DOMAIN ADAPTATION VIA TEACHER-STUDENT LEARNING FOR END-TO-END SPEECH RECOGNITION

Zhong Meng, Jinyu Li, Yashesh Gaur, Yifan Gong

Microsoft Corporation, Redmond, WA, USA

ABSTRACT

Teacher-student (T/S) has shown to be effective for domain adaptation of deep neural network acoustic models in hybrid speech recognition systems. In this work, we extend the T/S learning to large-scale unsupervised domain adaptation of an attention-based end-to-end (E2E) model through two levels of knowledge transfer: teacher's token posteriors as soft labels and one-best predictions as decoder guidance. To further improve T/S learning with the help of ground-truth labels, we propose adaptive T/S (AT/S) learning. Instead of conditionally choosing from either the teacher's soft token posteriors or the one-hot ground-truth label, in AT/S, the student always learns from both the teacher and the ground truth with a pair of adaptive weights assigned to the soft and one-hot labels quantifying the confidence on each of the knowledge sources. The confidence scores are dynamically estimated at each decoder step as a function of the soft and one-hot labels. With 3400 hours parallel close-talk and far-field Microsoft Cortana data for domain adaptation, T/S and AT/S achieves 6.3% and 10.3% relative word error rate improvement over a strong E2E model trained with the same amount of far-field data.

Index Terms— domain adaptation, teacher-student learning, end-to-end, encoder-decoder, speech recognition

1. INTRODUCTION

Recently, with the advancement of deep learning, great progress has been made in end-to-end (E2E) automatic speech recognition (ASR). With the goal of directly mapping a sequence of speech frames to a sequence of output tokens, an E2E ASR system incorporates the acoustic model, language model and pronunciation model of a conventional ASR system into a single deep neural network (DNN). The most dominant approaches for E2E ASR include connectionist temporal classification (CTC) [1, 2], recurrent neural network transducer (RNNT) [3] and attention-based encoder-decoder (AED) models [4, 5, 6].

However, the performance of E2E ASR degrades significantly when an acoustic mismatch exists between training and test conditions. An intuitive solution is domain adaptation where a well-trained source-domain E2E model is adapted to the data in the target domain. Different from speaker adaptation, domain adaptation allows for the usage of a large amount of adaptation data in both source and target domains.

There has been plenty of domain adaptation methods for hybrid systems that we can leverage for adapting E2E systems. One popular approach is the adversarial learning [7, 8, 9] in which a deep intermediate feature is learned to be discriminative in senone classification and invariant to the shifts between source and target domains. Adversarial domain adaptation is suitable for the situation where no transcription or *parallel* adaptation data in both domains are available. However, in speech area, a parallel sequence of target-domain data can be easily simulated from the source-domain data such that the speech from both domains are frame-by-frame synchronized. To take advantage of this, teacher-student (T/S) learning [10] was proposed for the *unsupervised* domain adaptation of acoustic models in DNN-hidden Markov model (HMM) hybrid systems [11]. In T/S learning, the Kullback-Leibler (KL) divergence between the output senone distributions of teacher and student acoustic models given parallel source and target domain data at the input is minimized by updating only the student model parameters. T/S training was shown to outperform the cross entropy training directly using the hard label in the target domain [11, 12, 13, 14, 15].

One drawback of unsupervised T/S learning is that, the teacher model is not perfect and will sometimes make inaccurate predictions that mislead the student model toward suboptimal directions. To overcome this, one-hot ground-truth labels are used to compensate for teacher's imperfections. Hinton et al. proposed interpolated T/S (IT/S) learning [16] to interpolate the teacher's soft class posteriors with one-hot ground truth using a pair of globally fixed weights. However, the optimal weights are data-dependent and can only be determined through careful tuning on a dev set. More recently, conditional T/S (CT/S) learning was proposed in [14] where the student model selectively chooses to learn from either the teacher or the ground truth depending on whether the teacher's prediction is correct or not. CT/S does not disturb the statistical relationships among classes naturally embedded in the class posteriors and achieves significant word error rate (WER) improvement over T/S for domain adaptation on CHiME-3 dataset [17].

In this work, we focus on the domain adaptation of AED

models for E2E ASR by using T/S learning which was previously applied to learn small-footprint AED models in [18, 19, 20] by distilling knowledge from a large powerful teacher AED. For unsupervised domain adaptation, we extend T/S learning to AED models by introducing a two-level knowledge transfer: in addition to learning from the teacher's soft token posteriors, the student AED also conditions its decoder on the one-best token sequence decoded by the teacher AED.

We further propose an *adaptive T/S (AT/S) learning* method to improve T/S learning using ground-truth labels. By taking advantage of both IT/S and CT/S, AT/S *adaptively* assigns a pair of weights to the teacher's soft token posteriors and the one-hot ground-truth label at each decoder step depending on the confidence scores on each of the labels. The confidence scores are dynamically estimated as a function of soft and one-hot labels. The student AED learns from an adaptive linear combination of both labels. AT/S inherits the linear interpolation of soft and one-hot labels from IT/S and borrows from CT/S the judgement on the credibility of both knowledge sources before merging them. It is expected to achieve improved performance over the other T/S methods for domain adaptation. As a general deep learning method, AT/S can be widely applied to the domain adaptation or model compression of any DNN.

With 3400 hours close-talk and far-field Microsoft Cortana data for domain adaptation, T/S learning achieves up to 24.9% and 6.3% relative WER gains over close-talk and far-field baseline AEDs, respectively. AT/S improves the close-talk and far-field AEDs by 28.2% and 10.3%, respectively, consistently outperforming IT/S and CT/S.

2. ATTENTION-BASED ENCODER-DECODER (AED) MODEL

In this work, we perform domain adaptation on AED models [4, 5, 6]. AED model was first introduced in [21, 22] for neural machine translation. Without any conditional independence assumption as in CTC [1], AED was successfully applied to to E2E ASR in [4, 5, 6] and has recently achieved superior performance to conventional hybrid systems in [23].

AED directly models the conditional probability distribution $P(\mathbf{Y}|\mathbf{X})$ over sequences of output tokens $\mathbf{Y} = \{y_1, \ldots, y_L\}$ given a sequence of input speech frames $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as below:

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{l=1}^{L} P(y_l|\mathbf{Y}_{0:l-1}, \mathbf{X}). \quad (1)$$

To achieve this, the AED model incorporates an encoder, a decoder and an attention network. The encoder maps a sequence of input speech frames $\mathbf{X}$ into a sequence of high-level features $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$ through an RNN. An attention network is used to determine which encoded features in $\mathbf{H}$

should be attended to predict the output label y_l and to generate a context vector $\mathbf{z}_l$ as a linear combination of $\mathbf{H}$ [4]. A decoder is used to model $P(\mathbf{Y}|\mathbf{H})$ which is equivalent to $P(\mathbf{Y}|\mathbf{X})$. At each time step t, the decoder RNN takes the sum of the previous token embedding $\mathbf{e}_{l-1}$ and the context vector $\mathbf{z}_{l-1}$ as the input to predict the conditional probability of each token, i.e., $P(u|\mathbf{Y}_{0:l-1}, \mathbf{H}), u \in \mathbb{U}$, at the decoder step l, where $\mathbb{U}$ is the set of all the output tokens:

$$\mathbf{q}_l = \text{RNN}^{\text{dec}}(\mathbf{q}_{l-1}, \mathbf{z}_{l-1}), \quad (2)$$

$$[P(u|\mathbf{Y}_{0:l-1}, \mathbf{X})]_{u \in \mathbb{U}} = \text{softmax}\left[K_y(\mathbf{q}_l + \mathbf{z}_l) + \mathbf{b}_y\right], \quad (3)$$

where $\mathbf{q}_l$ is the hidden state of the decoder RNN. bias $\mathbf{b}_y$ and the matrix K_y are learnable parameters.

An AED model is trained to minimize the following cross-entropy (CE) loss on the training corpus $\mathbb{T}_r$.

$$\mathcal{L}_{\text{CE}}(\theta, \mathbb{T}_r) = - \sum_{(\mathbf{X}, \mathbf{Y}^G) \in \mathbb{T}_r} \sum_{l=1}^{L^G} \log P(y_l^G|\mathbf{Y}_{0:l-1}^G, \mathbf{X}; \theta) \quad (4)$$

where $\mathbf{Y}^G = \{y_1^G, \ldots, y_{L^G}^G\}$ is the sequence of grouth-truth tokens, L^G represents the number of elements in $\mathbf{Y}^G$ and θ denotes all the model parameters in AED.

3. T/S LEARNING FOR UNSUPERVISED DOMAIN ADAPTATION OF AED

For unsupervised domain adaptation, we want to make use of a large amount of *unlabeled* data that is widely available. As shown in Fig. 1, with T/S learning, only two sequences of parallel data are required: an input sequence of source-domain speech frames to the teacher AED $\mathbf{X}^T = \{\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T\}$ and an input sequence of target-domain speech frames to the student model $\mathbf{X}^S = \{\mathbf{x}_1^S, \ldots, \mathbf{x}_N^S\}$. $\mathbf{X}^T$ and $\mathbf{X}^S$ are parallel to each other, i.e, each pair of $\mathbf{x}_n^S$ and $\mathbf{x}_n^T, \forall n \in \{1, \ldots, N\}$ are frame-by-frame synchronized. For most domain adaptation tasks in ASR, such as adapting from clean to noisy speech, close-talk to far-field speech, wide-band to narrow-band speech, the parallel data in the target domain can be easily simulated from the data in the source domain [11, 13].

Our goal is to train a student AED that can accurately predict the tokens of the target-domain data by forcing the student to emulate the behaviors of the teacher. To achieve this, we minimize the Kullback-Leibler (KL) divergence between the *token-level* output distributions of the teacher and the student AEDs given the parrallel data $\mathbf{X}^T$ and $\mathbf{X}^S$ are fed as the input to the AEDs. The KL divergence between the token-level output distributions of the teacher and student AEDs are formulated below

$$\sum_{l=1}^{L^T} \sum_{u \in \mathbb{U}} P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T; \theta^T) \log \left[\frac{P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T; \theta^T)}{P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^S; \theta^S)} \right],$$

$$(5)$$
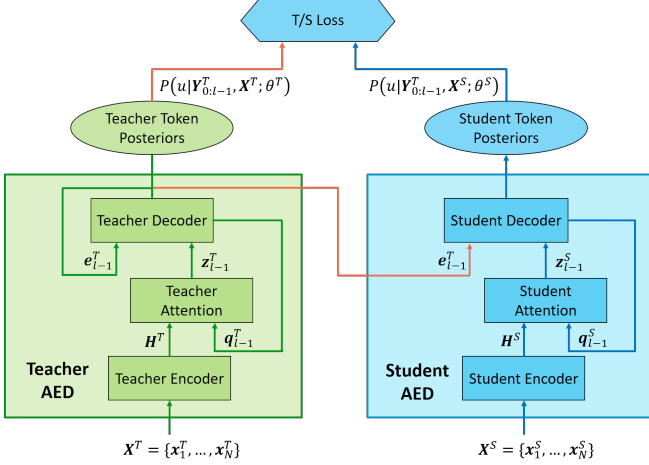
Fig. 1. T/S learning for unsupervised domain adaptation of AED model for E2E ASR. The two orange lines signify the two-level knowledge transfer.

where $\mathbf{Y}^T = \{y_1^T, \ldots, y_{L^T}^T\}$ is the sequence of *one-best* token sequence decoded by the teacher AED as follows

$$y_l^T = \arg\max_{u \in \mathbb{U}} P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T), \quad l = 1, \ldots, L^T, \quad (6)$$

where L^T is the number of tokens in $\mathbf{Y}^T$, and θ^T, θ^S denote all the parameters in the teacher and student AED models, respectively. Note that, for unsupervised domain adaptation, the teacher AED can only condition its decoder on the token y_{l-1}^T predicted at the previous step since the ground-truth labels $\mathbf{Y}^G$ are not available. We minimize the KL divergence with respect to θ^S while keeping θ^T fixed on the adaptation data corpus $\mathbb{A}$, which is equivalent to minimizing the token-level T/S loss function below:

$$\mathcal{L}_{\text{TS}}(\theta^S, \mathbb{A}) = - \sum_{(\mathbf{X}^T, \mathbf{X}^S) \in \mathbb{A}} \sum_{l=1}^{L^T} \sum_{u \in \mathbb{U}} P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T; \theta^T)$$
$$\log P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^S; \theta^S). \quad (7)$$

The steps of token-level T/S learning for unsupervised domain adaptation of AED model are summarized as follows:

1. Clone the student AED from a teacher AED well-trained with transcribed source-domain data by minimizing Eq. (4).

2. Forward-propagate the source-domain data $\mathbf{X}^T$ through the teacher AED, generate teacher's one-best token sequence $\mathbf{Y}^T$ using Eq. (6) and teacher's soft posteriors for each decoder step $P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T; \theta^T), u \in \mathbb{U}$ by Eqs. (2) and (3).

3. Forward-propagate the target-domain data $\mathbf{X}^S$ (*parallel* to $\mathbf{X}^T$) through the student AED, generate stu-

dent's soft posteriors for each teacher's decoder step $P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^S; \theta^S), u \in \mathbb{U}$ by Eqs. (2) and (3).

4. Compute error signal of the T/S loss function in Eq. (7) , back-propagate the error through student AED and update the parameters of the student AED.

5. Repeat Steps 2 to 4 until convergence.

After T/S adaptation, only the well-trained student AED is used for testing and the teacher AED is discarded after T/S learning.

As can be seen from Eqs. (6) and (7), to extend T/S learning to AED-based E2E models, two levels of knowledge transfer are involved: 1) the student learns from the teacher's soft token posteriors $P(u|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^T; \theta^T)$ at each decoder step; 2) the student AED conditions its decoder on the previous token y_{l-1}^T predicted by the teacher to make the current prediction.

Sequence-level T/S learning [18, 24] is another method for unsupervised domain adaptation in which a KL divergence between the *sequence-level* output distributions of the teacher and student AEDs are minimized. Equivalently, we minimize the sequence-level T/S loss function below with respect to θ^S

$$\mathcal{L}_{\text{TS-SEQ}}(\theta^S, \mathbb{A})$$
$$= - \sum_{(\mathbf{X}^T, \mathbf{X}^S) \in \mathbb{A}} \sum_{\mathbf{V} \in \mathbb{V}} P(\mathbf{V}|\mathbf{X}^T; \theta^T) \log P(\mathbf{V}|\mathbf{X}^S; \theta^S)$$
$$\approx - \sum_{(\mathbf{X}^T, \mathbf{X}^S) \in \mathbb{A}} \log P(\mathbf{Y}^T|\mathbf{X}^S; \theta^S)$$
$$= - \sum_{(\mathbf{X}^T, \mathbf{X}^S) \in \mathbb{A}} \sum_{l=1}^{L^T} \log P(y_l^T|\mathbf{Y}_{0:l-1}^T, \mathbf{X}^S; \theta^S), \quad (8)$$

where $\mathbb{V}$ is the set of all possible token sequences and the teacher's sequence-level output distribution $P(\mathbf{V}|\mathbf{X}^T)$ is approximated by $\mathbb{1}[\mathbf{V} = \mathbf{Y}^T]$ for easy implementation. $\mathbb{1}[\cdot]$ is an indicator function which equals to 1 if the condition in the squared bracket is satisfied and 0 otherwise.

From Eq. (8), we see that only one level of knowledge transfer exists in sequence-level T/S, i.e., the one-best token sequence $\mathbf{Y}^T$ decoded by the teacher AED. The student AED learns from $\mathbf{Y}^T$ and conditions its decoder on it at each step. Different from token-level T/S, in sequence-level T/S, one-hot labels in $\mathbf{Y}^T$ are used as training targets of the student AED instead of the soft token posteriors.

4. ADAPTIVE T/S (AT/S) LEARNING FOR SUPERVISED DOMAIN ADAPTATION OF AED

In this section, we want to make good use of the ground-truth labels of the adaptation data to further improve the T/S domain adaptation. Note that different from unsupervised T/S in Section 3, in supervised domain adaptation, the teacher AED

conditions its decoder on the ground-truth token instead of its previous decoding result because the token transcription $\mathbf{Y}^G$ is available in addition to $\mathbf{X}^S$ and $\mathbf{X}^T$.

One shortcoming of unsupervised T/S learning is that the teacher model can sporadically predict inaccurate token posteriors which misleads the student AED towards suboptimal performance. One-hot ground-truth labels can be utilized to alleviate this issue. One possible solution is the interpolated T/S (IT/S) learning [16] in which a weighted sum of teacher's soft posteriors and the one-hot ground truth is used as the target to train the student AED. A pair of *global* weights summed to be one is applied to each pair of soft and one-hot labels. However, the optimal global weights are hard to determine because they are data-dependent and need to be carefully tuned on a dev set.

To address this issue, conditional T/S learning (CT/S) [14] was proposed recently in which the student selectively chooses to learn from either the teacher AED or the ground truth conditioned on whether the teacher AED can correctly predict the ground-truth labels. CT/S have shown significant WER improvements over T/S and IT/S for both domain and speaker adaptation on CHiME-3 dataset. However, in CT/S, the student is still not "smart" enough because, for each token, the student AED solely relies on either the teacher's posteriors or the ground truth instead of dynamically extracting useful knowledge from both.

To further improve the effectiveness of knowledge transfer, we propose an *adaptive teacher-student (AT/S) learning* method by taking advantage of both CT/S and IT/S. As shown in Fig. 2, instead of assigning a fixed pair of soft weight w and one-hot weight $(1-w)$ for all the decoder steps, we adaptively weight the teacher's soft posteriors at the lth decoder step, $P(u|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T), u \in \mathbb{U}$, by $w_l \in [0,1]$ and the one-hot vector of the lth token in the ground-truth sequence $\mathbf{Y}^G$ by $(1-w_l)$. In order to quantify the value of the knowledge to be transferred, w_l should be positively correlated with a confidence score c_l on the teacher's prediction on token posteriors, while $(1 - w_l)$ should be positively correlated with a confidence score on the ground truth d_l. To achieve this, we compute w_l by normalizing c_l against its summation with d_l.

$$w_l = \frac{c_l}{c_l + d_l} \tag{9}$$

It is in general true that the higher posterior $P(y^G_l|\mathbf{Y}^G_{0:l-1}, \mathbf{X}^T;\theta^T)$ a teacher assigns to the correct (ground-truth) token y^G_l, the more accurate the teacher's soft posteriors are at this decoder step. Therefore, the confidence score c_l on teacher's soft posteriors $P(u|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T), u \in \mathbb{U}$ can be any *monotonically increasing* function of the correct token posterior predicted by the teacher $P(y^G_l|\mathbf{Y}^G_{0:l-1}, \mathbf{X}^T;\theta^T)$, while the confidence score d_l on the one-hot ground truth can be any monotonically increasing function of $(1 - P(y^G_l|\mathbf{Y}^G_{0:l-1}, \mathbf{X}^T;$
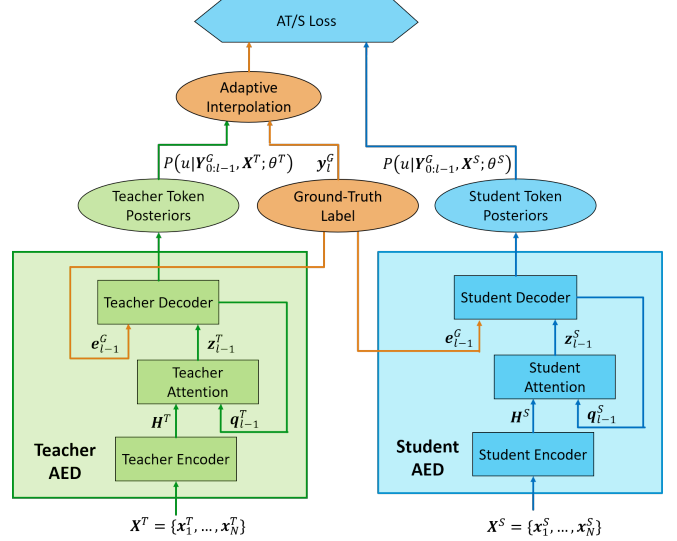
Fig. 2. Adaptive T/S (AT/S) learning for supervised domain adaptation of AED model for E2E ASR.

$\theta^T))$ as follow

$$c_l = f_1(P(y^G_l|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T)), \tag{10}$$

$$d_l = f_2(1 - P(y^G_l|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T)), \tag{11}$$

where both f_1 and f_2 are any monotonically increasing functions on the interval $[0,1]$. In this work, we simply assume that f_1 and f_2 are both power functions of the same form, i.e., $f_1(x) = f_2(x) = x^\lambda$, $\lambda > 0$. Note that w_l equals to $P(y^G_l|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T)$ when $\lambda = 1$.

In AT/S, a linear combination of the teacher's soft posteriors and the one-hot ground truth weighted by w_l and $(1-w_l)$, respectively, is used as the training target for the student AED at each decoder step. The loss function of AT/S is then formulated below

$$\mathcal{L}_{\text{ATS}}(\theta^S,\mathbb{A})=-\sum_{(\mathbf{X}^T,\mathbf{X}^S,\mathbf{Y}^G)\in\mathbb{A}} \sum_{l=1}^{L^G}\sum_{u\in\mathbb{U}} \left[w_l P(u|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T) \right.$$
$$\left. + (1-w_l)\mathbb{1}[u=y^G_l]\right] \log P(u|\mathbf{Y}^G_{0:l-1},\mathbf{X}^S;\theta^S). \tag{12}$$

The steps of AT/S learning for supervised domain adaptation of AED model are summarized as follows:

1. Perform token-level unsupervised T/S by following the steps in Section 3.

2. Forward-propagate the *parallel* source and target domain data $\mathbf{X}^T$ and $\mathbf{X}^S$ through the teacher and student AEDs, generate teacher and student's soft posteriors $P(u|\mathbf{Y}^G_{0:l-1},\mathbf{X}^T;\theta^T)$ and $P(u|\mathbf{Y}^G_{0:l-1}, \mathbf{X}^S;\theta^S), u \in \mathbb{U}$ for each decoder step by Eqs. (2) and (3).

3. Compute the confidence scores c_l and d_l for teacher's soft posteriors and one-hot vector of ground truth y_l^G by Eqs. (10) and (11), compute the adaptive weight w_l by Eq. (9).

4. Compute error signal of the AT/S loss function in Eq. (12) , back-propagate the error through student AED and update the parameters of the student AED.

5. Repeat Steps 2 to 4 until convergence.

AT/S is superior to IT/S in that the combination weights for soft and one-hot labels at each decoder step are adaptively assigned according to the confidence score on both labels. AT/S will degenerate to IT/S if the combination weights w_l are fixed globally. Compared to CT/S, in AT/S, the student always adaptively learns from both the teacher's soft posteriors and the one-hot ground truth rather than choosing either of them depending on the correctness of teacher's prediction.

5. EXPERIMENTS

In the experiment, we adapt a close-talk AED model to the far-field data through various T/S learning methods with parallel close-talk and far-field Microsoft Cortana data for E2E ASR.

5.1. Data Preparation

For both training and adaptation, close-talk data consisting of 3400 hours of Microsoft internal live US English Cortana utterances are collected through a number of deployed speech services including voice search and SMD. We simulate 3400 hours of far-field Microsoft Cortana data by convolving the close-talk signal with different room impulse responses and adding various environmental noise for both training and adaptation. The 3400 hours far-field data is *parallel* with the 3400 hours close-talk data. We collect 17.5k far-field utterances (about 19 hours) from Harman Kardon (HK) speaker as the test set.

80-dimensional log Mel filter bank features are extracted from the speech signal in all the training, adaptation and test set every 10 ms over a 25 ms window. We stack 3 consecutive frames and stride the stacked frame by 30 ms, to form a sequence of 240-dimensional input speech frames. We first generate 34k mixed units as in [25] based on the training transcription and then tokenize the training, adaptation transcriptions to generate mixed unit sequences serving as the training targets. A special token <space> in between every two adjacent words are inserted to indicate word boundaries and <sos>, <eos> are added to the beginning and the end of each label sequence, respectively, to represent sentence boundaries.

5.2. AED Baseline System

We first train an AED model predicting 34k mixed units with 3400 hours close-talk training data and it ground-truth labels for E2E ASR. The encoder is a bi-directional gated recurrent units (GRU)-recurrent neural network (RNN) [21, 26] with 6 hidden layers, each with 512 hidden units. Layer normalization [27] is applied for each encoder hidden layer. Each of the 34k mixed units is represented as a 512-dimensional embedding vector. The decoder is a uni-directional GRU-RNN with 2 hidden layers, each with 512 hidden units. The 34k-dimensional output layer of the decoder predicts the posteriors of all the mixed units in the vocabulary. During training, scheduled sampling [28] is applied to the decoder with a sampling probability starting at 0.0 and gradually increasing to 0.4 [23]. Dropout [29] with a probability of 0.1 is used in both encoder and decoder. A label-smoothed cross-entropy [30] loss is minimized during training. Greedy decoding is performed to generate the ASR transcription. We use PyTorch [31] toolkit for all the experiments. As shown in Table 1, the close-talk AED model achieves 7.58% and 17.39% WERs on a close-talk Cortana test set used in [32] and the far-field HK speaker test set, respectively.

Using the well-trained close-talk AED as the initialization, we then train a far-field AED with 3400 hours far-field data and its ground-truth labels by following the same procedure. When evaluated on the HK speaker test set, the baseline far-field AED achieves 13.93% WER for E2E ASR as in Table 1.

Adaptation	Method	WER	WERR
Direct	Far-Field CE	13.93	-
Unsupervised	Token T/S	13.06	6.3
	Seq T/S	14.00	-0.5
Supervised	IT/S ($w = 0.2$)	12.95	7.0
	IT/S ($w = 0.8$)	12.96	7.0
	CT/S	12.82	8.0
	AT/S ($\lambda = 0.10$)	12.56	9.8
	AT/S ($\lambda = 0.25$)	**12.49**	**10.3**
	AT/S ($\lambda = 1.0$)	12.66	9.1
	AT/S ($\lambda = 3.0$)	12.71	8.8

Table 1. The WER (%) performance of far-field AEDs trained with CE and AED models adapted by various T/S learning methods to 3400 hours far-field Microsoft Cortana data for E2E ASR on HK speaker test set. "Seq T/S" stands for sequence-level T/S and WERR (%) represents relative WER reduction.

5.3. Unsupervised Domain Adaptation with T/S Learning

We adapt the close-talk baseline AED to the 3400 hours far-field data using token and sequence level T/S learning as discussed in Section 3. To achieve this, we feed the 3400 hours

close-talk adaptation data as the input to the teacher AED and the 3400 hours *parallel* far-field adaptation data as the input to the student AED. The student AED conditions its decoder on one-best token sequences generated by the teacher AED through greedy decoding. In token-level T/S, the soft posteriors generated by the teacher serve as the training targets of the student while in sequence-level T/S, the one-best sequences decoded by the teacher are used the targets.

As shown in Table 1, the token-level T/S achieves 13.06% WER on HK speaker test set, which is 24.9% and 6.25% relative improvements over the close-talk and far-field AED models, respectively. The sequence-level T/S achieves 14.00% WER, which is 19.5% relative improvement over the close-talk AED model. The sequence-level T/S performs slightly worse than the far-field AED trained with ground-truth labels because the one-best decoding from the teacher AED is not always reliable to serve as the training targets for the student model. The sequence-level T/S can be improved by using multiple decoded hypotheses generated by the teacher AED as the training targets as in [19, 20]. We did not perform N-best decoding because it will drastically increase the computational cost and will consumes much more adaptation time than the other T/S methods. The 6.7% relative WER gain obtained by token-level T/S over sequence-level T/S shows the benefit of using soft posteriors generated by the teacher AED as the training target at each decoder step when a reliable ground-truth transcription is not available.

The 6.3% relative WER gain of token T/S over far-field AED baseline shows that the unsupervised T/S learning with no ground-truth labels can significantly outperform the supervised domain adaptation with such information available. Compared to the one-hot labels, the soft posteriors accurately models the inherent statistical relationships among different token classes in addition to the token identity encoded by a one-hot vector. It proves to be a more powerful target for the student to learn from which is consistent with what was observed in [11, 12, 13, 14, 15].

5.4. Supervised Domain Adaptation with AT/S Learning

As discussed in Section 4, we want to further improve the T/S learning by using one-hot ground-truth labels when they are available. As in [16], we perform IT/S learning for supervised domain adaptation by using the linear interpolation of soft posterior and one-hot ground truth as the training target of the student. The interpolation weights are globally fixed at 0.5 and 0.5 for all decoder steps. By following [14], we also conduct CT/S for supervised domain adaptation where soft posteriors are used as the training target of the student if the teacher's prediction is correct at the current decoder step, otherwise the one-hot ground truth is used as the target. Finally, AT/S domain adaptation is performed by adaptively adjusting the weights assigned to the soft and one-hot labels at each decoder step as in Eqs. (9) to (11). We explore us-

ing different power functions as $f_1(x)$ and $f_2(x)$ to compute the confidence scores by adjusting λ. For all the above supervised T/S learning methods, the 3400 hours close-talk and 3400 hours far-field parallel adaptation data is fed as the input to the teacher and student AEDs, respectively.

As shown in Table 1, IT/S with $w = 0.2$ achieves 13.95% WER on HK speaker test set which is 25.5%, 7.0% and 0.8% relative improvements over the close-talk, far-field and token-level T/S adapted AED models, respectively. With a 12.82% WER, CT/S relatively improves the close-talk, far-field and token-level T/S adapted AED models by 26.3%, 8.0% and 1.8% respectively. Among different λs for AT/S, the best WER is 12.49%, which is 28.2%, 10.3% and 4.4% relative gains over close-talk, far-field and token-level T/S adapted AEDs. The minimum WER is reached when $\lambda = 0.25$ and $c_l = P(y_l | \mathbf{Y}_{0:l-1}^G, \mathbf{X}^T; \theta^T)^{0.25}$. Compared to $\lambda > 1$, AT/S works better for $\lambda \in [0, 1]$ when confidence scores c_l, d_l are both concave functions of the correct token posterior and the sum of incorrect token posteriors, respectively. All the IT/S, CT/S and AT/S outperform the unsupervised T/S learning indicating that the one-hot ground truth can further improve T/S domain adaptation when it is properly used. AT/S achieves the largest gain in supervised domain adaptation methods showing the superiority of adaptively extracting useful knowledge from both the soft and one-hot labels depending on their confidence scores.

6. CONCLUSION

In this paper, we extend T/S learning to unsupervised domain adaptation of AED models for E2E ASR. T/S learning requires only unlabeled parallel source and target domain data as the input to the teacher and student AEDs, respectively. In T/S, the student AED conditions its decoder on the one-best token sequences generated by the teacher. The teacher's soft posteriors and decoded one-hot tokens are used as the training target of the student AED for token-level and sequence-level T/S learning, respectively.

For supervised domain adaption, we propose adaptive T/S learning in which the student always learns from a linear combination of the teacher's soft posteriors and the one-hot ground truth. The combination weights are adaptively computed at each decoder step based on the confidence scores on both knowledge sources.

Domain adaptation is conducted on 3400 hours close-talk and 3400 hours far-field Microsoft Cortana data. Token-level T/S achieves 6.3% relative WER improvement over the baseline far-field AED model trained with CE criterion. By making use of the ground-truth labels, AT/S further improves the token-level T/S by 4.4% relative and achieves a total 10.3% relative gain over the far-field AED. AT/S also consistently outperforms IT/S and CT/S showing the advantage of learning from both the teacher and the ground truth as well as the adaptive adjustment of the combination weights.

7. REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*. ACM, 2006.

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[3] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[4] J. K. Chorowski, D. Bahdanau, D. Serdyuk *et al.*, "Attention-based models for speech recognition," in *NIPS*, 2015, pp. 577–585.

[5] D. Bahdanau, J. Chorowski, D. Serdyuk *et al.*, "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4945–4949.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4960–4964.

[7] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, vol. 37. Lille, France: PMLR, 2015, pp. 1180–1189.

[8] S. Sun, B. Zhang, L. Xie *et al.*, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79 – 87, 2017.

[9] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, "Unsupervised adaptation with domain separation networks for robust speech recognition," in *Proc. ASRU*, 2017.

[10] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria." in *Proc. INTERSPEECH*, 2014, pp. 1910–1914.

[11] J. Li, M. L. Seltzer, X. Wang *et al.*, "Large-scale domain adaptation via teacher-student learning," in *Proc. INTERSPEECH*, 2017.

[12] S. Watanabe, T. Hori, J. L. Roux, and J. Hershey, "Student- teacher network learning with enhanced features," in *Proc. ICASSP*, 2017.

[13] J. Li, R. Zhao, Z. Chen *et al.*, "Developing far-field speaker system via teacher-student learning," in *Proc. ICASSP*, 2018.

[14] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional teacher-student learning," in *Proc. ICASSP*, 2019.

[15] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5185–5189.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[18] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *EMNLP*, 2016.

[19] R. M. Munim, N. Inoue, and K. Shinoda, "Sequence-level knowledge distillation for model compression of attention-based sequence-to-sequence speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6151–6155.

[20] R. Pang, T. Sainath, R. Prabhavalkar *et al.*, "Compression of end-to-end models," in *Proc. Interspeech 2018*, 2018, pp. 27–31.

[21] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] C.-C. Chiu, T. N. Sainath, Y. Wu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP*. IEEE, 2018, pp. 4774–4778.

[24] J. H. Wong and M. J. Gales, "Sequence student-teacher training of deep neural networks," 2016.

[25] J. Li, G. Ye, A. Das *et al.*, "Advancing acoustic-to-word ctc model," in *Proc. ICASSP*. IEEE, 2018, pp. 5794–5798.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[28] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 1171–1179.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *CoRR*, vol. abs/1612.02695, 2016. [Online]. Available: http://arxiv.org/abs/1612.02695

[31] A. Paszke, S. Gross, S. Chintala *et al.*, "Automatic differentiation in pytorch," 2017.

[32] Z. Meng, Y. Gaur, J. Li, and Y. Gong, "Speaker adaptation for attention-based end-to-end speech recognition," *Proc. Interspeech*, 2019.