

Homogeneous Speaker Features for On-the-Fly Dysarthric and Elderly Speaker Adaptation

Mengzhe Geng, Xurong Xie, Jiajun Deng, Zengrui Jin, Guinan Li, Tianzi Wang, Shujie Hu, Zhaoqing Li,
Helen Meng, Xunying Liu

arXiv:2407.06310v1 [cs.SD] 8 Jul 2024

Abstract—The application of data-intensive automatic speech recognition (ASR) technologies to dysarthric and elderly adult speech is confronted by their mismatch against healthy and non-aged voices, data scarcity and large speaker-level variability. To this end, this paper proposes two novel data-efficient methods to learn homogeneous dysarthric and elderly speaker-level features for rapid, on-the-fly test-time adaptation of DNN/TDNN and Conformer ASR models. These include: 1) speaker-level variance-regularized spectral basis embedding (VR-SBE) features that exploit a special regularization term to enforce homogeneity of speaker features in adaptation; and 2) feature-based learning hidden unit contributions (f-LHUC) transforms that are conditioned on VR-SBE features. Experiments are conducted on four tasks across two languages: the English UASpeech and TORG0 dysarthric speech datasets, the English DementiaBank Pitt and Cantonese JCCOCC MoCA elderly speech corpora. The proposed on-the-fly speaker adaptation techniques consistently outperform baseline iVector and xVector adaptation by statistically significant word or character error rate reductions up to 5.32% absolute (18.57% relative) and batch-mode LHUC speaker adaptation by 2.24% absolute (9.20% relative), while operating with real-time factors speeding up to 33.6 times against xVectors during adaptation. The efficacy of the proposed adaptation techniques is demonstrated in a comparison against current ASR technologies including SSL pre-trained systems on UASpeech, where our best system produces a state-of-the-art WER of 23.33%. Analyses show VR-SBE features and f-LHUC transforms are insensitive to speaker-level data quantity in test-time adaptation. T-SNE visualization reveals they have stronger speaker-level homogeneity than baseline iVectors, xVectors and batch-mode LHUC transforms.

Index Terms—Dysarthric Speech Recognition, Elderly Speech Recognition, Rapid Adaptation, Speaker Features

I. INTRODUCTION

IN spite of the remarkable advancement on automatic speech recognition (ASR) techniques for normal speech, accurate recognition of disordered speech, for example, voice recorded from speakers with dysarthria, remains a highly challenging task to date [1]–[11]. Millions of people worldwide are clinically diagnosed with speech disorders [12]. In addition, neurocognitive disorders, such as Alzheimer’s disease (AD), are often found among older adults and manifest

Mengzhe Geng, Jiajun Deng, Zengrui Jin, Guinan Li, Tianzi Wang, Shujie Hu and Zhaoqing Li are with the Chinese University of Hong Kong, China (email: {mzgeng, jjdeng, zrjin, gnli, twang, sjhu, zqli}@se.cuhk.edu.hk).

Xurong Xie is with Institute of Software, Chinese Academy of Sciences, Beijing, China (email: xurong@iscas.ac.cn).

Helen Meng is with the Chinese University of Hong Kong, China (email: hmmeng@se.cuhk.edu.hk).

Xunying Liu is with the Chinese University of Hong Kong, China and the corresponding author (email: xyliu@se.cuhk.edu.hk).

themselves in speech impairments [13], [14]. Given that large-scale pathological assessment of speech disorders among all older adults is practically difficult due to the limited availability of professional speech pathologists, the actual number of people affected by speech disorders is much larger. As aging presents enormous challenges to health care worldwide, there is a pressing need to develop suitable ASR-based assistive technologies customized for dysarthric and elderly speakers to improve their life quality and social inclusion [15]–[44].

Dysarthric and elderly speech bring challenges on three fronts to current ASR technologies predominantly targeting normal speech recorded from healthy, non-aged users. These challenges include: **a) substantial mismatch against normal speech** due to the underlying neuro-motor control conditions [45], [46] and aging, for example, imprecise articulation, hoarse voice and increased disfluency [47], [48]; **b) data scarcity** due to the difficulty in collecting large quantities of such data from dysarthric and elderly speakers with mobility limitations; and **c) large speaker-level diversity** among dysarthric and elderly talkers, when sources of variability commonly found in normal speech, e.g. accent or gender, are further aggregated with speech pathology severity and aging.

A key task for all ASR systems is to model the speech variability attributed to speaker-level characteristics. To this end, speaker adaptation techniques provide a key role in customizing ASR systems for users’ needs. For normal speech recognition tasks, three main categories of such techniques have been studied: **1) auxiliary embedding features** that are speaker-dependent (SD) [49]–[51]; **2) feature transformations** that produce speaker-independent (SI) features to remove such variability at the front-ends [52]–[54]; and **3) model-based adaptation** using specifically designed SD DNN parameters [55]–[58]. A detailed review of speaker adaptation techniques targeting normal speech is presented in Sec. II.

Modeling the large speaker-level heterogeneity in dysarthric and elderly speech requires powerful speaker adaptation techniques to be developed. However, only limited prior research in this direction has been conducted. Earlier works focused on deriving speaker adaptation methods for traditional HMM-based ASR systems with Gaussian mixture model (GMM) based hidden state densities, predominantly targeting dysarthric speech only. A combination of HMM state transition interpolation and maximum a posteriori (MAP) adaptation is utilized to account for dysarthric speaker diversity in [59]. The use of maximum likelihood linear regression (MLLR) and MAP adaptation was explored in [1], [60]–[62]. Combined use of MLLR with MAP adaptation in speaker adaptative training (SAT) of GMM-

HMM models was investigated in [2]. Feature-space MLLR (f-MLLR) based SAT training [63] and regularized speaker adaptation to dysarthric speakers using Kullback-Leibler (KL) divergence [64] were also developed for GMM-HMM systems. MLLR adaptation of GMM-HMM models to elderly speakers using cross-speaker statistics pooling was studied in [60].

More recent dysarthric and elderly speaker adaptation approaches applied to state-of-the-art hybrid and end-to-end (E2E) neural network based ASR systems include the following categories: 1) direct speaker-level parameter fine-tuning of hybrid LF-MMI trained TDNN [27], [28] and RNN-T [65] models; 2) the use of iVector speaker adaptation for dysarthric [66], [67] and elderly [34] speech recognition; 3) model-based speaker adaptation using, for example, learning hidden unit contributions (LHUC) for dysarthric [24], [29], [30], [38] and elderly [34], [38] speakers; 4) Bayesian domain [68] and speaker adaptation [29] methods that are more robust to dysarthric and elderly speech data scarcity; 5) dysarthric and elderly speaker-level averaged spectro-temporal basis embedding features [36], [38] for hybrid DNN/TDNN and E2E Conformer adaptation; 6) f-MLLR and xVector based dysarthric speaker adaptation of self-supervised learning (SSL) based Wav2vec 2.0 models [5]; and 7) dysarthric speaker adapter fusion of pre-trained E2E Transformer models [42].

Suitable dysarthric and elderly speaker adaptation methods that can meet with above-mentioned challenges should satisfy the following requirements: **1) high data efficiency** to model the very limited speaker-level data; **2) strong speaker homogeneity** to ensure the distinct speaker-level characteristics are consistently represented in the adapted model; and **3) low processing latency** to allow adaptation to be performed immediately on-the-fly from the onset of a new user's enrolment in the system to minimize their efforts and fatigue. In this regard, prior research only addressed some of the above. For example, the Bayesian model based adaptation using very limited speaker data [29], [34], [68] only addresses the aforementioned data scarcity issue, but the latency problem remains unvisited. Similarly, the spectro-temporal deep embedding features [36], [38] are averaged over all speaker-level data in an offline manner. This introduces considerable processing latency and is thus unsuitable for on-the-fly test-time adaptation.

To this end, one possible solution is to derive suitable rapid, on-the-fly feature-based dysarthric and elderly adaptation techniques. Such methods serve as multi-purpose solutions to handle not only the speaker-level data scarcity and diversity when representing speaker attributes, but also the processing latency incurred by model-based fine-tuning or adaptation to speaker data. In this paper, two novel forms of feature-based on-the-fly rapid speaker adaptation approaches are proposed. The first is based on speaker-level variance-regularized spectral basis embedding (VR-SBE) features. An extra variance regularization term is included when training spectral basis embedding neural networks [36], [38] to ensure speaker homogeneity of the embedding features. This in turn allows them to be applied on-the-fly during test-time adaptation to dysarthric or elderly speakers. The second approach utilizes on-the-fly feature-based LHUC (f-LHUC) transforms conditioned on VR-SBE features. Specially designed regression TDNN [69] predicting

dysarthric or elderly speaker-level LHUC transforms is constructed to directly generate and apply such SD parameters on the fly during test-time adaptation.

Experiments are conducted on four different tasks: 1) the English UASpeech [70] and TORG0 [71] dysarthric speech datasets; 2) the English DementiaBank Pitt [72] and the Cantonese JCCOCC MoCA [73] elderly speech corpora. Among these, UASpeech is by far the largest available and most extensively used dysarthric speech corpus, while DementiaBank Pitt is the largest publicly available elderly speech database. The performance of the proposed two feature-based on-the-fly speaker adaptation approaches: VR-SBE and VR-SBE feature conditioned f-LHUC, are compared against those of baseline iVector [50] or xVector [74] based on-the-fly adaptation and LHUC [57] based model adaptation on three fronts:

1) ASR performance in WER/CER: Our proposed on-the-fly speaker adapted hybrid DNN/TDNN and E2E Conformer systems consistently outperform the corresponding on-the-fly iVector/xVector adapted systems, with statistically significant¹ word/character error rate (WER/CER) reductions up to **5.32% absolute (18.57% relative)**. Our on-the-fly adaptation methods also outperform the comparable offline batch-mode model-based LHUC adaptation by statistically significant WER/CER reductions up to **2.24% absolute (9.20% relative)**.

2) Analysis on processing latency: Experiments on the benchmark UASpeech dysarthric and DementiaBank Pitt elderly speech corpora suggest that statistically significant WER/CER reductions are consistently obtained over the corresponding on-the-fly iVector/xVector adapted systems, while the proposed VR-SBE speaker adaptation operates with a real-time factor speeding up ratio up to 33.6 times against xVector. It incurs a minimal processing latency by using an input acoustic feature window **as short as 10 ms** in duration.

3) Analysis on speaker feature homogeneity: T-SNE visualization [76] of baseline iVectors/xVectors, offline model-based LHUC transforms, offline SBE [38] features, our proposed on-the-fly VR-SBE features and the associated f-LHUC transforms across varying amounts of adaptation data is conducted. This suggests more consistent, data quantity invariant dysarthric and elderly speaker characteristics can be learned via the proposed on-the-fly speaker adaptation approaches.

The main contributions of this paper are summarized below:

1) This paper presents **novel approaches to learn homogeneous speaker features tailored for rapid, on-the-fly dysarthric and elderly speaker adaptation**. In contrast, prior studies either considered adaptation techniques using all speaker-level data and operating in batch-mode, offline manner [24], [68], [30], [29], [34], [38], [42], or used existing iVector/xVector features not tailored for dysarthric/elderly speech and produced mixed results on such data [66], [34], [5], [67]. In particular, model-based adaptation methods not only use all speaker-level data, but also introduce multiple decoding passes and explicit parameter estimation or fine-tuning stages during test-time adaptation [68], [30], [29], [34], [42].

¹Matched pairs sentence-segment word error (MAPSSWE) based statistical significance test [75] is performed at a significance level $\alpha = 0.05$.

2) The proposed two on-the-fly feature-based speaker adaptation methods can consistently learn the latent features characterizing dysarthric and elderly speech, such as reduced speech volume and clarity, irrespective of the amount of speaker-level adaptation data available, as demonstrated in the t-SNE visualizations against baseline iVector or xVector adaptation and LHUC model-based adaptation approaches.

3) Our proposed on-the-fly feature-based adaptation methods produce statistically significant performance improvements over the baseline iVector/xVector adapted hybrid DNN/TDNN and E2E Conformer systems. Statistically significant WER/CER reductions up to **5.32% absolute (18.57% relative)** on four dysarthric/elderly speech corpora across two languages are obtained, while real-time factor speeding up ratio up to 33.6 times obtained against xVector. In addition, our approaches also produce statistically significant WER/CER reductions up to 2.24% absolute (9.20% relative) over LHUC adaptation. These confirm our proposed approaches' effectiveness and genericity in learning homogeneous speaker features for rapid, on-the-fly dysarthric/elderly speaker adaptation.

The rest of this paper is structured as follows. Sec. II reviews speaker adaptation targeting normal speech. Sec. III introduces the proposed on-the-fly VR-SBE features. The associated on-the-fly f-LHUC transforms driven by VR-SBE are presented in Sec. IV. Sec. V addresses several implementation issues impacting the performance of on-the-fly VR-SBE adaptation and estimation of f-LHUC transforms. The experimental results and analyses are presented in Sec. VI, while Sec. VII provides conclusions and future research directions.

II. REVIEW OF SPEAKER ADAPTATION

This section reviews three major categories of speaker adaptation techniques traditionally designed for normal non-aged, healthy speakers, respectively based on: a) auxiliary features, b) feature transformations, and c) model-based adaptation.

Among these, auxiliary speaker embedding features encode speaker-dependent (SD) characteristics via compact representations. The resulting SD features are used as auxiliary inputs to facilitate speaker adaptation during the training and evaluation of ASR systems. Such SD features can be either estimated independently of the back-end systems, e.g. using universal background models (UBMs) based on Gaussian mixture models (GMMs) to learn iVectors [50], [51], or jointly learned with the back-end systems, e.g. alternate updating speaker representations and the remaining parameters when learning speaker codes [49]. Auxiliary speaker features can be flexibly incorporated into both GMM-HMM or hybrid systems and more recent end-to-end E2E systems [77], [78], [5].

Speaker adaptative feature transformations are applied to acoustic front-ends to produce speaker-invariant canonical input features, for example, feature-space maximum likelihood linear regression (f-MLLR) [52] estimated at speaker-level from GMM-HMM systems. Speaker-level physiological differences motivated adaptation based on vocal tract length normalization (VTLN) [79], [80] can be further adopted by applying piecewise linear frequency warping factors [80].

Model-based speaker adaptation techniques apply compact forms of specially designed SD parameters in different net-

work layers, for example, linear input networks (LIN) [81], [56], linear hidden networks (LHN) [55], learning hidden unit contributions (LHUC) [57], [82], linear output networks (LON) [55], parameterized activation functions (PAct) [58], factorized linear transformation [83], and SD neural beamforming, encoder, attention or decoder modules in E2E multi-channel systems [84]. Speaker adaptive training (SAT) [85], [86] can be further applied during training to enable a joint optimization of both SD and SI parameters and to produce a canonical model that can be better adapted to unseen speakers during test-time adaptation. To alleviate the risk of overfitting to limited speaker data during model adaptation, a series of regularized speaker adaptation strategies have been developed. These include the use of L2 [87], [88] or Kullback–Leibler (KL) divergence regularization [89]–[92], and maximum a posterior (MAP) [93], [94] or Bayesian inspired adaptation [82]. Data augmentation on the target speaker via text-to-speech (TTS) synthesis has also been explored in [95], [96]. Alternative objective functions have been investigated for speaker adaptation, for example, adversarial learning [97], [98], [91] and multitask learning [99]–[101], [90], [91].

III. ON-THE-FLY SPEAKER ADAPTATION VIA VARIANCE REGULARIZED SPECTRAL BASIS EMBEDDING FEATURES

On-the-fly feature-based speaker adaptation techniques provide practical solutions to handle both speaker-level data scarcity and processing latency incurred by model-based fine-tuning or adaptation to user data. A key task in designing such techniques is to ensure the homogeneity of speaker-level features and consistent representation of speaker characteristics. To this end, the proposed variance-regularized spectral basis embedding (VR-SBE) features are extracted in three phases (Fig. 1), including: a) conducting singular value decomposition (SVD) on utterance-level dysarthric/elderly speech spectra to produce initial time-invariant spectral bases (Fig. 1 left, in yellow); b) feeding the spectral bases through a first embedding module trained using speaker ID and speech intelligibility/age information, which extracts latent embeddings more consistent and relevant to dysarthric/elderly speaker attributes (Fig. 1 top right, in light blue); and c) feeding the spectral bases through a second embedding module, with an additional output variance regularization cost using the averaged speaker embeddings obtained in stage b) as targets (Fig. 1 bottom right, in pink). Such a process ensures the maximum speaker homogeneity of the final VR-SBE features for on-the-fly adaptation.

A. Speech Spectrum Subspace Decomposition

SVD-based spectrum decomposition [36], [38] provides an intuitive approach to decouple the latent time-invariant spectral and time-variant temporal features in speech signals. In phase-1 of VR-SBE feature extraction (Fig. 1 left, in yellow), SVD is applied to the mel-filterbank log amplitude spectrum $\mathcal{O}_{C \times T_r}^r$ of utterance r with T_r frames and C filterbank channels:

$$\mathcal{O}_{C \times T_r}^r = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^T \quad (1)$$

where the set of column vectors of the $C \times C$ dimensional \mathbf{U}_r matrix (the left-singular vectors) and the set of row vectors of the $T_r \times T_r$ dimensional \mathbf{V}_r^T matrix (the right-singular

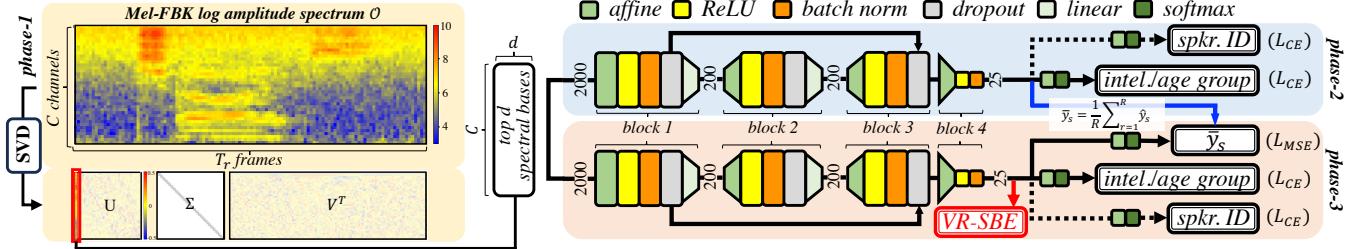


Fig. 1: Example extraction of variance-regularized spectral basis embedding (VR-SBE) features (bottom right) for on-the-fly speaker adaptation over 3 phases: 1) utterance-level SVD spectrum decomposition (left, in light yellow) with $C = 40$ mel-filterbank channels and retaining top $d = 2$ principal spectral bases; 2) multitask learning to derive utterance-level spectral basis embeddings using speaker (“spkr.”) IDs and speech intelligibility (“intel.”) or age groups; and 3) extraction of VR-SBE features with averaged speaker embeddings from phase-2 for computing the additional MSE cost.

vectors) are respectively the bases of the spectral and temporal subspaces. Σ_r is a $C \times T_r$ diagonal matrix containing the singular values in descending order. Its rank equals the number of non-zero singular values, i.e., $\text{rank}(\mathcal{O}_r^{C \times T_r}) \leq \min\{C, T_r\}$.

The top few, for example, $d = 2$, spectral bases are found to contain the most distinct time-invariant dysarthric or elderly speech features for speaker adaptation, such as overall reduced speech volume and clarity, when compared with retaining all spectral bases or further using the temporal bases [36], [38].

To tailor for the low latency requirement of on-the-fly adaptation tasks, instead of performing SVD on the complete spectrum of each utterance, such decomposition can also be performed on part of each utterance’s spectrum in a streaming mode via a sliding analysis widow, e.g. as short as 10 ms. Detailed ablation studies on the number of principal spectral bases and the size of the spectrum decomposition context window are shown in Sec. V-A and VI (Tables XIV and XVII).

B. Spectral Basis Deep Embedding

The above spectral bases are obtained in an unsupervised manner during SVD. To further extract latent information more consistent and relevant to speech impairment severity or age information, supervised feature learning is performed in phase-2 (Fig. 1 top right, in light blue) via developing a 4-block speech intelligibility or age embedding module with the selected $top d$ spectral bases as inputs. The first three blocks each contain 2000 hidden nodes, while the fourth has a 25-dim bottleneck for embedding extraction. Each block comprises the following internal structures in sequence: affine transformation (in green), rectified linear unit (ReLU) activation (in yellow), and batch normalization (in orange). Additionally, we apply linear bottleneck projection (in light green), dropout operation (in grey) and softmax activation (in dark green) respectively to the intermediate two blocks’ inputs, the first three blocks’ outputs and the output block. The first block is connected to the third via a skipping connection. Following [38], the training targets comprise speech intelligibility groups + speaker IDs for dysarthric speech and binary aged vs. non-aged annotations for elderly speech. We then extract 25-dim spectral basis embedding (SBE) features from the trained bottleneck block.

C. Variance-Regularized Deep Embedding

Feature-based adaptation techniques require speaker homogeneity to be consistently encoded in the spectral basis embed-

ding features obtained above in Sec. III-B. As such features are computed over individual utterances or portions of them, additional smoothing is required to ensure their homogeneity, e.g. an overall reduction of speech volume among dysarthric or elderly speakers. Prior research applies an offline speaker-level averaging [38] of all utterance-level produced features, which conflicts with the low processing latency and streaming objective of on-the-fly adaptation.

To this end, an alternative form of speaker-level spectral basis embedding features smoothing based on variance regularization is proposed. This requires a second embedding module to be constructed in phase-3 (Fig. 1 bottom right, in pink). During its training, an additional regression task is adopted using the speaker-level averaged spectral basis embeddings obtained from phase-2 as targets (Fig. 1 blue bold line), which minimizes the output features’ variance and thus further maximizes speaker homogeneity in the final embedding features for on-the-fly adaptation. The overall multitask learning cost interpolates: 1) the cross-entropy (CE) loss over speech intelligibility/age groups, optionally plus that over speaker IDs, and 2) the mean squared error (MSE) between the bottleneck hidden features and the averaged speaker representations obtained from phase-2, given as²:

$$\mathcal{L}_{MTL} = \beta_1 \cdot \mathcal{L}_{MSE} + \beta_2 \cdot \mathcal{L}_{CE_{group}} + \beta_3 \cdot \mathcal{L}_{CE_{ID}} \quad (2)$$

Here $\mathcal{L}_{MSE} = \frac{1}{R} \sum_{r=1}^R (y_s^r - \bar{y}_s)^2$, where y_s^r and \bar{y}_s respectively denote the r^{th} bottleneck hidden feature and averaged representation of speaker s from phase-2.

The 25-dim variance-regularized spectral basis embedding (VR-SBE) features are then extracted from the bottleneck of the second embedding module (Fig. 1 bottom right, red bold line) and concatenated with acoustic features before being fed into hybrid DNN/TDNN (Fig. 2(a)) or E2E Conformer (Fig. 2(b)) ASR systems to facilitate on-the-fly speaker adaptation during both model training and test-time adaptation.

IV. ON-THE-FLY SPEAKER ADAPTATION VIA VR-SBE FEATURE DRIVEN F-LHUC TRANSFORMS

Model-based speaker adaptation techniques offer powerful, fine-grained parameterization of speaker-level attributes when personalizing ASR systems. However, their application to

²The weights are empirically set as $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$ for dysarthric speech datasets, while for elderly speech corpora as $\beta_1 = \beta_2 = \frac{1}{2}, \beta_3 = 0$.

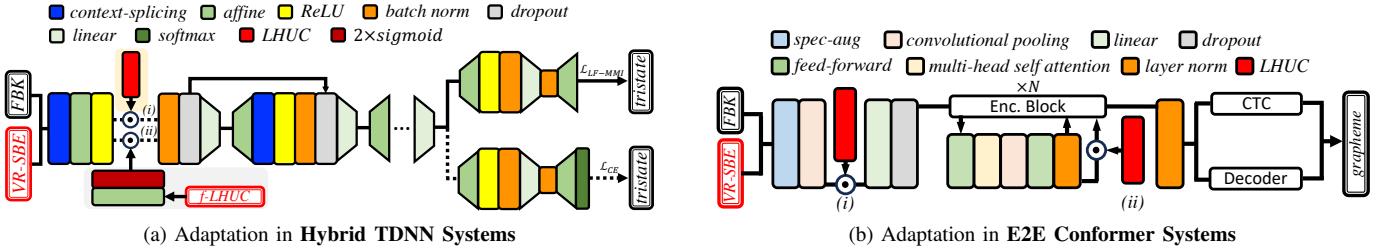


Fig. 2: Incorporating our on-the-fly VR-SBE adaptation at front-ends of (a) hybrid TDNN and (b) E2E Conformer Systems. In (a), path (i) leads to additional model-based LHUC adaptation (top) while path (ii) incorporates our on-the-fly f-LHUC adaptation (bottom). In (b), path (i) and (ii) respectively apply LHUC adaptation after convolution pooling and a particular encoder block.

dysarthric and elderly speech is often hindered by both the scarcity of speaker-level data and processing latency during user fine-tuning. In order to leverage the strengths of both model-based and feature-based adaptation techniques, a novel form of on-the-fly model-based adaptation approach using VR-SBE feature-based LHUC (f-LHUC) transformations is proposed in this paper. The following Sec. IV-A reviews standard model-based LHUC speaker adaptation before VR-SBE feature-based LHUC adaptation is introduced in Sec. IV-B.

A. Batch-Mode Model-Based LHUC Adaptation

The standard approach of applying LHUC speaker adaptation involves training speaker-dependent (SD) linear scaling transforms to adjust activation amplitudes on DNN nodes [102], [57]. Let \mathbf{A}^s denote such transform in a specific layer for speaker s . To reduce the risk of overfitting to limited speaker data, \mathbf{A}^s can be further constrained as a diagonal matrix, equivalent to applying vector \mathbf{v}^s to modify amplitudes of activations. The adapted outputs can be expressed as:

$$\mathbf{h}^s = \xi(\mathbf{v}^s) \odot \mathbf{h} \quad (3)$$

where \mathbf{h} and \odot denote the activated hidden vector and the Hadamard product operation. $\xi(\cdot)$ denotes the activation function, e.g. the element-wise $2 \times \text{sigmoid}(\cdot)$ function [57] ranging from 0 to 2. In this case, the SD LHUC parameters \mathbf{v}^s are typically initialized as 0. Such parameters can be estimated by finding the minimum cross entropy (CE) estimator $\tilde{\mathbf{v}}_{CE}^s$:

$$\tilde{\mathbf{v}}_{CE}^s = \arg \min_{\mathbf{v}^s} \{-\log P(\mathcal{H}^s | \mathcal{D}^s, \mathbf{v}^s)\} \quad (4)$$

where \mathcal{D}^s and \mathcal{H}^s denote the adaptation data recorded from speaker s and the corresponding supervision label, e.g. HMM states in hybrid systems or output tokens in E2E systems. Without loss of generality, we omit the SI model parameters.

During unsupervised test-time adaptation, the label \mathcal{H}^s is commonly produced by initially decoding the adaptation data \mathcal{D}^s . Such batch-mode adaptation process can be iteratively performed to refine both the quality of hypothesis supervision and resulting re-estimated SD parameters. The adapted model with the final estimated SD parameters undergoes another decoding pass to infer hypotheses $\tilde{\mathcal{H}}^s$ for test data $\tilde{\mathcal{D}}^s$.

When using LHUC-based speaker adaptive training (LHUC-SAT) [85], the SD parameters tied to training speakers are joint-optimized with the shared SI parameters. The resulting SI parameters provide a canonical model that more effectively learns speaker-invariant speech properties and can be better adapted to unseen speakers during test-time adaptation.

B. Feature-Based LHUC Transforms

All batch-mode speaker adaptation techniques, including those based on LHUC of Sec. IV-A, operate with multiple decoding and SD parameter estimation passes. This normally requires all speaker-level data to be used to ensure the robustness and fine-granularity of SD parameters. In order to improve data efficiency, reduce processing latency and minimize the impact of supervision errors, an alternative approach is to directly learn a homogeneous mapping between speaker data and associated SD parameters, bypassing the multi-pass decoding followed by unsupervised adaptation procedure entirely. This allows speaker-level feature-based LHUC (f-LHUC) transforms to be directly predicted on the fly during evaluation using the speech data arriving in a streaming mode. To ensure the homogeneity of such regression mapping, the VR-SBE features proposed in Sec. III are used as the regression inputs together with mel-scale filterbank (FBK).

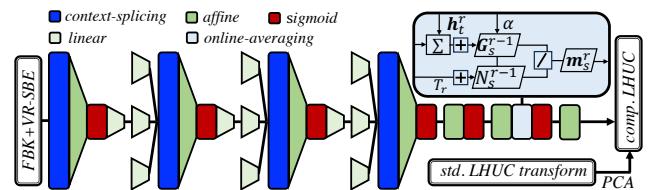


Fig. 3: Example f-LHUC regression network using FBK + VR-SBE input features with an online cross-utterance hidden context averaging layer (top right in light blue) to predict homogeneous speaker-level LHUC transforms on the fly.

As shown in Fig. 3, our TDNN LHUC regression network [69] contains four 500-dim context-splicing layers (in blue) and two 500-dim feed-forward layers comprising affine transformation (in green) and sigmoid activation (in red). Three 300-dim linear bottleneck projections (in light green) are inserted between the context-splicing layers to reduce the number of parameters, while a specifically designed online averaging layer (in light blue) is integrated before the sigmoid activation of the second feed-forward layer (Fig. 3 upper right). Such a layer stores the preceding speech utterances in memory via an accumulated history vector and a frame counter for each speaker. The output of this online averaging layer is given by:

$$\mathbf{m}_s^r = \frac{\sum_{t=1}^{T_r} \mathbf{h}_t^r + \alpha \times \mathbf{G}_s^{r-1}}{T_r + \alpha \times N_s^{r-1}} \quad (5)$$

where \mathbf{G}_s^{r-1} and N_s^{r-1} refer to the accumulated history vector and the frame counter until the $(r-1)^{\text{th}}$ utt. of speaker s . T_r

denotes the frame counts in the r^{th} utt., while \mathbf{h}_t^r denotes its t^{th} hidden vector. α is the history interpolation factor ranging from $[0, 1]$. The numerator of Eqn. (5) represents the weighted sum of hidden vectors until the r^{th} utt. of speaker s , while the denominator denotes the weighted frame counts. The output \mathbf{m}_s^r represents the averaged representation for speaker s until the r^{th} utt. For all speakers, we set $\mathbf{G}_s^0 = \mathbf{0}$ and $N_s^0 = 0$.

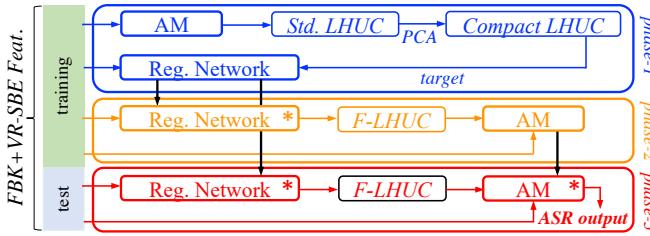


Fig. 4: Overall procedure of building VR-SBE feature driven f-LHUC SAT DNN/TDNN systems over three phases: **1**) f-LHUC regression network training (in blue); **2**) acoustic model (AM) fine-tuning with f-LHUC transforms (in orange); and **3**) on-the-fly speaker adaptation with f-LHUC transforms (in red). Here “AM” stands for DNN/TDNN acoustic model, while “Reg. Network” denotes the f-LHUC regression network. “*” indicates the model or network is frozen during inference.

Fig. 4 demonstrates the three-phase procedure to build our VR-SBE feature driven f-LHUC SAT DNN/TDNN systems:

Phase-1 f-LHUC regression network training involves DNN/TDNN acoustic models (Fig. 2a) to generate standard (std.) LHUC transforms for each training speaker, before being principal component analysis (PCA) projected to a more compact subset of most distinct bases to serve as the f-LHUC regression (reg.) network (Fig. 3) training targets.

Phase-2 acoustic model fine-tuning using training speaker f-LHUC transforms inferred by the well-trained f-LHUC regression network of Phase-1 that is now frozen.

Phase-3 on-the-fly test-time speaker adaptation of the well-trained DNN/TDNN (Phase-2) using test speaker f-LHUC transforms inferred by the f-LHUC regression network (Phase-1) to produce the speaker-dependent (SD) decoding outputs.

V. IMPLEMENTATION DETAILS

This section discusses several implementation details that affect the performance of our VR-SBE features and f-LHUC transforms, including: **1**) the number of principal spectral bases as inputs to VR-SBE embedding network; **2-4**) the inputs to f-LHUC regression network, its PCA compressed regression targets’ dimensionality and its history interpolation factor; and **5**) baseline speaker adaptation of E2E Conformer. Ablation studies are conducted on the benchmark UASpeech [70] dysarthric and DementiaBank Pitt [72] elderly speech datasets³. 40-dim mel-filterbank (FBK) log amplitude spectra serve as inputs to SVD. The history interpolation factor

³A 130.1h training set and a 9h evaluation set are used for UASpeech [24], while a 58.9h training set, a 2.5h development set and a 0.6h evaluation set are used for DementiaBank Pitt [34].

α for f-LHUC regression⁴ (Eqn. (5), Sec. IV-B) is empirically set to 0.9 unless otherwise stated.

A. Number of Spectral Bases for VR-SBE

TABLE I: Ablation study of the number of principal spectral bases fed into the VR-SBE embedding network (Fig. 1) on **UASpeech**. “ d ” refers to the number of principal spectral vectors. “VL”, “L”, “M” and “H” denote speech intelligibility “very low”, “low”, “mid” and “high”.

Sys.	Model (#Para.)	#Hrs	Embed. Input		WER%				
			d	Dim.	VL	L	M	H	All
1			1	40	62.83	29.86	20.50	9.18	28.08
2			2	80	62.54	30.22	18.54	8.59	27.54
3			3	120	63.42	29.94	19.92	8.80	27.98
4	DNN	130.1	4	160	63.79	29.13	20.96	9.25	28.20
5	(6M)		5	200	63.22	29.92	20.43	9.08	28.12
6			10	400	62.86	29.46	19.62	9.04	27.76
7			20	800	63.52	30.06	19.76	9.44	28.22
8			40	1600	64.29	31.02	22.47	9.54	29.18

When performing VR-SBE adaptation of Sec. III-C, Table I suggests selecting top-2 SVD derived spectral bases (Fig. 1, left) for UASpeech produces the best performance across most speech intelligibility groups, including the most challenging “VL”. In contrast, Table II shows that the optimal number of principal spectral bases for DementiaBank Pitt is 3. Such settings are respectively adopted for dysarthric and elderly speech datasets in the following VR-SBE adaptation experiments.

TABLE II: Ablation study of the number of principal spectral bases fed into the VR-SBE embedding network (Fig. 1) on **DementiaBank Pitt**. “Dev” and “Eval” refer to the development and evaluation sets. “INV” and “PAR” denote non-aged clinical investigators and elderly participants.

Sys.	Model (#Para.)	#Hrs	Embed. Input		WER%					
			d	Dim.	Dev		Eval		All (PAR)	
					INV	PAR	INV	PAR	INV	PAR
1			1	40	18.84	44.77	18.09	34.10	31.63	(41.61)
2			2	80	19.15	45.16	17.20	33.53	31.79	(41.72)
3			3	120	18.72	44.67	18.65	34.03	31.55	(41.52)
4	TDNN	58.9	4	160	18.91	44.79	18.31	34.68	31.78	(41.80)
5	(18M)		5	200	19.47	45.45	18.65	34.87	32.31	(42.32)
6			10	400	19.75	45.79	20.64	34.96	32.63	(42.58)
7			20	800	19.94	45.08	18.42	34.22	32.24	(41.86)
8			40	1600	19.50	45.17	18.42	33.53	31.98	(41.72)

B. Inputs to F-LHUC Regression Network

When performing f-LHUC online speaker adaptation of Sec. IV-B, the results in Tables III and IV suggest that on both the UASpeech and DementiaBank Pitt, the combined use of both FBK and VR-SBE features as the inputs to the f-LHUC regression network (Fig. 3, left) generally outperforms using either feature alone, especially when the most powerful configuration combining VR-SBE and f-LHUC is used (Sys.8 vs. Sys.6-7 in Tables III and IV). Hence, both FBK and VR-SBE features are used as the inputs to the f-LHUC regression network for all the following f-LHUC adaptation experiments.

⁴The context-splicing indices for f-LHUC regression are set as $\{-2, -1, 0, 1, 2\}, \{-2, 0, 2\}, \{-3, 0, 3\}$ and $\{-4, 0, 4\}$, following [69]

TABLE III: Ablation study on the input features to the f-LHUC regression network on **UASpeech**. “VR-SBE” denotes variance-regularized spectral basis embedding.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC	WER%				
				VL	L	M	H	All
1	Hybrid DNN (6M)	VR-SBE	FBK	66.45	28.95	20.37	9.62	28.73
2				62.54	30.22	18.54	8.59	27.54
3		VR-SBE	65.06	27.94	18.76	8.39	27.45	
4			63.13	28.65	18.33	8.19	27.07	
5		FBK+VR-SBE	63.22	28.40	18.21	8.23	27.02	
6		FBK	62.86	29.68	19.52	8.67	27.68	
7		VR-SBE	62.38	29.55	19.21	8.66	27.48	
8		FBK+VR-SBE	61.56	28.81	18.39	8.50	26.90	

TABLE IV: Ablation study on the input features to the f-LHUC regression network on **DementiaBank Pitt**.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC	WER%				
				Dev		Eval		All (PAR)
				INV	PAR	INV	PAR	
1	Hybrid TDNN (18M)	VR-SBE	FBK	19.91	47.93	19.76	36.66	33.80 (44.59)
2				18.72	44.67	18.65	34.03	31.55 (41.52)
3		VR-SBE	19.61	45.40	18.87	34.77	32.33 (42.25)	
4			19.60	46.91	18.09	34.98	32.94 (43.38)	
5		FBK+VR-SBE	19.54	46.15	18.20	35.88	32.77 (43.11)	
6		FBK	19.95	45.70	19.20	34.41	32.54 (42.36)	
7		VR-SBE	19.24	45.25	19.53	33.78	31.98 (41.85)	
8		FBK+VR-SBE	17.87	43.83	16.87	34.56	30.91 (41.09)	

C. Dimensionality of F-LHUC Regression Targets

Similarly, when performing f-LHUC online speaker adaptation of Sec. IV-B, Table V and VI suggest that setting the PCA projected speaker LHUC transform subspace dimensionality as 29 and 25 respectively for UASpeech and DementiaBank Pitt produce the best performance (Sys.6 vs. Sys.1-5 in Table V, Sys.1 vs. Sys.2-5 in Table VI) than other settings, for example, 5 on UASpeech intuitively representing the control and four dysarthric speech intelligibility based speakers grouping, or up to 400 for the DementiaBank Pitt which contains more speakers. These settings are used in all the following f-LHUC on-the-fly speaker adaptation experiments of this paper.

TABLE V: Ablation study on the f-LHUC regression network training targets dimensionality (Tar. Dim.) on **UASpeech**.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC	WER%					
				Inputs	Tar. Dim.	VL	L	M	
1	Hybrid DNN (6M)	VR-SBE	FBK	2	64.11	31.06	20.39	8.69	28.47
2				5	63.81	30.51	20.19	8.59	28.19
3			FBK + VR-SBE	15	63.68	30.29	19.90	8.66	28.08
4		VR-SBE	FBK	20	62.58	29.60	19.00	8.70	27.51
5			FBK + VR-SBE	25	61.99	29.16	18.58	8.69	27.19
6			FBK + VR-SBE	29	61.56	28.81	18.39	8.50	26.90

TABLE VI: Ablation study on f-LHUC regression training targets dimensionality (Tar. Dim.) on **DementiaBank Pitt**.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC		WER%						
			Inputs	Tar. Dim.	Dev		Eval		All (PAR)		
					INV	PAR	INV	PAR	INV	PAR	All
1	Hybrid TDNN (18M)	VR-SBE	FBK	25	17.87	43.83	16.87	34.56	30.91 (41.09)		
2				50	18.61	45.01	17.65	35.12	31.79	(42.08)	
3			VR-SBE	100	18.64	44.66	17.76	34.60	31.35	(41.68)	
4				200	18.71	44.68	17.65	34.77	31.65	(41.75)	
5			VR-SBE	400	18.54	44.64	17.31	35.23	31.62	(41.85)	

D. History Interpolation Factor of F-LHUC Regression

Tables VII and VIII suggest that setting the f-LHUC adaptation history interpolation factor (Eqn. (5), Sec. IV-B) as $\alpha = 0.9$ produces the best performance (Sys.7 vs. others in

Tables VII and VIII) on both UASpeech and DementiaBank Pitt. The DementiaBank Pitt task is more sensitive to α , which may be attributed to its longer average sentence duration and temporal contexts than the UASpeech task (3.4s vs. 1.1s).

TABLE VII: Ablation study on history interpolation factor α (Int. Fac.) in f-LHUC regression training on **UASpeech**.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC		WER%					
			Inputs	Int. Fac.	VL	L	M	H	All	
1	Hybrid DNN (6M)	VR-SBE	FBK		0.0	62.83	29.74	19.37	8.68	27.66
2					0.1	63.47	29.70	19.31	8.71	27.78
3		VR-SBE	FBK		0.3	62.60	29.62	19.11	8.69	27.53
4			FBK + VR-SBE		0.5	62.58	29.71	19.03	8.71	27.54
5		VR-SBE	+		0.7	62.62	29.65	18.98	8.61	27.49
6			VR-SBE		0.8	62.35	29.61	19.19	8.70	27.50
7		VR-SBE	FBK + VR-SBE		0.9	61.56	28.81	18.39	8.50	26.90
8			FBK + VR-SBE		1.0	62.29	29.58	19.13	8.64	27.45

TABLE VIII: Ablation study on history interpolation factor α in f-LHUC regression training on **DementiaBank Pitt**.

Sys.	Model (#Para.)	Adapt. Feat.	f-LHUC		WER%						
			Inputs	Int. Fac.	Dev		Eval		All (PAR)		
					INV	PAR	INV	PAR	INV	PAR	All
1	Hybrid TDNN (18M)	VR-SBE	FBK		0.0	19.44	48.71	19.42	36.93	34.57	(45.22)
2					0.1	21.44	47.39	19.53	37.16	34.28	(44.36)
3		VR-SBE	FBK		0.2	22.51	50.56	21.86	40.85	36.66	(47.69)
4			FBK + VR-SBE		0.3	22.66	50.16	20.19	39.63	36.31	(47.04)
5		VR-SBE	+		0.4	22.74	50.12	20.64	39.68	36.35	(47.03)
6			VR-SBE		0.5	21.48	48.32	17.42	37.79	34.71	(45.20)
7		VR-SBE	iVector		0.6	21.77	49.38	18.09	37.83	35.27	(45.96)
8			xVector		0.7	20.83	47.39	17.64	36.34	33.84	(44.12)

Baseline adaptation of E2E Conformer models of this paper uses feature-based on-the-fly adaptation approaches only that are based on, e.g. iVectors or xVectors. Such a choice is based on the ablation studies in Tables IX and X on UASpeech⁵ and

⁵Block 2 data of the control speech and its speed-perturbed versions are also used for Conformer system training.

DementiaBank Pitt. The performance of iVector, xVector and VR-SBE adaptation (Sys.6-8) are found generally more competitive than those obtained using LHUC adaptation performed in supervised mode at different Conformer sublayers (Sys.2-5), serving as the upper bound performance of model-based adaptation approaches including both LHUC and f-LHUC.

VI. EXPERIMENTS AND RESULTS

In this section, we investigate the performance⁶ of our feature-based on-the-fly adaptation approaches on four tasks: the English UASpeech [70] and TORG0 [71] dysarthric speech datasets, and the English DementiaBank Pitt [72] and Cantonese JCCOCC MoCA [73] elderly speech corpora. All the settings determined in the ablation studies of Sec. V are adopted. SI and SD speed perturbations [24], [34] based data augmentation are applied to all tasks. The extraction of 100-dim iVector⁷ and 25-dim xVector⁸ follow the Kaldi recipes using the same training data as for ASR systems⁹.

A. Experiments on Dysarthric Speech

1) The UASpeech Dataset: As the largest publicly available dysarthric speech dataset, UASpeech [70] is an isolated word recognition task in English with 103 hours of speech from 16 dysarthric and 13 control speakers recorded using a 7-channel microphone array. The block based training-evaluation data partitioning protocol [70], [1]-[3], [17] is adopted. For each speaker, the data is split into B1, B2 and B3, each with the same 155 common words and a different 100 uncommon words. The training set combines B1 and B3 data of all 29 speakers, while the evaluation set contains the B2 data from the 16 dysarthric speakers. Silence stripping [29] leads to a 30.6h training set (99195 utt.) and a 9h evaluation set (26520 utt.). Further data augmentation [24], [29] produces a 130.1h training set (399110 utt.). As E2E systems are sensitive to training data coverage, B2 of the control speech and its speed-perturbed versions are also used for Conformer system training, creating a 173h training set (538292 utt.).

2) The TORG0 Dataset: The English TORG0 [71] dysarthric dataset contains 13.5h of speech from 8 dysarthric and 7 control speakers based on short sentences and single words. We adopt a 3-block based training-evaluation data partition similar to UASpeech. The control speech and two-thirds of the speech of each impaired speaker are used for training, while the remaining one-third is for evaluation. Silence stripping produces a 6.5h training set (14541 utt.) and a 1h evaluation set (1892 utt.). Further speed perturbation [24], [104] produces a 34.1h augmented training set (61813 utt.).

3) Experimental Setup for UASpeech: Following [24], [29], the hybrid DNN systems with six 2000-dim and one 100-dim layers are implemented using extended Kaldi [105]. Each hidden layer contains linear bottleneck projection, affine

⁶For all speech recognition results measured in WER/CER, a matched pairs sentence-segment word error (MAPSSWE) statistical significance test [103] is performed at the significance level $\alpha = 0.05$.

⁷Kaldi: egs/wsj/s5/local/nnet3/run_ivector_common.sh

⁸Kaldi: egs/sre16/v1/local/nnet3/xvector/tuning/run_xvector_1a.sh

⁹Changing the dimensionality of iVector and xVector to other settings produced marginal and statistically non-significant differences in performance.

transformation, rectified linear unit (ReLU) activation, batch normalization and dropout operations, while Softmax activation is used in the output layer. Two skipping connections respectively feed the first layer's outputs to the third and those of the fourth to the sixth. Multi-task learning is adopted, using frame-level triphone state and mono-phone alignments as targets. The E2E Conformers are built using ESPnet [106]¹⁰ to predict grapheme (letter) sequence outputs. The inputs to both systems are 80-dim FBK + Δ features, while a 9-frame context is used in hybrid DNN. A uniform language model with a word grammar network is adopted during evaluation [1].

4) Experimental Setup for TORG0: The hybrid factorized TDNN systems with 7 context slicing layers are constructed following Kaldi [105] chain recipe. The setup of E2E graphemic Conformers is similar to that for UASpeech¹¹. 40-dim FBK features are used as input for both systems, with a 3-frame context for the hybrid system. A 3-gram language model (LM) trained using all TORG0 transcripts with a word recognition vocabulary of 1.6k is utilized during evaluation.

5) Performance Analyses: The performance contrast between baseline iVector, xVector, batch-mode LHUC or SBE speaker adaptation, and the proposed VR-SBE on-the-fly speaker adaptation on the UASpeech task is shown in Table XI. Several trends can be observed:

i) Our proposed on-the-fly VR-SBE adapted systems (Sys.6,18,31) consistently outperform iVector (Sys.2,14,27) and xVector adaptation (Sys.3,15,28) with varying amounts of training data on both hybrid DNN and E2E Conformer systems. Statistically significant overall word error rate (WER) reductions by up to **2.48% abs. (7.92% rel.) and 2.85% abs. (8.06% rel.)** can be respectively achieved on DNN (Sys.6 vs. Sys.2) and Conformer (Sys.31 vs. Sys.27) models.

ii) Our on-the-fly VR-SBE adaptation (Sys.6,18,31) consistently outperforms the comparable variance-regularized iVector (iVR) adaptation (Sys.5,17,30).

iii) The improvements obtained by batch-mode LHUC adaptation (Sys.7,19) over the SI systems (Sys.1,13) are largely retained (by 82%) and comparable to those obtained using on-the-fly VR-SBE adaptation (Sys.6,18).

iv) On-the-fly VR-SBE adaptation (Sys.6,18) also produces performance comparable to batch-mode SBE adaptation [38] that requires speaker-level embedding averaging (Sys.4,16) and incurs additional processing latency.

v) The combined use of both input speaker features and f-LHUC adaptation leads to the most powerful form of on-the-fly adaptation configurations (Sys.10-12, Sys.22-24). Among these, on-the-fly VR-SBE plus FBK+VR-SBE driven f-LHUC (Sys.12,24) adaptation not only consistently outperforms the comparable baselines by replacing VR-SBE with iVector or xVector (Sys.10-11, Sys.22-23), but also produces further WER reductions over using VR-SBE adaptation alone by up to 0.64% abs. (2.32% rel.) (Sys.12,24 vs. Sys.6,18).

vi) The best performance is obtained by combining the system using VR-SBE plus FBK+VR-SBE driven f-LHUC adaptation (Sys.24) and using only VR-SBE adaptation alone

¹⁰12 encoder + 12 decoder layers, feed-forward layer dim = 2048, attention heads = 4, dim of attention heads = 256, interpolated CTC+AED cost.

¹¹8 encoder + 4 decoder layers, feed-forward layer dim = 1024

TABLE XI: Performance comparison between baseline iVector, xVector, batch-mode LHUC or SBE [38] speaker adaptation and the proposed VR-SBE on-the-fly speaker adaptation on the **UASpeech** test set of 16 dysarthric speakers. “CONF.” denotes Conformer. “VL”, “L”, “M” and “H” refer to speech intelligibility “very low”, “low”, “mid” and “high”. “On Fly” indicates using on-the-fly adaptation. \dagger , \ddagger and $*$ denote a stat. significant improvement ($\alpha = 0.05$) obtained over iVector (Sys.2,10,14,22,27), xVector (Sys.3,11,15,23,28), or both. “A \rightarrow B” denotes system A produces N-best outputs in the 1st decoding pass before the 2nd pass rescoring by system B.

Sys.	Model (#Para.)	#Hrs	Adapt. Feat.	LHUC SAT	f-LHUC	On Fly	WER%				
							VL	L	M	H	All
1	Hybrid DNN (6M)	30.6	X			-	69.82	32.61	24.53	10.40	31.45
2			iVector			✓	69.46	33.75	22.58	10.45	31.33
3			xVector	X	X	✓	68.07	33.56	24.07	9.49	30.93
4			SBE [36]			X	64.43	29.71	19.84	8.57	28.05
5			iVR			✓	68.66	33.72	22.84	9.83	30.99
6			VR-SBE			✓	65.04*	30.90*	20.70*	10.15*	28.85*
7			X	X	X	X	64.39	29.88	20.27	8.95	28.29
8			SBE [36]	✓	X	X	63.40	28.90	18.64	8.13	27.24
9			X	(FBK)		✓	66.47†	29.55†	21.00†	8.99†	28.80†
10			iVector	X	(+iVector)	✓	64.86	36.44	21.17	9.03	30.29
11			xVector	(+xVector)		✓	68.28	30.79	21.98	9.31	29.80
12			VR-SBE	(+VR-SBE)		✓	65.75†	29.80*	19.07*	8.99‡	28.31*
6+12			-			✓	64.36*	29.68*	18.96*	8.89*	27.96*
13	Hybrid TDNN (6M)	130.1	X			-	66.45	28.95	20.37	9.62	28.73
14			iVector			✓	65.73	30.10	20.21	9.03	28.65
15			xVector	X	X	✓	65.40	29.32	20.72	8.85	28.41
16			SBE [36]			X	61.55	27.52	17.31	8.22	26.26
17			iVR			✓	66.02	29.52	19.56	9.32	28.53
18			VR-SBE			✓	62.54*	30.22‡	18.54*	8.59‡	27.54*
19			X		X	X	62.50	27.26	18.41	8.04	26.55
20			SBE [36]	✓	X	X	59.83	27.16	16.80	7.91	25.60
21			X	(FBK)		✓	65.06†	27.94	18.76†	8.39†	27.45†
22			iVector	X	(+iVector)	✓	63.63	32.56	18.52	8.31	28.28
23			xVector	(+xVector)		✓	65.70	29.22	20.13	8.90	28.35
24			VR-SBE	(+VR-SBE)		✓	61.56*	28.81*	18.39*	8.50‡	26.90*
18+24			-			✓	60.80*	28.19*	17.72*	8.23*	26.36*
25	CONF. (52M)	130.1	X			-	73.88	53.12	49.92	42.03	53.17
26			X			-	65.70	40.63	33.39	9.53	34.07
27			iVector			✓	69.05	42.45	33.60	9.74	35.37
28			xVector	X	X	✓	67.70	40.13	30.80	8.49	33.52
29			SBE [36]			X	65.18	34.90	24.21	5.00	29.19
30			iVR			✓	68.94	42.00	32.19	8.52	34.55
31			VR-SBE			✓	67.52†	38.85*	28.60*	7.88*	32.52*
32		18+24 \rightarrow 31					57.33*	24.79*	14.27*	5.99*	23.33*

(Sys.18) via frame-level joint decoding (Sys.18+24)¹², before the resulting 50-best outputs are further rescored by the VR-SBE adapted Conformer¹³. Using this combined system (Sys.32), statistically significant overall WER reductions by up to **5.32% abs. (18.57% rel.)** can be achieved over iVector and xVector adaptation (Sys.32 vs. Sys.14).

Performance comparison between our best on-the-fly speaker-adapted system (Sys.32, Table XI) and recently published systems on UASpeech is presented in Table XII. To our best knowledge, our system (Sys.32, Table XI) gives the best performance among all systems in Table XII without using out-of-domain data and self-supervised learning (SSL). It produces performance comparable or superior to the fine-tuned SSL systems (cross-lingual XLRS, WavLM, or Wav2vec2.0), particularly on the most challenging “VL” intelligibility subset.

Similar trends can be observed on a comparable set of experiments conducted on TORG [71] (Table XIII). Our on-the-fly VR-SBE adaptation (Sys.6,18) outperforms iVector (Sys.2,14) or xVector (Sys.3,15) adaptation by up to **1.75% abs. (12.38% rel.)** WER reduction (Sys.18 vs. Sys.14), while outperforming batch-mode LHUC adaptation by **1.39% abs. (14.85% rel.)** WER reduction (Sys.6 vs. Sys.7). Combining on-the-fly VR-SBE and FBK+VR-SBE driven f-LHUC adaptation leads to up to **1.34% abs.(14.77% rel.)** WER reduction over iVector and xVector adaptation (Sys.6+12 vs. Sys.2).

¹²System weights empirically set as 11:9.

¹³1st and 2nd pass system weights empirically set as 1:9.

TABLE XII: Performance against recently published systems on **UASpeech**. “DA” denotes data augmentation. “SSL” refers to incorporating fine-tuned self-supervised learning foundation speech models or features.

	Online Adapt	SSL	VL	All
Sheffield-2015 speaker adaptive training [2]	X	X	70.78	34.85
Sheffield-2020 fine-tuning CNN-TDNN speaker adaptation [27]	✓	X	68.24	30.76
CUHK-2020 DNN + DA + LHUC SAT [24]	X	X	62.44	26.37
CUHK-2021 QuartzNet + CTC meta-learning + SAT [32]	X	X	69.30	30.50
Sheffield-2022 DA + source filter features + f-Vector adapt [67]	✓	X	-	30.30
FAU-2022 cross-lingual XLRS + Conformer [6]	-	✓	62.00	26.10
Nagoya-2022 WavLM [7]	-	✓	71.50	51.80
BUT-2022 Wav2vec2.0 + fMLLR + vectors [5]	✓	✓	57.72	22.83
JHU-2023 DuTa-VC (Diffusion) + Conformer [107]	-	X	63.70	27.90
DA + SVR adapt + f-LHUC system combination + 2-pass rescoring (Table XI, Sys.32)	✓	X	57.33	23.33

TABLE XIII: Performance comparison between baseline iVector, xVector, batch-mode LHUC or SBE [38] adaptation and the proposed VR-SBE on-the-fly speaker adaptation on **TORG**. “Seve.”, “Mod.” and “Mild” refer to the speech intelligibility. \dagger and $*$ denote a stat. significant improvement ($\alpha = 0.05$) obtained over iVector adaptation (Sys.2,10,14), or further over xVector adaptation (Sys.3,11,15).

Sys.	Model (#Para.)	Adapt. Feat.	LHUC	f-LHUC	On Fly	WER%				
						Seve.	Mod.	Mild	All	
1	Hybrid TDNN (10M)	130.1	X		-	12.80	8.78	3.64	9.47	
2			iVector			✓	13.82	5.92	2.40	9.07
3			xVector	X	X	✓	13.86	4.39	3.02	8.94
4			SBE [38]			X	11.67	4.59	2.86	7.80
5			iVR			✓	13.62	5.31	2.86	8.96
6			VR-SBE			✓	12.07*	4.69†	2.63	7.97*
7			X		X	✓	12.60	8.78	3.64	9.36
8			SBE [38]	✓	X	X	11.71	4.29	2.86	7.76
9			X	(FBK)		✓	12.80	5.00	2.94	8.50
10			iVector	X	(+iVector)	✓	13.94	5.61	2.55	9.11
11			xVector	(+xVector)		✓	14.35	5.82	2.55	9.36
12			VR-SBE	(+VR-SBE)		✓	12.07*	4.90*	2.79	8.05*
6+12			-			✓	11.59*	4.69*	2.71	7.73*
13	CONF. (18M)	130.1	X		-	21.66	6.22	4.10	13.67	
14			iVector			✓	22.15	7.44	3.94	14.13
15			xVector	X	X	✓	21.38	9.08	4.02	14.09
16			SBE [38]			X	18.69	6.83	3.71	12.15
17			iVR			✓	21.34	7.65	3.63	13.67
18			VR-SBE			✓	19.22*	6.32*	3.94	12.38*

rel.) WER reduction (Sys.18 vs. Sys.14), while outperforming batch-mode LHUC adaptation by **1.39% abs. (14.85% rel.)** WER reduction (Sys.6 vs. Sys.7). Combining on-the-fly VR-SBE and FBK+VR-SBE driven f-LHUC adaptation leads to up to **1.34% abs.(14.77% rel.)** WER reduction over iVector and xVector adaptation (Sys.6+12 vs. Sys.2).

6) *Processing Latency Analyses:* We then conduct further analysis on the processing latency and feature homogeneity of our on-the-fly adaptation techniques. The size of the analysis sliding window imposed on the VR-SBE feature extraction is gradually reduced to as short as 10 ms of acoustic inputs. As shown in Table XIV, the performance of our on-the-fly VR-SBE adaptation is largely insensitive to the duration of analysis sliding windows from 1 utt. to 10 ms (Sys.3-10). This suggests that VR-SBE features extracted on the fly can instantaneously capture homogeneous characteristics of dysarthric speakers. The real-time factor (RTF) represents the overall delay incurred by waiting for data and model processing. By using only 10 ms of acoustic inputs, our VR-SBE adaptation achieves both stat. significant improvements over iVector or xVector adaptation and lower processing latency with a real-time factor speeding up by 33.6 times (Sys.10 vs. Sys.1-2).

TABLE XIV: Ablation study of the analysis sliding window (Slid. Wind.) size for on-the-fly VR-SBE feature extraction on **UASpeech**. ‘‘RTF’’ denotes real-time factor. \dagger and $*$ denote a stat. significant improvement ($\alpha = 0.05$) obtained over iVector (Sys.1), or both iVector and xVector (Sys.1-2).

Sys.	Model (#Para.)	Adapt. Feat.	Slid. Wind.	RTF	WER%				
					VL	L	M	H	All
1	Hybrid DNN (6M)	iVector	100ms	0.10	65.73	30.10	20.21	9.03	28.65
2			utt.	1.01	65.40	29.32	20.72	8.85	28.41
3		xVector	utt.	1.02	62.54*	30.22	18.54 \dagger	8.59†	27.54*
4			250ms	0.23	64.26*	28.62*	19.90*	9.04	27.89*
5		VR-SBE	150ms	0.14	64.18*	28.71*	19.82*	9.29	27.97*
6			100ms	0.10	64.49*	28.36*	19.47*	9.27	27.87*
7			50ms	0.06	63.99*	28.68*	19.17*	8.99	27.70*
8			30ms	0.04	64.20*	29.29 \dagger	19.39*	9.21	28.02*
9			20ms	0.03	64.38*	28.99 \dagger	19.47*	9.19	27.98*
10			10ms	0.03	64.77*	29.03 \dagger	19.21*	9.09	28.00*

B. Experiments and Results on Elderly Speech

1) *The DementiaBank Pitt Dataset*: The English DementiaBank Pitt [72] dataset contains 33h of cognitive impairment assessment interviews between 292 elderly participants and clinical investigators. The training set includes 688 speakers (244 elderly and 444 investigators), while the development and evaluation sets¹⁴ respectively contain 119 (43 elderly and 76 investigators) and 95 speakers (48 elderly and 47 investigators). There is no overlap speaker between the training and the development or evaluation sets. Silence stripping [34] leads to a 15.7h training set (29682 utt.), a 2.5h development set (5103 utt.) and a 0.6h evaluation set (928 utt.), while data augmentation [34] produces a 58.9h training set (112830 utt.).

2) *The JCCOCC MoCA Dataset*: The Cantonese JCCOCC MoCA dataset comprises cognitive impairment assessment interviews between 256 elderly participants and clinical investigators [73]. The training set contains 369 speakers (158 elderly and 211 investigators), while the development and evaluation sets each contain speech from 49 elderly other than those in the training set. Silence stripping leads to a 32.1h training set (95448 utt.), a 3.5h development set (13675 utt.) and a 3.4h evaluation set (13414 utt.). Further data augmentation [36] produces a 156.9h training set (389049 utt.).

3) *Experiment Setup for the DementiaBank Pitt Corpus*: Following the Kaldi [105] chain setup, the hybrid factorized TDNN systems contain 14 context-slicing layers with a 3-frame context. The setup of E2E graphemic Conformers follows that for UASpeech. 40-dim FBK features are used as inputs. A word-level 4-gram LM with Kneser-Ney (KN) smoothing is trained [34], with a 3.8k vocabulary covering all words in DementiaBank Pitt adopted during evaluation.

4) *Experiment Setup for the JCCOCC MoCA Corpus*: The setup of the hybrid TDNN and E2E character Conformer systems are the same as those for DementiaBank Pitt. 40-dim FBK features are adopted as inputs. A word-level 4-gram LM with KN smoothing is trained using the transcription of JCCOCC MoCA (610k words), with a 5.2k recognition vocabulary covering all words in JCCOCC MoCA adopted.

¹⁴The evaluation set contains the same 48 speakers’ Cookie theft picture description recordings as the test set of ADReSS [108] while the development set contains recordings of these speakers in other tasks if available.

TABLE XV: Performance comparison between baseline iVector, xVector, batch-mode LHUC or SBE [38] adaptation and the proposed VR-SBE on-the-fly speaker adaptation on the **DementiaBank Pitt** corpus. ‘‘Dev’’ and ‘‘Eval’’ stand for the development and evaluation sets. ‘‘INV’’ and ‘‘PAR’’ refer to clinical investigator and elderly participant. $*$ denotes a stat. significant improvement ($\alpha = 0.05$) obtained over both iVector (Sys.2,10,14) and xVector (Sys.3,11,15) adaptation.

Sys.	Model (#Para.)	Adapt. Feat.	LHUC SAT	f-LHUC	On Fly	WER%				
						Dev	Eval	INV	PAR	All (PAR)
1					-	19.91	47.93	19.76	36.66	33.80 (44.59)
2		iVector			✓	19.97	46.76	18.20	37.01	33.37 (43.87)
3		xVector			✓	19.75	47.58	19.31	35.57	33.40 (44.02)
4		SBE [38]			✗	18.61	43.84	17.98	33.82	31.12 (40.87)
5	Hybrid	iVR			✓	19.19	47.64	18.65	35.80	33.26 (44.13)
6	TDNN	VR-SBE			✓	18.72*	44.67*	18.65	34.03*	31.55* (41.52)*
7	(18M)				✗	19.26	45.49	18.42	35.44	32.33 (42.51)
8		SBE [38]	✓		✗	17.41	40.94	17.98	31.89	29.16 (38.26)
9			(FBK)		✓	19.61	45.40	18.87	34.77	32.33 (42.25)
10		iVector	✗	(+iVector)	✓	18.75	47.07	17.98	36.11	32.85 (43.83)
11		xVector	✗	(+xVector)	✓	19.50	46.23	18.98	35.53	32.75 (43.06)
12		VR-SBE	✗	(+VR-SBE)	✓	17.87*	43.83*	16.87*	34.56*	30.91* (41.09)*
6+12					-	17.66*	43.48*	16.09*	33.68*	30.51* (40.58)*
13					-	20.97	48.71	19.42	36.93	34.57 (45.22)
14		iVector			✓	21.48	48.32	17.42	37.79	34.71 (45.20)
15	CONF.	xVector			✓	21.77	49.38	18.09	37.83	35.27 (45.96)
16	(52M)	SBE [38]			✗	20.44	47.70	17.31	36.11	33.76 (44.27)
17		iVR			✓	22.09	49.56	19.64	38.58	35.65 (46.31)
18		VR-SBE			✓	20.83*	47.39*	17.64	36.34*	33.84* (44.12)*

5) *Performance Analyses*: The performance comparison between our on-the-fly VR-SBE adaptation, iVector [50] or xVector [74] adaptation and the batch-mode LHUC or SBE adaptation on DementiaBank Pitt is demonstrated in Table XV. Several trends can be observed:

- i) The proposed on-the-fly VR-SBE adapted systems (Sys.6,18) consistently outperform the iVector (Sys.2,14) and xVector adaptation (Sys.3,15) on hybrid TDNN and E2E Conformer systems. Statistically significant overall WER reductions by up to **1.85% abs. (5.54% rel.)** and **1.43% abs. (4.05% relative)** are respectively obtained on TDNN (Sys.6 vs. Sys.3) and Conformer (Sys.18 vs. Sys.15) models.
- ii) On-the-fly VR-SBE adaptation (Sys.6,18) consistently outperforms the comparable variance-regularized iVector (iVR) adaptation (Sys.5,17).
- iii) On-the-fly VR-SBE adaptation also outperforms batch-mode LHUC adaptation with a WER reduction of **0.78% abs. (2.41% rel.)** (Sys.6 vs. Sys.7).
- iv) On-the-fly VR-SBE adaptation (Sys.6,18) produces performance comparable to batch-mode SBE adaptation [38] using time-consuming speaker embedding averaging (Sys.4,16).
- v) On-the-fly VR-SBE plus FBK+VR-SBE driven f-LHUC adaptation (Sys.12) not only outperforms the comparable baselines replacing VR-SBE with iVector or xVector (Sys.10-11), but also gives further WER reductions over VR-SBE adaptation alone by **0.64% abs. (2.03% rel.)** (Sys.12 vs. Sys.6).
- vi) Combining VR-SBE plus FBK+VR-SBE feature driven f-LHUC adaptation (Sys.12) and VR-SBE adaptation alone (Sys.6) via frame-level joint decoding (Sys.6+12)¹⁵ gives the best performance, with statistically significant overall WER reductions by up to **2.89% abs. (8.65% rel.)** over iVector and xVector adaptation (Sys.6+12 vs. Sys.3).

¹⁵System weights empirically set as 11:9.

TABLE XVI: Performance comparison between baseline iVector, xVector, batch-mode LHUC or SBE [38] adaptation and the proposed VR-SBE on-the-fly speaker adaptation on the JCCOCC MoCA development (Dev) and evaluation (Eval) sets containing elderly speakers only. * denotes a stat. significant improvement ($\alpha = 0.05$) obtained over both iVector (Sys.2,10,14) and xVector (Sys.3,11,15) adaptation.

Sys.	Model (#Para.)	Adapt. Feat.	LHUC	f-LHUC	On Fly	CER%		
						Dev	Eval	All
1	Hybrid TDNN (18M)	X	X	X	-	26.87	23.71	25.28
2		iVector			✓	25.46	22.80	24.12
3		xVector			✓	26.72	23.80	25.25
4		SBE [38]			X	24.43	21.68	23.05
5		iVR			✓	25.22	22.74	23.97
6		VR-SBE			✓	24.86*	22.18*	23.51*
7	SBE [38]	X	X	(FBK)	X	25.77	22.94	24.35
8		✓			X	23.59	21.42	22.50
9		X			✓	24.93	21.82	23.37
10		iVector			✓	25.03	22.58	23.80
11		xVector			✓	25.42	22.28	23.84
12		VR-SBE			✓	23.59*	21.12*	22.35*
6+12	CONF. (53M)	-	X	X	✓	23.55*	20.70*	22.11*
13		X			-	33.08	31.24	32.15
14		iVector			✓	33.76	31.83	32.79
15		xVector			✓	33.43	32.10	32.76
16		SBE [38]			X	32.08	30.75	31.41
17		iVR			✓	34.79	32.48	33.63
18		VR-SBE			✓	32.42*	31.01*	31.71*

Experiments conducted on JCCOCC MoCA [73] (Table XVI) demonstrate similar trends. On-the-fly VR-SBE adaptation (Sys.6,18) outperforms iVector (Sys.2,14) or xVector (Sys.3,15) adaptation by up to **1.74% abs. (6.89% rel.)** CER reduction (Sys.6 vs. Sys.3), and batch-mode LHUC adaptation by **0.84% abs. (3.45% rel.)** CER reduction (Sys.6 vs. Sys.7). Combining on-the-fly VR-SBE and FBK+VR-SBE driven f-LHUC adaptation leads to CER reductions of up to **3.14% abs. (12.44% rel.)** over iVector and xVector adaptation (Sys.6+12 vs. Sys.3) and **2.24% abs.(9.20% rel.)** over batch-mode LHUC adaptation (Sys.6+12 vs. Sys.7).

TABLE XVII: Ablation study of the analysis sliding window (Slid. Wind.) size for on-the-fly VR-SBE feature extraction on **DementiaBank Pitt**. “RTF” denotes real-time factor. †, ‡ and * denote a stat. significant improvement ($\alpha = 0.05$) is obtained over iVector (Sys.1), xVector (Sys.2), or both.

Sys.	Model (#Para.)	Adapt. Feat.	Slid. Wind.	RTF	WER%					
					Dev		Eval		All (PAR)	
					INV	PAR	INV	PAR		
1	Hybrid TDNN (18M)	iVector	100ms	0.08	19.97	46.76	18.20	37.01	33.37 (43.87)	
2		xVector	utt.	1.02	19.75	47.58	19.31	35.57	33.40 (44.02)	
3			utt.	1.03	18.72*	44.67*	18.65	34.03*	31.55* (41.52)*	
4				0.16	18.88*	44.91*	18.98	34.37*	31.78* (41.79)*	
5			250ms	0.11	19.21†	44.83*	17.87	33.53*	31.71* (41.48)*	
6			150ms	0.08	19.08*	45.03*	19.42	34.41*	31.93* (41.89)*	
7			50ms	0.06	19.38	44.87*	18.31‡	34.52*	31.97* (41.81)*	
8			30ms	0.05	18.93*	44.57*	18.53	34.52*	31.68* (41.59)*	
9			20ms	0.04	18.91*	45.45*	18.20‡	34.10*	31.94* (42.09)*	
10			10ms	0.04	19.33	44.93*	19.09	34.35*	31.97* (41.80)*	

6) *Processing Latency Analyses:* The results in Table XVII show trends similar to those previously found in Table XIV on UASpeech, and further confirm that the performance of on-the-fly VR-SBE adaptation is largely insensitive to changes in the size of input feature analysis sliding window.

C. Further Analyses on Feature Homogeneity

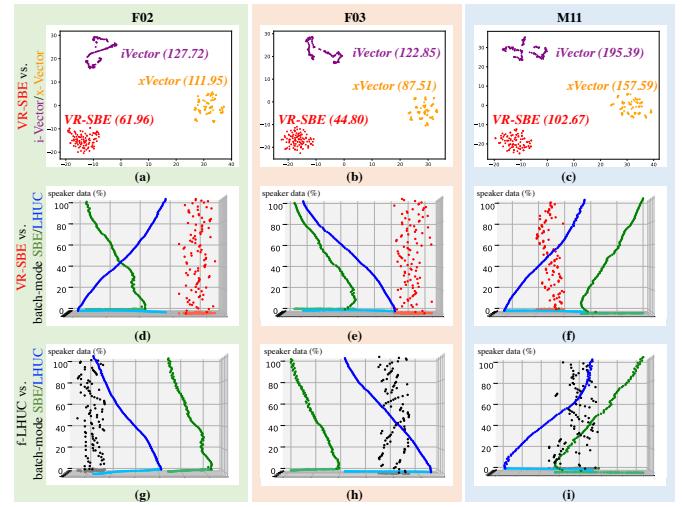


Fig. 5: T-SNE visualization illustrating speaker feature homogeneity measured by covariance determinants after applying t-SNE projection: (a)-(c) on-the-fly VR-SBE features vs. iVectors and xVectors (d)-(f) VR-SBE vs. batch-mode SBE features and LHUC transforms and (g)-(i) VR-SBE feature driven f-LHUC vs. batch-mode SBE features and LHUC transforms obtained from speaker F02, F03 and M11 of **UASpeech**.

We further analyze the homogeneity of the proposed on-the-fly adaptation features. T-distributed stochastic neighbor embedding (t-SNE) [76] visualization is conducted on the proposed VR-SBE features against baseline iVectors and x-Vectors in 2-D plots (Fig. 5(a)-(c)), and also on VR-SBE or FBK+VR-SBE feature driven f-LHUC transforms against batch-mode spectral basis embedding (SBE) features and LHUC transforms in 3-D plots (Fig. 5(d)-(i)). Three UASpeech¹⁶ dysarthric speakers of mixed genders are selected from the very low (F03), low (F02) and mid (M11) speech intelligibility groups. For each speaker, 101 distinct speaker-level adaptation data quantity operating points are used for computing the above speaker features or LHUC transforms. These points correspond to using only one utterance or 1%, 2%, ..., 99%, and up to 100% of the speaker-level data. Fig. 5 shows that the proposed on-the-fly VR-SBE features and their associated f-LHUC transforms consistently exhibit stronger speaker homogeneity measured in covariance determinants after applying t-SNE projection than baseline iVectors, x-Vectors and batch-mode SBE features and LHUC transforms.

We further compare the performance of the proposed on-the-fly VR-SBE adaptation with batch-mode LHUC and SBE adaptation when using limited amounts of speaker-level adaptation data. As depicted in Fig. 6(a) and Fig. 6(b), VR-SBE adaptation is more robust to changes in the amount of speaker-level data than batch-mode LHUC and SBE adaptation. Specifically, VR-SBE adaptation outperforms SBE adaptation when using less than 20% of the speaker data and surpasses LHUC adaptation when using less than 40% of speaker data.

¹⁶The amounts of speaker-level data of test speakers of UASpeech (34 min, including all channels) are larger than those of TORG (7.8 min), DementiaBank Pitt (0.4 min) and JCCOCC MoCA (4.2 min). Given this, t-SNE visualizations are conducted using the UASpeech data.

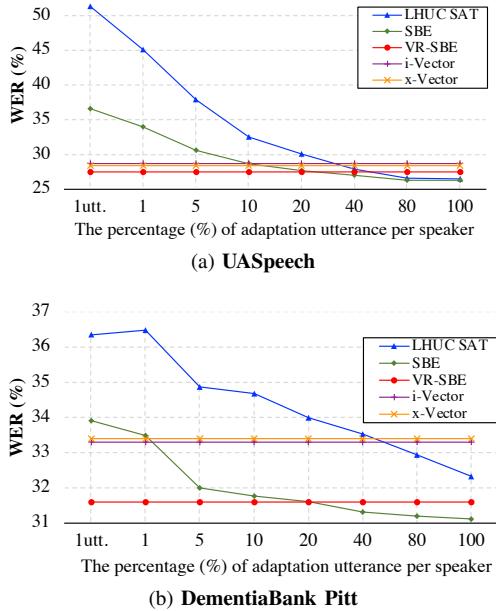


Fig. 6: Performance comparison of hybrid DNN/TDNN systems between batch-mode LHUC/SBE adaptation, on-the-fly iVector/x-Vector adaptation, and the proposed on-the-fly VR-SBE adaptation with various percentages of speaker-level adaptation data on (a) UASpeech and (b) DementiaBank Pitt.

VII. DISCUSSION AND CONCLUSION

This paper presents two novel methods to learn homogeneous dysarthric and elderly speaker features for on-the-fly test-time adaptation of TDNN and Conformer models, including: 1) speaker-level variance-regularized spectral basis embedding (VR-SBE) features that ensure speaker-level feature consistency via specially designed regularization; and feature-based LHUC (f-LHUC) transforms driven by VR-SBE. Experiments conducted on four dysarthric and elderly speech corpora across English and Cantonese suggest our proposed approaches achieve statistically significant WER or CER reductions of up to 5.32% absolute (18.57% relative) over baseline iVector/xVector adaptation, and 2.24% absolute (9.20% relative) over batch-mode offline LHUC adaptation. Further processing latency analyses and t-SNE visualization show that our VR-SBE features and f-LHUC transforms are robust to speaker-level data quantity in test-time adaptation, while exhibiting stronger speaker-level homogeneity than iVectors, xVectors and batch-mode LHUC transforms. Future research will focus on rapid adaptation of pre-trained ASR systems.

REFERENCES

- [1] H. Christensen *et al.*, “A comparative study of adaptive, automatic recognition of disordered speech,” in *INTERSPEECH*, 2012.
- [2] S. Sehgal *et al.*, “Model adaptation and adaptive training for the recognition of dysarthric speech,” in *SLPAT*, 2015.
- [3] S. Hu *et al.*, “The CUHK Dysarthric Speech Recognition Systems for English and Cantonese,” in *INTERSPEECH*, 2019.
- [4] S. R. Shahamiri, “Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system,” *IEEE T NEUR SYS REH*, 2021.
- [5] M. K. Baskar *et al.*, “Speaker adaptation for Wav2vec2 based dysarthric ASR,” in *INTERSPEECH*, 2022.
- [6] A. Hernandez *et al.*, “Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition,” in *INTERSPEECH*, 2022.
- [7] L. P. Violeta *et al.*, “Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition,” in *INTERSPEECH*, 2022.
- [8] S. Hu *et al.*, “Exploiting Cross-domain And Cross-Lingual Ultrasound Tongue Imaging Features For Elderly And Dysarthric Speech Recognition,” in *INTERSPEECH*, 2023.
- [9] A. Almadhor *et al.*, “E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition,” *EXPERT SYST APPL*, 2023.
- [10] J.-W. Kim *et al.*, “Unsupervised Representation Learning with Task-Agnostic Feature Masking for Robust End-to-End Speech Recognition,” *Mathematics*, 2023.
- [11] Z. Jin *et al.*, “Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition,” *IEEE TAUDIO SPEECH*, 2023.
- [12] W. Lanier, *Speech disorders*. Greenhaven Publishing LLC, 2010.
- [13] K. C. Fraser *et al.*, “Linguistic features identify Alzheimer’s disease in narrative speech,” *J. Alzheimer’s Dis.*, 2016.
- [14] A. Association, “2019 Alzheimer’s disease facts and figures,” *Alzheimer’s & dementia*, 2019.
- [15] V. Young *et al.*, “Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review,” *ASSIST TECHNOL*, 2010.
- [16] R. Vipperla *et al.*, “Ageing voices: The effect of changes in voice parameters on ASR performance,” *EURASIP J AUDIO SPEE*, 2010.
- [17] H. Christensen *et al.*, “Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech,” in *INTERSPEECH*, 2013.
- [18] L. Zhou *et al.*, “Speech Recognition in Alzheimer’s Disease and in its Assessment,” in *INTERSPEECH*, 2016.
- [19] B. Vachhani *et al.*, “Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition,” in *INTERSPEECH*, 2017.
- [20] M. J. Kim *et al.*, “Dysarthric Speech Recognition Using Convolutional LSTM Neural Network,” in *INTERSPEECH*, 2018.
- [21] N. M. Joy *et al.*, “Improving acoustic models in TORG0 dysarthric speech database,” *IEEE T NEUR SYS REH*, 2018.
- [22] S. Liu *et al.*, “Exploiting Visual Features Using Bayesian Gated Neural Networks for Disordered Speech Recognition,” in *INTERSPEECH*, 2019.
- [23] J. Shor *et al.*, “Personalizing ASR for Dysarthric and Accented Speech with Limited Data,” in *INTERSPEECH*, 2019.
- [24] M. Geng *et al.*, “Investigation of Data Augmentation Techniques for Disordered Speech Recognition,” in *INTERSPEECH*, 2020.
- [25] Y. Lin *et al.*, “Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription,” in *INTERSPEECH*, 2020.
- [26] I. Kodrasi *et al.*, “Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection,” *IEEE T AUDIO SPEECH*, 2020.
- [27] F. Xiong *et al.*, “Source Domain Data Selection for Improved Transfer Learning Targeting Dysarthric Speech Recognition,” in *ICASSP*, 2020.
- [28] R. Takashima *et al.*, “Two-step acoustic model adaptation for dysarthric speech recognition,” in *ICASSP*, 2020.
- [29] S. Liu *et al.*, “Recent Progress in the CUHK Dysarthric Speech Recognition System,” *IEEE T AUDIO SPEECH*, 2021.
- [30] Z. Jin *et al.*, “Adversarial Data Augmentation for Disordered Speech Recognition,” in *INTERSPEECH*, 2021.
- [31] S. Hu *et al.*, “Bayesian Learning of LF-MMI Trained Time Delay Neural Networks for Speech Recognition,” *IEEE T AUDIO SPEECH*, 2021.
- [32] D. Wang *et al.*, “Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization,” in *ISCSLP*, 2021.
- [33] J. R. Green *et al.*, “Automatic Speech Recognition of Disordered Speech: Personalized models outperforming human listeners on short phrases,” in *INTERSPEECH*, 2021.
- [34] Z. Ye *et al.*, “Development of the CUHK Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the Dementia-bank Corpus,” in *ICASSP*, 2021.
- [35] Y. Pan *et al.*, “Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer’s Dementia Detection Through Spontaneous Speech,” in *INTERSPEECH*, 2021.
- [36] M. Geng *et al.*, “Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition,” in *INTERSPEECH*, 2021.
- [37] J. Harvill *et al.*, “Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary,” in *ICASSP*, 2021.

- [38] M. Geng *et al.*, "Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition," *IEEE T AUDIO SPEECH*, 2022.
- [39] S. Hu *et al.*, "Exploring Self-supervised Pre-trained ASR Models For Dysarthric and Elderly Speech Recognition," in *ICASSP*, 2023.
- [40] M. Geng *et al.*, "On-the-Fly Feature Based Rapid Speaker Adaptation for Dysarthric and Elderly Speech Recognition," in *INTERSPEECH*, 2023.
- [41] Z. Jin *et al.*, "Adversarial Data Augmentation Using VAE-GAN for Disordered Speech Recognition," in *ICASSP*, 2023.
- [42] J. Qi *et al.*, "Parameter-efficient Dysarthric Speech Recognition Using Adapter Fusion and Householder Transformation," in *INTERSPEECH*, 2023.
- [43] M. Geng *et al.*, "Use of Speech Impairment Severity for Dysarthric Speech Recognition," in *INTERSPEECH*, 2023.
- [44] H. Wang *et al.*, "Enhancing Pre-trained ASR System Fine-tuning for Dysarthric Speech Recognition using Adversarial Data Augmentation," in *ICASSP*, 2024.
- [45] P. Jerntorp *et al.*, "Stroke registry in Malmö, Sweden," *STROKE*, 1992.
- [46] T. L. Whitehill *et al.*, "Speech errors in Cantonese speaking adults with cerebral palsy," *CLIN LINGUIST PHONET*, 2000.
- [47] M. P. Caligiuri, "The influence of speaking rate on articulatory hypokinesia in Parkinsonian dysarthria," *BRAIN LANG*, 1989.
- [48] S. J. Peppé, "Why is prosody in speech-language pathology so difficult?" *INT J SPEECH-LANG PA*, 2009.
- [49] O. Abdel-Hamid *et al.*, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *ICASSP*, 2013.
- [50] G. Saon *et al.*, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*, 2013.
- [51] A. Senior *et al.*, "Improving DNN speaker independence with i-vector inputs," in *ICASSP*, 2014.
- [52] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *COMPUT SPEECH LANG*, 1998.
- [53] L. F. Uebel *et al.*, "An investigation into vocal tract length normalisation," in *EUROSPEECH*, 1999.
- [54] F. Seide *et al.*, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011.
- [55] R. Gemello *et al.*, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *SPEECH COMMUN*, 2007.
- [56] B. Li *et al.*, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *INTERSPEECH*, 2010.
- [57] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE T AUDIO SPEECH*, 2016.
- [58] C. Zhang *et al.*, "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions," in *ICASSP*, 2016.
- [59] H. V. Sharma *et al.*, "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition," in *SLPAT*, 2010.
- [60] A. Baba *et al.*, "Elderly acoustic model for large vocabulary continuous speech recognition," in *EUROSPEECH*, 2001.
- [61] K. T. Mengistu *et al.*, "Adapting acoustic and lexical models to dysarthric speech," in *ICASSP*, 2011.
- [62] M. J. Kim *et al.*, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models," in *INTERSPEECH*, 2013.
- [63] C. Bhat *et al.*, "Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation," in *INTERSPEECH*, 2016.
- [64] M. Kim *et al.*, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE T NEUR SYS REH*, 2017.
- [65] J. Tobin *et al.*, "Personalized automatic speech recognition trained on small disordered speech datasets," in *ICASSP*, 2022.
- [66] C. Espana-Bonet *et al.*, "Automatic speech recognition with deep neural networks for impaired speech," in *IberSPEECH*, 2016.
- [67] Z. Yue *et al.*, "Acoustic Modelling From Raw Source and Filter Components for Dysarthric Speech Recognition," *IEEE T AUDIO SPEECH*, 2022.
- [68] J. Deng *et al.*, "Bayesian Parametric and Architectural Domain Adaptation of LF-MMI Trained TDNNs for Elderly and Dysarthric Speech Recognition," in *INTERSPEECH*, 2021.
- [69] X. Xie *et al.*, "Fast DNN Acoustic Model Speaker Adaptation by Learning Hidden Unit Contribution Features," in *INTERSPEECH*, 2019.
- [70] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.
- [71] F. Rudzicz *et al.*, "The TORG0 database of acoustic and articulatory speech from speakers with dysarthria," *LANG RESOUR EVAL*, 2012.
- [72] J. T. Becker *et al.*, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *ARCH NEUROL CHICAGO*, 1994.
- [73] S. S. Xu *et al.*, "Speaker Turn Aware Similarity Scoring for Diarization of Speech-Based Cognitive Assessments," in *APSIPA ASC*, 2021.
- [74] D. Snyder *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.
- [75] D. S. Pallet *et al.*, "Tools for the analysis of benchmark speech recognition tests," in *ICASSP*, 1990.
- [76] L. Van der Maaten *et al.*, "Visualizing data using t-SNE," *J MACH LEARN RES*, 2008.
- [77] Z. Tüske *et al.*, "On the limit of English conversational speech recognition," in *INTERSPEECH*, 2021.
- [78] M. Zeineldeen *et al.*, "Improving the training recipe for a robust conformer-based hybrid model," in *INTERSPEECH*, 2022.
- [79] E. Eide *et al.*, "A parametric approach to vocal tract length normalization," in *ICASSP*, 1996.
- [80] L. Lee *et al.*, "Speaker normalization using efficient frequency warping procedures," in *ICASSP*, 1996.
- [81] J. Neto *et al.*, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *EUROSPEECH*, 1995.
- [82] X. Xie *et al.*, "Bayesian learning for deep neural network adaptation," *IEEE T AUDIO SPEECH*, 2021.
- [83] Y. Zhao *et al.*, "Low-rank plus diagonal adaptation for deep neural networks," in *ICASSP*, 2016.
- [84] T. Ochiai *et al.*, "Speaker adaptation for multichannel end-to-end speech recognition," in *ICASSP*, 2018.
- [85] T. Anastasakos *et al.*, "A compact model for speaker-adaptive training," in *ICSLP*, 1996.
- [86] T. Ochiai *et al.*, "Speaker adaptive training using deep neural networks," in *ICASSP*, 2014.
- [87] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *ICASSP*, 2013.
- [88] M. Kitza *et al.*, "Comparison of BLSTM-Layer-Specific Affine Transformations for Speaker Adaptation," in *INTERSPEECH*, 2018.
- [89] D. Yu *et al.*, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, 2013.
- [90] K. Li *et al.*, "Speaker adaptation for end-to-end CTC models," in *SLT*, 2018.
- [91] Z. Meng *et al.*, "Speaker adaptation for attention-based end-to-end speech recognition," in *INTERSPEECH*, 2019.
- [92] Y. Huang *et al.*, "Acoustic model adaptation for presentation transcription and intelligent meeting assistant systems," in *ICASSP*, 2020.
- [93] Z. Huang *et al.*, "Maximum a posteriori adaptation of network parameters in deep models," in *INTERSPEECH*, 2015.
- [94] Z. Huang *et al.*, "Bayesian unsupervised batch and online speaker adaptation of activation function parameters in deep models for automatic speech recognition," *IEEE T AUDIO SPEECH*, 2016.
- [95] Y. Huang *et al.*, "Rapid RNN-T Adaptation Using Personalized Speech Synthesis and Neural Language Generator," in *INTERSPEECH*, 2020.
- [96] K. C. Sim *et al.*, "Personalization of end-to-end speech recognition on mobile devices for named entities," in *ASRU*, 2019.
- [97] Z. Meng *et al.*, "Speaker-invariant training via adversarial learning," in *ICASSP*, 2018.
- [98] T. Tsuchiya *et al.*, "Speaker invariant feature extraction for zero-resource languages with adversarial learning," in *ICASSP*, 2018.
- [99] R. Price *et al.*, "Speaker adaptation of deep neural networks using a hierarchy of output layers," in *SLT*, 2014.
- [100] P. Swietojanski *et al.*, "Structured output layer with auxiliary targets for context-dependent acoustic modelling," in *INTERSPEECH*, 2015.
- [101] Z. Huang *et al.*, "Rapid adaptation for deep neural networks through multi-task learning," in *INTERSPEECH*, 2015.
- [102] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *SLT*, 2014.
- [103] L. Gillick *et al.*, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, 1989.
- [104] S. Hu *et al.*, "Exploiting Cross Domain Acoustic-to-articulatory Inverted Features for Disordered Speech Recognition," in *ICASSP*, 2022.
- [105] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.
- [106] S. Watanabe *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *INTERPSEECH*, 2018.

- [107] H. Wang *et al.*, “DuTa-VC: A Duration-aware Typical-to-atypical Voice Conversion Approach with Diffusion Probabilistic Model,” in *INTERSPEECH*, 2023.
- [108] S. Luz *et al.*, “Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge,” in *INTERSPEECH*, 2020.