# Spectro-Temporal Deep Features
# for Disordered Speech Assessment and Recognition

*Mengzhe Geng[1], Shansong Liu[1], Jianwei Yu[1], Xurong Xie[2], Shoukang Hu[1], Zi Ye[1], Zengrui Jin[1],*
*Xunying Liu[1], Helen Meng[1]*

[1]The Chinese University of Hong Kong
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{mzgeng,ssliu,jwyu,skhu,zye,zrjin,xyliu,hmmeng}@se.cuhk.edu.hk, xr.xie@siat.ac.cn

## Abstract

Automatic recognition of disordered speech remains a highly challenging task to date. Sources of variability commonly found in normal speech including accent, age or gender, when further compounded with the underlying causes of speech impairment and varying severity levels, create large diversity among speakers. To this end, speaker adaptation techniques play a vital role in current speech recognition systems. Motivated by the spectro-temporal level differences between disordered and normal speech that systematically manifest in articulatory imprecision, decreased volume and clarity, slower speaking rates and increased dysfluencies, novel spectro-temporal subspace basis embedding deep features derived by SVD decomposition of speech spectrum are proposed to facilitate both accurate speech intelligibility assessment and auxiliary feature based speaker adaptation of state-of-the-art hybrid DNN and end-to-end disordered speech recognition systems. Experiments conducted on the UASpeech corpus suggest the proposed spectro-temporal deep feature adapted systems consistently outperformed baseline i-Vector adaptation by up to 2.63% absolute (8.6% relative) reduction in word error rate (WER) with or without data augmentation. Learning hidden unit contribution (LHUC) based speaker adaptation was further applied. The final speaker adapted system using the proposed spectral basis embedding features gave an overall WER of 25.6% on the UASpeech test set of 16 dysarthric speakers.

**Index Terms**: Speech Disorders, Speech Recognition, Speaker Adaptation, Speech Assessment, Subspace-based Learning

## 1. Introduction

In spite of the swift progress of automatic speech recognition (ASR) technologies targeting normal speech in the past few decades [1–9], accurate recognition of disordered speech remains a demanding task to date [10–16]. The underlying causes of speech disorders include a wide range of neuro-motor conditions, such as cerebral palsy, Parkinson disease, amyotrophic lateral sclerosis and stroke or traumatic brain injuries [17]. Despite the reduced intelligibility of disordered speech, hands-free and speech-enabled assistive technologies are among natural alternatives [18] to aid people with speech disorders since they often suffer from the co-occurring physical disabilities.

A key challenge in current ASR system development is to systematically model the latent variations among diverse speech data. A wide range of sources of variability commonly found in normal speech, including speaker specific idiosyncrasy such as accent and physiological differences brought by age or gender, when further compounded with the underlying inducements of speech impairment and varying severity levels, create large di-

versity among disordered speech. Due to the associate difficulties in controlling the muscles and articulators used in speech production [19], abnormalities including articulation imprecision, reduced intensity and clarity, slower speaking rates and increased disfluencies are observed in disordered speech [20]. Furthermore, temporal or spectral perturbation based data augmentation techniques widely used in both state-of-the-art ASR systems for normal speech [21–24] and recently those designed for impaired speech [25–27] introduce extra diversity. To this end, speaker adaptation techniques play a crucial role in current ASR systems for both normal and disordered speech.

Previous researches of speech adaptation targeting normal speech recognition can be divided into three broad categories: 1) auxiliary speaker embedding feature based approaches that encodes speaker-dependent (SD) characteristics in a compact vector representation, such as speaker codes [28], i-Vectors [29] and bottleneck features [30]; 2) feature transformation based methods applied at acoustic front-ends that produce speaker invariant input features in a canonical representation, such as feature-space maximum likelihood linear regression (f-MLLR) [31]; 3) model based adaptation techniques that exploits specially designed SD transformations in model parameters to handle the speaker level variability [32,33].

In contrast, so far there has been limited research on speaker adaptation targeting disordered speech recognition, particularly those suitable for state-of-the-art ASR systems. Many of the earlier researches were conducted in the context of traditional hidden Markov models (HMMs) with Gaussian mixture model (GMM) state density distributions. In [10, 34, 35], maximum likelihood linear regression (MLLR) and maximum a posterior (MAP) were applied to speaker-independent (SI) GMM-HMM systems. In [12], a combination of MLLR and MAP adaptation were used in speaker adaptive training (SAT) of SI GMM-HMM models. In [36], f-MLLR based SAT was studied. In [37], regularized speaker adaptation on Kullback-Leibler divergence-based HMMs (KL-HMMs) was conducted. More recent researches investigated model adaptation of state-of-the-art deep neural network (DNN) based systems. Dysarthric speaker adaptation of recurrent neural network transducers (RNN-Ts) [4] and lattice-free MMI trained time delay neural networks (TDNNs) [5] via direct model parameter fine-tuning were studied in [38, 39]. Learning hidden unit contributions based (LHUC) SAT [33] was investigated in [13, 27]. The majority of prior researches on disordered speech adaptation focused on feature transformation and model based adaptation. On the contrary, very limited research has been conducted on auxiliary speaker embedding feature based adaptation approaches, particularly those motivated by the underlying spectra-temporal variation of impaired speech of diverse causes and severity levels.

This paper proposes novel deep spectro-temporal subspace basis embedding features to facilitate both accurate speech intelligibility assessment and auxiliary feature based speaker adaptation for disordered speech recognition. Spectral and temporal basis vectors derived by singular value decomposition (SVD) [40] of speech spectrum were used to structurally represent the spectro-temporal level key attributes found in disordered speech [41–43], such as the overall decrease in speaking rate and speech volume as well as changes in spectral envelope. These two form of basis vectors were then used to construct a DNN speech intelligibility classifier. More compact, lower dimensional speaker specific spectral and temporal DNN embedding features produced by the bottleneck layer of the resulting intelligibility classifier were further employed as auxiliary features to adapt start-of-the-art hybrid DNN [27] and CTC end-to-end [3] disordered speech recognition systems. Experiments were conducted on the largest available and most widely used UASpeech [44] dysarthric speech corpus and LHUC based speaker adaptation was further applied.

The main contributions of the paper are summarized below:

1) To the best of our knowledge, our novel spectro-temporal deep feature based adaptation approach is the first work to exploit auxiliary speaker embedding features in disordered speech adaptation. In contrast, prior works [10, 12, 13, 27, 34–39] focused on feature transformation and model based adaptation. Speaker embedding features, e.g. i-Vector [45, 46], were used in speech assessment rather than ASR adaptation tasks.

2) The spectro-temporal deep features are intuitively related to the underlying diversity of disordered speech. The spectral basis embedding features are designed to learn characteristics such as volume reduction, changes of formant position, imprecise articulation and hoarse voice while the temporal ones to capture patterns such as increased disfluencies and pauses. Experiments conducted on UASpeech suggest that our proposed spectro-temporal deep feature adapted systems consistently outperformed comparable baseline i-Vector adaptation [29, 47] by up to 2.63% absolute (8.6% relative) reduction in word error rate (WER) with or without data augmentation. The final speaker adapted system using the proposed spectral basis embedding features gave an overall WER of 25.6% on the UASpeech test set of 16 dysarthric speakers, which is the best ASR performance so far published as far as we know.

The rest of this paper is organized as follows. The derivation of spectro-temporal basis vectors via speech spectrum subspace decomposition is presented in Sec.2. The subspace basis vector based DNN speech intelligibility classifier and embedding features for speaker adaptation are proposed in Sec.3. Sec.4 presents experiments and results of both speech intelligibility assessment and speech recognition on UASpeech. The last section concludes and discuss possible future works.

## 2. Speech Spectrum Subspace Decomposition

To systematically reveal the patterns contained in disordered speech, we conduct SVD on the mel-filterbank log amplitude spectrum following [40] to derive basis vectors of the spectral and the temporal subspaces. Let $\mathbf{S_r}$ represent a $C \times T$ dimensional mel-spectrogram of utterance $r$ with $C$ mel-filterbank channels and $T$ frames. The SVD of $\mathbf{S_r}$ is given by

$$\mathbf{S_r} = \mathbf{U_r}\boldsymbol{\Sigma_r}\mathbf{V_r^T} \tag{1}$$

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 $T \times T$ dimensional $\mathbf{V_r^T}$ matrix (the right-singular vectors) are respectively the bases of the spectral and the temporal subspaces, and $\boldsymbol{\Sigma_r}$ is a $C \times T$ diagonal matrix containing the singular values in descending order [40]. The rank of $\mathbf{S_r}$ is equal to the number of non-zero singular values, i.e. $rank(\mathbf{S_r}) \leq min\{C, T\}$. Motivated by low-rank approximation [48], we select the top $d$ principal spectral and temporal basis vectors for all the experiments of this paper.
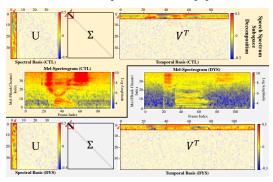


Figure 1: *Example subspace decomposition of mel-spectrograms of normal (CTL, upper) and dysarthric (DYS, lower) utterances of word "choice" to obtain top $d$ spectral and temporal basis vectors (circled in red in $\mathbf{U}$ and $\mathbf{V}^T$).*

As shown in Fig.1, when compared with that of the normal speaker, the mel-spectrogram of the dysarthric speaker contains reduced energy especially in the lower portions of mel-scale frequencies. The overall spectral envelope of the dysarthric spectrum is less visible with weakened formants. Spectral basis vectors are designed to capture these patterns as well as imprecise articulation and hoarse voice. As the dysarthric speaker speaks more slowly and less fluently, the extracted temporal basis vectors encode the speaking rate in their dimensionality in addition to other patterns such as increased disfluencies and pauses.

Traditional disordered speech assessment methods often require the contents spoken by different speakers to be the same [43, 49], which allows normal and dysarthric speech of identical contents but varying durations to be compared after an alignment procedure [50]. In order to facilitate a more practical assessment scheme applicable to unrestricted speech contents of unknown durations and derive speaker-level embedding vectors of consistent dimensionality for ASR system adaptation in this paper, when processing the temporal basis vectors of each utterance, a frame-level sliding window of 25 dimensions was applied to the top $d$ selected temporal basis vectors. Their 25 dimensional mean and standard deviation vectors were then computed to serve as the "average" temporal basis representations of fixed dimensionality, as shown in Fig.2.
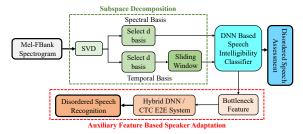


Figure 2: *Framework of the proposed speech intelligibility assessment and auxiliary feature based speaker adaptation founded on spectro-temporal subspace basis vectors derived from speech spectrum subspace decomposition.*

## 3. Spectro-Temporal Deep Features

The overall framework of our proposed subspace decomposition based speech intelligibility assessment and spectro-temporal deep feature based speaker adaptation for disordered speech recognition is shown in Fig.2. A DNN speech intelligibility classifier is trained using the spectro-temporal basis vectors discussed in Sec.2. More compact, lower dimensional features obtained from the DNN classifier are used as auxiliary speaker embedding features for disordered speech adaptation.

### 3.1. DNN Speech Intelligibility Classifier

The DNN speech intelligibility classifier shares similar structure with the hybrid DNN acoustic model for disordered speech recognition except that the DNN classifier has four hidden layers while the hybrid DNN has seven. As illustrated in Fig.3, each hidden layer of the DNN classifier contains a basic set of neural operations performed in sequence, i.e. affine transformation (in green), rectified linear unit (ReLU) activation (in yellow) and batch normalization (in orange). Apart from this, linear bottleneck projections (in light green) are applied to the inputs of the intermediate two hidden layers and dropout operations (in grey) are applied to the outputs of the first three hidden layers. A skip connection connects the output of the first layer to that of the third layer. The first three layers are of 2000 dimensions while the 25-dimensional fourth layer serves at the bottleneck layer. Softmax activation (in dark green) is applied in the output layer. Multi-task learning (MTL) [51] is implemented and the labels for the two tasks are the speaker's intelligibility group and speaker ID respectively.
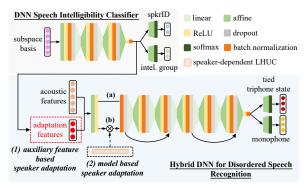


Figure 3: *Architecture of the DNN speech intelligibility classifier (upper) and the hybrid-DNN system for disordered speech recognition (lower). Selecting connection (a) leads to systems using auxiliary feature based adaptation only, while selecting connection (b) leads to systems with additional LHUC SAT.*

### 3.2. Auxiliary Feature Based Speaker Adaptation

As shown in Fig.2 and Fig.3, the output of the 25 dimensional bottleneck layer, after a speaker-level averaging, is concatenated to the acoustic features at the front-end of the hybrid DNN (shown in Fig.3) or CTC end-to-end system[1] for disordered speech recognition. The hybrid DNN acoustic model shares the same structure as our previous work [27] except for the front-end. The MTL labels for speech recognition are tied triphone state and monophone alignments. In Fig.3, **1)** Excluding both the adaptation features concatenated at input and the speaker dependent LHUC [33] model parameters leads to the baseline

---

[1]The CTC E2E system consists of four 2D convolutional layers and three BLSTM layers.

speaker independent (SI) systems. **2)** Keeping the concatenated adaptation features but excluding the LHUC parameters leads to the spectro-temporal deep feature adapted systems. **3)** Keeping both adaptation features LHUC parameters leads to the spectro-temporal deep feature adapted systems with LHUC SAT.

## 4. Experiments and Results

### 4.1. Task Description

The UASpeech [44] corpus is the largest available and most widely used corpus for research on speech disorders. It is an isolated word recognition task consisting of speech recordings from 16 dysarthric and 13 control speakers on 155 common words and 300 uncommon words. There're three blocks per speaker, each containing all 155 common words and one third of the uncommon words. Speech intelligibility rating given to each speaker lies in five groups, i.e. "very low", "low", "mid", "high" and "control". Block 1 and block 3 of all 29 speakers are treated as training set while block 2 of the 16 dysarthric speakers as test set. Silence stripping was performed using a HTK [52] trained GMM-HMM system as described in our previous work [13], which results in a 30.6-hour training set (99195 utterances) and a 9-hour test set (26520 utterances).

### 4.2. Experiment Setup

In our experiments, the DNN speech intelligibility classifier and the hybrid DNN acoustic model were implemented using an extension to the Kaldi toolkit [53] while the CTC end-to-end system was implemented using PyTorch. We select top 2 and top 5 principal spectral and temporal basis vectors based on the experimental results [42]. The inputs to the acoustic models were 80-dimensional filter bank (FBank) + $\Delta$ features plus 25-dimensional DNN bottleneck spectro-temporal features or 100-dimensional i-Vectors[2]. The other settings of the hybrid DNN were exactly the same as our previous work [27]. Following [10], a uniform language model was used in decoding.

### 4.3. Performance of Speech Intelligibility Assessment

Table 1: *Utterance-level accuracy of the DNN speech intelligibility classifier on 49717 utterances from block 2 of all 29 speakers. Here "DYS" is short for dysarthric. "SB" and "TB" denote "spectral basis vectors" and "temporal basis vectors" respectively. "CTL v.s. DYS" and "5-way intel." stand for binary speech intelligibility (normal / dysarthric) assessment and finer classification among DYS intelligibility subgroups including very low (VL), low (L), mid (M), high (H) and control (CTL).*

| DNN Input | DNN Label | Assess | Accuracy(%) | | | | | CTL | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | | | DYS | | | | | | |
| | | | VL | L | M | H | Avg | | |
| SB | Intel. | 5-way Intel. | 92.8 | 95.6 | 94.5 | 95.8 | 94.9 | 99.3 | 96.9 |
| SB+TB | | | 91.1 | 94.9 | 94.4 | 95.5 | 94.2 | 99.3 | 96.6 |
| SB | Intel. + spkrID | | 93.4 | 96.7 | 96.5 | 96.3 | 95.9 | 99.5 | 97.6 |
| TB | | | 78.2 | 69.4 | 61.8 | 58.6 | 66.2 | 93.6 | 79.0 |
| SB+TB | | | **93.9** | 96.8 | 97.2 | 95.7 | **95.9** | 99.7 | **97.7** |
| SB | Intel. | DYS v.s. CTL | - | - | - | - | 98.9 | 99.3 | 99.1 |
| SB+TB | | | - | - | - | - | 98.8 | 99.3 | 99.0 |
| SB | Intel. + spkrID | | - | - | - | - | **99.3** | 99.5 | **99.4** |
| TB | | | - | - | - | - | 87.2 | 93.6 | 90.2 |
| SB+TB | | | - | - | - | - | 98.9 | 99.7 | 99.3 |

For speech intelligibility assessment, apart from block 2 of the 16 dysarthric speakers, block 2 of the 13 control speakers

---

[2]We follow Kaldi: egs/wsj/s5/steps/nnet/ivector/extract_ivectors.sh

are also used for testing the DNN speech intelligibility classifier, making a total of 49717 utterances. As shown in Table 1, for the 5-way intelligibility assessment, the best overall accuracy of 97.7% is achieved when both spectral and temporal basis vectors are given as input to the DNN classifier (line 5). For the binary assessment which classifies a speaker as control or dysarthric, the best overall accuracy of 99.4% is achieved with spectral basis as input (line 8). Ablation study which removes speaker ID from the DNN label (line 1-2, 6-7) indicates that there is only small degradation in assessment accuracy if the intelligibility group labels only are provided during training.

### 4.4. Performance of Speaker Adaptation

Table 2: *WER performance comparison of the proposed spectro-temporal deep feature adaptation, i-Vector adaptation and LHUC adaptation on the UASpeech test set of 16 dysarthric speakers. "6M" and "26M" refer to # of network parameters. "DYS" and "CTL" in "Data Aug" column denote perturbing the disordered and the normal speech respectively for data augmentation. "SBE" and "TBE" denote spectral basis and temporal basis embedding features. "SBE⋆" and "STBE⋆" mean only intelligibility group labels were used in the previous DNN speech intelligibility classifier for ablation study. "VL / L / M / H" refer to intelligibility groups. "† and "‡ denote a statistically significant improvement is obtained over the comparable baseline i-Vector or SI systems and LHUC adaptation respectively.*

| Sys. | Model (# Par.) | Data Aug. | # Hrs | Adapt. Feat. | LHUC SAT | WER% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | VL | L | M | H | Avg |
| 1 | Hybrid DNN (6M) | ✗ | 30.6 | ✗ | ✗ | 69.82 | 32.61 | 24.53 | 10.40 | 31.45 |
| 2 | | | | i-Vector | | 67.25 | 32.70 | 22.56 | 10.11 | 30.46 |
| 3 | | | | TBE | | 70.87 | 35.06 | 21.11 | 10.84 | 31.81 |
| 4 | | | | SBE | | 64.43 | 29.71 | 19.84 | 8.57 | **28.05**† |
| 5 | | | | SBE+TBE | | 64.54 | 29.13 | 18.90 | 8.69 | **27.83**† |
| 6 | | | | SBE⋆ | | 66.23 | 29.71 | 19.29 | 9.07 | 28.49† |
| 7 | | | | SBE+TBE⋆ | | 65.77 | 30.06 | 20.76 | 8.70 | 28.64† |
| 8 | | | | ✗ | ✓ | 64.39 | 29.88 | 20.27 | 8.95 | 28.29 |
| 9 | | | | SBE | | 63.40 | 28.90 | 18.64 | 8.13 | **27.24**‡ |
| 10 | | | | SBE+TBE | | **63.01** | 28.10 | 18.25 | 8.22 | **26.90**‡ |
| 11 | Hybrid DNN (6M) | DYS | 65.9 | ✗ | ✗ | 68.43 | 29.60 | 21.37 | 10.44 | 29.79 |
| 12 | | | | i-Vector | | 66.06 | 31.16 | 20.27 | 8.86 | 28.95 |
| 13 | | | | SBE | | 62.56 | 28.33 | 18.21 | 9.01 | **27.13**† |
| 14 | | | | SBE+TBE | | **61.40** | 28.93 | 18.74 | 9.00 | **27.14**† |
| 15 | | | | ✗ | ✓ | 60.99 | 28.20 | 18.86 | 8.41 | 26.69 |
| 16 | | | | SBE | | 60.98 | 27.29 | 17.96 | 8.54 | **26.32** |
| 17 | | | | SBE+TBE | | **60.23** | 27.87 | 18.27 | 8.67 | 26.41 |
| 18 | Hybrid DNN (6M) | DYS + CTL | 130.1 | ✗ | ✗ | 66.45 | 28.95 | 20.37 | 9.62 | 28.73 |
| 19 | | | | i-Vector | | 65.52 | 30.63 | 19.27 | 8.60 | 28.42 |
| 20 | | | | SBE | | 61.55 | 27.52 | 17.31 | 8.22 | **26.26**† |
| 21 | | | | SBE+TBE | | **61.24** | 27.77 | 17.45 | 8.31 | **26.32**† |
| 22 | | | | ✗ | ✓ | 62.50 | 27.26 | 18.41 | 8.04 | 26.55 |
| 23 | | | | SBE | | **59.83** | 27.16 | 16.80 | 7.91 | **25.60**‡ |
| 24 | | | | SBE+TBE | | 60.35 | 27.11 | 17.19 | 7.95 | **25.79**‡ |
| 25 | CTC (26M) | DYS + CTL | 130.1 | ✗ | ✗ | 78.25 | 53.18 | 46.56 | 34.09 | 50.79 |
| 26 | | | | SBE | | **68.16** | 47.08 | 40.68 | 32.40 | **45.37**† |
| 27 | | | | SBE+TBE | | 68.32 | 47.81 | 39.70 | 32.75 | **45.52**† |

Table 2 shows the WER performance comparison between our proposed spectro-temporal deep feature based speaker adaptation, i-Vector based adaptation and LHUC based adaptation. We follow [27] for data augmentation. Sys.1-10 are without data augmentation, Sys.11-17 are with partial data augmentation of perturbing dysarthic speech only and Sys. 18-27 are with full data augmentation of perturbing both dysarthric and control speech. Several trends can be observed[3]:

**1)** Our proposed spectro-temporal deep feature adapted systems consistently outperform the comparable baseline SI systems (Sys.4-5 *vs* Sys.1, Sys.13-14 *vs* Sys.11 and Sys.20-21 *vs* Sys.18) by up to 3.62% (Sys.5 *vs* Sys.1) absolute (11.5% relative) WER reduction. **2)** Our proposed spectro-temporal deep feature adapted systems consistently outperform the comparable baseline i-Vector adaptation (Sys.4-5 *vs* Sys.2, Sys.13-14 *vs* Sys.12 and Sys.20-21 *vs* Sys.19) by up to and 2.63% (Sys.5 *vs* Sys.2) absolute (8.6% relative) WER reduction. **3)** When further compounded with LHUC adaptation, our proposed spectro-temporal deep feature adapted systems consistently outperform the comparable baseline LHUC adaptation (Sys.9-10 *vs* Sys.8, Sys.16-17 *vs* Sys.15 and Sys.23-24 *vs* Sys.22) by up to 1.39% (Sys.10 *vs* Sys.8) absolute (4.9 % relative) WER reduction. **4)** Temporal basis embedding features contribute to adaptation of speakers with low intelligibility (Sys.10,14,17,21) especially when the amount of training data is small. **5)** Experiments on the CTC E2E system (Sys.25-27) suggests our proposed spectro-temporal deep feature adapted systems also reduced the WER by 5.42% (Sys.26 *vs* Sys.25) absolute (10.67% relative) over the comparable baseline unadapted system.

A comparison between previously published ASR systems on UASpeech and our system is shown in Table 3.

Table 3: *A comparison between published systems on UASpeech and our system. Here "DA" refers to data augmentation.*

| Systems | WER% |
|---|---|
| Sheffield-2013 Cross domain augmentation [11] | 37.50 |
| Sheffield-2015 Speaker adaptive training [12] | 34.80 |
| CUHK-2018 DNN System Combination [13] | 30.60 |
| Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation [39] | 30.76 |
| CUHK-2020 DNN + DA + LHUC SAT [27] | 26.37 |
| CUHK-2021 LAS + CTC + Meta-learning + SAT [16] | 35.00 |
| CUHK-2021 QuartzNet + CTC + Meta-learning + SAT [16] | 30.50 |
| **DA + SBE Adapt + LHUC SAT (Table 2, Sys.23)** | **25.60** |

## 5. Conclusions

This paper presents novel spectro-temporal deep features for both accurate speech intelligibility assessment and auxiliary feature based speaker adaptation. Experiments conduct on UASpeech suggest that our proposed spectro-temporal deep feature adaptation consistently outperform i-Vector adaptation. Future research will focus on improving spectral and temporal basis features extraction for variable length speech data and application to more advanced neural network systems.

## 6. Acknowledgement

## 7. References

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

[2] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, 2002.

[3] A. Graves, S. Fernández, F. Gomez *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

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

---

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

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

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

[8] S. Hu, X. Xie, S. Liu *et al.*, "Neural Architecture Search For LF-MMI Trained Time Delay Neural Networks," in *ICASSP*, 2021.

[9] S. Hu, X. Xie, S. Liu *et al.*, "Bayesian Learning of LF-MMI Trained Time Delay Neural Networks for Speech Recognition," *IEEE T AUDIO SPEECH*, vol. 29, 2021.

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

[11] H. Christensen, M. Aniol, P. Bell *et al.*, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *INTERSPEECH*, 2013.

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

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

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

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

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

[17] W. Lanier, *Speech disorders*. Greenhaven Publishing LLC, 2010.

[18] M. S. Hawley, P. Green, P. Enderby *et al.*, "Speech technology for e-inclusion of people with physical disabilities and disordered speech," in *EUROSPEECH*, 2005.

[19] T. J. Hixon and J. C. Hardy, "Restricted motility of the speech articulators in cerebral palsy," *J SPEECH HEAR DISORD*, vol. 29, no. 3, 1964.

[20] R. D. Kent, J. F. Kent, G. Weismer *et al.*, "What dysarthrias can tell us about the neural control of speech," *J PHONETICS*, vol. 28, no. 3, 2000.

[21] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *ICML*, vol. 117, 2013.

[22] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *ASRU*, 2013.

[23] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE T AUDIO SPEECH*, vol. 23, no. 9, 2015.

[24] T. Ko, V. Peddinti, D. Povey *et al.*, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[25] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition," in *INTERSPEECH*, 2018.

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

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

[28] 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.

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

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

[31] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *COMPUT SPEECH LANG*, 1998.

[32] T. Anastasakos, J. McDonough, R. Schwartz *et al.*, "A compact model for speaker-adaptive training," in *ICSLP*, 1996.

[33] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE T AUDIO SPEECH*, vol. 24, no. 8, 2016.

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

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

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

[37] M. Kim, Y. Kim, J. Yoo *et al.*, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE T NEUR SYS REH*, vol. 25, no. 9, 2017.

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

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

[40] A.-J. Van Der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace-based signal analysis using singular value decomposition," *P IEEE*, vol. 81, no. 9, 1993.

[41] K. M. Rosen, R. D. Kent, A. L. Delaney *et al.*, "Parametric Quantitative Acoustic Analysis of Conversation Produced by Speakers With Dysarthria and Healthy Speakers," *J SPEECH LANG HEAR R*, vol. 49, no. 2, 2006.

[42] A. Kacha, F. Grenez, J. R. Orozco-Arroyave *et al.*, "Principal component analysis of the spectrogram of the speech signal: interpretation and application to dysarthric speech," *COMPUT SPEECH LANG*, vol. 59, 2020.

[43] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based Learning for Automatic Dysarthric Speech Detection," *IEEE SIGNAL PROC LET*, 2020.

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

[45] G. An, D. G. Brizan, M. Ma *et al.*, "Automatic recognition of unified Parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *INTERSPEECH*, 2015.

[46] N. Garcia, J. C.-V. Correa, J. R. Orozco-Arroyave *et al.*, "Multimodal I-vectors to Detect and Evaluate Parkinson's Disease," in *INTERSPEECH*, 2018.

[47] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *ICASSP*, 2014.

[48] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *NEURAL COMPUT*, vol. 21, no. 3, 2009.

[49] T. Bocklet, S. Steidl, E. Nöth *et al.*, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues," in *INTERSPEECH*, 2013.

[50] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD*, 1994.

[51] R. Caruana, "Multitask learning," *MACH LEARN*, vol. 28, no. 1, 1997.

[52] S. Young, G. Evermann, M. Gales *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, 2006.

[53] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.