)	✓	18.75	47.07	17.98	36.11	32.85
10			SVR		(+SVR)	✓	17.87†	43.83†	16.87†	34.56†	**30.91**†
5+10			-			✓	17.66†	**43.48**†	16.09†	**33.68**†	**30.51**†
11	Conformer (52M)	58.9	✗	✗	✗	-	20.97	48.71	19.42	36.93	34.57
12			iVector			✓	21.48	48.32	17.42	37.79	34.71
13			SBE [16]			✗	20.44	47.70	17.31	36.11	33.76
14			iVR			✓	22.09	49.56	19.64	38.58	35.65
15			SVR			✓	20.83	**47.39**‡	17.64	**36.34**‡	**33.84**‡

Result Analysis: Table 2 shows the performance of the proposed on-the-fly SVR adaptation, f-LHUC adaptation, iVector adaptation and LHUC adaptation on DementiaBank Pitt. Trends similar to those on UASpeech in Table 1 are observed: **1)** On-the-fly SVR adaptation (Sys.5,15) statistically significantly outperform iVector adaptation (Sys.2,12) on both TDNN and Conformer systems by up to **1.82% absolute (5.45% relative)** WER reduction (Sys.5 vs. Sys.2). **2)** On-the-fly SVR adaptation outperforms offline LHUC adaptation by 0.78% absolute (2.41% relative) overall WER reduction (Sys.5 vs. Sys.6). **3)** On-the-fly SVR adaptation (Sys.5) largely retains (by 84%) the improvements obtained by offline SBE adaptation [16] (Sys.3) over the SI system (Sys.1). **4)** The proposed FBK+SVR driven f-LHUC adapted system (Sys.10) outperforms offline LHUC adaptation (Sys.6) while also outperforming FBK driven f-LHUC adaptation (Sys.8) and SVR adaptation alone (Sys.5). **5)** Frame-level score combination between SVR and FBK+SVR driven f-LHUC on-the-fly adaptation leads to **1.82% absolute (5.63% relative)** WER reduction over the offline LHUC adaptation (Sys.5+10 vs. Sys.6). **6)** Our proposed on-the-fly SVR (Sys.5,15) and FBK+SVR driven f-LHUC adaptation (Sys.10) consistently outperform the comparable iVR (Sys.4,14) and FBK+iVector driven f-LHUC adaptation (Sys.9).

4.3. Further Ablation Studies

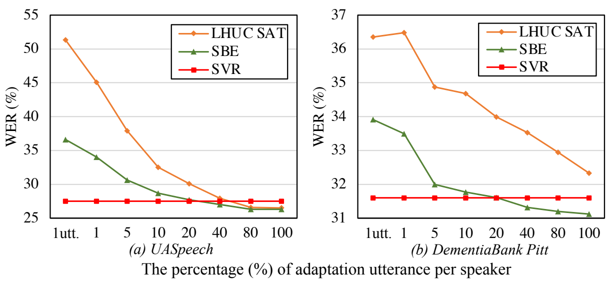

Figure 6: *Performance (WER%) of offline LHUC [11], offline SBE adaptation [10, 16] and on-the-fly SVR adaptation of Sec.2 on varying percentage of test set speaker-level adaptation data on* ***UASpeech*** *and* ***DementiaBank Pitt****.*

As expected, the ablation study in Fig. 6 confirms on-the-fly SVR adaptation is more robust to varying amounts of speaker-level data used in adaptation than offline LHUC and SBE adaptation, and consistently outperforms both when less than 40% of speaker-level data is used. A further ablation study in Table 3 suggests the performance of on-the-fly SVR adaptation is largely insensitive to the length of analysis sliding windows (from 1 utt. down to 10ms). Homogeneous dysarthric and elderly speaker characteristics can be instantaneously extracted in the SVR features on the fly. The real-time (R.T.) factor indicates the total delay of waiting for data and model processing.

Table 3: *Ablation study on the augmented* ***UASpeech*** *(UA.) and* ***DementiaBank Pitt*** *(DBK.) corpora with various sizes of sliding window (Slid. Wind.) for on-the-fly SVR feature extraction. "R.T." is short for real time.* † *denotes a statistically significant improvement (*$\alpha = 0.05$*) is obtained over the comparable on-the-fly iVector adapted systems.*

UA. Sys.	Model (#Para.)	#Hrs	Adapt. Feat.	Slid. Wind.	On Fly	R.T. Factor	WER% VL	WER% L	WER% M	WER% H	WER% All
1	Hybrid DNN (6M)	130.1	iVector	100ms	✓	0.10	65.73	30.10	20.21	9.03	28.65
2			SVR	utt.		1.02	62.54†	30.22	18.54†	8.59†	27.54†
3				300ms		0.27	63.74†	29.01†	19.56†	9.09	27.84†
4				200ms		0.18	63.81†	29.15†	20.19	9.31	28.08†
5				100ms		0.10	64.49†	28.36†	19.47†	9.27	27.87†
6				50ms		0.06	63.99†	28.68†	19.17†	8.99	27.70†
7				10ms		0.03	**64.77**†	**29.03**†	**19.21**†	9.09	**28.00**†

DBK. Sys.	Model (#Para.)	#Hrs	Adapt. Feat.	Slid. Wind.	On Fly	R.T. Factor	WER% Dev INV	WER% Dev PAR	WER% Eval INV	WER% Eval PAR	WER% All
1	Hybrid TDNN (18M)	58.9	iVector	100ms	✓	0.08	19.97	46.76	18.20	37.01	33.37
2			SVR	utt.		1.03	18.72†	44.67†	18.65	34.03†	31.55†
3				300ms		0.19	19.29†	45.01†	19.09	33.28†	31.81†
4				200ms		0.14	19.36	45.15†	19.20	33.89†	32.01†
5				100ms		0.08	19.08†	45.03†	19.42	34.41†	31.93†
6				50ms		0.06	19.38	44.87†	18.31	34.52†	31.97†
7				10ms		0.04	19.33†	**44.93**†	19.09	**34.35**†	**31.97**†

Table 4: *Performance comparison against recently published systems on* ***UASpeech****. "DA" denotes data augmentation.*

Sys.	Online	VL	All
Sheffield-2015 Speaker adaptive training [2]	✗	70.78	34.85
Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation (15spk) [4]	✓	68.24	30.76
CUHK-2020 DNN + DA + LHUC SAT [6]	✗	62.44	26.37
CUHK-2021 LAS + CTC + Meta-learning + SAT [48]	✗	68.70	35.00
CUHK-2021 QuartzNet + CTC + Meta-learning + SAT [48]	✗	69.30	30.50
Sheffield-2022 DA + Source Filter Features + iVector adapt [17]	✓	-	30.30
DA + SVR Adapt + f-LHUC system combination (Table 1, Sys.15+20)	✓	**60.80**	**26.36**

5. Conclusions

This paper proposes two novel forms of feature-based on-the-fly speaker adaptation approaches: speaker-level variance-regularized spectral basis embedding (SVR) features adaptation and spectral feature driven f-LHUC adaptation. Experiments conducted on benchmark UASpeech dysarthric and DementiaBank Pitt Elderly datasets suggest both methods can efficiently encode homogeneous dysarthric and elderly speaker specific characteristics and outperform both online iVector and offline model based LHUC adaptation. Future research will focus on rapid speaker adaptation of pre-trained ASR systems.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200220, 14200021, TRS T45-407/19N, Innovation Technology Fund grant No. ITS/218/21, National Natural Science Foundation of China (NSFC) Grant 62106255, and Youth Innovation Promotion Association CAS Grant 2023119.

7. References

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

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

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

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

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

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

[7] S. Hu, X. Xie *et al.*, "Bayesian Learning of LF-MMI Trained Time Delay Neural Networks for Speech Recognition," *IEEE T AUDIO SPEECH*, 2021.

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

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

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

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

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

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

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

[15] T. Wang, J. Deng *et al.*, "Conformer based elderly speech recognition system for alzheimer's disease detection," in *INTERSPEECH*, 2022.

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

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

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

[19] A. Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & dementia*, 2019.

[20] I. Kodrasi and H. Bourlard, "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection," *IEEE T AUDIO SPEECH*, 2020.

[21] B. L. Smith, J. Wasowicz *et al.*, "Temporal characteristics of the speech of normal elderly adults," *J SPEECH LANG HEAR R*, 1987.

[22] B. Vachhani, C. Bhat *et al.*, "Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition," in *INTERSPEECH*, 2018.

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

[24] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *ICASSP*, 2013.

[25] G. Saon, H. Soltau *et al.*, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*, 2013.

[26] H. Huang and K. C. Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for dnn-based speech recognition," in *ICASSP*, 2015.

[27] V. V. Digalakis, D. Rtischev *et al.*, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE T SPEECH AUDI P*, 1995.

[28] F. Seide, G. Li *et al.*, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011.

[29] C. Zhang and P. C. Woodland, "Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling," in *INTERSPEECH*, 2015.

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

[31] A. BABA, S. YOSHIZAWA *et al.*, "Elderly Acoustic Models for Large Vocabulary Continuous Speech Recognition," *IEICE T INF SYST*, 2002.

[32] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *ICASSP*, 2011.

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

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

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

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

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

[38] J. R. Green, R. L. MacDonald *et al.*, "Automatic Speech Recognition of Disordered Speech: Personalized models outperforming human listeners on short phrases," in *INTERSPEECH*, 2021.

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

[40] X. Xie, X. Liu *et al.*, "Bayesian learning for deep neural network adaptation," *IEEE T AUDIO SPEECH*, 2021.

[41] X. Xie, X. Liu *et al.*, "Fast DNN Acoustic Model Speaker Adaptation by Learning Hidden Unit Contribution Features," in *INTERSPEECH*, 2019.

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

[43] J. T. Becker, F. Boiler *et al.*, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *ARCH NEUROL-CHICAGO*, 1994.

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

[45] A.-J. Van Der Veen, E. F. Deprettere *et al.*, "Subspace-based signal analysis using singular value decomposition," *P IEEE*, 1993.

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

[47] S. Luz, F. Haider *et al.*, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH*, 2020.

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