INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Factorised Speaker-environment Adaptive Training of Conformer Speech Recognition Systems

Jiajun Deng[1], *Guinan Li*[1], *Xurong Xie*[2], *Zengrui Jin*[1], *Mingyu Cui*[1], *Tianzi Wang*[1], *Shujie Hu*[1], *Mengzhe Geng*[1], *Xunying Liu*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Institute of Software, Chinese Academy of Sciences, Beijing, China

{jjdeng,gnli,zrjin,mycui,twang,sjhu,mzgeng,xyliu}@se.cuhk.edu.hk, xurong@iscas.ac.cn

Abstract

Rich sources of variability in natural speech present significant challenges to current data intensive speech recognition technologies. To model both speaker and environment level diversity, this paper proposes a novel Bayesian factorised speaker-environment adaptive training and test time adaptation approach for Conformer ASR models. Speaker and environment level characteristics are separately modeled using compact hidden output transforms, which are then linearly or hierarchically combined to represent any speaker-environment combination. Bayesian learning is further utilized to model the adaptation parameter uncertainty. Experiments on the 300-hr WHAM noise corrupted Switchboard data suggest that factorised adaptation consistently outperforms the baseline and speaker label only adapted Conformers by up to 3.1% absolute (10.4% relative) word error rate reductions. Further analysis shows the proposed method offers potential for rapid adaption to unseen speaker-environment conditions.

Index Terms: Speech recognition, Conformer, Factorised adaptation, Bayesian learning

1. Introduction

The majority of end-to-end (E2E) automatic speech recognition (ASR) systems [1], including those based on state-of-the-art Conformer models [2], are usually trained and evaluated on found speech data collected from a wide range of real-world scenarios. Such naturalistic speech data is generally highly non-homogeneous. Rich sources of variability are brought by multiple acoustic factors [3], for example, speaker characteristics, background noise and recording channel conditions. The resulting high degree of speech heterogeneity presents significant challenges to current data intensive speech recognition technologies in multiple stages. These include both the construction of speaker and environment independent ASR systems, and their fine-grained adaptation to individual users' voice recorded in diverse acoustic environments.

Prior researches in this direction to date have been largely spearheaded into two separate areas with their respective focuses on either speaker adaptation [4], or speech enhancement and environment compensation only [5]. In the first area, auxiliary speaker-aware features that are based on i-vector [6–8], x-vector [8, 9], feature-space maximum likelihood linear regression (f-MLLR) [9, 10], or extracted from speaker-aware modules [11–13] are incorporated into various ASR models. Model-based speaker adaptation methods estimate speaker-dependent (SD) parameters, which are implemented as either internal DNN components [14, 15], or additional parameters such as learning hidden unit contributions (LHUC) [16–18], using the target speaker data during speaker adaptive training and test time adaptation [19, 20]. In the second area, single-channel based environment compensation [21–25] or multi-channel based speech enhancement front-ends [26–31] are separately constructed and optionally further integrated with the recognition back-end. Back-end model adaptation methods that aim to compensate for the modelling mismatch against the unseen target environment have also been studied [32, 33].

A simple approach to handle the multifaceted data heterogeneity in natural speech is to separately model each user's voice recorded in diverse environments as different speakers. However, this fails to account for the homogeneity over speaker-level characteristics, leading to fragmentation of data and poor generalization to unseen speaker-environment combinations.

An alternative and more general solution to such a problem is to structurally represent different factors of variability in ASR systems [34–38]. For example, during the adaptive training stage [19, 39], speaker and environment characteristics are "factored out" into their respective separately designed modelling components (e.g., vector Taylor series [40], MLLR or CMLLR [10, 41], or LHUC [16] transforms), thus the backbone ASR model can focus more on learning speaker and environment invariant speech representations and their mapping to spoken contents. During the test time adaptation stage, these sources of variabilities can be flexibly "factored in" to model any seen or unseen speaker-environment combination. Prior researches in this direction were mainly conducted for conventional GMM-HMM [34–36, 42, 43] and hybrid DNN-HMM [37, 38] ASR systems. In contrast, existing researches on E2E ASR systems represented by Conformer largely focus on modelling only one source of variability, for example, speaker characteristics [7, 8, 18, 44], or environmental mismatch [31, 45–47].

To this end, a novel factorised speaker-environment adaptive training approach is proposed in this paper to facilitate both adaptive training and test time unsupervised adaptation of E2E Conformer models. Speaker and environment level characteristics are separately modelled using compact LHUC [16] or hidden unit bias (HUB) [18, 48] transformations. These are linearly or hierarchically combined to represent any speaker-environment combination, observed in the training data or otherwise. Bayesian estimation of the speaker or environment factor specific transforms is also utilized to mitigate the risk of overfitting during test time unsupervised adaptation to the limited speaker or environment data. The acquired speaker and environment homogeneity can be exploited for rapid adaptation to the unseen speaker-environment combination. For example, speaker specific transforms estimated in one environment can be cached and reused in another environment. The main contributions of the paper are summarized below:

1) To the best of our knowledge, this paper presents the first work to investigate the model-based factorised adaptation

10.21437/Interspeech.2023-583

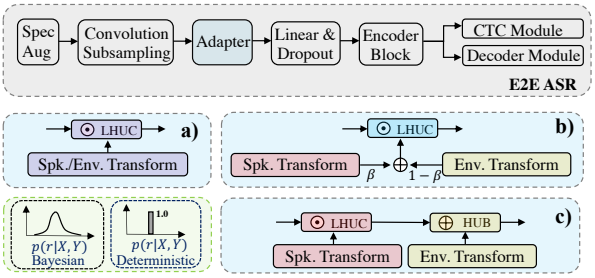

Figure 1: *Examples of Conformer E2E ASR models (grey box, top), together with three model-based adaptation methods:* ***a)*** *Conformer speaker adaptation using LHUC transforms;* ***b)*** *linear (superposition) factorised adaptation; and* ***c)*** *cascaded factorised adaptation. Bayesian and deterministic estimations of adaptation parameters are shown in the green box (bottom left).*

for E2E Conformer models by structurally representing speaker and environment factors of variability. In contrast, prior researches on E2E ASR systems largely focus on modelling only one source of variability [7, 8, 12, 15, 18, 31, 44–47].

2) The efficacy of the proposed Bayesian factorised adaptation approaches is consistently demonstrated on the 300-hr WHAM noise corrupted Switchboard task. Experimental results suggest that our approach consistently outperforms the unadapted baseline and speaker label only adapted Conformer systems by up to 3.1%, 2.7% and 2.9% absolute (10.4%, 8.1%, and 8.2% relative) word error rate (WER) reductions on the noise corrupted Hub5'00, RT02, and RT03 test sets respectively, before and after external language model rescoring is applied.

3) Further analysis shows that the proposed method offers the potential for rapid adaptation to the unseen speaker-environment combination by flexibly "factoring in" the already estimated speaker and environment specific transforms. Furthermore, their generic nature and the implementation details described in this paper allow their further application to handle two or more sources of variability in other E2E ASR tasks.

2. Conformer E2E ASR System

The Conformer [2] ASR model consists of an encoder module and a decoder module, which are both based on multi-blocked stacked architectures. The encoder module comprises a convolutional subsampling module, a linear layer with dropout operation, and stacked encoder blocks. Layer normalization and residual connections are performed on all encoder blocks. More details of Conformer components can be found in [49]. Fig. 1 shows an example of Conformer E2E ASR system.

For training the Conformer model, the following multi-task criterion interpolation between connectionist temporal classification (CTC) and attention error cost is adopted [50].

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc}, \tag{1}$$

where $\lambda \in [0, 1]$ is a tunable hyper-parameter and empirically set as 0.2 for training and 0.3 for recognition in this paper.

3. Conformer Speaker Adaptation

The key idea of LHUC adaptation [16, 18] is to use the SD scaling vector to modify the amplitudes of activation outputs. Let $\boldsymbol{r}^{l,s}$ denote the SD parameters for speaker s in the l-th hidden layer, the speaker adapted hidden outputs can be given by

$$\boldsymbol{h}^{l,s} = \boldsymbol{h}^{l} \odot \xi(\boldsymbol{r}^{l,s}), \tag{2}$$

where $\boldsymbol{h}^{l}$ is the hidden activation outputs in the l-th hidden layer, $\odot$ is the Hadamard product operation, and $\xi(\cdot)$ is the element-wise $2\times$ Sigmoid$(\cdot)$ function.

Alternatively the SD transform that is added to the hidden output as a bias vector [18, 48] leads to the HUB adaptation. Let $\zeta(\boldsymbol{r}^{l,s})$ denote the bias vector for speaker s in the l-th hidden layer. The speaker HUB adapted hidden outputs are given as

$$\boldsymbol{h}^{l,s} = \boldsymbol{h}^{l} + \zeta(\boldsymbol{r}^{l,s}), \tag{3}$$

where $\zeta(\cdot)$ is the identity activation function.

LHUC and HUB can be further used for environment adaptation by learning an environment specific LHUC or HUB transform, and a single joint speaker-environment transform to model a particular combination of these two factors.

4. Factorised Conformer Speaker-environment Adaptation

4.1. Linear Factorised Adaptation

Linear factorised adaptation (LFA) models the two acoustic factors using a linear interpolation between a speaker-dependent (SD) transform and an environment-dependent (ED) transform, as is shown in Fig. 1(b). Let $\boldsymbol{n}^{l,e}$ denote the ED parameters for environment e in the l-th hidden layer. The factorised adapted hidden outputs for speaker s in environment e can be derived by

$$\boldsymbol{h}^{l,s,e} = \boldsymbol{h}^{l} \odot (\beta\xi(\boldsymbol{r}^{l,s}) + (1-\beta)\xi(\boldsymbol{n}^{l,e})), \tag{4}$$

where $\beta \in [0, 1]$ is a hyper-parameter that balances the weighting between the speaker and environment factors. For example, $\beta = 1$ and $\beta = 0$ lead to the LHUC speaker only adaptation and environment only adaptation, respectively.

4.2. Cascaded Factorised Adaptation

In the cascaded factorised adaptation (CFA), the SD transform and ED transform, which serve as either an LHUC scaling vector or a HUB bias vector, are cascaded into the Conformer hidden layers. This leads to the following four cases: 1) both transforms are LHUC scaling vectors; 2) both transforms are HUB bias vectors; 3) the SD transform is an LHUC scaling vector while the ED transform is a HUB bias vector; and 4) the SD transform is a HUB bias vector while the ED transform is an LHUC scaling vector. For example, without loss of generality, assuming that both transforms are applied at the same layer, the above third case is shown in Fig. 1(c) and the factorised adapted hidden outputs can be given by

$$\boldsymbol{h}^{l,s,e} = \boldsymbol{h}^{l} \odot \xi(\boldsymbol{r}^{l,s}) + \zeta(\boldsymbol{n}^{l,e}). \tag{5}$$

4.3. Estimation of Factorised Adaptation Parameters

Let $\mathcal{D}^{s,e} = \{\boldsymbol{X}^{s,e}, \boldsymbol{Y}^{s,e}\}$ denote the data set for speaker s in the environment e, where $\boldsymbol{X}^{s,e}$ and $\boldsymbol{Y}^{s,e}$ are the acoustic features and the corresponding supervision token sequences, respectively. During unsupervised test time adaptation, the supervision $\boldsymbol{Y}^{s,e}$ of unseen test data need to be generated by initially decoding the corresponding utterances using an unadapted baseline Conformer model, before serving as the target token labels in the subsequent adaptation. The SD and ED parameters can be estimated by using the loss in Eqn. (1),

$$\{\hat{\boldsymbol{r}}^{s}, \hat{\boldsymbol{n}}^{e}\} = \underset{\{\boldsymbol{r}^{s}, \boldsymbol{n}^{e}\}}{\arg\min}\{\mathcal{L}(\bar{\mathcal{D}}^{s,e}; \boldsymbol{r}^{s}, \boldsymbol{n}^{e})\}, \tag{6}$$

where $\bar{\mathcal{D}}^{s,e}$ is the union of all speaker's adaptation data in a given environment e, $\cup_{i\in\mathcal{S}}\mathcal{D}^{i,e}$, and all environment's adaptation data associated with a speaker s, $\cup_{i\in\mathcal{E}}\mathcal{D}^{s,i}$.

During adaptive training, the SD and ED parameters associated with the training data are jointly optimized with the "canonical" model parameters $\boldsymbol{\Theta}$ that are independent of speaker or environment characteristics. This is given as

$$\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\theta}}_E\} = \underset{\{\boldsymbol{\Theta}, \boldsymbol{\theta}_S, \boldsymbol{\theta}_E\}}{\arg\min} \sum_{s \in \mathcal{S}} \sum_{e \in \mathcal{E}} \mathcal{L}(\mathcal{D}^{s,e}; \boldsymbol{\Theta}, \boldsymbol{\theta}_S, \boldsymbol{\theta}_E), \quad (7)$$

where $\boldsymbol{\theta}_S = \{\boldsymbol{r}^s\}_{s \in \mathcal{S}}$ and $\boldsymbol{\theta}_E = \{\boldsymbol{n}^e\}_{e \in \mathcal{E}}$ are the SD and ED parameter sets associated with training data, respectively.

4.4. Bayesian Learning of Factorised Adaptation

Bayesian learning [51] is adopted to model adaptation parameter uncertainty. Given limited adaptation data $\bar{\mathcal{D}}^{s,e}$, the Bayesian predictive distribution for a test utterance $\tilde{\boldsymbol{X}}^{s,e}$ is $\int\int p(\tilde{\boldsymbol{Y}}^{s,e}|\tilde{\boldsymbol{X}}^{s,e}, \boldsymbol{r}^s, \boldsymbol{n}^e) p(\boldsymbol{r}^s, \boldsymbol{n}^e|\bar{\mathcal{D}}^{s,e}) d\boldsymbol{r}^s d\boldsymbol{n}^e$, where $\hat{\boldsymbol{Y}}^{s,e}$ is the predicted token sequence and $p(\boldsymbol{r}^s, \boldsymbol{n}^e|\bar{\mathcal{D}}^{s,e})$ is the joint posterior distribution of the SD and ED parameters learned from the adaptation data. Using variational inference, a variational distribution $q(\boldsymbol{r}^s, \boldsymbol{n}^e)$ is used to approximate the joint posterior distribution $p(\boldsymbol{r}^s, \boldsymbol{n}^e|\bar{\mathcal{D}}^{s,e})$, and inferred by optimizing the hybrid attention plus CTC loss marginalization $\mathcal{L}(\bar{\mathcal{D}}^{s,e}) = (\lambda - 1) \log \int\int p_a(\boldsymbol{r}^s, \boldsymbol{n}^e|\bar{\mathcal{D}}^{s,e}) d\boldsymbol{r}^s d\boldsymbol{n}^e - \lambda \log \int\int p_c(\boldsymbol{r}^s, \boldsymbol{n}^e|\bar{\mathcal{D}}^{s,e}) d\boldsymbol{r}^s d\boldsymbol{n}^e$ over the uncertain parameters, $\boldsymbol{r}^s, \boldsymbol{n}^e$. The variational bound is given by

$$\begin{aligned}
&\mathcal{L}(\bar{\mathcal{D}}^{s,e}) \leq \iint q(\boldsymbol{r}^s, \boldsymbol{n}^e)\{(\lambda - 1) \log p_a(\bar{\mathcal{D}}^{s,e}|\boldsymbol{r}^s, \boldsymbol{n}^e) - \\
&\lambda \log p_c(\bar{\mathcal{D}}^{s,e}|\boldsymbol{r}^s, \boldsymbol{n}^e)\} d\boldsymbol{r}^s d\boldsymbol{n}^e + \mathrm{KL}(q(\boldsymbol{r}^s, \boldsymbol{n}^e)||p(\boldsymbol{r}^s, \boldsymbol{n}^e)) \\
&\triangleq \mathcal{L}_{int}(\bar{\mathcal{D}}^{s,e}; \boldsymbol{r}^s, \boldsymbol{n}^e) + \mathcal{L}_{KL}, \quad (8)
\end{aligned}$$

where p_a and p_c are the attention and CTC based sequence probabilities respectively, $p(\boldsymbol{r}^s, \boldsymbol{n}^e)$ is the joint prior distribution of the SD and ED parameters. $\mathrm{KL}(\cdot)$ is the KL divergence. Since the SD and ED latent variables $\{\boldsymbol{r}^s, \boldsymbol{n}^e\}$ are independent of each other, the joint variational and prior distributions can be modeled independently. The structured variational distributions $\{q(\boldsymbol{r}^s) = \mathcal{N}(\boldsymbol{\mu}_r^s, \boldsymbol{\sigma}_r^s), q(\boldsymbol{n}^e) = \mathcal{N}(\boldsymbol{\mu}_n^e, \boldsymbol{\sigma}_n^e)\}$ and the prior distributions $\{p(\boldsymbol{r}^s) = \mathcal{N}(\bar{\boldsymbol{\mu}}_r, \bar{\boldsymbol{\sigma}}_r), p(\boldsymbol{n}^e) = \mathcal{N}(\bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\sigma}}_n)\}$ are assumed to be standard normal distributions. Then the KL divergence term $\mathcal{L}_{KL}$ can be computed as closed form [18]. To ensure that the loss $\mathcal{L}_{int}$ is differentiable, the Monte Carlo sampling method is used to approximate it, which is given by

$$\mathcal{L}_{bayes} \approx \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{int}(\bar{\mathcal{D}}^{s,e}; \boldsymbol{r}_k^s, \boldsymbol{n}_k^e) + \mathcal{L}_{KL}, \quad (9)$$

where $\boldsymbol{r}_k^s = \boldsymbol{\mu}_r^s + \boldsymbol{\sigma}_r^s \odot \boldsymbol{\epsilon}_k^s$, $\boldsymbol{n}_k^e = \boldsymbol{\mu}_n^e + \boldsymbol{\sigma}_n^e \odot \boldsymbol{\epsilon}_k^e$, $\boldsymbol{\epsilon}_k^s$ and $\boldsymbol{\epsilon}_k^e$ are the k-th sample drawn from standard normal distributions. In this paper, $\mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\mathcal{N}(\mathbf{0}, \mathbf{0.001})$ are empirically selected as the priors for LHUC and HUB parameters respectively. The location of speaker and environment transforms is empirically selected and fixed at the convolution subsampling module. During adaptation, only one sample is drawn in Eqn. (9). The Bayesian predictive inference integral is efficiently approximated by the expectation of the posterior distribution as $p(\tilde{\boldsymbol{Y}}^{s,e}|\tilde{\boldsymbol{X}}^{s,e}, \boldsymbol{\mu}_r^s, \boldsymbol{\mu}_n^e)$ during recognition.

5. Experiments

5.1. Experimental Setup

The widely used 300-hr Switchboard-1 conversational telephone corpus (LDC97S62) [53] containing 4804 speakers is utilized for training. The NIST 3.8-hr Hub5'00 (LDC2002S09, LDC2002T43), 6.4-hr RT02 (LDC2004S11), and 6.2-hr RT03

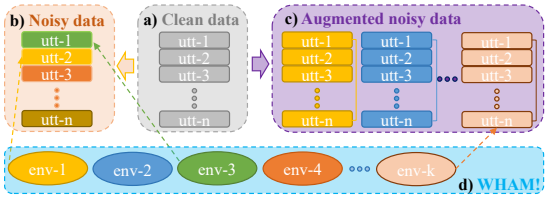

Figure 2: *Noisy data simulation: a) Original Switchboard clean data, b) Non-augmented noise simulation, and c) Augmented noise simulation of Sec. 5.1. d) Noise types in WHAM.*

(LDC2007S10) test sets containing 80, 120, and 144 speakers respectively are adopted for performance evaluation. The publicly available noise WHAM database [54] which is recorded in non-stationary ambient environments such as restaurants, coffee shops, bars, parks, and office buildings is used as the noise source. Two protocols to simulate noise corrupted data are: **1) Non-augmented noise simulation** whereby each utterance is randomly exposed to one of multiple environments with a uniform distribution, as is shown in Fig. 2(b). **2) Augmented noise simulation** whereby each utterance is exposed to all different environments independently, as is shown in Fig. 2(c). For the noise corrupted training data, the 300-hr Switchboard-1 data is mixed with ten types of noise at signal-to-noise ratios (SNR) uniformly sampled from $\{-5, 0, 5, 10, 20\}$dB by the non-augmented simulation. For the noise corrupted evaluation sets, the test data is mixed with ten types of noise at SNRs uniformly sampled from $\{-15, -10, -5, 0, 5, 10, 20\}$dB. Three additive noise types used for the evaluation sets are also used in the training data simulation. Non-augmented noise simulation is applied to all three NIST Hub5'00, RT02 and RT03 sets for the first experiment presented in Table 1. To further analyse the improvements from Bayesian factorised adaptation, the experiments in Table 2 used 38-hr noise corrupted data derived by applying augmented noise simulation to the Hub5'00 set.

The ESPnet recipe [2] configured Conformer model comprised 12 encoder and 6 decoder layers, each with 256-dim 4-head attention and 2048 feed-forward hidden nodes. 80-dim Mel-filter bank plus 3-dim pitch parameters were used as input features, with byte-pair-encoding (BPE) tokens of size 2000 serving as decoder outputs. Two 2-D convolutional layers with stride 2 were included in the convolution subsampling module. SpecAugment [55] was used for Conformer training. The initial learning rate of the Noam optimizer was 5.0. The dropout rate was set to 0.1, and the recognition model was averaged over the last ten epochs. The log-linearly interpolated external Transformer and Bi-LSTM language models (LMs), which were trained on the Switchboard and Fisher transcripts using cross-utterance contexts [56], were used for LM rescoring.

5.2. Experimental Results and Analysis

Performance of Bayesian factorised adaptation evaluated on the **non-augmented** noise corrupted test sets are shown in Table 1. Several trends can be observed. **a)** Both the proposed linear factorised adaptation (LFA) (sys.8-11) and cascaded factorised adaptation (CFA) (sys.12-15) consistently outperformed the unadapted baseline (sys.1) and the adapted baselines (sys.2-4) considering only speaker or environment variability across all three test sets. The best operating point for LFA is $\beta = 0.7$ (sys.10), while the best adaptation configuration of CFA is the "HUB-HUB" combination (sys.13). **b)** When Bayesian learning was further used to model LHUC and HUB parameters uncertainty, additional WER reductions of up to 0.8% absolute (Hub5'00, sys.18 vs. sys.10) were consistently obtained using Bayesian factorised adaptations (sys.18,19) over that of the

Table 1: *Performance (WER%) of adapted Conformer systems with/without Bayesian learning evaluated on the noise corrupted, non-augmented Hub5'00, RT02 and RT03 sets, before and after external Transformer plus LSTM LM rescoring. "CHE", "SWBD", "FSH" and "O.V." stand for "CallHome", "Switchboard", "Fisher" and "Overall" respectively. † and ∗ denote a statistically significant (MAPSSWE, α=0.05) WER difference [52] obtained over the baseline (sys. 1, 20) and the speaker adapted systems (sys. 3, 16, 21) respectively. The SNR and noise type combinations that appear in the train set are defined as "Seen" data, otherwise "Unseen" data.*

ID	Method	Adaptation Speaker	Adaptation Env.	Adapt. Param.	Language Model	Hub5'00 CHE	Hub5'00 SWBD	Hub5'00 O.V.	RT02 SWBD1	RT02 SWBD2	RT02 SWBD3	RT02 O.V.	RT03 FSH	RT03 SWBD	RT03 O.V.	ALL Seen	ALL Unseen	ALL O.V.
1	Baseline	✗	✗	-		36.4	24.8	30.6	28.2	34.8	38.7	34.3	33.3	38.8	36.1	22.9	37.3	34.2
2		LHUC	✗			35.0†	24.2†	29.6†	27.1†	33.8†	37.9†	33.3†	31.9†	38.5	35.3†	22.5†	36.3†	33.3†
3		HUB	✗			35.2†	23.8†	29.5†	27.2†	33.8†	37.7†	33.3†	32.2†	38.2†	35.3†	22.4†	36.3†	33.3†
4	Single Transform	✗	LHUC			35.9†	23.8†	29.9†	27.2†	34.2†	38.5	33.7†	32.5†	38.1†	35.4†	22.3†	36.6†	33.5†
5		✗	HUB			36.1	24.0†	30.1†	27.6†	34.2†	38.4	33.8†	32.9†	38.1†	35.6†	22.4†	36.8†	33.7†
6		Joint LHUC				36.3	24.6	30.5	28.1	34.6	38.7	34.2	33.3	39.7	36.6	23.0	37.5	34.4
7		Joint HUB				36.3	24.7	30.5	28.1	34.9	39.0	34.4	33.1	39.7	36.5	23.1	37.5	34.4
8	Linear Factorised Adaptation (LFA)	LHUC ($\beta = 0.3$)		Deterministic	✗	34.6$^{\dagger *}$	23.2$^{\dagger *}$	28.9$^{\dagger *}$	**26.2**$^{\dagger *}$	33.5†	36.8$^{\dagger *}$	32.5$^{\dagger *}$	31.9†	37.6$^{\dagger *}$	34.9$^{\dagger *}$	21.7$^{\dagger *}$	35.6$^{\dagger *}$	32.6$^{\dagger *}$
9		LHUC ($\beta = 0.5$)				34.5$^{\dagger *}$	23.0$^{\dagger *}$	28.8$^{\dagger *}$	26.3$^{\dagger *}$	33.2$^{\dagger *}$	36.9$^{\dagger *}$	32.5$^{\dagger *}$	31.3$^{\dagger *}$	37.3$^{\dagger *}$	34.4$^{\dagger *}$	21.8$^{\dagger *}$	35.4$^{\dagger *}$	32.5$^{\dagger *}$
10		LHUC ($\beta = 0.7$)				34.4$^{\dagger *}$	**22.8**$^{\dagger *}$	**28.6**$^{\dagger *}$	**26.2**$^{\dagger *}$	**33.1**$^{\dagger *}$	**36.5**$^{\dagger *}$	**32.3**†	**31.2**$^{\dagger *}$	**37.2**$^{\dagger *}$	**34.3**$^{\dagger *}$	**21.4**$^{\dagger *}$	**35.3**$^{\dagger *}$	**32.3**$^{\dagger *}$
11		LHUC ($\beta = 0.9$)				**34.3**$^{\dagger *}$	23.1$^{\dagger *}$	28.7$^{\dagger *}$	26.9†	**33.1**$^{\dagger *}$	36.8$^{\dagger *}$	32.6$^{\dagger *}$	**31.2**$^{\dagger *}$	37.6$^{\dagger *}$	34.5$^{\dagger *}$	21.7$^{\dagger *}$	35.5$^{\dagger *}$	32.5$^{\dagger *}$
12	Cascaded Factorised Adaptation (CFA)	LHUC	LHUC			34.5$^{\dagger *}$	23.6†	29.1$^{\dagger *}$	**26.5**$^{\dagger *}$	33.3$^{\dagger *}$	37.3$^{\dagger *}$	32.8$^{\dagger *}$	31.4$^{\dagger *}$	37.6$^{\dagger *}$	34.6$^{\dagger *}$	21.7$^{\dagger *}$	35.7$^{\dagger *}$	32.7$^{\dagger *}$
13		HUB	HUB			**33.7**$^{\dagger *}$	**23.0**$^{\dagger *}$	**28.4**$^{\dagger *}$	26.7$^{\dagger *}$	**33.0**$^{\dagger *}$	**36.2**$^{\dagger *}$	**32.3**$^{\dagger *}$	31.1$^{\dagger *}$	**36.7**$^{\dagger *}$	**34.0**$^{\dagger *}$	**21.2**$^{\dagger *}$	**35.1**$^{\dagger *}$	**32.1**$^{\dagger *}$
14		LHUC	HUB			34.3$^{\dagger *}$	23.1$^{\dagger *}$	28.7$^{\dagger *}$	26.8$^{\dagger *}$	33.4$^{\dagger *}$	37.0$^{\dagger *}$	32.7$^{\dagger *}$	**30.9**$^{\dagger *}$	37.3$^{\dagger *}$	34.2$^{\dagger *}$	**21.2**$^{\dagger *}$	35.5$^{\dagger *}$	32.4$^{\dagger *}$
15		HUB	LHUC			33.8$^{\dagger *}$	23.2$^{\dagger *}$	28.5$^{\dagger *}$	26.6$^{\dagger *}$	33.6†	37.0$^{\dagger *}$	32.8$^{\dagger *}$	31.2$^{\dagger *}$	37.5$^{\dagger *}$	34.4$^{\dagger *}$	21.5$^{\dagger *}$	35.5$^{\dagger *}$	32.5$^{\dagger *}$
16	Single	HUB	✗	Bayesian	✗	34.4†	23.1†	28.8†	27.0†	33.3†	37.1†	32.8†	31.3†	37.6†	34.6†	21.8†	35.6†	32.6†
17		Joint LHUC				35.0†	24.3†	29.7†	28.1	33.7†	38.1†	33.6†	32.2†	38.3†	35.4†	22.2†	36.5†	33.4†
18	LFA	LHUC ($\beta = 0.7$)				33.1$^{\dagger *}$	**22.4**$^{\dagger *}$	**27.8**$^{\dagger *}$	**25.9**$^{\dagger *}$	32.8$^{\dagger *}$	36.1$^{\dagger *}$	32.0$^{\dagger *}$	**30.4**$^{\dagger *}$	36.9$^{\dagger *}$	33.8$^{\dagger *}$	20.9$^{\dagger *}$	34.8$^{\dagger *}$	31.8$^{\dagger *}$
19	CFA	HUB	HUB			**33.0**$^{\dagger *}$	22.6$^{\dagger *}$	**27.8**$^{\dagger *}$	26.1$^{\dagger *}$	**32.0**$^{\dagger *}$	**35.7**$^{\dagger *}$	**31.6**$^{\dagger *}$	30.6$^{\dagger *}$	**36.6**$^{\dagger *}$	**33.7**$^{\dagger *}$	**20.7**$^{\dagger *}$	**34.6**$^{\dagger *}$	**31.6**$^{\dagger *}$
20	Baseline	✗	✗	-		35.9	23.8	29.9	27.3	33.9	37.3	33.2	32.5	38.0	35.3	21.9	36.4	33.3
21	Single	HUB	✗	Bayesian	✓	34.0†	22.4†	28.2†	26.2†	32.5†	35.9†	31.9†	30.3†	36.8†	33.7†	20.7†	34.8†	31.8†
22		Joint LHUC				34.5†	23.5	29.0†	26.8†	33.0†	37.2	32.7†	31.5†	37.5†	34.6†	20.9†	35.8†	32.7†
23	LFA	LHUC ($\beta = 0.7$)				**31.6**$^{\dagger *}$	21.8$^{\dagger *}$	**26.8**$^{\dagger *}$	25.4$^{\dagger *}$	31.6$^{\dagger *}$	34.8$^{\dagger *}$	30.9$^{\dagger *}$	29.7$^{\dagger *}$	**35.3**$^{\dagger *}$	32.6$^{\dagger *}$	19.7$^{\dagger *}$	33.6$^{\dagger *}$	30.6$^{\dagger *}$
24	CFA	HUB	HUB			32.0$^{\dagger *}$	**21.5**$^{\dagger *}$	**26.8**$^{\dagger *}$	**24.9**$^{\dagger *}$	**31.5**$^{\dagger *}$	**34.1**$^{\dagger *}$	**30.5**$^{\dagger *}$	**29.4**$^{\dagger *}$	**35.3**$^{\dagger *}$	**32.4**$^{\dagger *}$	**19.4**$^{\dagger *}$	**33.3**$^{\dagger *}$	**30.3**$^{\dagger *}$

comparable non-Bayesian adapted systems (sys.10,13). **c)** Consistent WER reductions were retained after external LM rescoring. Overall statistically significant WER reductions of **3.1%**, **2.7%**, **2.9%** absolute (**10.4%**, **8.1%**, and **8.2%** relative) were obtained by the proposed Bayesian CFA (sys.24) over the baseline Conformer (sys.20) on the noise corrupted Hub5'00, RT02 and RT03 test sets respectively. **d)** Joint speaker-environment adaptation (sys.6,7) using a single transform performed less well due to the lack of factorization between speaker and environment, and fragmentation of adaptation data.

Table 2: *Performance (WER%) of adapted Conformer systems evaluated on the 38-hr noise corrupted and augmented Hub5'00 sets. † and ∗ denote statistically significant WER differences [52] (MAPSSWE, α=0.05) over the baselines (sys. 1, 20) and joint speaker-environment adaptation (sys. 7, 17, 22).*

ID	Method	Adaptation Speaker	Adaptation Env.	Adapt. Param.	LM	38-hr Augmented Hub5'00 CHE	SWBD	Seen	Unseen	O.V.
1	Baseline	✗	✗	-		36.6	24.1	29.3	30.7	30.4
2		LHUC	✗			35.4†	24.0	28.8†	30.0†	29.7†
3		HUB	✗			35.4†	23.4†	28.4†	29.7†	29.4†
4	Single Transform	✗	LHUC			36.4	23.8†	29.1†	30.4†	30.1†
5		✗	HUB			35.9†	23.8†	28.8†	30.3†	29.9†
6		Joint LHUC				35.1†	23.3†	28.3†	29.5†	29.2†
7		Joint HUB				34.9†	23.2†	28.0†	29.3†	29.0†
8	LFA	LHUC ($\beta = 0.3$)		Deterministic	✗	34.3$^{\dagger *}$	22.8$^{\dagger *}$	27.7$^{\dagger *}$	28.8$^{\dagger *}$	28.6$^{\dagger *}$
9		LHUC ($\beta = 0.5$)				34.1$^{\dagger *}$	22.4$^{\dagger *}$	27.2$^{\dagger *}$	28.6$^{\dagger *}$	28.3$^{\dagger *}$
10		LHUC ($\beta = 0.7$)				34.1$^{\dagger *}$	**22.1**$^{\dagger *}$	**27.0**$^{\dagger *}$	**28.4**$^{\dagger *}$	**28.1**$^{\dagger *}$
11		LHUC ($\beta = 0.9$)				**34.0**$^{\dagger *}$	22.4$^{\dagger *}$	27.3$^{\dagger *}$	28.5$^{\dagger *}$	28.3$^{\dagger *}$
12	CFA	LHUC	LHUC			34.3$^{\dagger *}$	23.0†	27.6$^{\dagger *}$	29.0$^{\dagger *}$	28.7$^{\dagger *}$
13		HUB	HUB			**33.8**$^{\dagger *}$	**22.5**$^{\dagger *}$	**27.2**$^{\dagger *}$	**28.4**$^{\dagger *}$	**28.2**$^{\dagger *}$
14		LHUC	HUB			34.0$^{\dagger *}$	22.9$^{\dagger *}$	27.5$^{\dagger *}$	28.8$^{\dagger *}$	28.5$^{\dagger *}$
15		HUB	LHUC			34.0$^{\dagger *}$	23.1†	27.6$^{\dagger *}$	28.8$^{\dagger *}$	28.6$^{\dagger *}$
16	Single	HUB	✗	Bayes	✗	34.9†	23.2†	28.1†	29.3†	29.0†
17		Joint HUB				34.5†	22.8†	27.6†	28.9†	28.6†
18	LFA	LHUC ($\beta = 0.7$)				33.4$^{\dagger *}$	**21.9**$^{\dagger *}$	**26.6**$^{\dagger *}$	**28.0**$^{\dagger *}$	**27.7**$^{\dagger *}$
19	CFA	HUB	HUB			**33.3**$^{\dagger *}$	22.3$^{\dagger *}$	27.0$^{\dagger *}$	**28.0**$^{\dagger *}$	27.8$^{\dagger *}$
20	Baseline	✗	✗	-		36.0	23.3	28.5	30.1	29.7
21	Single	HUB	✗	Bayes	✓	34.4†	22.4†	27.6†	28.5†	28.4†
22		Joint HUB				33.7†	21.8†	26.5†	28.1†	27.7†
23	LFA	LHUC ($\beta = 0.7$)				32.4$^{\dagger *}$	**21.0**$^{\dagger *}$	**25.6**$^{\dagger *}$	27.0$^{\dagger *}$	**26.7**$^{\dagger *}$
24	CFA	HUB	HUB			**32.3**$^{\dagger *}$	21.1$^{\dagger *}$	26.0$^{\dagger *}$	**26.9**$^{\dagger *}$	**26.7**$^{\dagger *}$

Performance of Bayesian factorised adaptation evaluated on the **augmented** noise corrupted Hub5'00 set (38-hr) are shown in Table 2. The trends found in Table 1 were still retained. Overall absolute WER reductions of **3.0%** and **1.0%** were obtained by the Bayesian CFA (sys.24) over the baseline (sys.20) and the joint speaker-environment adapted (sys.22) systems.

Potential for rapid adaptation: In the experiments of Table 3 where either the speaker transforms are estimated using speaker level data in mismatched environments (sys.5), or the environment transforms are learned using environment level data with mismatched speakers (sys.6), or both being mismatched against the test data being adapted to (sys.7,8), factorised speaker-environment adaptation consistently produced absolute WER reductions of **1.1%-2.5%** over the baseline un-adapted Conformer (sys.1). In particular, the CFA factorised adaptation with both speaker and environment mismatches produced performance comparable to speaker only adaptation using the matched environment (sys.7 vs. sys.2). These results suggest that the proposed flexible factorization framework allows the separately acquired speaker and environment homogeneity by factorization to be exploited for rapid adaptation to unseen speaker-environment combinations.

Table 3: *Performance (WER%) of Bayesian factorised adaptation using matched or mismatched transforms evaluated on the 3.8-hr subset of 38-hr noise corrupted Hub5'00 set. Five mismatched conditions are randomly selected for each utterance.*

ID	Method	Speaker Transform	Env. Transform	Hub5'00 ($mean \pm std$) CHE	SWBD	O.V.
1	Baseline	-	-	37.7	24.2	31.0
2	HUB	Matched Env.	-	35.9	23.4	29.7
3	(Spk. adapt)	Mismatched Env.	-	37.0$_{\pm 0.27}$	24.3$_{\pm 0.31}$	30.7$_{\pm 0.23}$
4		Matched Env.	Matched Spk.	34.0	22.1	28.1
5	CFA	Mismatched Env.	Matched Spk.	34.4$_{\pm 0.22}$	22.4$_{\pm 0.25}$	28.5$_{\pm 0.15}$
6	(HUB-HUB)	Matched Env.	Mismatched Spk.	34.9$_{\pm 0.20}$	22.8$_{\pm 0.15}$	28.9$_{\pm 0.19}$
7		Mismatched Env.	Mismatched Spk.	35.8$_{\pm 0.23}$	23.4$_{\pm 0.26}$	29.6$_{\pm 0.20}$
8	LFA ($\beta = 0.7$)	Mismatched Env.	Mismatched Spk.	36.3$_{\pm 0.26}$	23.6$_{\pm 0.32}$	29.9$_{\pm 0.22}$

6. Conclusions

The paper proposed a novel Bayesian factorised speaker-environment adaptive training and test time unsupervised adaptation approach for Conformer models. Compact transformations were used to model speaker and environment level characteristics separately, which were linearly or hierarchically combined to represent any seen or unseen speaker-environment combination. Bayesian learning was further utilized to model the adaptation parameter uncertainty. Experiments on the 300-hour WHAM noise corrupted Switchboard corpus showed that the proposed Bayesian factorised adaptation produced up to 3.1% absolute (10.4% relative) WER reductions over the un-adapted baseline Conformer system.

7. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200021, 14200220, Innovation & Technology Fund grant No. ITS/254/19 and ITS/218/21, National Natural Science Foundation of China (NSFC) Grant 62106255, and Youth Innovation Promotion Association CAS Grant 2023119.

8. References

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA TRANS SIGNAL*, 2022.

[2] P. Guo *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *ICASSP*, 2021.

[3] M. Benzeghiba *et al.*, "Impact of variabilities on speech recognition," in *SPECOM*, 2006.

[4] P. Bell *et al.*, "Adaptation algorithms for neural network-based speech recognition: An overview," *Open Journal of Signal Processing*, 2020.

[5] J. Li *et al.*, "An overview of noise-robust automatic speech recognition," *IEEE/ACM TASLP*, 2014.

[6] M. Karafiát *et al.*, "ivector-based discriminative adaptation for automatic speech recognition," in *ASRU*, 2011.

[7] Z. Tüske *et al.*, "On the limit of english conversational speech recognition," in *INTERSPEECH*, 2021.

[8] M. Zeineldeen *et al.*, "Improving the training recipe for a robust conformer-based hybrid model," in *INTERSPEECH*, 2022.

[9] M. K. Baskar *et al.*, "Speaker adaptation for wav2vec2 based dysarthric asr," in *INTERSPEECH*, 2022.

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

[11] M. Delcroix *et al.*, "Auxiliary feature based adaptation of end-to-end asr systems." in *INTERSPEECH*, 2018.

[12] Y. Zhao *et al.*, "Speech transformer with speaker aware persistent memory." in *INTERSPEECH*, 2020.

[13] L. Sarı *et al.*, "Unsupervised speaker adaptation using attention-based speaker memory for end-to-end asr," in *ICASSP*, 2020.

[14] T. Ochiai *et al.*, "Speaker adaptation for multichannel end-to-end speech recognition," in *ICASSP*, 2018.

[15] Y. Huang *et al.*, "Rapid speaker adaptation for conformer transducer: Attention and bias are all you need." in *INTERSPEECH*, 2021.

[16] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM TASLP*, 2016.

[17] X. Xie *et al.*, "Bayesian learning for deep neural network adaptation," *IEEE/ACM TASLP*, 2021.

[18] J. Deng *et al.*, "Confidence score based speaker adaptation of conformer speech recognition systems," *IEEE/ACM TASLP*, 2023.

[19] T. Anastasakos *et al.*, "A compact model for speaker-adaptive training," in *ICSLP*, 1996.

[20] T. Ochiai *et al.*, "Speaker adaptive training using deep neural networks," in *ICASSP*, 2014.

[21] P. J. Moreno *et al.*, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*, 1996.

[22] V. Stouten *et al.*, "Model-based feature enhancement with uncertainty decoding for noise robust asr," *Speech Communi.*, 2006.

[23] D. Yu *et al.*, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *ICASSP*, 2008.

[24] T. Yoshioka *et al.*, "Environmentally robust asr front-end for deep neural network acoustic models," *Computer Speech & Language*, 2015.

[25] M. Ravanelli *et al.*, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP*, 2020.

[26] M. L. Seltzer *et al.*, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on speech and audio processing*, 2004.

[27] X. Anguera *et al.*, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM TASLP*, 2007.

[28] Y. Xu *et al.*, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR," in *ICASSP*, 2019.

[29] J. Heymann *et al.*, "Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR," in *ICASSP*, 2019.

[30] J. Yu *et al.*, "Audio-visual multi-channel integration and recognition of overlapped speech," *IEEE/ACM TASLP*, 2021.

[31] W. Zhang *et al.*, "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *ICASSP*, 2021.

[32] M. L. Seltzer *et al.*, "Acoustic model adaptation via linear spline interpolation for robust speech recognition," in *ICASSP*, 2010.

[33] X. Chen *et al.*, "An initial investigation of long-term adaptation for meeting transcription," in *INTERSPEECH*, 2014.

[34] M. Gales, "Acoustic factorisation," in *ASRU*, 2001.

[35] Y. Wang *et al.*, "An explicit independence constraint for factorised adaptation in speech recognition." in *INTERSPEECH*, 2013.

[36] M. Seltzer *et al.*, "Factored adaptation using a combination of feature-space and model-space transforms," in *INTERSPEECH*, 2012.

[37] J. Fainberg *et al.*, "Factorised representations for neural network adaptation to diverse acoustic environments." in *INTERSPEECH*, 2017.

[38] M. Kitza *et al.*, "Cumulative adaptation for blstm acoustic models," *INTERSPEECH*, 2019.

[39] M. Gales, "Adaptive training for robust asr," in *ASRU*, 2001.

[40] A. Acero *et al.*, "Hmm adaptation using vector taylor series for noisy speech recognition." in *INTERSPEECH*, 2000.

[41] C. J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, 1995.

[42] M. L. Seltzer *et al.*, "Separating speaker and environmental variability using factored transforms," in *INTERSPEECH*, 2011.

[43] Y. Wang *et al.*, "Speaker and noise factorization for robust speech recognition," *IEEE/ACM TASLP*, 2012.

[44] M. Zeineldeen *et al.*, "Conformer-based hybrid asr system for switchboard dataset," in *ICASSP*, 2022.

[45] X. Chang *et al.*, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP*, 2020.

[46] R. Kumar *et al.*, "End-to-end speech recognition with joint dereverberation of sub-band autoregressive envelopes," in *ICASSP*, 2022.

[47] V. N. Sukhadia *et al.*, "Domain adaptation of low-resource target-domain models using well-trained asr conformer models," in *SLT*, 2023.

[48] 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.

[49] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *INTERSPEECH*, 2020.

[50] S. Watanabe *et al.*, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE J-STSP*, 2017.

[51] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *NEURAL COMPUT*, 1992.

[52] L. Gillick *et al.*, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, 1989.

[53] J. J. Godfrey *et al.*, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, 1992.

[54] G. Wichern *et al.*, "Wham!: Extending speech separation to noisy environments," in *INTERSPEECH*, 2019.

[55] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[56] G. Sun *et al.*, "Transformer language models with lstm-based cross-utterance information representation," in *ICASSP*, 2021.