CONTINUAL LEARNING FOR ON-DEVICE SPEECH RECOGNITION USING DISENTANGLED CONFORMERS

Anuj Diwan[*,1], *Ching-Feng Yeh*[2], *Wei-Ning Hsu*[2], *Paden Tomasello*[2],
Eunsol Choi[1], *David Harwath*[1], *Abdelrahman Mohamed*[2]

[1] University of Texas at Austin   [2] Meta Inc.

{anuj.diwan,eunsol,harwath}@utexas.edu
{cfyeh,wnhsu,padentomasello,abdo}@meta.com

ABSTRACT

Automatic speech recognition research focuses on training and evaluating on static datasets. Yet, as speech models are increasingly deployed on personal devices, such models encounter user-specific distributional shifts. To simulate this real-world scenario, we introduce LibriContinual, a continual learning benchmark for speaker-specific domain adaptation derived from LibriVox audiobooks, with data corresponding to 118 individual speakers and 6 train splits per speaker of different sizes. Additionally, current speech recognition models and continual learning algorithms are not optimized to be compute-efficient. We adapt a general-purpose training algorithm NetAug for ASR and create a novel Conformer variant called the DisConformer (Disentangled Conformer). This algorithm produces ASR models consisting of a frozen 'core' network for general-purpose use and several tunable 'augment' networks for speaker-specific tuning. Using such models, we propose a novel compute-efficient continual learning algorithm called DisentangledCL. Our experiments show that the DisConformer models significantly outperform baselines on general ASR i.e. LibriSpeech (15.58% rel. WER on test-other). On speaker-specific LibriContinual they significantly outperform trainable-parameter-matched baselines (by 20.65% rel. WER on test) and even match fully finetuned baselines in some settings.

Index Terms— Continual Learning, ASR, On-Device, Domain Adaptation

1. INTRODUCTION

Today, speech recognition models are deployed on millions of personal devices. Such deployed models encounter an ever-changing distributional shift associated with their user's environment (e.g. speaker characteristics). Models should continually learn and adapt to their environment in a tractable, compute-efficient manner. While doing so, models should still perform well for other speakers without suffering from catastrophic forgetting [1]. Measuring such a continual-learning ability is not possible with current static ASR datasets. Therefore, we introduce the **LibriContinual** benchmark, a continual learning dataset for speaker-specific adaptation. This new benchmark is derived from LibriVox audiobooks and consists of training, validation and test datasets corresponding to 118 different speakers, with 6 training splits per speaker ranging from 10 min to 10 hr of speaker-specific data. Our benchmark measures the ability of models to continually adapt to new speakers in a compute-efficient manner, while preserving performance on the training dataset.

Furthermore, current speech recognition models do not inherently support compute-efficient techniques for on-device continual
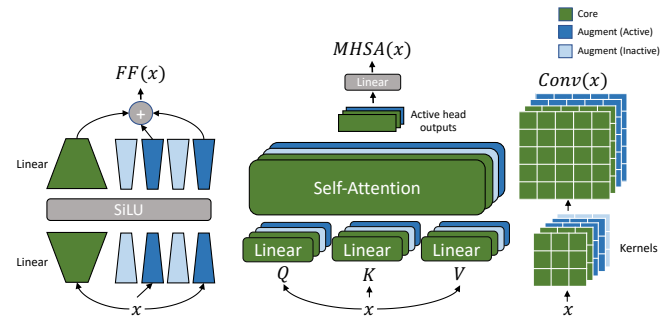
*Work done at Meta Inc.

Fig. 1: The DisConformer architecture depicting disentanglement in the Feedforward, Self-Attention and Convolution modules.

learning. Current continual learning techniques for ASR from prior work [2, 3, 4] require finetuning the entire model, which is not compute efficient. We propose (a) a novel general-purpose ASR algorithm derived from NetAug [5] to train ASR models that consist of 'core' and 'augment' networks and (b) **DisentangledCL**, a novel continual learning algorithm inspired by adapter networks [6] that only requires finetuning a small subset of these 'augment' networks and is compute-efficient. We apply our disentanglement approach to the Conformer [7] architecture to obtain 'DisConformers'. We describe the DisConformer and DisentangledCL in Section 4.

We find DisConformer models significantly outperform baselines on speaker-independent LibriSpeech by 15.58% relative WER on test-other with n-gram LM decoding; further, on speaker-specific LibriContinual, they significantly outperform trainable-parameter-matched baselines (by 20.65% relative on test set with n-gram LM) and sometimes even match fully finetuned baselines (in the DisConformer-Att and -Conv settings).

2. RELATED WORK

Continual Learning for Speech. [2, 3, 4] all explore continual learning in the context of ASR using regularization-based (e.g. EWC [8])) and data-replay based (e.g. GEM [9]) approaches. Other work explores settings such as SSL [10] and online learning [11].

Disentangled Models and On-Device ASR. Our DisConformer uses the NetAug algorithm [5] proposed for CNNs which we adapt for Transformers and ASR. While the original paper only uses the 'core' network at inference, we repurpose and also use the 'augment' networks for performing continual learning. Other on-device ASR settings include adaptation to rare words [12] and other prior work such as [13, 14] that train several subnets within a network to decrease model size while preserving high accuracies.

3. THE LIBRICONTINUAL BENCHMARK

Real-world speech models encounter user-specific distributional shift and must adapt to this domain shift. To measure this ability, we present the LibriContinual benchmark, a continual learning speaker adaptation benchmark. The same model should be capable of efficient speaker adaptation while still maintaining general-purpose ASR performance (e.g. to transcribe audio not spoken by the user like videos, phone calls, etc.). Our evaluation framework reflects these three requirements: a) efficient adaptation b) high speaker performance and c) high general-purpose performance.

Dataset Creation: LibriContinual is sourced from the LibriVox project: open-sourced speech from thousands of open-domain audiobooks. We first remove speakers already in the Librispeech [15] dataset. Then, we select a subset of the remaining speakers that have at least 2-hrs of data and 2 audiobooks each to make val and test sets and at least 10-hr of data to create a training set. Thus, we select a subset of 118 speakers that have sufficient data in order to create a 10-hr training set and validation and test sets of at least 2-hr, ensuring that there is no overlap between the audiobooks used in each set. We apply a Voice Activity Detector (VAD) to segment each audiobook into utterances of max duration 16 s. Finally, subsets of the 10 hr training set are constructed to obtain 5 hr, 2 hr, 1 hr, 30 min and 10 min training splits, such that each split is a superset of the next split. We obtain synthetic text transcriptions by running ASR using a wav2vec 2.0 Large [16] model pretrained and self-trained on LibriLight [17] and finetuned on Librispeech [15] and decode with a word Transformer LM. [1] Since the transcriptions are not human-derived, progress on this benchmark should only be interpreted as making better wav2vec2.0 model-like predictions. The avg. utt duration is 5.3s. While train set durations are fixed (e.g. 10h), the val and test sets of each speaker have variable durations (2-14h).

Evaluation Framework: Given a general ASR model $\mathcal{M}$ trained on an ASR dataset $\mathcal{D}_{orig}$ (LibriSpeech in all experiments) and a continual learning algorithm $\mathcal{A}(\mathcal{M}, \mathcal{D})$ to finetune it on a dataset $\mathcal{D}$, we run $\mathcal{A}$ on $\mathcal{M}$ for every speaker s to obtain 118 speaker-specific models $\mathcal{M}^{(s)} = \mathcal{A}(\mathcal{M}, \mathcal{D}^{(s)}_{LC,train})$, where $\mathcal{D}^{(s)}_{LC,train}$ is the LibriContinual (LC) train data for speaker s. We report:

Number of trainable params #CL-Params available during continual learning, a proxy for measuring compute-efficiency of $\mathcal{A}$.
Speaker-aggregate WER$_{LC}$. Each model $\mathcal{M}^{(s)}$ is evaluated on s's val/test sets $\mathcal{D}^{(s)}_{LC,val/test}$ to compute 118 different WERs WER$^{(s)}_{LC}$. Their median is taken to define a single number, WER$_{LC}$.
Original-aggregate WER$_{orig}$. Each model $\mathcal{M}^{(s)}$ is evaluated on the $\mathcal{D}_{orig}$ test set to obtain 118 different WERs and then their median is taken to compute WER$_{orig}$, measuring the ability to retain performance on the original $\mathcal{D}_{orig}$.

The above evaluation is repeated for every train split. Our benchmark contains data for 6 train splits but in our experiments we only report results for the 1 hr and 10 hr splits to be concise.

4. THE DISCONFORMER MODEL

We propose the DisConformer model (Fig 1) based on a *disentangled* approach designed to achieve a good tradeoff between adapting to new speakers and minimizing catastrophic forgetting by training two different types of model parameters: 'core' W_c and 'augment' W_a. Given an input x, the parameters used for the forward pass are dynamically constructed from W_c and a (potentially random) subset

---

of W_a. Given a (randomized) 'selector' function $S(W_a, x) \subseteq W_a$, the forward pass uses W_c and $S(W_a, x)$: $\mathcal{M}([W_c, S(W_a, x)], x)$. The core is always active while only a subset of augment params are. Then, the core is used for general-purpose ASR while the augment params are finetuned on speaker-specific data.

Our approach can be applied to any neural network, but we focus on the Conformer [7] model. We dub these versions as 'DisConformers' and propose disentangling the three types of modules (Feedforward, Self-Attention and Convolution), giving rise to DisConformer-FF, -MHSA, and -Conv. For e.g., in DisConformer-FF, the FF module is disentangled while the MHSA and Conv modules only have core parameters like a standard Conformer.

DisConformer-FF: The FF module in a Conformer consists of a sequence of layers: a linear layer, a non-linearity, and another linear layer. In the DisConformer-FF, we disentangle the feedforward dimension f into core and augment dimensions. The first linear layer has a core module with parameters $W_{1,c} \in \mathbb{R}^{d \times f_c}, b_{1,c} \in \mathbb{R}^{d \times f_c}$ and n_a augment *experts*, each with parameters $W^i_{1,a} \in \mathbb{R}^{d \times f_a}, b^i_{1,a} \in \mathbb{R}^{d \times f_a}$, where $f = f_c + n_a f_a$ is the feedforward dimension in the vanilla Conformer. Similarly, the second linear layer has a core module with weight parameters $W_{2,c} \in \mathbb{R}^{f_c \times d}$ and n_a augment experts with parameters $W^i_{2,a} \in \mathbb{R}^{f_a \times d}, b_2 \in \mathbb{R}^d$. Given an input x and a subset of r active augment experts with indices $i_1, i_2, \ldots, i_r$, the output y is computed as in eqs. (1) to (3):

$$h_c = \sigma(W_{1,c}x + b_{1,c}) \tag{1}$$

$$h_a = \sigma([W^{i_1}_{1,a}, \ldots, W^{i_r}_{1,a}]x + [b^{i_1}_{1,a}, \ldots, b^{i_r}_{1,a}]) \tag{2}$$

$$y = W_{2,c}h_c + [W^{i_1}_{2,a}, \ldots, W^{i_r}_{2,a}]h_a + b_2 \tag{3}$$

DisConformer-Att: The Att module in a Conformer performs multi-head self-attention with h different heads. In DisConformer-Att, we first disentangle the heads into h_c core heads and h_a augment heads, where $h = h_c + h_a$. Given a subset of r active augment heads, we perform multi-head self-attention as usual, but using just the h_c core heads and r augment heads, not all the h_a augment heads.

DisConformer-Conv: The Conv module in a standard Conformer consists of a sequence of layers: a Pointwise Conv PC_1, a 1D Depthwise Conv DC, Layer Norm LN, another Pointwise Conv PC_2. Each layer is parametrized by the number of intermediate conv channels, d_{conv}. For e.g., PC_1 maps the input from d to d_{conv} channels. In the DisConformer-Conv, we disentangle the d_{conv} channels into d_c core channels and d_a augment channels. Given a subset of r active augment channels, we index into each layer's kernels to create new kernels with $d'_{conv} = d_c + r$ intermediate channels and compute convolutional operations normally using this new kernel.

General ASR Training using NetAug: We train the DisConformer as a general-purpose ASR model on $\mathcal{D}_{orig}$ i.e. Librispeech, using *NetAug training* inspired by [5]. Let the DisConformer-FF (/Att/Conv) model have n_{ffn} (/n_{att}/n_{conv}) augment experts (/heads/channels). For ease of explanation, we describe the approach using DisConformer-FF but the approach is analogously applied to DisConformer-Att and DisConformer-Conv. Given a training example (x, y), we first uniformly sample a number n from $\{1, 2, 4, \ldots, n_{ffn}\}$. Then, we uniformly sample n FF augment experts from the total n_{ffn} experts, whose parameters one can denote as $W_{aug,ffn}$. That is, we sample a random-sized random subset of augment params. Denoting the core parameters by W_{core}, we can define the training loss $L(\mathcal{M}, x, y)$ as in eq. (4):

$$L(\mathcal{M}, x, y) = \text{CTC}(\mathcal{M}(W_{core}, x), y) \tag{4}$$
$$+ \alpha \, \text{CTC}(\mathcal{M}([W_{core}, W_{aug,ffn}], x), y)$$

---

[1] beam=100;beamthres=20;lmweight=1.51;wordscore=2.06;silweight=-3

where CTC is the Connectionist Temporal Classification loss [18]) and α is a hyperparameter; in practice, we always set it to 1.0 as that performed best on the Librispeech dev-other validation set in initial experiments. This loss encourages the model to train the core parameters in isolation (term 1) as well as in conjunction with a random subset of augment parameters (term 2).

DisentangledCL: We introduce a novel compute-efficient continual learning algorithm *DisentangledCL*. Again, for brevity, we describe the approach using DisConformer-FF. We first start with a general-purpose ASR model trained using NetAug. To finetune on a training dataset $\mathcal{D}$, we randomly select a subset of $k_{ffn} < n_{ffn}$ augment experts, denoting their params by $W_{aug,ffn}^k$, such that $|W_{aug,ffn}^k| << |W_{core}|$ i.e no. of trainable augment params is a small fraction of core params; at most 13% in all experiments. We then finetune these $W_{aug,ffn}^k$ parameters while W_{core} is frozen, using the regular CTC loss $\text{CTC}(\mathcal{M}([W_{core}, W_{aug,ffn}^k], x), y)$. We use W_{core} i.e. the core parameters for general-purpose inference (on $\mathcal{D}_{orig}$). For speaker-specific inference, we use $[W_{core}, W_{aug,ffn}^k]$. Thus, we get the best of both worlds; performance is retained on the original dataset via the core parameters, while speaker-specific improvements can come from the finetuned augment parameters.

5. EXPERIMENTS

5.1. Experimental Setup

We use the standard Conformer-CTC [7] architecture, but with a time reduction layer (similar to [19]) instead of 2 learnable CNN layers for efficiency. All models share the following hyperparams: 256 model dim, 30 output dim, 16 layers, 64 FF dim per expert, 31 depthwise conv kernel, and 0.1 dropout. The aspects in which they differ are summarized in Table 1. The output vocabulary consists of the English alphabet (26 letters), space, apostrophe and CTC blank.

	DisCo-FF	DisCo-Att	DisCo-Conv
# FF (core,aug)	(8,12)	(20,0)	(20,0)
# Att (core,aug) heads	(4,0)	(2,2)	(4,0)
Conv channels/expert	16	16	8
# Conv (core,aug)	(16,0)	(16,0)	(16,16)

Table 1: Summary of DisConformer model architectures.

NetAug ASR Training: Details. We train on the Librispeech [15] 960-hr training set. We use SpecAugment with 2 27-channel freq masks and 2 100-frame time masks. We use Adam with an lr of $0.0004, \beta_1=0.9, \beta_2=0.98$ and train for 200k steps on 16 GPUs with a 4-stage linear LR schedule: warmup 8%, const 32%, decay 40%, const 20%. We use a per-GPU batch size of 32 subject to a max of 320 s. We choose the checkpoint with the min WER on Librispeech dev-clean + dev-other.

DisentangledCL: Details. We set $k_{ffn}=2$ for DisCo-FF, $k_{att}=2$ for DisCo-Att, and $k_{conv}=12$ for DisCo-Conv. We use Adam with an lr of $0.0001, \beta_1=0.9, \beta_2=0.98$. In this paper, we run experiments for only the 1 hr and 10 hr subsets. For 10 hr, we train for 30k steps while for the 1 hr subset, we train for 10k steps on 1 GPU. These numbers were chosen to ensure overall model convergence. We use the same 3-stage LR schedule for both; 40% const, 40% decay, 20% const. Other hyperparams are same as NetAug training. We report eval results using both Viterbi decoding and n-gram LM decoding. We use a 4-gram LM trained on the Librispeech book corpus with beam=20, lmweight=1.74, wordscore=-0.52.

5.2. Baselines

Baseline Models: For each of the three DisConformer models, we construct corresponding Conformer baselines dubbed Base-FF, Base-Att and Base-Conv. Base-FF is a Conformer with an FFN dimension of $512 = 64 \times 8$ and otherwise identical to DisCo-FF. Thus, it has the same architecture as a DisCo-FF with only its core (8 experts each with dim 64). Similarly, Base-Att is a Conformer with 2 heads (identical to DisCo-Att with only its 2 core heads) and Base-Conv is a Conformer with $128 = 8 \times 16$ channels (identical to DisCo-Conv with only its 16 core experts). We perform general-purpose ASR training on Librispeech using the regular CTC loss with the same optimizer hyperparams and number of steps as the DisConformers and choose the best-performing model on Librispeech dev-clean + dev-other.

Baseline Continual Learning algorithms: All ASR continual learning techniques investigated in prior work [2, 3, 4] finetune the entire model, which is more computationally expensive than our DisentangledCL which only finetunes a small subset of parameters. We first investigate two existing baseline CL algorithms (which finetune the whole model). Further, for a fairer comparison with our approach, we analyze simple, efficient variants of both algorithms.
(1) Full-FT: We fully finetune the baseline models using CTC loss. We use the same hyperparameters as the DisConformers, except that we use a more stable learning rate of 0.00005 for the 1 hr subset.
(2) KD (Knowledge Distillation): Following previous work [2, 20, 21], to prevent catastrophic forgetting, this approach adds an auxiliary loss to minimize the KL Divergence between the model being trained ($\mathcal{M}$) and the original initialization ($\mathcal{M}^*$) as in eq. (5):

$$\mathcal{L}(\mathcal{M}, x, y) = \text{CTC}(\mathcal{M}, x, y) + \lambda \, \text{KL-div}(p(x), p^*(x)) \quad (5)$$

where $p(x) = \text{softmax}(\mathcal{M}(x)/T)$ and $p^*(x) = \text{softmax}(\mathcal{M}^*(x) /T)$ i.e. temperature-scaled logits. We set $\lambda = 8.0$ and ablate this choice in Section 5.4. We set $T = 1.0$. The other hyperparameters are the same as Full-FT. This approach is even more computationally expensive than Full-FT, because it involves an extra forward pass.
(3) Full-FT-Efficient: This is an efficient variant of Full-FT that only fine-tunes the top few layers such that the number of parameters being fine-tuned is approximately equal to that in DisentangledCL. We finetune 2 layers for FF and 1 layer for Att and Conv.
(4) KD-Efficient: This is an efficient variant of the Knowledge Distillation approach, similar to Full-FT-Efficient.

5.3. Results

Evaluating general ASR-trained models. We first investigate the setting where there is no continual learning performed (i.e. no speaker data is available). Thus, we directly compare a NetAug-trained DisConformer with a baseline model, both trained on LibriSpeech. We report results for LibriSpeech and LibriContinual in Table 2 where all WERs are median WERs across speakers. We run inference on just the DisConformer core (W_{core}) discarding all augment experts. This is fair since the DisConformer core and the baseline have the exact same architecture. We observe that all 3 DisConformer models consistently outperform the baselines. With LM decoding, DisConformers achieve an average relative WER reduction of 5.6% on LibriSpeech test-clean, 3.7% on test-other and 5.5% on LibriContinual test. This shows that NetAug ASR training is well-suited for obtaining better general-purpose ASR models, even outside the context of continual learning.

Evaluating Continual Learning. Table 3 presents the evaluation results on the LibriContinual benchmark. We exhaustively report n-gram LM decoding results and refer to aggregate Viterbi de-

	Viterbi				n-gram LM			
	LibriSpeech		LibriContinual		LibriSpeech		LibriContinual	
Model	test-c	test-o	val	test	test-c	test-o	val	test
Base-FF	5.71	14.35	11.46	12.14	4.02	10.16	7.92	8.36
DisCo-FF	**5.38**	**13.69**	**10.8**	**11.22**	**3.75**	**9.82**	**7.41**	**7.82**
Base-Att	4.33	11.23	8.94	9.52	3.42	8.54	6.40	6.76
DisCo-Att	**4.02**	**10.76**	**8.31**	**8.74**	**3.29**	**8.22**	**6.08**	**6.34**
Base-Conv	4.28	11.31	9.48	9.80	3.50	8.62	6.88	7.22
DisCo-Conv	**4.13**	**10.83**	**8.93**	**9.36**	**3.28**	**8.19**	**6.66**	**6.94**

Table 2: Results on LibriSpeech and LibriContinual for general ASR-training without Continual Learning.

			1hr			10hr		
			LS		LC	LS		LC
Model	CL Algo	# CL-θ	test-c	test-o	test	test-c	test-o	test
Base-FF (16.1M)	Full-FT	16.1M	5.4	14.0	**6.3**	7.2	20.1	**5.6**
	KD	16.1M	5.3	13.5	6.7	5.0	17.7	5.8
	Full-FT-Eff	2.0M	4.8	11.8	8.2	5.0	12.3	7.5
	KD-Eff	2.0M	4.8	11.8	8.3	5.0	12.3	7.7
DisCo-FF (18.3M)	DisCL	2.1M	__3.8__	__9.8__	__6.9__	__3.8__	__9.8__	__6.2__
Base-Att (26.6M)	Full-FT	26.6M	4.6	12.0	**5.4**	6.1	17.8	**4.8**
	KD	26.6M	4.3	11.5	5.6	5.3	14.9	4.9
	Full-FT-Eff	1.8M	3.8	9.4	6.8	3.9	9.6	6.4
	KD-Eff	1.8M	3.8	9.4	6.8	3.8	9.6	6.5
DisCo-Att (28.7M)	DisCL	2.1M	__3.3__	__8.2__	__5.5__	__3.3__	__8.2__	__4.9__
Base-Conv (27M)	Full-FT	27M	4.7	12.3	**5.5**	6.2	18.1	**4.9**
	KD	27M	4.5	11.9	5.7	5.4	15.4	5.1
	Full-FT-Eff	1.7M	3.9	9.5	7.1	3.9	9.8	6.7
	KD-Eff	1.7M	3.9	9.6	7.2	3.9	9.8	6.8
DisCo-Conv (28.3M)	DisCL	1.3M	__3.3__	__8.2__	__5.5__	__3.3__	__8.2__	__4.9__

Table 3: Results on LibriSpeech and LibriContinual with n-gram LM. (xM) next to model denotes total model params. # CL-θ is the # trainable params for CL. LS=LibriSpeech, LC=LibriContinual. **Bold** numbers are the best WERs across all approaches while Underlined are the best across # CL-θ-matched approaches.

coding results in-text. As the amount of speaker data increases (0-hr from Table 2 to 1 hr to 10 hr), the WERs on the LibriContinual val/test sets decrease as expected.

We start by analyzing preservation of performance on the original dataset, Librispeech, after continual learning. The performance of all baselines degrades considerably, with the effect more pronounced for the 10 hr split, exhibiting catastrophic forgetting. KD performs better than Full-FT (likely due to the KL-divergence term), and the Efficient variants perform better than the vanilla approaches (likely since only a subset of parameters are finetuned). In contrast, our DisentangledCL has the same performance as the general ASR model from Table 2 , resulting in no catastrophic forgetting at all and it significantly outperforms the best baseline. Averaged across all settings, over the best baseline, this results in relative WER gains of 23.94% with Viterbi and 17.66% with n-gram LM for test-clean, and 17.06% and 15.58% respectively for test-other.

On LibriContinual, in all settings, the DisConformer models with at most 13% extra total params significantly outperform the #CL-Param-matched Efficient baselines. Averaged across all settings, over the best Efficient baseline, this is a relative WER gain of 21.16% with Viterbi and 18.26% with n-gram LM for the LibriContinual validation set, and 21.85% and 20.65 respectively for the test set. Surprisingly, in both the 1 hr and 10 hr settings, despite only finetuning a much smaller fraction of parameters (7% at maximum)

the DisConformer-Att and DisConformer-Conv are within ± 0.1 WER of fully finetuned baselines (Full-FT and KD). In contrast, the DisConformer-FF model performs much worse than the fully finetuned baselines, likely owing to the much smaller number of trainable params. It has a max abs. WER difference of $+0.9$ with the best baseline Full-FT across all settings. On the other hand, on LibriSpeech, Full-FT has a much worse WER performance; min -4.7 abs. WER across all settings. This is a tradeoff between speaker-specific and general performance. Depending on the end use-case, the magnitude of acceptable degradation of general vs. speaker-specific performance will vary.

Overall, this analysis reveals that with a small number of available parameters for finetuning (at most 13% of baselines), the DisConformer models offer superior performance on Librispeech and on speaker-specific LibriContinual, they perform better than trainable-parameter-matched baselines, and are sometimes able to match even fully-finetuned baselines (for Att and Conv, but not FF). This also suggests that DisConformers may be more effective when applied to Att or Conv layers than FF.

5.4. Ablations

Using DisentangledCL on baseline models. We analyze whether NetAug disentangled training is necessary by training baseline models in the disentangled 'core+augment' framework. We perform LM-decoded LibriContinual test set eval with the 1 hr train set for these 4 settings which all have the same architectures:

Base-Conv + Random. We use the Base-Conv model as the core and randomly initialize $k_{conv} = 12$ augment experts for finetuning.

Base-Conv + Base-Conv. We train a Base-Conv Conformer with 224 conv channels with CTC loss and treat its first $128 = 8 \times 16$ channels as the core and the last $96 = 8 \times 12$ channels as the augment experts.

DisCo-Conv (core) + Random. We take the trained DisConformer-Conv and randomly re-initialize its augment parameters.

DisCo-Conv (core) + Disco-Conv. This is our DisCo-Conv model. We find that these 4 settings achieve a test set WER of (a) 5.9, (b) 5.7, (c) 5.7 and (d) 5.5 respectively. NetAug-trained core experts are better [(a) vs. (c)] and NetAug-trained augment experts are better [(c) vs. (d)], showing NetAug is important. (b) shows that baselines can also be trained in our DisentangledCL framework with a good 5.7 WER, although it is worse than our model, (d) 5.5.

Ablating KD hyperparam λ. The λ hyperparam in the KD baseline loss controls the tradeoff between learning on new data (CTC loss) and staying close to the old model (KL div loss). We tune λ over the set $\{0, 1, 2, 4, 8, 16, 32\}$ using the Base-FF model finetuned on the 10 hr split and decoded with Viterbi. As λ increases, performance on LibriSpeech monotonically improves (from 26.4 at λ=0 to 20.2 at λ=32); however, while LibriContinual performance is also improved from 8.2 at λ=0 to 7.9 at λ=8, it significantly worsens (to 8.6 at λ=32). Thus, we choose $\lambda = 8$ for the KD baseline.

6. CONCLUSION

We introduced LibriContinual, a new continual learning benchmark for efficient speaker-specific domain adaptation. We also proposed DisConformers and novel ASR training (NetAug) and continual learning (DisentangledCL) algorithms which use different parts of the same model to achieve strong general ASR performance and speaker-specific performance in a parameter-efficient manner. For future work, we plan to extend the LibriContinual benchmark to the unlabelled setting (via weak supervision for ASR) and add more speech tasks. We also plan to extend our NetAug algorithm to build speaker-specialized experts.

7. REFERENCES

[1] Michael McCloskey and Neal J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," vol. 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989.

[2] Heng-Jui Chang, Hung yi Lee, and Lin shan Lee, "Towards Lifelong Learning of End-to-End ASR," in *Proc. Interspeech 2021*, 2021, pp. 2551–2555.

[3] Steven Vander Eeckt and Hugo Van Hamme, "Continual Learning for Monolingual End-to-End Automatic Speech Recognition," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 459–463.

[4] Samik Sadhu and Hynek Hermansky, "Continual Learning in Automatic Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 1246–1250.

[5] Han Cai, Chuang Gan, Ji Lin, and song han, "Network Augmentation for Tiny Deep Learning," in *International Conference on Learning Representations*, 2022.

[6] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov, Eds. 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, PMLR.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, mar 2017.

[9] David Lopez-Paz and Marc' Aurelio Ranzato, "Gradient Episodic Memory for Continual Learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[10] Samuel Kessler, Bethan Thomas, and Salah Karout, "An Adapter Based Pre-Training for Efficient and Scalable Self-Supervised Speech Representation Learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3179–3183.

[11] Muqiao Yang, Ian Lane, and Shinji Watanabe, "Online Continual Learning of End-to-End Speech Recognition Models," in *Proc. Interspeech 2022*, 2022, pp. 2668–2672.

[12] Tsendsuren Munkhdalai, Khe Chai Sim, Angad Chandorkar, Fan Gao, Mason Chua, Trevor Strohman, and Françoise Beaufays, "Fast contextual adaptation with neural associative memory for on-device personalized speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6632–6636.

[13] Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, and Haizhou Li, "LightHuBERT: Lightweight and Configurable Speech Representation Learning with Once-for-All Hidden-Unit BERT," in *Proc. Interspeech 2022*, 2022, pp. 1686–1690.

[14] Haichuan Yang, Yuan Shangguan, Dilin Wang, Meng Li, Pierce Chuang, Xiaohui Zhang, Ganesh Venkatesh, Ozlem Kalinli, and Vikas Chandra, "Omni-Sparsity DNN: Fast Sparsity Optimization for On-Device Streaming E2E ASR Via Supernet," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8197–8201.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[17] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, https://github.com/facebookresearch/libri-light.

[18] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[19] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein, "Streaming End-to-end Speech Recognition for Mobile Devices," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[20] Zhizhong Li and Derek Hoiem, "Learning without Forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[21] Jiabin Xue, Jiqing Han, Tieran Zheng, Xiang Gao, and Jiaxing Guo, "A Multi-Task Learning Framework for Overcoming the Catastrophic Forgetting in Automatic Speech Recognition," 2019.