COUPLED TRAINING OF SEQUENCE-TO-SEQUENCE MODELS FOR ACCENTED SPEECH RECOGNITION

Vinit Unni[*], *Nitish Joshi*[*], *Preethi Jyothi*

Department of Computer Science and Engineering, IIT Bombay

ABSTRACT

Accented speech poses significant challenges for state-of-the-art automatic speech recognition (ASR) systems. Accent is a property of speech that lasts throughout an utterance in varying degrees of strength. This makes it hard to isolate the influence of accent on individual speech sounds. We propose *coupled training* for encoder-decoder ASR models that acts on pairs of utterances corresponding to the same text spoken by speakers with different accents. This training regime introduces an L2 loss between the attention-weighted representations corresponding to pairs of utterances with the same text, thus acting as a regularizer and encouraging representations from the encoder to be more accent-invariant. We focus on recognizing accented English samples from the Mozilla Common Voice corpus. We obtain significant error rate reductions on accented samples from a large set of diverse accents using coupled training. We also show consistent improvements in performance on heavily accented samples (as determined by a standalone accent classifier).

Index Terms— Accented speech recognition, sequence-to-sequence models with attention, coupled training

1. INTRODUCTION

Automatic speech recognition (ASR) technologies have achieved remarkable progress in recent years and are gaining widespread adoption in various applications. Despite these impressive advances, ASR performance is sub-par on speech that is not “typical”; for example, ASR performance degrades when evaluated on heavily accented speech. Labeled speech is plentiful for certain standard accents and limited for many underrepresented accents. How do we adapt an end-to-end ASR system trained on large amounts of a standard accent, using relatively smaller amounts of underrepresented accents, such that ASR performance on the latter improves while performance on the standard accent does not deteriorate? How can we effectively leverage the same text spoken by different speakers with different accents? These are the main questions we tackle in this work.

State-of-the-art sequence-to-sequence ASR systems, that directly learn transformations from speech to text, have surpassed the performance of traditional cascaded ASR systems in recent years [1]. Our proposed solution builds on top of an end-to-end sequence-to-sequence model with attention [2]. In this work, we focus on improved recognition for different accents of English. We use Mozilla’s Common Voice dataset [3] that consists of diverse speech samples spanning multiple speech accents. This dataset also has a substantial amount of overlap in content. That is, the same text is spoken by multiple speakers in a number of different accents. We exploit this feature and introduce a *coupled training* paradigm.

During coupled training, we feed pairs of utterances with the same underlying text as inputs to a sequence-to-sequence model. The encoder weights are shared across both utterances. Apart from the standard cross-entropy (CE) loss that drives the decoder to produce a character sequence for each utterance, we additionally impose an L2 loss between the context vectors at each decoder time-step. This encourages the context vectors for pairs of utterances corresponding to the same text to be close to each other despite varying accents. Having access to speech samples corresponding to the same text from diversely accented speakers allows us to be agnostic to the content and encourage representations from the encoder to be more invariant to accents.

To summarize, our contributions in this work are:

- We introduce a coupled training paradigm where pairs of utterances with the same underlying text are fed as inputs and an L2 loss is imposed between the context vectors for each utterance at each decoder time-step.
- We demonstrate the utility of coupled training by showing significant improvements in word error rates (WERs) on more than ten different accents.
- We present a thorough discussion analyzing how our coupled training benefits accented speech recognition.

2. RELATED WORK

Improving speech recognition for accented speech has remained a fairly active area of research. Traditional ASR systems tackled this problem by either changing the pronunciation dictionary [4, 5] or the acoustic model [6, 7, 8, 9]. There have also been attempts to augment the feature vector using accent specific input [10, 11]. More recent work has

[*] Joint first authors

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

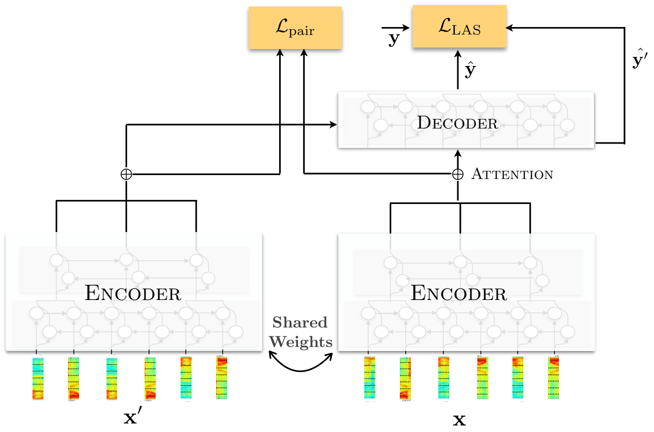

Fig. 1. Schematic diagram illustrating coupled training.

focused on using end-to-end ASR models to recognize accented speech. [12] integrated a hierarchical loss based on both phonemes and graphemes into an end-to-end model and showed improvements on multiple English dialects. There have also been attempts at exploring this problem within a multi-task framework [13, 14] where the auxiliary task involves predicting accents or accent-specific features. Features from these auxiliary tasks could also be fed back to the primary task in the form of accent embeddings [14, 15, 16]. Another interesting direction explored for accented speech recognition has been to use adversarial training to learn representations that are accent-invariant [17, 18]. [19] explored the use of a mixture of feature extractors where each individual extractor focused on a particular phone or accent class. [20] demonstrated how large error rate reductions can be obtained by fine-tuning only the initial encoder layers when very limited amount of accented speech is available.

3. METHODOLOGY

A key contribution of this work is incorporating coupled training within a basic LAS model, as illustrated in Figure 1. An LAS model consists of three components: an encoder, an attention module and a decoder. For a speech sample, $\mathbf{x} = \{x_1, \ldots, x_T\}$ with a label sequence $\mathbf{y} = \{y_1, \ldots, y_M\}$, LAS models the conditional probability $P(\mathbf{y}|\mathbf{x})$ as follows:

$$\begin{aligned}\Pr(\mathbf{y}|\mathbf{x}) &= \prod_{j=1}^{M} \Pr(y_j|\mathbf{x}, y_{1:j-1}) \\ &= \prod_{j=1}^{M} \Pr(y_j|\mathbf{h}, s_i, c_i) \qquad (1)\end{aligned}$$

where $\mathbf{h} = \textsc{Enc}(\mathbf{x})$ and $\{s_i, c_i\} = \textsc{Dec-Att}(\mathbf{h}, y_{1:j-1})$; we will define both these functions shortly. In the LAS framework, the encoder is a stacked Bidirectional Long Short Term Memory (BiLSTM) network and the decoder consists of unidirectional LSTM-based recurrent layers. The encoder, denoted by $\textsc{Enc}(\mathbf{x})$, is layered in a pyramidal structure as originally proposed for LAS: The bottommost encoder layer unfolds across the length of the utterance for T time-steps and each additional encoder layer on top reduces the number of effective time-steps by a factor of 2. $\textsc{Enc}(\mathbf{x})$ returns a sequence of encoder states $\mathbf{h} = \{h_1, \ldots, h_K\}$ over which the attention and decoder modules interact. (Here, K is the number of time-steps in the topmost encoder layer; in our implementation with 3 encoder layers, $K \approx \frac{T}{4}$ as there is no subsampling in the first layer.) For each decoder state s_{i-1}, LAS learns an attention distribution $\{\alpha_{i1}, \ldots, \alpha_{iK}\}$ that is used to linearly interpolate $\{h_1, \ldots, h_K\}$ to form a context vector c_i for every i^{th} decoder time-step:

$$c_i = \sum_{k=1}^{K} \alpha_{ik} h_k$$

Each attention weight α_{ik} is a function of the encoder state h_k and the decoder state s_{i-1} and is learned using an MLP network. $\textsc{Dec-Att}(\mathbf{h}, y_{1:j-1})$ returns the context vector c_i as well as the decoder state $s_i = LSTM(s_{i-1}, c_i, y_{i-1})$. The training objective to be maximized in LAS, with $\Pr(\mathbf{y}|\mathbf{x})$ defined as in Eq. (1), is:

$$\mathcal{L}_{\text{LAS}} = \log \Pr(\mathbf{y}|\mathbf{x}, y^*_{1:j-1})$$

where $y^*_{1:j-1}$ corresponds to the ground-truth of previous characters.

Coupled Training: In coupled training, we use a pair of utterances $\mathbf{x}$ and $\mathbf{x}'$ from different speakers with the same underlying label sequence $\mathbf{y}$. Let $\mathbf{h} = \textsc{Enc}(\mathbf{x})$ and $\mathbf{h}' = \textsc{Enc}(\mathbf{x}')$ and let the corresponding context vectors at each decoder time-step i be c_i and c'_i, respectively. Since the character sequence $\mathbf{y}$ is identical, both utterances will produce the same number of decoding time-steps. We introduce an L2 loss, $\mathcal{L}_{\text{pair}}$, across the context vectors:

$$\mathcal{L}_{\text{pair}} = \frac{1}{K} \sum_{i} ||c_i - c'_i||_2$$

We now optimize a linear combination of the LAS and coupled objectives:

$$\mathcal{L}_{\text{final}} = (1 - \lambda)\mathcal{L}_{\text{LAS}} + \lambda\mathcal{L}_{\text{pair}}$$

where $\lambda \in [0, 1]$ is a tunable hyperparameter.

Coupled training can be invoked with different training schedules. We could start with a fully trained LAS-model, followed by a pass of coupled training. Or, we could start from the very beginning with the combined training objective $\mathcal{L}_{\text{comb}}$ so that the L2 regularization over context vectors is effective right from the start. We will discuss which training schedule is more effective in Section 5.

Dataset	No. of sentences	Duration (hrs)
TRAIN-US	99933	115.32
0.25-NONUS	25764	30.40
0.5-NONUS	51339	60.66
TRAIN-IN	15766	20.13
TEST-US	11527	16.42
TEST-NONUS	13126	14.35
TEST-IN	2198	2.84

Table 1. Statistics for all the datasets.

4. DATASET CONSTRUCTION

We used the Mozilla Common Voice dataset (version 3) [3] for all our experiments. The original corpus contained a significant amount of redundant content with the same text rendered in speech by a number of different speakers. It also contained many samples that were not tagged with accent labels. We extracted a subset of the corpus and created train/test splits that were disjoint in speakers and sentences, and the train set contained many instances of the same text spoken by different speakers. (40.5% of the sentences in 0.5-NONUS are spoken by multiple speakers.) Our resulting splits are detailed in Table 1. We present both US-based and NONUS-based datasets. The latter has speech in the following accents: African (AFK), Australian (AU), Bermuda (BM), Canadian (CA), Great Britain (GB), Hongkong(HK), Indian (IN), Ireland (IR), Malaysia (ML), New-Zealand (NZ), Philippines (PHI), Scotland (SC), Singapore (SG), South-Atlantic (ST) and Wales (WL). GB, IN, AU and CA are the dominant accents and contribute towards 36%, 15%, 14% and 13% of the datasets, respectively, with the remaining accents contributing less than 4% each.[1]

We created 0.25-NONUS and 0.5-NONUS that contain speech samples from 15 different non-US accents and are approximately 25% and 50% the size of the TRAIN-US corpus respectively. We use these datasets to show how performance of coupled training varies with different amounts of accented speech. We also created TRAIN-IN consisting of speech samples only in the Indian accent. The latter was chosen because compared to all the other non-US accents, the Indian accented samples were particularly difficult for a baseline system trained only on TRAIN-US to recognize.

5. EXPERIMENTS AND RESULTS

5.1. Implementation Details

We utilized the ESPnet toolkit [21] for all our experiments and added new code to support coupled training. Our base LAS model consists of 2 VGG-ish convolutional layers followed by three 1024-sized BiLSTM layers, location-based attention and two 1024-sized decoder layers. We regularized the model using a dropout rate of 0.5. The model was optimized using Adadelta [22] with a starting learning rate of 1 and an epsilon value of $1e^{-8}$. The scaling factor λ for the coupled loss $\mathcal{L}_{\text{pair}}$ was tuned on a held-out dataset and set to 0.0001. We used 150 sub-word units and a scheduled sampling rate of 0.3.

5.2. How much does coupled training help?

We use the TRAIN-US and 0.25-NONUS datasets to train all the systems defined in this section. We define four baseline LAS systems that are all trained by optimizing $\mathcal{L}_{\text{LAS}}$: 1) **US**: Only TRAIN-US is used during training. 2) **MIXED-1.25**: A mixture of TRAIN-US and 0.25-NONUS is used during training. 3) **US+FT-0.25**: A fully trained US model is fine-tuned using 0.25-NONUS. 4) **US+FT-1.25**: A fully trained US model is fine-tuned using a mixture of 0.25-NONUS and TRAIN-US.

We present five systems that use coupled training (i.e. by optimizing $\mathcal{L}_{\text{comb}}$): 1) **MIXED$_2$+C-1.25**: The MIXED-1.25 baseline trained for two epochs is optimized with coupled training using a mixture of TRAIN-US + 0.25-NONUS data. 2) **C-1.25**: A mixture of TRAIN-US and 0.25-NONUS is used for coupled training from scratch. 3) **US+C-0.25**: A fully trained US model is fine-tuned with coupled training using 0.25-NONUS. 4) **US+C-1.25**: A fully trained US model is fine-tuned with coupled training using a mixture of 0.25-NONUS and TRAIN-US. 5) **C-1.25+C-0.25**: A fully trained C-1.25 model is fine-tuned with coupled training using 0.25-NONUS.

Table 2 shows CERs and WERs for all nine systems. (Utterance pairs for coupled training were constructed by distributing utterances corresponding to the same text into pairs; any odd sample left out was not considered for coupled training.) We see significant WER reductions (at $p < 0.001$ using the MAPSSWE test [23]) on both TEST-US and TEST-NONUS using our best form of coupled training (C-1.25+C-0.25) compared to the best baseline system (i.e. US+FT-0.25 for TEST-NONUS and US+FT-1.25 for TEST-US).

System	TEST-US	TEST-NONUS
US	19.29/36.4	37.25/58.93
MIXED-1.25	21.09/39.06	35.44/57.19
US+FT-0.25	20.08/37.99	31.90/53.05
US+FT-1.25	17.81/34.45	32.15/53.21
MIXED$_2$+C-1.25	18.95/36.80	33.0/55.14
US+C-0.25	20.94/39.68	31.58/53.25
C-1.25	17.34/33.84	31.06/51.96
US+C-1.25	16.92/33.45	31.18/52.47
C-1.25+C-0.25	**16.84/33.28**	**28.39/48.82**

Table 2. CER/WER on TEST-US and TEST-NONUS from baseline and coupled training systems.

[1] More details about our data splits are available at www.cse.iitb.ac.in/~vinit/MCV_splits/.

System	TEST-US	TEST-NONUS
US+FT-0.25	20.08/37.99	31.90/53.05
C-1.25	17.34/33.84	31.06/51.96
MIXED-1.5	22.52/39.92	36.07/56.00
US+FT-0.5	19.61/37.12	30.62/51.45
US+FT-1.5	17.54/34.24	30.43/51.27
C-1.5	16.84/33.50	29.07/49.88
C-1.5+C-0.5	**16.20/32.41**	**26.19/45.64**

Table 3. CER/WER for coupled training using 0.5-NONUS.

5.3. Varying amounts of accented speech

Similar to the systems shown in Table 2, we train both baseline systems and coupled training systems using the larger 0.5-NONUS dataset. This highlights the effect of using more accented data with the coupled training paradigm. From Table 3, we see that C-1.5+C-0.5 significantly outperforms the best baseline system US+FT-1.5 (at $p < 0.001$). Coupled training provides consistent performance benefits even with larger amounts of accented speech.

5.4. Focusing on a single accent

We focus on the effect of coupled training when a small amount of accented speech is available for a single accent (IN).Table 4 shows the performance of two fully trained US models fine-tuned with TRAIN-IN using CE and coupled loss, respectively. The last row denotes a model that is trained on TRAIN-US+TRAIN-IN from scratch using coupled loss, followed by fine-tuning on TRAIN-IN using coupled loss. Even with limited data (i.e. $\approx$ 20 hours of speech), this model gives significant improvements in WERs (at $p < 0.001$) compared to a standard fine-tuning pass using CE loss.

5.5. Discussion

Breakdown across accents: In Fig. 2, we analyze improvements for each individual accent by comparing US+FT-1.5 vs. C-1.5+C-0.5. Our model consistently outperforms the baseline systems across accents and gets notably larger improvements on strong accents such as Indian and Malaysian.
Clearly accented samples: We use a standalone BiLSTM-based accent classifier, that was trained on all the non-US training data, to identify samples with clearly discernible accents.Test samples that were correctly predicted by this classifier belonging to accents with an average confidence score of 0.8 or higher are referred to as being "heavily accented" and the remaining test samples are "non heavily accented". Fig. 3 shows the absolute improvements in WER for five different accents. Improvements on the heavily accented samples in IN and GB accents are larger than on the less heavily accented samples; these are also the two accents whose test samples are correctly predicted with the highest confidence scores.

System	TEST-US	TEST-IN
US+FT-IN	26.92/47.51	35.3/57.42
US+C-IN	26.57/47.21	34.60/56.67
US+IN+C-IN	**17.74/35.08**	**31.55/53.28**

Table 4. CER/WERs from baseline and coupled training using only Indian accented samples.

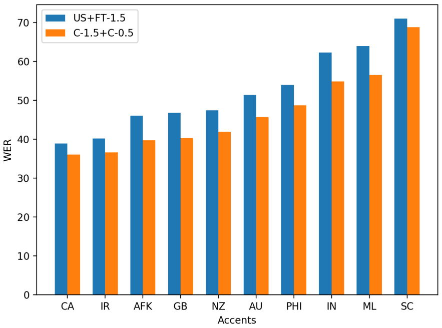

Fig. 2. WERs for US+FT1.5 and C-1.5+C-0.5 systems across different accents.

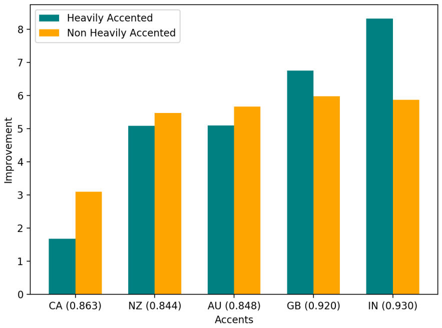

Fig. 3. WER reductions for heavy and non-heavy accented samples. Confidence scores are within "()" for each accent.

6. CONCLUSION

In this work, we proposed a new coupled training paradigm which imposes an L2 regularization between the context vectors for two utterances with the same text. We showed significant improvements in WERs across diverse accented samples and data settings. In future work, we will devise coupled training paradigms for CTC and hybrid end-to-end models.

7. ACKNOWLEDGEMENTS

The last author gratefully acknowledges support from a Google Faculty Research Award for this project.

8. REFERENCES

[1] Chung-Cheng Chiu, Tara Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani, "State-of-the-art speech recognition with Sequence-to-Sequence models," in *Proceedings of ICASSP*, 2017.

[2] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, Attend and Spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of ICASSP*, 2016.

[3] "Mozilla common voice," `https://voice.mozilla.org/en`, 2017.

[4] J J Humphries and P C Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition," in *Proceedings of EUROSPEECH*, 1997.

[5] Felix Weninger, Yang Sun, Junho Park, Daniel Willett, and Puming Zhan, "Deep learning based Mandarin accent identification for accent robust ASR," in *Proceedings of Interspeech*, 2019.

[6] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Dan Jurafsky, Rebecca Starr, and Su-Youn Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin.," in *Proceedings of Interspeech*, 2005.

[7] Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel, "Speech recognition of multiple accented english data using acoustic model interpolation," in *Proceedings of EUSIPCO*, 2014.

[8] Herman Kamper and Thomas Niesler, "Multi-accent speech recognition of Afrikaans, Black and White varieties of South African English," in *Proceedings of Interspeech*, 2011.

[9] Sanghyun Yoo, Inchul Song, and Y. Bengio, "A highly adaptive acoustic model for accurate multi-dialect speech recognition," in *Proceedings of ICASSP*, 2019.

[10] Mingming Chen, Zhanlei Yang, Jizhong Liang, Yanpeng Li, and Wenju Liu, "Improving deep neural networks based multi-accent Mandarin speech recognition using i-vectors and accent-specific top layer," in *Proceedings of Interspeech*, 2015.

[11] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation," in *Proceedings of Interspeech*, 2014.

[12] Kanishka Rao and Haim Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *Proceedings of ICASSP*, 2017.

[13] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *Proceedings of ICASSP*, 2018.

[14] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved accented speech recognition using accent embeddings and Multi-task learning," in *Proceedings of Interspeech*, 2018.

[15] Thibault Viglino, Petr Motlicek, and Milos Cernak, "End-to-End accented speech recognition," in *Proceedings of Interspeech*, 2019.

[16] Han Zhu, Li Wang, Pengyuan Zhang, and Yonghong Yan, "Multi-accent adaptation based on gate mechanism," in *Proceedings of Interspeech*, 2019.

[17] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," in *Proceedings of ICASSP*, 2018.

[18] Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh, "Adversarial learning of raw speech features for domain invariant speech recognition," in *Proceedings of ICASSP*, 2018.

[19] Abhinav Jain, Vishwanath P. Singh, and Shakti P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," in *Proceedings of Interspeech*, 2019.

[20] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias, "Personalizing ASR for dysarthric and accented speech with limited data," in *Proceedings of Interspeech*, 2019.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End speech processing toolkit," in *Proceedings of Interspeech*, 2018.

[22] Matthew D. Zeiler, "ADADELTA: An adaptive learning rate method," *ArXiv*, vol. abs/1212.5701, 2012.

[23] "The NIST scoring toolkit (SCTK)," `https://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm`, 2018.