SPEAKER AND LANGUAGE AWARE TRAINING FOR END-TO-END ASR

Shubham Bansal, Karan Malhotra, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, 560012.

ABSTRACT

The end-to-end (E2E) approach to automatic speech recognition (ASR) is a simplified and an elegant approach where a single deep neural network model directly converts the acoustic feature sequence to the text sequence. The current approach to end-to-end ASR uses the neural network model (trained with sequence loss) along with an external character/word based language model (LM) in a decoding pass to output the text sequence. In this work, we propose a new objective function for end-to-end ASR training where the LM score is explicitly introduced in the attention model loss function without any additional training parameters. In this manner, the neural network is made LM aware and this simplifies the model training process. We also propose to incorporate an attention based sequence summary feature in the ASR model which allows the system to be speaker aware. With several E2E ASR experiments on TED-LIUM, WSJ and Librispeech datasets, we show that the proposed speaker and LM aware training improves the ASR performance significantly over the state-of-art E2E approaches. We achieve the best published results reported for WSJ dataset.

Index Terms— End-to-end speech recognition, Language modeling, Speaker adaptation.

1. INTRODUCTION

The conventional modular approach to automatic speech recognition (ASR) consists of several modules such as acoustic model, lexical model and language model. The end-to-end (E2E) ASR models have been shown to overcome limitations of the conventional approach using a single deep recurrent neural network model which could directly be trained on words [1, 2, 3], sub-words [4, 5, 6] or character targets thereby eliminating the need for the hand-crafted pronunciation dictionary. The earliest approach to E2E ASR used the connectionist temporal classification (CTC) [7, 8] cost to optimize the recurrent neural network model (RNN). The CTC based modeling allows the network to be trained for speech recognition task without the need of any prior alignment between the speech and text sequence but makes the strong conditional independence assumptions between the frame level posterior probability estimates.

The attention based models proposed recently [9, 10, 11] does not make any conditional independence assumption and they attempt to learn an implicit language model using an encoder-decoder attention framework. This model however suffers from the problem of irregular alignments due to ill constrained attention mechanism and also suffers from premature end-of-sentence issues. In order to combine the advantages of the two models (CTC and attention), Watanabe et al. proposed a hybrid CTC-Attention model [12]. The hybrid model presents a combination of both the CTC and attention approaches with a shared encoder. The output of the E2E models in the form of posterior probabilities are processed using a beam-search decoding process. This model has been shown to outperform both CTC and attention based models and has provided the state-of-art E2E results for many ASR tasks [13].

The E2E ASR models, being sequence-to-sequence (Seq2Seq), are limited by requirement of parallel training data comprising of both speech and transcripts. The use of an external language model (LM) can effectively leverage large text data and is extensively used [11, 14, 15, 16, 17] to improve the E2E ASR performance by integrating it either during training or in the decoding phase. Multi-level LM integration [18] involving both the character-based RNN-LM and word-based RNN-LM have also been shown to be effective in improving the end-to-end ASR performance. In many cases, the integration of LM in the E2E ASR involves training additional parameters [15].

In this paper, we propose a novel LM-aware objective function for training the hybrid CTC-Attention ASR model. We motivate the approach using a mathematical formulation where the modified loss function allows the ASR model to explicitly model only the acoustics while the implicit character level LM is provided by an external LM. The proposed approach yields significant improvements over the hybrid CTC-Attention E2E model.

Speaker adaptation for E2E ASR [19] based on sequence summary network [20] appends an auxiliary feature which is learned during the training keeping the E2E ASR pipeline simple. In our work, we also propose an improvement over the sequence summary network by replacing the average by an attention layer. We show that the attention network is essentially capturing the speaker information and hence we refer to this approach as the speaker-aware training.

The rest of the paper is organized as follows. The prior work on LM integration and auxiliary feature adaptation for E2E ASR is highlighted in Section 2. The proposed LM-aware objective function for hybrid CTC-Attention ASR model training is outlined in Section 3. The proposed speaker-aware training is described in Section 4 The ASR experiments and results are reported in Section 5. The paper concludes with a summary in Section 6.

2. RELATED PRIOR WORK

Early work on integrating LM, also known as shallow fusion [14], combines the log-probability score of Seq2Seq model and LM in a linear way during beam search decoding with a fixed LM fusion weight. A deep fusion approach [11] attempts to fuse hidden states of pre-trained external RNN-LM and pre-trained ASR with the help of gating mechanism and trains the fused model further on the ASR training data. A recent method, called cold fusion [15], overcomes the limitation of deep fusion by using a pre-trained external LM during ASR training. The authors also suggest that the learning of an implicit LM is demanding on the decoder network in the attention model and also makes it difficult to fuse an external LM of different domain. Toshniwal et al. [16] compares different LM fusion ap-

proaches for E2E ASR and shows that the simple approach of shallow fusion performs the best on several E2E ASR tasks. Hori et al. [18] uses multiple LM during shallow fusion where the character LM is used to score until the word boundary comes followed by a rescoring with the word LM. Hori et al. [17] further omits the need of the character LM and uses only the word LM to provide look-ahead probability at the character level and achieves the state-of-art results for several english E2E ASR tasks.

Motivated by a prior work on LM integration in neural machine translation (NMT) [21], we propose a new LM-aware training approach for E2E ASR model. The proposed LM-aware training simplifies the role of the decoder in the attention model as the language modeling objective is performed by an external LM. We show that our proposed approach increases the efficiency of the attention based encoder-decoder model and achieves faster convergence. During testing, the proposed approach achieves up to 9% in relative WER improvement over the state-of-art hybrid CTC-Attention model combined with the look-ahead word LM during decoding. [17].

For speaker adaptation in the conventional hidden Markov model - deep neural network (DNN) based ASR, several methods have been investigated like feature transformation based approaches [22, 23, 24] and i-vector feature augmentation methods [25, 26]. For E2E ASR, Delcroix et al. [19] recently proposed to use the sequence summary network [20]. The sequence summary network uses a feed-forward neural network and its output is averaged to a single vector. However, for tasks like speaker verification, it has been shown that sequence summary based on attention improves over simple statistical average [27], [28].

In our work, we propose to use a sequence summary network with an attention layer in the E2E ASR model. We refer to this approach as speaker-aware E2E model. In our ASR experiments, the speaker-aware training shows relative improvements of up to 5% over the simple average approach [19] and up to 10% over the the state-of-art hybrid CTC-Attention model .

3. LM AWARE E2E MODEL

We describe the E2E model using hybrid CTC-Attention framework in Section 3.1 and the existing LM fusion approaches in Section 3.2. This is followed by description of the proposed LM aware training in Section 3.3.

3.1. Hybrid CTC-Attention Architecture

For a given word sequence $\mathbf{W}$ and corresponding speech feature sequence $\mathbf{X}$, the aim of automatic speech recogniton (ASR) is to estimate the posterior distribution $P(\mathbf{W}|\mathbf{X})$.

3.1.1. Attention Model

Let $\mathbf{C}$ be a sequence of characters/wordpieces of length L corresponding to a word sequence $\mathbf{W}$. The attention model directly estimates $P(\mathbf{C}|\mathbf{X})$ as follows:

$$P(\mathbf{C}|\mathbf{X}) \triangleq P_{\text{att}}(\mathbf{C}|\mathbf{X}) = \prod_{\text{l}=1}^{\text{L}} P(c_{\text{l}}|c_1, ..., c_{\text{l}-1}, \mathbf{X}) \tag{1}$$

To obtain $P(c_{\text{l}}|c_1, ..., c_{\text{l}-1}, \mathbf{X})$, we have an encoder network, an attention mechanism and a decoder network.

Encoder network consists of stacked Bi-LSTM layers which converts an input speech feature sequence $\mathbf{X}$ to a high-level feature sequence $\{\mathbf{h}_{\text{t}}\}_{\text{t}=1}^{\frac{T}{S}}$. We can describe encoder network as follows:

$$\mathbf{h}_{\text{t}} = Encoder(\mathbf{X}) \triangleq \text{BLSTM}_{\text{t}}(\mathbf{X}) \quad \forall t \in \left\{1, 2, 3, ...\frac{T}{S}\right\},$$

where, T is the length of the input speech feature sequence and S is the sub-sampling factor used in the encoder network.

Attention Mechanism is location-aware [9] and generates a context vector $\mathbf{r}_{\text{l}}$ which is a function of the decoder hidden state $\mathbf{q}_{\text{l}-1}^{\text{dec}}$ and the high-level feature sequence $\{\mathbf{h}_{\text{t}}\}_{\text{t}=1}^{\frac{T}{S}}$ as follows:

$$\mathbf{r}_{\text{l}} = Attention(\{\mathbf{h}_{\text{t}}\}_{\text{t}=1}^{\frac{T}{S}}, \mathbf{q}_{\text{l}-1}^{\text{dec}}) \quad \forall l \in \{1, 2, 3, ...L\}$$

Decoder network consists of stacked one or more uni-directional LSTM layers and is a function of the previous output $c_{\text{l}-1}$ during prediction or previous ground truth character c_{l-1}^{*} during training, context vector $\mathbf{r}_l$ and the hidden state $\mathbf{q}_{\text{l}-1}^{\text{dec}}$ as follows:

$$P(c_{\text{l}}|c_1, ., c_{\text{l}-1}, \mathbf{X}) \triangleq \text{SM}(\mathbf{W}_{\text{d}}\mathbf{q}_{\text{l}}^{\text{dec}} + \mathbf{b}_{\text{d}}), \tag{2}$$

$$\mathbf{q}_{\text{l}}^{\text{dec}} = \text{LSTM}_{\text{l}}^{\text{att}}(\mathbf{r}_{\text{l}}, \mathbf{q}_{\text{l}-1}^{\text{dec}}, c_{\text{l}-1}) \quad \forall l \in \{1, 2, 3, ...L\}, \tag{3}$$

where, SM represents the softmax operator and $\mathbf{W}_{\text{d}}$, $\mathbf{b}_{\text{d}}$ represents the learnable linear layer parameters. The decoder LSTM uses previous character as an embedding vector learned during training.

3.1.2. Probabilistic Interpretation of Attention Model

We interpret the probability distribution learned by the attention model. Left-hand side expression of Equation (2) can be simplified as follows:

$$\begin{aligned} P(c_{\text{l}}|c_1, ., c_{\text{l}-1}, \mathbf{X}) &= \frac{P(c_1, ., c_{\text{l}-1}, c_{\text{l}}|\mathbf{X})}{P(c_1, ., c_{\text{l}-1}|\mathbf{X})} \\ &= P(c_{\text{l}}|c_1, ., c_{\text{l}-1}) \frac{P(\mathbf{X}|c_1, ., c_{\text{l}-1}, c_{\text{l}})}{P(\mathbf{X}|c_1, ., c_{\text{l}-1})}, \end{aligned} \tag{4}$$

where, $P(c_{\text{l}}|c_1, ., c_{\text{l}-1})$ represents the implicit LM learning aspect and $\frac{P(\mathbf{X}|c_1, ., c_{\text{l}-1}, c_{\text{l}})}{P(\mathbf{X}|c_1, ., c_{\text{l}-1})}$ represents how more likely the source speech feature sequence becomes when a particular token c_{l} is revealed. Right-hand side expression of Equation (2) could be simplified as follows:

$$\text{SM}(\mathbf{W}_{\text{d}}\mathbf{q}_{\text{l}}^{\text{dec}} + \mathbf{b}_{\text{d}}) \propto \exp(\mathbf{W}_{\text{d}}\mathbf{q}_{\text{l}}^{\text{dec}} + \mathbf{b}_{\text{d}}) \tag{5}$$

From Equations (4) and (5),

$$\exp(\mathbf{W}_{\text{d}}\mathbf{q}_{\text{l}}^{\text{dec}} + \mathbf{b}_{\text{d}}) \propto \underbrace{P(c_{\text{l}}|c_1, ., c_{\text{l}-1})}_{\hookrightarrow \text{implicit LM}} \frac{P(\mathbf{X}|c_1, ., c_{\text{l}-1}, c_{\text{l}})}{P(\mathbf{X}|c_1, ., c_{\text{l}-1})} \tag{6}$$

From Equation (6), it is observed that the attention model also learns an implicit LM during the ASR training [10].

3.1.3. CTC Model

The connectionist temporal classification (CTC) is an objective function that allows the RNN/LSTM network to be trained for a sequence transcription task without the need of any prior alignment between the input and target sequences. The output layer emits probability for each of the character/wordpiece and an extra unit (referred to as the blank) which corresponds to a null emission. More details about the formulation of CTC likelihood $P_{\text{ctc}}(\mathbf{C}|\mathbf{X})$ in the context of hybrid CTC-Attention architecture is given in Watanabe et. al [12].

3.1.4. Multiobjective Learning

The attention and CTC models share the encoder network and the hybrid model tries to maximize the linear combination of log-

likelihood of both the models as follows:

$$L_{\text{joint}} = \lambda \log(P_{\text{ctc}}(\mathbf{C}|\mathbf{X})) + (1-\lambda)\log(P_{\text{att}}(\mathbf{C}|\mathbf{X})),$$

where, λ is a hyperparameter.

3.1.5. *One Pass Joint Decoding*

The hybrid model uses the beam search decoding and for each partial hypothesis, a score is calculated as a linear combination of scores from attention and CTC models as follows:

$$\alpha_{rmp} = \gamma\alpha_{\text{ctc}}(\mathbf{p}, \mathbf{X}) + (1-\gamma)\alpha_{\text{att}}(\mathbf{p}, \mathbf{X}), \quad (7)$$

$$\alpha_{\text{ctc}}(\mathbf{p}, \mathbf{X}) \triangleq \log(P_{\text{ctc}}(\mathbf{p}|\mathbf{X})), \;\; \alpha_{\text{att}}(\mathbf{p}, \mathbf{X}) \triangleq \log(P_{\text{att}}(\mathbf{p}|\mathbf{X})),$$

where, $\mathbf{p}$ represents the partial hypothesis and γ is the CTC weight during decoding.

3.2. LM Fusion Approaches

3.2.1. *Shallow Fusion*

In shallow fusion [14], the LM is fused only during inference time to guide beam-search. Equation (7) which calculates the score of partial hypothesis is modified as follows:

$$\alpha_{\text{p}} = \gamma\alpha_{\text{ctc}}(\mathbf{p}, \mathbf{X}) + (1-\gamma)\alpha_{\text{att}}(\mathbf{p}, \mathbf{X}) + \beta\log(P_{\text{LM}}(\mathbf{p})), \quad (8)$$

where β is the LM fusion weight. The shallow fusion has been shown to outperform other fusion approaches [16] on several benchmarks.

3.2.2. *Shallow Fusion with Look Ahead Word LM*

Decoding with the word LM is limited by the presence of out-of-vocabulary words but can effectively model long sequences of characters. Hori et. al [17] proposes look-ahead word-based RNN-LM which enables to predict probability of the next character using a look-ahead mechanism over the word probabilities. Equation (7) which calculates the score of partial hypothesis is modified as follows:

$$\alpha_{\text{p}} = \gamma\alpha_{\text{ctc}}(\mathbf{p}, \mathbf{X}) + (1-\gamma)\alpha_{\text{att}}(\mathbf{p}, \mathbf{X}) + \beta\log(P^{\text{la}}_{\text{LM}}(\mathbf{p})), \quad (9)$$

where, $\log(P^{\text{la}}_{\text{LM}}(\mathbf{p}))$ represents the look-ahead external word-based RNN-LM score. Detailed formulation of look-ahead mechanism is given in [17].

3.2.3. *Cold Fusion*

Cold Fusion [15] enables learning of contribution from the pre-trained LM during ASR training with the help of the gating mechanism. The authors also suggests that the early training integration of the pre-trained LM with the ASR training enables the decoder to focus its capacity on ASR learning without worrying about the language modelling aspect. Cold fusion works as follows:

$$\mathbf{h}_{\text{l}}^{\text{LM}} = \text{DNN}(\mathbf{s}_{\text{l}}^{\text{LM}}) \quad (10)$$

$$\mathbf{g}_{\text{l}} = \sigma(\mathbf{W}_{\text{g}}[\mathbf{q}_{\text{l}}^{\text{dec}}; \mathbf{h}_{\text{l}}^{LM}] + \mathbf{b}_{\text{g}})$$

$$\mathbf{q}_{\text{l}}^{\text{cold}} = [\mathbf{q}_{\text{l}}^{\text{dec}}; \mathbf{g}_{\text{l}}\mathbf{h}_{\text{l}}^{\text{LM}}]$$

Equation (2) is modified as follows:

$$P(c_{\text{l}}|c_1, ., c_{\text{l}-1}, \mathbf{X}) \triangleq \text{SM}(\mathbf{W}_{\text{c}}\mathbf{q}_{\text{l}}^{\text{cold}} + \mathbf{b}_{\text{c}}),$$

where, $\mathbf{s}_{\text{l}}^{\text{LM}}$ represents the soft-max output of RNN-LM, DNN can be of any number of layers and $\mathbf{g}_{\text{l}}$ represents the gating mechanism for cold fusion. $\mathbf{W}_{\text{g}}$, $\mathbf{b}_{\text{g}}$, $\mathbf{W}_{\text{c}}$ and $\mathbf{b}_{\text{c}}$ are the linear layer parameters learned during ASR training. The soft-max output $\mathbf{s}_{\text{l}}^{\text{LM}}$ is used because the distribution of RNN-LM hidden state $\mathbf{q}_{\text{l}}^{\text{LM}}$ can vary significantly across different datasets and LM.

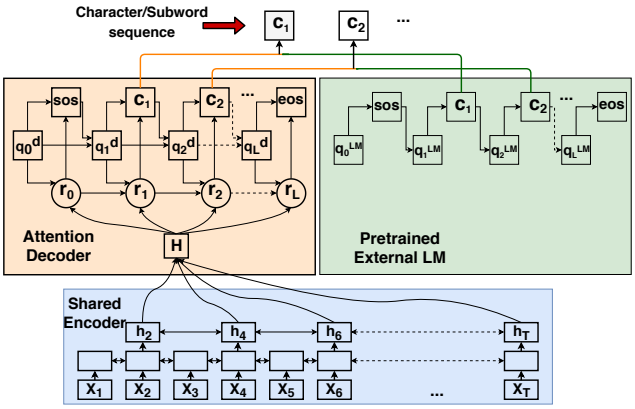

Fig. 1. LM-aware training for attention model in end-to-end ASR training.

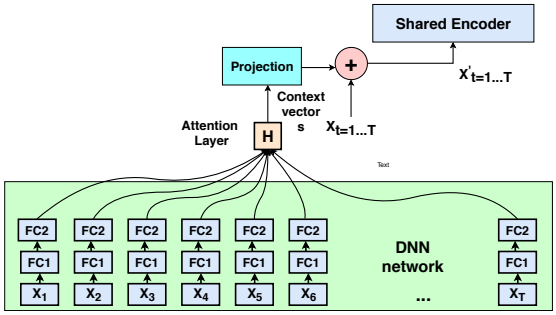

Fig. 2. Attention based auxiliary feature adaptation in end-to-end modeling.

3.3. Proposed LM-aware Training

The proposed work is motivated from pre-norm method proposed in NMT by Stahlberg et. al [21], which greatly simplifies the fusion architecture by removing the need of any gating mechanism. Figure 1 illustrates the LM-aware training block diagram for attention model. LM-aware training also uses a pre-trained external LM for ASR training and only modifies the training of attention model.

3.3.1. *Training*

The attention model under LM-aware training estimates $P(\mathbf{C}|\mathbf{X})$ as follows:

$$P(\mathbf{C}|\mathbf{X}) \triangleq P_{\text{LMaware}}(\mathbf{C}|\mathbf{X}) = \prod_{l=1}^{L} P(c_{\text{l}}|c_1, ., c_{\text{l}-1}, \mathbf{X})$$

Equation (2) is modified as:

$$P(c_{\text{l}}|c_1, ., c_{\text{l}-1}, \mathbf{X}) \triangleq \text{SM}(\mathbf{W}_{\text{d}}\mathbf{q}_{\text{l}}^{\text{dec}} + \mathbf{b}_{\text{d}} + LM_{\text{score}}), \quad (11)$$

$$LM_{\text{score}} = \log(P_{\text{LM}}(c_{\text{l}}|c_1, ., c_{\text{l}-1})),$$

where, $\log(P_{\text{LM}}(c_{\text{l}}|c_1, ., c_{\text{l}-1}))$ is the log-probability score obtained from the pre-trained RNN-LM. Multi-objective likelihood is modified as follows:

$$L_{\text{joint}} = \lambda \log(P_{\text{LMaware}}(\mathbf{C}|\mathbf{X})) + (1-\lambda)\log(P_{\text{ctc}}(\mathbf{C}|\mathbf{X})),$$

It is observed that the number of learnable parameters in LM-aware training does not increase unlike in the cold fusion method.

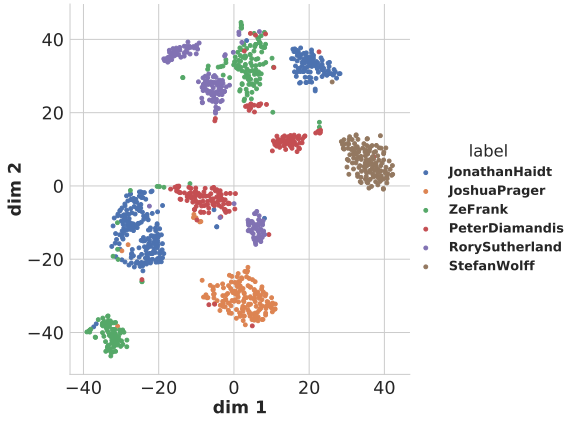

Fig. 3. tSNE plot of the auxiliary features obtained from the attention based network for the utterances of the TED-LIUM corpus.

3.3.2. Probabilistic Interpretation of LM-aware Training

We interpret the probability distribution being learned by the attention model under LM-aware training. The right-hand side expression of Equation (11) could be simplified as follows:

$$\begin{aligned}&\mathrm{SM}(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}}+\log(P_{\mathrm{LM}}(c_{\mathrm{l}}|c_1,.,c_{\mathrm{l}-1})))\\&\propto\exp(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}})\exp(\log P_{\mathrm{LM}}(c_{\mathrm{l}}|c_1,.,c_{\mathrm{l}-1}))\\&\propto P_{\mathrm{LM}}(c_{\mathrm{l}}|c_1,.,c_{\mathrm{l}-1})\exp(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}})\end{aligned}\tag{12}$$

From Equations (4), (11) and (12),

$$\begin{aligned}&\underbrace{P_{\mathrm{LM}}(c_{\mathrm{l}}|c_1,.,c_{\mathrm{l}-1})}_{\hookrightarrow\text{ Pretrained external LM}}\exp(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}})\\&\propto\underbrace{P(c_{\mathrm{l}}|c_1,.,c_{\mathrm{l}-1})}_{\hookrightarrow\text{ Implicit LM}}\frac{P(\mathbf{X}|c_1,.,c_{\mathrm{l}-1},c_{\mathrm{l}})}{P(\mathbf{X}|c_1,.,c_{\mathrm{l}-1})}\end{aligned}\tag{13}$$

Equation (13) could be further approximated as follows:

$$\exp(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}})\propto\frac{P(\mathbf{X}|c_1,.,c_{\mathrm{l}-1},c_{\mathrm{l}})}{P(\mathbf{X}|c_1,.,c_{\mathrm{l}-1})}\tag{14}$$

Comparing equations (6) and (14), we observe that, under LM-aware training, the decoder network of attention model focuses its entire capacity for conditioning on the source speech sequence and the implicit LM is modeled with the help of the pre-trained external LM.

3.3.3. Decoding with Look Ahead Word LM

Shallow fusion in equation (9), which calculates the score of partial hypothesis $\mathbf{p}$ of length N, is modified as follows:

$$\alpha_p=\gamma\alpha_{\mathrm{ctc}}(\mathbf{p},\mathbf{X})+(1-\gamma)\alpha_{\mathrm{LMaware}}(\mathbf{p},\mathbf{X})+\beta\log(P_{\mathrm{LM}}^{\mathrm{la}}(\mathbf{p})),\tag{15}$$

where, $\log(P_{\mathrm{LM}}^{\mathrm{la}}(\mathbf{p}))$ represents the look-ahead external word-based RNN-LM score and

$$\alpha_{\mathrm{LMaware}}(\mathbf{p},\mathbf{X})\triangleq\log(P_{\mathrm{LMaware}}(\mathbf{p}|\mathbf{X})),$$

$$P_{\mathrm{LMaware}}(\mathbf{C}|\mathbf{X})=\prod_{n=1}^{N}P(p_{\mathrm{n}}|p_1,.,p_{\mathrm{n}-1},\mathbf{X}),$$

$$P(p_{\mathrm{n}}|p_1,.,p_{\mathrm{n}-1},\mathbf{X})\triangleq\mathrm{SM}(\mathbf{W}_{\mathrm{d}}\mathbf{q}_{\mathrm{l}}^{\mathrm{dec}}+\mathbf{b}_{\mathrm{d}}+\zeta LM_{\mathrm{score}}),$$

$$LM_{\mathrm{score}}=\log(P_{\mathrm{LM}}^{\mathrm{c}}(p_{\mathrm{n}}|p_1,.,p_{\mathrm{n}-1})),$$

where, $\log(P_{\mathrm{LM}}^{\mathrm{c}}(p_{\mathrm{n}}|p_1,.,p_{\mathrm{n}-1}))$ represents the external character-based RNN-LM score. ζ and β represents the tunable scaling factor for character LM score and look-ahead word LM score respectively. These parameters enable us to control the contribution from both the LMs during the decoding stage.

4. SPEAKER AWARE TRAINING

We describe the existing sequence summary based auxiliary feature adaptation in Section 4.1. This is followed by description of the proposed attention based feature adaptation in Section 4.2.

4.1. Sequence Summary Network

Delcroix et. al [19] proposed to augment an auxiliary feature as a bias term to the input speech feature just before the encoder in the hybrid CTC-Attention model. The modified input feature sequence is given as follows:

$$\mathbf{x}_{\mathrm{t}}^{'}=\mathbf{x}_{\mathrm{t}}+\mathbf{P}\mathbf{s}\quad\forall t\in\{1,2,3,...T\},\tag{16}$$

where, $\mathbf{x}_{\mathrm{t}}$ represents the speech feature for t^{th} frame, T is the length of the speech feature sequence, $\mathbf{P}$ is the projection matrix and $\mathbf{s}$ represents the utterance level auxiliary feature learned with the help of the sequence summary network [20]. For each utterance, the sequence summary network averages over the DNN output of the corresponding speech feature sequence to compute its auxiliary feature as follows:

$$\begin{aligned}\mathbf{y}_{\mathrm{t}}&=\mathrm{DNN}(\mathbf{x}_{\mathrm{t}})\quad\forall t\in\{1,2,3,...T\},\\\mathbf{s}&=\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_{\mathrm{t}}\quad\forall t\in\{1,2,3,...T\}\end{aligned}\tag{17}$$

4.2. Proposed Attention for Auxiliary Feature Adaptation

In Figure 2, we show the block schematic of attention based auxiliary feature adaptation (referred to as speaker aware training). Unlike averaging, attention models [29] have been shown to learn a task-specific representation for a given sequence for several Seq2Seq tasks in NLP and in speaker recognition [27, 28]. In our work, we propose to replace the average in the sequence summary network given by Equation (17) by an additive attention layer as follows:

$$e_{\mathrm{t}}=\mathbf{g}^{\top}\tanh(\mathbf{W}\mathbf{y}_{\mathrm{t}}+\mathbf{b})\quad\forall t\in\{1,2,3,...T\}$$

where, e_{t} represents the attention score for DNN output of input feature at time t and $\mathbf{W},\mathbf{b},\mathbf{g}$ represents the learnable attention parameters. The attention weights are given by,

$$c_{\mathrm{t}}=\frac{\exp(e_{\mathrm{t}})}{\sum_{k=1}^{T}\exp(e_{\mathrm{k}})},$$

where, c_{t} represents the attention weight after normalization. Auxiliary feature $\mathbf{s}$ is now computed as follows:

$$\mathbf{s}=\sum_{t=1}^{T}c_{\mathrm{t}}\mathbf{y}_{\mathrm{t}}$$

For the TED-LIUM dataset (dataset details described in experiments section), we plot the learned auxiliary feature $\mathbf{s}$ for 6 different speakers in Figure 3. In this figure, we plot the first two dimensions of t-distributed stochastic neighborhood embedding (tSNE) for the attention based embedding $\mathbf{s}$. From the plot, we observe that our attention network effectively learns to capture the speaker information where similar speakers are clustered together in the embedding space. We

Table 1. Description about the datasets being used in our experiments

Dataset	*Train*	*Dev*	*Test*
WSJ [12]	81 h (283 spk.)	1.1 h (10 spk.)	0.7 h (8 spk.)
TED-LIUM [30]	210 h (5079 talks)	1.6 h (8 spk.)	2.6 h (10 spk.)
LIBRISPEECH clean [31]	463.7 h (1172 spk.)	5.4 h (40 spk.)	5.4 h (40 spk.)
LIBRISPEECH other [31]	496.7 h (1166 spk.)	5.3 h (33 spk.)	5.1 h (33 spk.)

hypothesize that using these embeddings would enable the end-to-end model to be more robust to speaker variations in speech

5. RESULTS AND DISCUSSION

All our experiments are performed with the End-to-End Speech Processing Toolkit (ESPNET) [13]. We have considered Hybrid CTC-Attention model training combined with the look-ahead word LM fusion as a baseline for our setup. The details about the datasets in the training, development and test conditions are described in Table 1. For all the datasets, we have used the 80-dimensional mel scale filterbank feature combined with 3 pitch features as our input feature. The external character LM network is a 2 layer LSTM network with 650 units in each layer and the external word LM is a 1 layer LSTM network with 1024 units. For cold fusion, the DNN used is a single affine layer followed by a $\tanh$ non-linearity as originally proposed [15]. The number of units in the DNN are same as the number of units in decoder. The sequence summary network is a 2 layer DNN with 1024 units in each layer and followed by the average and a projection layer. For the proposed speaker aware training, the attention based speaker adaptation replaces average using a 2048 dimensional additive attention layer.

5.1. TED-LIUM

The TED-LIUM [30] corpus (release 2) was made from audio talks and their transcriptions available on the TED website. The encoder used is a 6-layer VGG-BLSTMP network with 320 cells in each layer and direction and 320 units in the projection layer. The attention used is location attention and the decoder network is a 1-layer LSTM network with 300 cells. During training, CTC-weight λ is fixed at 0.5 and during decoding, CTC-weight γ and beam-size are fixed at 0.3 and 20 respectively. Both the external LMs are trained on the external LM training text available in TED-LIUM corpus combined with the transcripts of TED-LIUM training speech data.

Figure 4 shows the learning curve of the character accuracy for the training set as a function of the number of epochs. The total number of epochs for training convergence is significantly less with LM-aware training than the hybrid CTC-attention training. From the learning curve behaviour, we can re-emphasize the claim that with LM-aware training, the attention model decoder is not burdened with implicit LM learning and this leads to faster convergence. The cold fusion also shows fast convergence. However, this comes with the cost of more number of training parameters and leveraging the gating mechanism for early LM integration. The performance of various E2E approaches (in terms of Word Error Rate (WER) %) is reported in Table 2. As seen in the table, there is a relative WER improvement with LM-aware training and speaker-aware training respectively for test data over the state-of-art hybrid CTC-Attention model when combined without and with the look-ahead word LM. We obtain average relative improvements on test data of about 10 % and 6 % over the hybrid CTC-Attention model without the LM for the LM-aware and speaker-aware training respectively. For decoding with word LM, the proposed approaches of LM-aware and speaker aware training yield relative improvements of about 7 % over the hybrid CTC-Attention model. Using the combination of the LM-aware and speaker-aware training, we also achieve about 10% relative WER improvement over the baseline. It is also worth noting that the LM-aware approach improves over the cold fusion method (with word LM based decoding) because the scaling contribution factor of both the LMs are tuned during decoding. Also, speaker-aware training improves over the sequence summary approach.

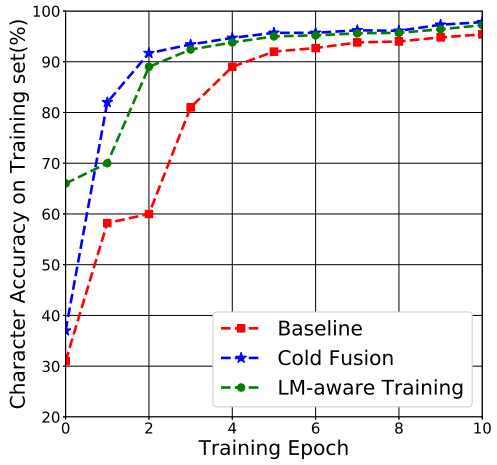

Fig. 4. Learning curve comparison of LM-aware training with baseline and cold fusion for TED-LIUM dataset

Table 2. WER(%) Performance comparison for different approaches with TED-LIUM dataset

Method	WER		ζ	β
	Test	*Dev*		
Without LM in decoding				
Hybrid CTC-Attention	22	22.4	-	-
Cold fusion (Character LM)	19.7	20.5	-	-
Sequence summary features	20.8	21.6	-	-
Prop. LM-aware (Character LM)	19.8	20.9	0.9	-
Prop. Speaker-aware	20.6	21.6	-	-
Prop. Speaker and LM-aware	19.6	20.8	-	-
With LM-based decoding (Look-ahead word LM)				
Hybrid CTC-Attention	15.1	15.6	-	0.7
Cold fusion (Character LM)	14.4	15.5	-	0.6
Sequence summary features	14.6	15.4	-	0.7
Prop. LM-aware (Character LM)	14.0	**15.3**	0.6	0.7
Prop. Speaker-aware	14.1	15.4	-	0.6
Prop. Speaker and LM-aware	**13.7**	**15.3**	0.6	0.8

5.2. WSJ

We use the Wall Street Journal (WSJ1) [32] and WSJ0 [33] for training our model, "dev93" for hyperparameter tuning and "eval92" for

Table 3. WER(%) Performance comparison for different approaches with WSJ dataset

Method	WER		ζ	β
	Eval	*Dev*		
Without LM in decoding				
Hybrid CTC-Attention	15.5	20.3	-	-
Cold fusion (Character LM)	11.3	15.6	-	-
Sequence summary features	15.0	19.8	-	-
Prop. LM-aware (Character LM)	8.2	12.0	1.0	-
Prop. Speaker-aware	14.9	19.5	-	-
With LM-based decoding (Look-ahead word LM)				
Hybrid CTC-Attention	4.4	7.4	-	1.0
Cold fusion (Character LM)	4.1	7.1	-	0.9
Sequence summary features	4.0	7.1	-	1.0
Prop. LM-aware (Character LM)	4.1	**6.4**	0.3	0.8
Prop. Speaker-aware	**4.0**	6.5	-	1.0

evaluation. The encoder used is a 4-layer BLSTMP network with 320 cells in each layer and direction and 320 units in the projection layer. The attention used is location attention and the decoder network is a 1-layer LSTM network with 300 cells. During training, CTC-weight λ is fixed at 0.2 and during decoding, CTC-weight γ and beam-size are fixed at 0.3 and 30 respectively. Both the external character and word LMs are trained on the external LM training text available with WSJ1 combined with the transcripts of WSJ training speech data.

The Word Error Rate (WER) results reported in Table 3 show that we achieve 10% and 8% relative WER improvement with LM-aware training and speaker-aware training respectively for Eval92 task over the state-of-art hybrid CTC-Attention model when combined with the look-ahead word LM. To the best of our knowledge, these results are the best reported WER for the WSJ dataset with E2E ASR. We also see that the results shown in Table 3 are consistent with those obtained for TED-LIUM dataset (Table 2).

5.3. LIBRISPEECH

In Librispeech experiments, we have used the vocabulary size of 1000 Byte Pair Encodings (BPE) to compare the ASR performance. The encoder used is a 4-layer VGG-BLSTMP network with 1024 cells in each layer and direction and 1024 units in the projection layer. The attention used is location attention and the decoder network is a 1-layer LSTM network with 1024 cells. During training, CTC-weight λ is fixed at 0.5 and during decoding, CTC-weight γ and beam-size are fixed at 0.3 and 20 respectively. The external sub-word based LM network is a 2 layer LSTM network with 1024 cells in each layer and is trained on the external normalized LM training text released along with Librispeech corpus combined with the transcripts of Librispeech training speech data. The word-LM in Librispeech setup is complicated due to the use of sub-word units (1000 BPE) as targets in the E2E model. Thus, in the Librispeech experiments, we use the sub-word based LM twice, once during the training of LM-aware model and additionally during the decoding to guide the beam search. As seen in previous two datasets, we observe about 6% and 5% relative WER improvement with the LM-aware training and speaker-aware training respectively over the state-of-art hybrid CTC-Attention model. These experiments show that the proposed approaches are effective in improving the state-of-art E2E systems.

Table 4. WER(%) Performance comparison for different approaches on Librispeech clean and other dataset.

Method	WER		ζ	β
	Dev	*Test*		
Librispeech Clean - With sub-word LM based decoding				
Hybrid CTC-Attention	4.2	4.2	-	0.9
Cold fusion (Sub-word LM)	4.2	4.1	-	0.8
Prop. LM-aware (Sub-word LM)	**4.0**	**4.1**	0.5	1.0
Prop. Speaker-aware	**4.0**	**4.1**	-	-
Librispeech Other - With sub-word LM-based decoding				
Hybrid CTC-Attention	12.5	12.7	-	0.9
Cold fusion (Sub-word LM)	12.2	12.6	-	0.8
Prop. LM-aware (Sub-word LM)	**11.8**	12.6	0.5	1.0
Prop. Speaker-aware	12.3	**12.5**	-	-

5.4. Relationship between LM-aware Training and Cold Fusion

LM-aware training is much simpler method compared to the cold fusion. In LM-aware training, an external LM is used during training to improve the decoder efficiency but learning the contribution from this external LM happens during the decoding stage with just an additional tunable parameter ζ. This avoids incorporating the additional training parameters and also makes the learning more efficient for smaller datasets. We observe that for single pass LM-scoring, the LM-aware training matches or exceeds the performance of the cold fusion. Also, unlike cold fusion, the proposed approach provides us with an unbiased way for tuning the contribution of character LM and word LM jointly during decoding. As also shown in our results, we observe that the LM-aware training(character LM) shows relative improvements of up to 9% over the cold fusion (character LM) when both of them were combined with the look ahead word LM in second pass LM-scoring leading to the state-of-art results for WSJ dataset. We also observe that both the LM-aware training and cold fusion benefits from using the look-ahead word LM during decoding.

6. CONCLUSIONS

In this paper, we have proposed two novel approaches to incorporate speaker adaptation and language model awareness in the end-to-end ASR training. The language model is incorporated directly in the cost function of the attention encoder-decoder model without any increase in the number of parameters. The speaker aware training is performed by using an attention based projection layer that captures the speaker information. With several speech recognition experiments, we have shown that the proposed methods improve over the state-of-art E2E ASR. We also show the effectiveness of the LM-aware training approach in simplifying the model training process. We have also shown for TED-LIUM dataset that the combined speaker and LM-aware training further improves the ASR performance.

7. ACKNOWLEDGEMENT

We would like to thank Jithendra Vepa, Shatrughan Singh and our colleagues from LEAP lab for their valuable guidance and support.

8. REFERENCES

[1] Hagen Soltau, Hank Liao, and Hasim Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[2] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," *arXiv preprint arXiv:1703.07754*, 2017.

[3] Kartik Audhkhasi, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Michael Picheny, "Building competitive direct acoustics-to-word models for english conversational speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4759–4763.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[5] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[6] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv preprint arXiv:1805.03294*, 2018.

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

[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

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

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "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)*. IEEE, 2016, pp. 4960–4964.

[11] Jan Chorowski and Navdeep Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

[12] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[13] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[14] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[15] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[16] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," *arXiv preprint arXiv:1807.10857*, 2018.

[17] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based rnn language models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018, pp. 389–396.

[18] T. Hori, S. Watanabe, and J. R. Hershey, "Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 287–293.

[19] Marc Delcroix, Shinji Watanabe, Atsunori Ogawa, Shigeki Karita, and Tomohiro Nakatani, "Auxiliary feature based adaptation of end-to-end asr systems," in *Proc. Interspeech*, 2018, pp. 2444–2448.

[20] Karel Veselỳ, Shinji Watanabe, Katerina Žmolíková, Martin Karafiát, Lukáš Burget, and Jan Honza Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5315–5319.

[21] Felix Stahlberg, James Cross, and Veselin Stoyanov, "Simple fusion: Return of the language model," in *WMT*, 2018.

[22] Jordan Cohen, Terri Kamm, and Andreas G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," 1995.

[23] Mark JF Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[24] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 24–29.

[25] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černocký, "ivector-based discriminative adaptation for automatic speech recognition," in *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, Dec 2011, pp. 152–157.

[26] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 55–59, 2013.

[27] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, "Self-attentive speaker embeddings for text-independent speaker verification," .

[28] Hossein Zeinali, Luka Burget, Johan Rohdin, Themos Stafylakis, and Jan Honza Cernocky, "How to improve your speaker embeddings extractor in generic toolkits," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6141–6145.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] Anthony Rousseau, Paul Deléglise, and Yannick Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus.," in *LREC*, 2012, pp. 125–129.

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[32] Linguistic Data Consortium et al., "CSR-II (WSJ1) complete," *Linguistic Data Consortium, Philadelphia, vol. LDC94S13A*, 1994.

[33] John Garofalo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.