Speech Technology for Everyone: Automatic Speech Recognition for Non-Native English with Transfer Learning

Toshiko Shibano[*]   **Xinyi Zhang**[*]   **Mia Taige Li**[*]   **Haejin Cho**[*]
Peter Sullivan[*]   **Muhammad Abdul-Mageed**[*]
The University of British Columbia

tshibano@student.ubc.ca, {jeremyzxy0803, mia.taige.li, haejin2909}@gmail.com,
prsull@student.ubc.ca, muhammad.mageed@ubc.ca

Abstract

To address the performance gap of English ASR models on L2 English speakers, we evaluate fine-tuning of pretrained wav2vec 2.0 models (Baevski et al., 2020; Xu et al., 2021) on L2-ARCTIC, a non-native English speech corpus (Zhao et al., 2018) under different training settings. We compare **(a)** models trained with a combination of diverse accents to ones trained with only specific accents and **(b)** results from different single-accent models. Our experiments demonstrate the promise of developing ASR models for non-native English speakers, even with small amounts of L2 training data and even without a language model. Our models also excel in the zero-shot setting where we train on multiple L2 datasets and test on a blind L2 test set.

Index Terms: Automatic Speech Recognition (ASR), ASR for L2 English speakers

1 Introduction

Although non-native (L2) English speakers outnumber native (L1) English speakers (Crystal, 2003), major challenges contribute to a gap between performance of ASR systems on L2 speech, mainly due to the influence of L1 pronunciation on the learned language, and the lack of annotated L2 speech data (Radzikowski et al., 2021; Viglino et al., 2019). To meet these challenges, previous studies have exhibited two distinct approaches. The first is to make L2 speech representations more closely match those of L1 speech (Radzikowski et al., 2021). The second approach leverages L2 speech data to improve model robustness. Due to L2 data scarcity, and hence the challenge of training L2 models from scratch, this second approach necessitates employment of transfer learning or domain adaptation (Shi et al., 2021; Sun et al., 2018).

Figure 1: The various data splits we use in our experiments. Shade represents a different run of our training, with the gradient blocks in Split 4 being present in all runs. For cross validation splits, we show a single fold as an example, where number indicates the participants included.

State-of-the-art ASR models based on unsupervised/self-supervised pre-training such as wav2vec (Schneider et al., 2019) and wav2vec 2.0 (Baevski et al., 2020)[1] offer a tantalizing starting point for applying the second approach we list above, especially due to their strong performance on ASR even without a language model. However, challenges remain in identifying how best to apply models such as wav2vec 2.0 in L2 fine-tuning scenarios. For this reason, our objective in the current work is to investigate a rich set of conditions under which we can fine-tune ASR models for optimal L2 performance. More concretely, we attempt to achieve the following:

1. Evaluate fine-tuning strategies for adapting

---

[*]All authors contributed equally.

[1]Although sometimes referred to as 'unsupervised', these models employ a self-supervised objective.

pre-trained L1 English ASR models to L2 English;

2. Explore impact of non-native (L2) accents on performance of these fine-tuned ASR models, comparing multi-accent training to single-accent training; and

3. Quantify the impact of L2 fine-tuning on model performance for L1 English speech recognition.

Although external language models are often used in improving ASR performance (Nakatani, 2019; Xu et al., 2020), models trained with great quantities of data can potentially internalize this linguistic information (Graves and Jaitly, 2014). In particular, some of the wav2vec 2.0 models perform nearly as well with and without a language model on difficult speech such as LibriSpeech Test-Other (Xu et al., 2021). We thus use this robust pre-trained model as our starting point, and carry out our work without use of an external language model to see if this performance is retained through the fine-tuning process.

The rest of the paper is organized as follows: Section 2 is an overview of related works. We describe our data in Section 3. Section 4 is about our experiments and results. We conclude in Section 5.

2 Related Work

Because of the difficulty in linguistically annotating corpora for Hidden Markov Model (HMM)-based ASR (Graves and Jaitly, 2014), researchers have broadly embraced End-to-End (E2E) deep learning architectures either based on Connectionist Temporal Classification (CTC) (Graves et al., 2006; Graves and Jaitly, 2014), Attention (Chorowski et al., 2015; Chan et al., 2016; Gulati et al., 2020), or hybrids of the two (Watanabe et al., 2017; Wang et al., 2020). Recent efforts inspired by work such as BERT (Devlin et al., 2019) have improved on these purely supervised learning baselines through self-supervised pre-training (Schneider et al., 2019; Baevski et al., 2019, 2020) and self-training (Xu et al., 2021). These self-supervised wav2vec models represent one line of research in speech representation. Other works include models similar to wav2vec that also use a contrastive loss (Oord et al., 2018), models using an autoregressive loss function (Ling et al., 2020; Chung et al., 2019), as well as models using a masked language model closer to the original BERT (Liu et al., 2020a).

With these efforts, ASR technologies for native languages have evolved significantly. However, we still observe problems in many applications. In particular, several researchers have emphasized how performance of ASR models drops when the input speech is from non-native speakers whose native languages are different from the models' target languages (Radzikowski et al., 2021; Livescu and Glass, 2000; Wang et al., 2003; Ping, 2008). For systems developed for English ASR, this can be a real issue. The reason, as observed earlier, is that large populations of English language speakers are non-native (Crystal, 2003). In line with this argument, Ping (2008), for example, pointed out the necessity to improve speech recognition technology for L2 speakers given that many people speak more than one language for economic and social reasons, especially considering human migration is becoming more common these days. It is hoped that continued efforts aiming at improving ASR for non-native speakers will eventually lead to improved results for many as voice recognition technology becomes increasingly pervasive in our daily lives (Ping, 2008).

As we explained in Section 1, there are two distinct approaches to improve current ASR performance on L2 speech: 1) accent conversion as an extension to the active area of research of voice conversion; and 2) incorporation of L2 speech data, which is often limited in quantity and quality, during the model training process.

The first approach takes inspiration from voice conversion, but instead of focusing on modifying the pitch, it modifies the pronunciation to reduce accents. Additionally, voice conversion models aim to generate results that are speaker-dependent, while accent conversion models deal with generalizing accents from a group of speakers, hence being speaker-independent. With this approach, the resulting model can be used as a pre-processing step to remove accents in the data prior to feeding these data into an ASR model. Bearman et al. (2017) adopt this approach but focus on L1 English accents, while Radzikowski et al. (2021) work on L2 English accents with speakers' L1 being Japanese. Liu et al. (2020b) took a step further and turned Hindi-accented English to native American English without utilizing native utterances.

The second approach often employs techniques such as domain adversarial training and transfer learning in order to utilize as much available ac-

cented speech data as possible. Domain adversarial training (DAT) is a popular approach as it encourages models to learn accent-invariant features (Sun et al., 2018; Hou et al., 2019; Hu et al., 2021). Transfer learning is another popular approach in L2 speech recognition, as it possibly allows a model to gain knowledge from both the base task and the new task, even when the new task has limited data (Matassoni et al., 2018; Das et al., 2021; Shi et al., 2021). In the Accented English Speech Recognition Challenge 2020 (AESRC2020), many teams utilize transfer learning to tackle the L2 accent recognition task (Shi et al., 2021). In a recent work, Das et al. (2021) combine both DAT and transfer learning to achieve robust accented speech recognition performance. We now introduce our data.

3 Data

3.1 Corpus Information

We choose **L2-ARCTIC**, a non-native English speech corpus (Zhao et al., 2018), for L2 fine-tuning. The recordings are from 24 non-native speakers of English with a total of six different L1s, and each of the L1s consists of two female speakers and two male speakers. The L1s we use for our experiments are Arabic (AR), Hindi (HI), Korean (KO), Mandarin (ZH), Spanish (ES), and Vietnamese (VI). Because L2-ARCTIC is based on the original L1 English corpus, CMU ARCTIC (Kominek et al., 2003) (henceforth **L1-ARCTIC**, for simplicity), we can easily evaluate performance from fine-tuning on same-domain L1 data.

Each speaker in L2-ARCTIC contributed approximately one hour of phonetically-balanced read speech based on the L1-ARCTIC prompts, which consist of carefully selected sentences ($1,132$ sentence prompts) from Project Gutenberg (Kominek et al., 2003). We note this, as the pretrained wav2vec 2.0 model we use was first pre-trained on LibriSpeech[2] (Panayotov et al., 2015) and then self-trained on Libri-Light[3] (Kahn et al., 2020). Both corpora rely on audiobooks from the LibriVox project,[4] much of which comes from Project Gutenberg.[5] This minimizes discrepancies between domains of the text.

We also evaluate our fine-tuned models on **1) LibriSpeech** to compare the fine-tuning with the original performance of self-trained wav2vec 2.0 Large (LV-60) model (Xu et al., 2021), which we will refer to as *Wav2Vec 2.0-ST*. In addition, we evaluate on **2) L1-ARCTIC**, identical to our L2-ARCTIC corpus but spoken by four native US English speakers, allowing us to identify any degradation in performance on L1 speech. Each of L1-ARCTIC speakers' datasets contain approximately the same number of utterances ($n = \sim 1,132 * 4$) as each of L2-ARCTIC speakers' datasets.

For the purpose of our experiments, we define *native (L1) accents* as those represented in the LibriSpeech and L1-ARCTIC, and *non-native (L2) accents* as those represented in L2-ARCTIC.

3.2 Data Splits

For both L2-ARCTIC and L1-ARCTIC, we split the data into three distinct Train, Dev, and Test sets with an 80:10:10 ratio. Importantly, we ensure there is *no overlap between utterances*. For L2-ARCTIC, we split the data across the following settings (see Fig. 1).

- **Split-1** *(speaker-dependent, multi-accent split)*: All speakers from all accents in the Train set are also included in the Dev and Test sets; however, no utterances are shared between Train, Dev, and Test.

- **Split-2** *(speaker-independent cross-validation splits with multiple accents)*: A speaker from each accent[6] is removed from the Train and Dev sets, but other speakers with the same accent remain in the Train and Dev sets.

- **Split-3** *(speaker-independent zero-shot splits with multiple accents)*: All speakers from one of the accents are entirely removed from the Train and Dev sets. The removed speakers are included in Test.

- **Split-4** *(all-speaker, single-accent split)*: Speakers are broken down by accents (six accents in total) and all speakers in a given accent are split into the Train, Dev, and Test sets (3 data splits x 6 accents).

- **Split-5** *(speaker-independent cross-validation splits with single accent)*: One speaker in each

---

[2]http://www.openslr.org/12/
[3]https://github.com/facebookresearch/libri-light
[4]https://librivox.org
[5]http://www.gutenberg.org

---

[6]We use the term 'accent' here to loosely refer to variation in speakers with L1 other than English.

		Accent dependency		Speaker dependency	
		Dependent	**Independent**	**Dependent**	**Independent**
Multi-accent	Model-1 (Split 1)	x		x	
	Model-2 (Split 2)	x			x
	Model-3 (Split 3)		x		x
Single-accent	Model-4 (Split 4)	x	x	x	x
	Model-5 (Split 5)	x			x

Table 1: Summary of data splits, fine-tuning, and evaluation setups.

accent is removed from the Train and Dev sets, but the other speakers with the same accent remain in the Train and Dev sets. As there are four speakers per accent, four splits are created for each accent, which are further split into the Train, Dev, and Test sets (3 data splits x 6 accents x 4 speakers).

4 Experiments

For all our wav2vec 2.0 models, we use Fairseq [7] fine-tuning default settings as a reference and convert the hyper-parameters to align with Huggingface's implementation. We train each model with three random seeds and take average over three WERs, one each from the three seeds.

4.1 Model Architecture, Fine-tuning, Baselines, and Evaluation

For our model development, we use the wav2vec 2.0 architecture (Baevski et al., 2020) which is composed of a multi-layer convolutional neural network feature extractor and a Transformer context network. It takes in raw audio and converts it into representations of the input sequence. The encoder consists of multiple blocks of temporal convolution followed by a layer normalization and a GELU activation function. The relative positional embedding in the Transformer is accomplished by a convolutional layer.

Fine-tuning of pre-trained wav2vec 2.0 is performed with CTC and the transcriptions of the audio segments. For each model, we identify the optimal hyper-parameters on the respective Dev set. We choose hyper-parameters as follows: For `mask_feature_prob`, we pick from {*0.25, 0.5*}, for `mask_feature_length`, we choose from {*15, 30*}, for `mask_time_prob` we use {*0.5, 0.75*}, and a batch size of 16. To mimic the tri-state learning rate schedule (Baevski et al., 2020), we set different learning rates for different stages:

warm-up (1e-5, 3e-5), constant stage (1e-5, 3e-5), and decay (1e-5, 3e-5, 5e-6). The decay stage is followed by another constant stage (1e-5, 2e-6, 5e-6) to simulate the Fairseq's fine-tuning configuration. We evaluate all our models in terms of word error rate (WER). *All* our results are the average of three runs, and we use the following baselines:

- **Baseline-I:** Wav2Vec 2.0-ST (Xu et al., 2021),[8] a self-trained version of wav2vec 2.0 (Baevski et al., 2020) exploiting a Transformer large architecture and pre-training on 960 hours of speech data from LibriSpeech (Panayotov et al., 2015). The self-training is performed on 60K hours of Libri-Light (Kahn et al., 2020). We believe this as an already strong baseline. We use the model released via HuggingFace. [9]

- **Baseline-II:** This is Wav2Vec 2.0-ST, the same as Baseline-I, fine-tuned on L1-ARCTIC described earlier. The purpose of Baseline-II is to allow for measuring the trade-off of L1 English ASR performance by fine-tuning the English pre-trained model on L2 accents.

4.2 Multi-Accent Models

With our multi-accent models, we examine performance using multiple accents during training. We introduce each of our models here, and present the results acquired with each. We provide a summary of our different data splits and models across accent and speaker dependency categories in Table 1.

Model-1 (speaker- and accent-dependent): The model is fine-tuned with Split-1 data to identify any speaker-dependent training impact, as well as an upper limit on performance. In addition to

Model	L2-ARCTIC		L1-ARCTIC		LS$_{dev}$		LS$_{test}$	
	Dev	Test	Dev	Test	Clean	Other	Clean	Other
Baseline-I	13.47	12.47	2.30	2.23	**1.69**	**3.55**	**1.86**	**3.89**
Baseline-II	17.29	15.95	**1.26**	**1.30**	2.19	5.13	2.32	5.00
Model-1	**9.78**	**9.27**	1.94	1.86	2.75	5.55	2.82	6.36

Table 2: Model-1 performance in word error rate (WER) (lower is better) on non-native accents (L2-ARCTIC) and native accents (L1-ARCTIC, LS$_{dev}$ and LS$_{test}$). Baseline-I and Baseline-II are reported on the same Dev and Test sets of each corpus for comparison.
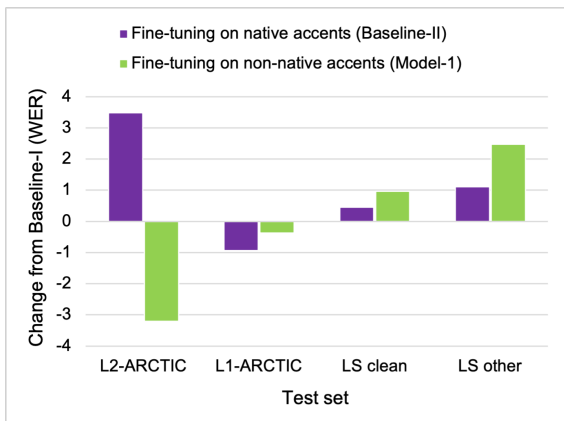

Figure 2: Trade-offs of fine-tuning on native accents (Baseline-II) vs. non-native accents (Model-1). As we evaluate model accuracy by error rate, the bars extending into the negative values mean that the model gains accuracy by fine-tuning.

Model	Dev$_{L2}$		Test$_{L2}$	
	Mean	SD	Mean	SD
Baseline-I	13.47	0.23	12.47	0.84
Baseline-II	17.29	0.41	15.96	1.58
Model-2	**9.57**	0.19	**9.96**	0.64

Table 3: Model-2 cross validated performance on L2-ARCTIC Dev and Test sets, alongside Baseline-I and Baseline-II performance on the same cross validation splits. Mean refers to the average WER over the four runs and SD refers to the standard deviation.

Figure 3: HI-specific Model-4 evaluated on individual accents. As we evaluate model accuracy by error rate, the bars extending downwards represent the performance gain by fine-tuning. HI-specific fine-tuning benefits HI but hinders performance on all the other accents.
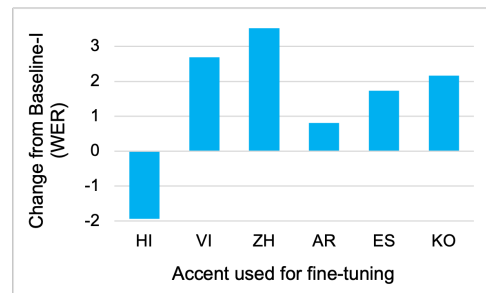

Figure 4: Individual Model-4s evaluated on the HI accent. All the bars except HI extend upwards, meaning that all the other single-accent models hinder performance on the HI accent.

evaluating on L2-ARCTIC Test, we evaluate on L1-ARCTIC Test and LibriSpeech in order to observe any changes in model performance on L1 English.

As Table 2 shows, our Model-1 achieves best performance on both Dev and Test of **L2-ARCTIC** as compared to our two baselines. On Test, our Model-1 acquires 25.66% improvement over our Baseline-I wav2vec 2.0 system on L2-ARCTIC (9.27 WER for our model vs. 12.47 WER for Baseline-I). This gain is not surprising and simply means that a model with access to L2 data for fine-tuning will improve over models fine-tuned

with L1 data (Baseline-II, which is fine-tuned on L1-ARCTIC) or not-fine-tuned at all (Baseline-I). Nor is performance on **L1-ARCTIC** surprising: a model fine-tuned with native data (Baseline-II) outperforms one fine-tuned with accented data (our Model-1), both of which outperform a model without fine-tuning (Baseline-I). These results, however, show that in absence of L1 data, L2 data can be valuable for improving ASR model performance even on L1. For **LibriSpeech**, Baseline-I, which is trained on LibriSpeech data, outperforms the two fine-tuned models (our Model-1 and Baseline-II). The reason is that these two latter

L1$_{removed}$	Baseline-I Test$_{zeroshot}$	Baseline-II Test$_{zeroshot}$	Model-3 Test$_{zeroshot}$	Model-3 Test$_{all}$
VI	23.30	28.81	18.81	9.43
ZH	14.85	19.32	12.13	9.08
AR	10.95	14.82	10.10	9.13
ES	10.48	13.48	8.89	**8.98**
KO	8.18	10.22	6.95	9.01
HI	**6.93**	**8.93**	**6.67**	9.11

Table 4: Model-3 setting, where a different accent is removed each run. Test$_{all}$ refers to Test of *all* 24 speakers, and Test$_{zeroshot}$ refers to Test of those four speakers who have L1$_{removed}$ accent. Baseline-I acquires 12.47 on Test$_{all}$, while Baseline-II acquires 15.95 on the same test set (i.e., Test$_{all}$).

L1	Baseline-I Test$_{L2}$	Baseline-II Test$_{L2}$	Model-1 Test$_{L2}$	Model-4 Test$_{L2}$	Model-4 Test$_{L1}$	Model-4 LS$_{Clean}$	Model-4 LS$_{Other}$
VI	23.30	28.81	15.14	**12.12**	2.02	3.08	6.96
ZH	14.85	19.32	11.49	**8.95**	1.82	2.84	6.22
AR	10.95	14.82	8.90	**6.92**	1.55	2.66	6.24
ES	10.48	13.48	8.92	**6.68**	1.56	2.53	6.11
KO	8.18	10.22	6.60	**4.99**	1.71	2.51	5.63
HI	6.93	8.93	5.51	**4.99**	1.52	2.36	6.05
Mean	12.45	15.93	9.43	7.44	1.70	2.66	6.20
SD	5.97	7.30	3.49	2.72	0.20	0.26	0.43

Table 5: Model-4 performance on L2 accent (Test$_{L2}$) and native accent (Test$_{L1}$, LS$_{Clean}$, LS$_{Other}$), compared with Baseline-I, Baseline-II, and Model-1. SD refers to the standard deviation.

models are fine-tuned on a domain that is different from LibriSpeech. That is, fine-tuning models on out-of-domain data will, and as we see here does, result in deterioration of performance on in-domain data. We also note that our Model-1's performance on LibriSpeech is worse than that of Baseline-II on both the 'Clean' (LS$_{Clean}$, native speech under quite recording environments), and 'Other' (LS$_{Other}$, both noisy environment and accented recordings), Dev and Test splits. This may be because LibriSpeech is mostly comprised of L1 data and the greater variability on our L2-ARCTIC Train set (24 non-native speakers in our Model-1 vs. 4 native speakers in Baseline-II).

Model-2 (speaker-independent, accent-dependent): While Model-1 mimics a situation where we have some training data from speakers that we serve (i.e., test on), this is rarely a realistic scenario. We instead switch to a speaker-independent (but still *accent-dependent*) setting, Split-2. We carry out four-fold cross-validation with the 24 speakers in the data, every time using 18 speakers (three speakers per accent) in Train[10]

and six speakers in Test (one per accent). We report the average of the four folds/runs, along with standard deviation.

As Table 3 shows, Model-2 performance is consistent with Model-1. Our Model-2 outperforms the two baselines on both Dev and Test, reaching 9.96 WER on Test compared to 12.47 for Baseline-I and 15.96 for Baseline-II. These results demonstrate that fine-tuning with multiple accents improves the accented ASR system without access to test speaker data.

Model-3 (speaker- and accent-independent): To evaluate performance on *unseen* accents, we adopt a zero-shot strategy by removing one accent at a time from both Train and Dev sets and evaluating on the Test set of the removed accent, Split-3. To evaluate model performance on each accent, we conduct six runs in total with one accent removed at a time.

As Table 4 shows, fine-tuning on accented speech benefits unseen accents and speakers (Model-3 setting). All the multi-accent, zero-shot models outperform Baseline-I and Baseline-II, which means each of the six accents benefit from other accents through this process of transfer

	VI	**ZH**	**AR**	**ES**	**KO**	**HI**
Baseline-I	23.30	14.85	10.95	10.48	8.18	6.93
VI-specific	12.12	13.62	13.01	9.95	8.55	9.62
ΔWER	-11.18	-1.23	2.06	-0.53	0.37	2.69
Δ%	-48.00	**-8.31**	18.84	-5.03	4.52	38.77
ZH-specific	20.37	8.95	11.42	9.79	6.82	10.91
ΔWER	-2.93	-5.90	0.47	-0.69	-1.36	3.98
Δ%	-12.58	-39.75	4.26	-6.62	**-16.67**	57.43
AR-specific	23.88	14.86	6.92	9.86	9.16	7.74
ΔWER	0.58	0.01	-4.03	-0.62	0.98	0.81
Δ%	2.47	0.07	-36.83	**-5.92**	11.94	11.69
ES-specific	20.71	13.99	11.00	6.68	7.92	8.66
ΔWER	-2.59	-0.86	0.05	-3.80	-0.26	1.73
Δ%	**-11.13**	-5.81	0.43	-36.23	-3.22	25.01
KO-specific	20.07	12.12	11.66	10.04	4.99	9.09
ΔWER	-3.23	-2.73	0.71	-0.44	-3.19	2.16
Δ%	-13.88	**-18.38**	6.45	-4.23	-39.04	31.17
HI-specific	26.18	18.39	13.51	11.90	10.72	4.99
ΔWER	2.88	3.54	2.56	1.42	2.54	-1.94
Δ%	**12.37**	23.82	23.35	13.55	31.01	-27.99

Table 6: Model-4 performance in the zero-shot setting. Bold fonts represent the accent whose WER drops the most in the zero-shot setting. For example, compared with Baseline-I, the VI-specific fine-tuning not only improves performance on VI (i.e., a drop in WER), but also improves on ZH despite ZH being the unseen accent. One notable pattern is that HI-specific fine-tuning only benefits HI-accented speech recognition while all the other fine-tuning hinder performance on the HI accent.

learning. Our results also show that, in absence of in-accent data, some unseen accents are easier for the model than others. For example, on Test$_{\text{zeroshot}}$, Vietnamese (VI) is the most challenging (with 18.81 WER) and Hindi (HI) is the least challenging (with only 6.67 WER).

	Test$_{\text{all}}$		**Test$_{\text{zeroshot-speaker}}$**	
L1	**Mean**	**SD**	**Mean**	**SD**
VI	12.67	0.38	14.28	4.87
ZH	9.65	0.31	11.26	3.03
AR	7.28	0.29	8.56	2.28
ES	6.95	0.26	7.76	3.99
KO	5.22	0.18	5.69	2.20
HI	5.27	0.11	5.79	1.12

Table 7: Model-5 performance on L2 accent. Test$_{\text{all}}$ contains utterances by all speakers within each L1 whereas Test$_{\text{zeroshot-speaker}}$ contains utterances by a single speaker that is absent in the training phase. Mean refers to the average WER over four folds for each L1, and SD refers to the standard deviation.

4.3 Accent-Specific Models

We evaluate the accent-dependent performance by fine-tuning our models on a single type of L1-specific accent at a time.

Model-4 (speaker-dependent, accent-dependent): The model is fine-tuned with Split-4 data to identify any accent-dependent training impact on downstream performance, as well as an upper bound on performance when the model is optimized for a single accent. In addition to evaluating on L2-ARCTIC Test, we test the model on L1-ARCTIC Test and LibriSpeech as a means to identify any degradation on L1 English data.

As Table 5 shows, while the multi-accent model (Model-1) outperforms Baseline-I for all six accents, all of the accent-specific models (Model-4 setting) outperform Model-1 on the Test$_{\text{L2}}$ setting despite the small amount of data (roughly five hours) used for fine-tuning each of the versions of Model-4. On average, Model-4 setting is two points WER better than Model-1. In addition, Model-4 type models (each of which is fine-tuned on one non-native accent) perform reasonably well

Model	Model output
Ref	at lake linderman i had one canoe very good peterborough canoe
VI	at LAY LINDEMAN i had one canoe very good PETERBORROUG CANOES
	A lake LNDER MAN i had one canoe very good BIET OF ROCK canoe
ZH	at lake LINGERMAN i had ONCE canoe very good PETERBROUGH canoe
	at lake LINERMAN i had one canoe very good PETERE BROUGHTA canoe
AR	at lake LUNDERBOGH i had one canoe very good BITTERBOROUGH canoe
	at lake LUNDERMAN i had one canoe very good BETTER BORT canoe
ES	at lake linderman i had one canoe a very good PETERBOURN canoe
	at lake linderman i had ONCE canoe very good PIERREBOROUGH canoe
KO	at lake linderman i had one canoe very good peterborough canoe
	at lake LINDEMAN i had ONCE canoe very good PITTEBRAUG canoe
HI	at lake LINDEMAN i had one canoe very good PETERBURGH canoe
	at lake linderman i had one canoe A very good PEACHERBROROU canoe

Table 8: Examples of transcription output of selected utterances from the Test set of Model-4 among all six L1s without a language model. Capitalized words indicate errors. We show samples from two speakers per accent.

on L1 data (Test$_{L1}$, LS$_{Clean}$, and LS$_{Other}$). Further, large accent-specific variability is observed across different model types on Test$_{L2}$ (SD = $[2.72 - 7.30]$), compared with native counterparts such as Test$_{L1}$ (SD = $[0.20 - 0.43]$). An interesting result is the apparent difficulty difference between different accents (HI and KO easiest, VI hardest), regardless of model types. We provide sample outputs from Model-4 in Table 8.

As shown in Table 6, we also perform accent-wise zero-shot evaluation. Results of this set of experiments reveal an interesting pattern: while fine-tuning on a single accent generally benefits *at least one other accent*, fine-tuning on the Hindi accent only benefits Hindi (the same accent) and hinders performance on *all the other accents*. Figure 3 and Figure 4 illustrate this observation.

Model-5 (speaker-independent and accent-dependent): This setup simulates a more realistic scenario where we target a single accent, without access to all speakers during development time. Thus, we use Split-5 data which mimics a speaker-independent setting. We cross-validate each L1 subset with one of the four speakers per fold. The hyper-parameters we use are those identified for Model-4. To evaluate the performance on each speaker, we conduct 24 folds in total with one speaker removed at a time, and report the average and standard deviation of the four folds per each accent.

As Table 7 shows, speaker-dependent variability is small for Test$_{all}$ (SD = $[0.11 - 0.38]$) but large for Test$_{zeroshot-speaker}$ (SD = $[1.12 - 4.87]$). These results suggest that individual speaker's differences may play an important role in how much performance gain can be obtained by fine-tuning.[11]

---
[11]For those speakers whose TOEFL scores are known (Zhao

5 Conclusion

We demonstrated potential of developing accent-independent and accent-dependent models that improve non-native speech recognition simply by fine-tuning the pre-trained wav2vec 2.0 model on a small amount of labeled data. Both the multi- and single-accent models improve performance on L2 English speakers. However, each accent benefits differently: results of the multi-accent, zero-shot experiments suggest that transfer learning on accent is possible and single-accent models improve the most for the target L2 accents.

As to future work, while we chose a language model-free setting to focus specifically on wav2vec 2.0's acoustic capacity, comparison with language model decoding would be a useful direction to explore as a way to gauge any further potential improvements a language model can bring. In addition, finding the optimal combination of accented speech datasets when there is no available dataset for a target accent (Model-3) may constitute another interesting direction. Finally, although we have offered a number of sample transcriptions from one of our models, a thorough error analysis on each experiment would help advance the research into improving ASR models for non-native English speakers. Since L2 English speakers have specific accent characteristics influenced by their native languages, an error analysis focused on each language as well as on groups or families of languages will likely aid effective model development. Future directions could also investigate different strategies for developing ASR systems for challenging languages such as Vietnamese.

---
et al., 2018), a strong negative correlation was observed between speaker-specific WERs of Baseline-I and speaker's TOEFL scores, $r(8) \approx -.77$, $p < .01$.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Amy Bearman, Kelsey Josund, and Gawan Fiore. 2017. Accent conversion using artificial neural networks. Technical report, Stanford University, Tech. Rep, Tech. Rep.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*.

David Crystal. 2003. *English as a global language*. Ernst Klett Sprachen.

Nilaksh Das, Sravan Bodapati, Monica Sunkara, Sundararajan Srinivasan, and Duen Horng Chau. 2021. Best of both worlds: Robust accented speech recognition with adversarial transfer learning. *arXiv preprint arXiv:2103.05834*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Jingyong Hou, Pengcheng Guo, Sining Sun, Frank K Soong, Wenping Hu, and Lei Xie. 2019. Domain adversarial training for improving keyword spotting performance of esl speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8122–8126. IEEE.

Hu Hu, Xuesong Yang, Zeynab Raeesy, Jinxi Guo, Gokce Keskin, Harish Arsikere, Ariya Rastrow, Andreas Stolcke, and Roland Maas. 2021. Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6408–6412. IEEE.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. 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)*, pages 7669–7673. IEEE.

John Kominek, Alan W Black, and Ver Ver. 2003. Cmu arctic databases for speech synthesis.

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433. IEEE.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020a. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Songxiang Liu, Disong Wang, Yuewen Cao, Lifa Sun, Xixin Wu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, et al. 2020b. End-to-end accent conversion without using native utterances. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293. IEEE.

Karen Livescu and James Glass. 2000. Lexical modeling of non-native speech for automatic speech recognition. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1683–1686. IEEE.

Marco Matassoni, Roberto Gretter, Daniele Falavigna, and Diego Giuliani. 2018. Non-native children speech recognition through transfer learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6229–6233. IEEE.

Tomohiro Nakatani. 2019. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech 2019*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Tan Tien Ping. 2008. *Automatic speech recognition for non-native speakers*. Ph.D. thesis, Université Joseph-Fourier-Grenoble I.

Kacper Radzikowski, Le Wang, Osamu Yoshie, and Robert Nowak. 2021. Accent modification for speech recognition of non-native speakers using neural style transfer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–10.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. 2021. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922. IEEE.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. 2018. Domain adversarial training for accented speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4854–4858. IEEE.

Thibault Viglino, Petr Motlicek, and Milos Cernak. 2019. End-to-end accented speech recognition. In *Interspeech*, pages 2140–2144.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. 2020. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE.

Zhirong Wang, Tanja Schultz, and Alex Waibel. 2003. Comparison of acoustic model adaptation techniques on non-native speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Haihua Xu, Yerbolat Khassanov, Zhiping Zeng, Eng Siong Chng, Chongjia Ni, Bin Ma, Haizhou Li, et al. 2020. Independent language modeling architecture for end-to-end asr. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7059–7063. IEEE.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE.

Guanlong Zhao, Sinem Sonsaat, Alif O Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. *Perception Sensing Instrumentation Lab*.