Improved training strategies for end-to-end speech recognition in digital voice assistants

Hitesh Tulsiani, Ashtosh Sapru, Harish Arsikere, Surabhi Punjabi, Sri Garimella

Amazon Alexa, Bangalore, India

{hittul, sapru, arsikere, spunjabi, srigar}@amazon.com

Abstract

The speech recognition training data corresponding to digital voice assistants is dominated by wake-words. Training end-to-end (E2E) speech recognition models without careful attention to such data results in sub-optimal performance as models prioritize learning wake-words. To address this problem, we propose a novel discriminative initialization strategy by introducing a regularization term to penalize model for incorrectly hallucinating wake-words in early phases of training. We also explore other training strategies such as multi-task learning with listen-attend-spell (LAS), label smoothing via probabilistic modelling of silence and use of multiple pronunciations, and show how they can be combined with the proposed initialization technique. In addition, we show the connection between cost function of proposed discriminative initialization technique and minimum word error rate (MWER) criterion. We evaluate our methods on two E2E ASR systems, a phone-based system and a word-piece based system, trained on 6500 hours of Alexa's Indian English speech corpus. We show that proposed techniques yield 20% word error rate reductions for phone based system and 6% for word-piece based system compared to corresponding baselines trained on the same data.

Index Terms: acoustics-to-word, automatic speech recognition, connectionist temporal classification, end-to-end, initialization, voice assistant

1. Introduction

Recently, many studies have focused on end-to-end (E2E) automatic speech recognition (ASR) technology, which directly converts audio to sequence of words. These systems offer great simplification as they overcome complexities associated with maintaining a traditional hybrid ASR technology. In literature, the work on E2E technology utilizes connectionist temporal classification (CTC) framework (includes recurrent neural network transducer) [1–7], attention based encoder-decoder architecture [8–10], or both [11, 12]. Among these, CTC introduces an additional blank label to learn intermediate frame-level alignments between audio and output labels. The ability of models trained with CTC objective to provide frame synchronous predictions make them suitable for streaming applications, such as digital voice assistants [13].

Note that unidirectional long short term memory (LSTM) networks trained with CTC objective require careful initialization and training to achieve good accuracies [4, 14, 15]. Otherwise, networks may fail to find a good local optima. A number of initialization and training strategies have been proposed. Most of the initialization techniques focus on training an intermediate model. For hybrid ASR systems, this intermediate model is typically a cross entropy trained model with context dependent phones as units [14, 16] whereas for CTC based acoustics-to-word systems, initialization from the phone based system has been found to be useful [4]. Apart from initialization techniques, studies in [6,15,17–19] explore hierarchical training and curriculum learning to stabilize training with unidirectional networks. In [15,20], authors demonstrate joint training of CTC models with frame-wise cross-entropy loss also provides significant improvements.

Training end-to-end ASR for voice assistants poses further challenges. This is mainly because of frequent occurrences of wake-words (such as "alexa", "okay google", "siri"), usually at the start of utterances, in training data owing to typical command/response interaction mode. Since unidirectional LSTM network with CTC prioritizes learning simple patterns in the data [21], it learns only wake-word during initial phase of training, which leads to sub-optimal results. To the best of our knowledge, this problem has not been addressed for building an accurate end-to-end ASR system. Although bootstrapping from a well-trained cross-entropy model may mitigate this issue, it is cumbersome to maintain a traditional cross-entropy hybrid system (i.e. frame-level alignments, force alignment model, state tying decision trees etc.).

In this work, we systematically investigate this issue for CTC based end-to-end speech recognition systems. This paper proposes a novel discriminative training strategy to penalize overconfident hallucinations of wake-words during initial phase of training by introducing a new regularization term. We also establish its connection to minimum word error rate (MWER) training. To improve the system further, we also explore various training strategies that can be interpreted as variants of regularization: (i) Label smoothing via probabilistic modelling of silence and use of multiple pronunciations (ii) Multi-task learning with listen-attend-spell (LAS). We evaluate two CTC based end-to-end frameworks, acoustics-to-word (A2W) and phone based systems, on Alexa's Indian English data. For A2W system, we directly model words (and sub-words) whereas for phone based system, we model phones and use pronunciation lexicon and language model for evaluation. We refer to our phone based system as E2E (similar to [22]) because we do not make use of any previously trained models, forced alignments, or build state-tying decision trees.

2. Focusing on wake-word: Not the ideal way to start learning

Deep networks have been shown to prioritize learning simple patterns in the data [21]. For a voice assistants dataset, due to presence of wake-word in many of the utterances, the network could learn to focus on wake-word. In this section, we empirically explore learning prioritization by network and discuss why learning wake-word at the start of training may not be the optimal strategy.

In order to understand the learning prioritization, we carry out a couple of experiments with both phone based and

acoustics-to-word (A2W) system (system details mentioned later in section 4.1): (i) Uniform sampling: In the first experiment, to train our model, we uniformly sample the training data and thereby end up using natural distribution of data (ii) In the second experiment, we use a well known curriculum learning strategy - sortaGrad [4,6,19] which sorts the utterance based on length. With the use of sortaGrad, we expect a lot of utterances containing only wake-word to be sampled first.

Unit	Sampling strategy	WERR (%)
Phone	Uniform	0
	sortaGrad	- 8.27
A2W	Uniform	0
	sortaGrad	- 6.50

Table 1: *Comparison of phone based and A2W system trained using uniform sampling (Uniform) and sortaGrad. Negative relative word error rate (WERR) indicates degradation over uniform sampling baseline.*

We make two important observations: (i) The results in Table 1 shows that sortaGrad is significantly worse (6-8% relative) than uniform sampling for both the systems. This is in stark contrast to what has been reported in literature [4, 6] (ii) On looking at the training progress of models trained with uniform sampling and sortaGrad, we found that the model trained with sortaGrad hallucinates wake-word 'Alexa' (i.e. its exact phone sequence or its representation as single word-piece) for first 6k steps whereas model trained with uniform sampling hallucinates 'Alexa' for first 2.5k steps before gradually learning new words. This shows that model indeed starts learning by focusing on most common word in the dataset, which for our case is wake-word 'Alexa'. Both the observations when combined together tell us that by focusing on learning wake-word for a longer time, the performance degrades. In other words, lesser focus on learning wake-word initially may lead to significant improvements in performance.

3. Improved training of E2E systems

3.1. Discriminative initialization

In section 2, we found that the network starts learning by remembering wake-word. Based on this observation, we propose a new initialization technique to explicitly penalize the network for incorrectly hallucinating wake-word during training. We formulate it within the framework of multi-task learning as minimization of following:

$$L(X,Y) = L_{ctc}(X,Y) - \lambda * L_{ctc}(X, Y = ww) \quad (1)$$

where λ is a hyper-parameter to control contribution of various terms and is set to 0 for utterances that contain wake-word. $L_{ctc}(X,Y)$ is CTC loss for label sequence corresponding to word sequence Y and acoustic features X and $L_{ctc}(X, Y = ww)$ is CTC loss assuming label sequence corresponding to wake-word ww. Here, additional penalty term $L_{ctc}(X, Y = ww)$ acts as regularizer for the network and can be thought of as regularization via network architecture [23]. We use the above formulation only in the initial stages of training and then use the usual CTC loss formulation. We refer to this technique as discriminative initialization (DI).

Minimization of the loss can then be interpreted as (using $L_{ctc}(X,Y) = -lnP(Y/X)$):

$$\max(lnP(Y/X) - \lambda * lnP(Y = ww/X)) \quad (2)$$

As seen from the equation 2, we now try to maximize the likelihood of correct sequence and minimize the likelihood of predicting wake-word ww, thus adding an explicit penalty for incorrectly hallucinating it. We apply this formulation only to utterances which don't contain wake-word ww in transcription and use the usual CTC loss formulation $\min L_{ctc}(X,Y)$ for utterances containing wake-word ww. This ensures we don't penalize the correct predictions of wake-word but only incorrect hallucinations. Although this technique is proposed for wake-words, any other word can be substituted for it. It can also be easily adapted for any number of words by summing over them to compute the additional loss term.

3.1.1. Relation of discriminative initialization with minimum word error rate training

MWER training criteria is a discriminative training criteria to minimize the expected number of word errors over all possible hypothesis [24] [25]. It is formulated as minimization of following:

$$L_{mwer}(X,Y) = \mathbb{E}[W(Y,Y')] = \sum_{Y'} P(Y'/X)W(Y,Y') \quad (3)$$

where $W(Y,Y')$ is the WER between transcription Y and hypothesis Y'. However, computation of above loss is intractable as it involves summation over all possible label sequences. Hence the loss is approximated by limiting the summation over N-best list.

Typically, for E2E systems, MWER loss is added as an extra term in addition to the task specific loss (e.g. cross-entropy for LAS) [25] [26]. For our case, this becomes:

$$\begin{aligned} &\min(L_{ctc}(X,Y) + \lambda * L_{mwer}(X,Y)) \\ &= \min(-lnP(Y/X) + \lambda * \sum_{i=1}^{N} P(Y_i'/X)W(Y,Y_i')) \end{aligned} \quad (4)$$

Assuming that the probability mass (in N-best list) is concentrated over one recognition hypothesis ww (say 'Alexa', which holds in initial part of training as mentioned in section 2), the equation becomes:

$$\begin{aligned} &\min(-lnP(Y/X) + \lambda * P(Y_1' = ww/X) * W(Y, Y_1' = ww)) \\ &= \max(lnP(Y/X) - \lambda * P(Y_1' = ww/X) * W(Y, Y_1' = ww)) \end{aligned} \quad (5)$$

From equations 5 and 2, it is clear that both the formulations are identical except for unequal weighing of examples by constant factor $W(Y, Y_1' = ww)$ in case of MWER training. Note that the presence of monotonically increasing ln function in equation 2 doesn't affect optimization. Hence the proposed initialization technique can be considered as a special case of jointly minimizing CTC and MWER loss. Though, for MWER training one starts from a well trained model whereas we propose to start training with this formulation. With this formulation in place, we expect to mitigate the problem arising out of over representation of wake-word.

3.2. Multi-task learning with listen-attend-spell

Joint CTC-Attention based architectures have been proposed [11,12] to improve the alignment mechanism of attention based systems. For such joint systems, encoder is typically shared between the two and CTC is considered as an auxiliary task. These joint systems have been found to perform better than both the

systems individually but a major drawback of such systems is their reliance on attention based decoding and hence inapplicability to streaming recognition tasks. Note that attention based systems are also referred as LAS in literature and we use the two terms interchangeably.

We explored the simplest possible modification of joint CTC-Attention based system to make it suitable for streaming tasks. Instead of using attention based decoder and bidirectional network for encoder, we considered LAS as an auxiliary task, discarded the attend and spell networks of LAS after jointly training the network and used unidirectional network in encoder. At run time, we obtained frame synchronous encoder representation and used them for decoding similar to any other CTC system. A similar system capable of streaming recognition has been recently proposed in [27] where they have developed a CTC-triggered attention decoder.

The formulation incorporating both the CTC loss and cross-entropy of from LAS can be established within the framework of multi-task learning as:

$$L_{MTL}(X,Y) = (1-\alpha) * L_{ctc}(X,Y) + \alpha * L_{las}(X,Y) \quad (6)$$

Recently, it was empirically shown that LAS system achieves best performance with word-piece units [10]. Hence, in order to reap maximum benefit out of the auxiliary LAS task, we keep the auxiliary units fixed as word-pieces. For our phone based system, this meant use of two very different units for primary and auxiliary task. To the best of our knowledge, this is the first study on streaming recognition system, employing LAS as an auxiliary task and exploiting the complementary information of different unit types. By virtue of multi-tasking, this technique can also be considered as a form of regularization via network architecture [23].

3.3. Label smoothing via probabilistic modelling of silence and use of multiple pronunciation

Label smoothing was proposed as a regularization technique to improve generalization [28, 29]. It proposes to penalize low entropy output distribution by adding noise to the labels [19, 28]. In this work, we introduce a simple yet powerful technique for changing label distribution. To penalize over-confident predictions by network, while training we: (i) randomly use one of the multiple pronunciations from the lexicon (as also done in [10, 22]) (ii) randomly add silence label at start and end of the label sequence and in between words. This has the effect of changing label distribution for a fixed word sequence, thereby providing the regularization effect. For this work, we use probability of 0.8 for inserting silence label at start and end and 0.2 for inserting a silence label between words.

4. Experiments and Results

4.1. Dataset and systems

We investigate the performance of various techniques using Alexa's Indian English data. The training dataset in our experiments consists of 6500 hours of anonymized utterances. We augment the dataset by adding a noisy and reverberated copy of utterances. The dev and test used consists around 10 hours of anonymized utterances each. We use 256 dimensional STFT features [16] and employ a frame skipping approach where three consecutive frames are stacked to obtain 768 dimensional features.

We explored two different systems, depending on the output units: (i) Word and word-pieces combination: We have a total of 4000 classes, consisting of 3520 most frequent words (e.g. 'Alexa', 'play', 'stop') in our dataset and 480 word-pieces. Word-pieces are either single character or sub-words [30]. Any word outside of top 3520 frequent word list is represented as combination of these word-pieces. We call this as acoustics-to-word (A2W) system [5] (ii) Phones: We have a total of 54 phones including a silence phone. To obtain label sequence for training, unless stated otherwise, we convert transcription to phone sequence using a fixed pronunciation from lexicon.

4.2. Network architecture, training and evaluation details

Architecture: We use FLSTM architecture [31], with 2 bidirectional frequency LSTM layers and 5 unidirectional LSTM layers. Frequency LSTM layers operate with a window size of 48 frames, hop size of 15 frames and have 16 units in one direction whereas time LSTM layers have 768 units. We use the same network architecture for baseline and across all our experiments except for the setup of multi-task learning with LAS. In that case, encoder is the same FLSTM architecture and decoder has 2 unidirectional LSTM layers with 256 units each.
Training: We used synchronous distributed training on 16 GPUs with per GPU batch size of 128 and 64 utterances for phone-based and A2W systems respectively. We use adam optimizer with a warmup-hold-decay learning rate (LR) schedule. All the hyperparameters like LR, λ, α were individually tuned for each of the system to obtain best performance on dev set.
Evaluation: For A2W system, we use CTC prefix search decoding [1] without the use of any pronunciation lexicon or language model whereas for phone-based system we use pronunciation lexicon and language model (LM) in a FST based framework and rescore the N-best recognition list using RNN-LM. We use standard relative word error rate (WERR) metric for comparing various techniques throughout the paper. A positive value of WERR indicates improvement over baseline whereas a negative value indicates degradation.

4.3. Results with phone based system

We present results below using all the proposed techniques for end-to-end phone based system. The baseline system doesn't use any of the proposed training strategies and is indicated as having 0 WERR.

System	Training strategy	WERR (%)
Phone	-	0
	DI ($\lambda = 0.1$)	+11.50
	MP	+9.23
	PS	+10.00
	LAS - WP ($\alpha = 0.5$)	+15.15

Table 2: *Comparison of phone based system in terms of WERR for various training strategies: (i) DI: discriminative initialization (ii) MP: using multiple pronunciation (iii) PS: probabilistic modelling of silence (iv) LAS - WP: multi-task learning with LAS using WP as units. Hyperparameters λ and α in brackets.*

As seen from Table 2, DI improves relative WER of phone based system by 11% on top of baseline with randomly initialized parameters. For a phone based system, 'Alexa' is represented as a sequence of six phones. When observing the training progress, we found that because of DI it would now hypothesize only partial phone sequence of 'Alexa' instead of mem-

orizing the exact phone sequence (as observed earlier and explained in section 2). This shouldn't come as a surprise because by simply dropping a couple of phones from the phone sequence of 'Alexa', it could still minimize sequence probability $P(Y = ww/X)$ (here $ww = Alexa$) and hence achieve the objective described by equation 2. In order to verify that there was no negative effect of DI on recognition of 'Alexa', we looked at the percentage of 'Alexa' errors (insertions, deletions and substitutions) and found them to be similar with and without the usage of DI. For experiments, we tuned value of λ (equation 2) on dev set and found setting it to 0.1 for first 25k training steps and 0 afterwards provided best performance.

It can also be seen from the above table, that each of the proposed training strategy helps improve the performance considerably. The gains obtained by introduction of multi-task learning with LAS (+15.15%) indicate different unit types (phones, word-pieces) provide complementary information to the learner. Further, on observing training logs when using PS and PS+MP, we saw: (i) instead of prioritizing learning 'Alexa', the network focused on both 'Alexa' and silence (ii) instead of predicting only one pronunciation of 'Alexa', it also starts predicting other pronunciations. This is indicative of regularization effect introduced via label smoothing.

4.4. Results with acoustics-to-word system

In this section, we evaluate the efficacy of various techniques on acoustics-to-word system. It is a much harder task since it involves modelling words directly and may also suffer from data sparsity issues [4].

Unit	Training strategy	WERR (%)
A2W	-	0
	DI ($\lambda = 0.2$)	+3.48
	PS	+3.84
	LAS - WP ($\alpha = 0.3$)	+5.49

Table 3: *Comparison of A2W system in terms of WERR for (i) DI: discriminative initialization (ii) PS: probabilistic modelling of silence (ii) LAS - WP: multi-task learning with LAS using WP as units. Hyperparameters λ and α in brackets.*

For A2W system, 3520 most frequent words in the dataset are represented as a single word-piece, one of which is 'Alexa'. For DI, this makes the task of regularization more challenging compared to phone based system where 'Alexa' is represented as sequence of six phones and it could minimize $P(Y = ww/X)$ by dropping a couple of phones (as explained in section 4.3). This is also supported by the observation that we needed to use higher λ (0.2) for A2W system. We believe this inherent difficulty makes DI less effective for A2W system compared with phone based system (Table 3 vs Table 2), although still providing significant improvements (3.48%) over baseline. All the other techniques like use of probabilistic silence and multi-task learning with LAS seem to provide intended regularization effect and individually provide gains of 4-5% over baseline which doesn't incorporate these strategies.

4.5. Interaction of proposed discriminative initialization and training strategies for phone-based system

We carry out this study only for phone based system. However, the improvements obtained from combination of multiple training strategies are expected to hold for acoustics-to-word system.

For combining DI with other techniques, we first train the network with loss described by equation 2 for 25k steps and $\lambda = 0.1$ but without incorporating any of the techniques like PS, MP. We then use it as an initialization for tasks incorporating various training strategies with usual CTC loss. For multi-task learning with LAS, we only initialized encoder with discriminatively trained network and then used the joint loss described in equation 6.

No.	Training strategy	WERR (%)	DI impact as WERR(%)
1	-	0	+11.50
2	DI	+11.50	
3	PS	+10.00	+7.07
4	PS + DI	+17.07	
5	PS + MP	+13.07	+4.23
6	PS + MP + DI	+17.30	
7	PS + MP + LAS	+17.69	+2.31
8	PS + MP + LAS + DI	+20.00	

Table 4: *Performance of phone based system with combination of various training strategies. Last column quantifies the contribution of DI in presence of other training strategies*

From Table 4, comparing consecutive rows (1-2, 3-4, 5-6, 7-8 as shown in last column of Table 4) to assess the impact of discriminative initialization, we see that: (i) we gain the most (11.50%) by adding regularization via DI when there is no other form of regularization present (row 1,2) (ii) although the benefit of discriminative initialization reduces (+11.50%, +7.07%, +4.23%, +2.31%) as other terms contribute to regularization (from PS, PS+MP, PS+MP+LAS), it still provides significant gains in combination with every other technique (iii) the gains obtained from each of the techniques are complementary. This is evident from the fact that by combining multiple techniques, performance improves (+10.0, +13.07, +17.69) (rows 1,3,5,7).

5. Conclusions

The over representation of keyword in the dataset affects generalization of networks. We studied one such case of naturally high presence of wake-word in the dataset of voice assistant. In this work, we demonstrated the ill effects of training an end-to-end speech recognition system without careful attention to such data. To mitigate it, an initialization technique to penalize model for incorrectly hallucinating keywords was proposed. We showed that the proposed initialization has an effect of regularization and is effective against overconfident predictions of such keywords. In addition, training strategies like multi-task learning with listen-attend-spell and label smoothing via probabilistic modelling of silence and multiple pronunciation were explored to further improve the performance. We studied the efficacy of each of the proposed technique independently as well as in combination and established that (i) each of the technique independently provides significant word error rate reductions (ii) gains from combination of multiple training strategies are complementary. In future, we plan to leverage the proposed techniques to improve other E2E architectures like recurrent neural network transducer.

6. References

[1] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

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

[3] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[4] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. 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.

[5] J. Li, G. Ye, A. Das, R. Zhao, and Y. Gong, "Advancing acoustic-to-word ctc model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5794–5798.

[6] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[7] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving rnn transducer modeling for end-to-end speech recognition," *arXiv preprint arXiv:1909.12415*, 2019.

[8] W. Chan, N. Jaitly, Q. Le, and O. 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.

[9] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. 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.

[10] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen, "On the choice of modeling unit for sequence-to-sequence speech recognition," *Proc. Interspeech 2019*, pp. 3800–3804, 2019.

[11] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[12] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[13] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. K. Chin *et al.*, "Acoustic modeling for google home." in *Interspeech*, 2017, pp. 399–403.

[14] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4280–4284.

[15] C. Yu, C. Zhang, C. Weng, J. Cui, and D. Yu, "A multistage training framework for acoustic-to-word model." in *Interspeech*, 2018, pp. 786–790.

[16] H. Sak, A. W. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *CoRR*, vol. abs/1507.06947, 2015. [Online]. Available: http://arxiv.org/abs/1507.06947

[17] K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4815–4819.

[18] R. Sanabria and F. Metze, "Hierarchical multitask learning with ctc," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 485–490.

[19] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," *arXiv preprint arXiv:1711.02212*, 2017.

[20] T.-S. Nguyen, S. Stueker, and A. Waibel, "Learning shared encoding representation for end-to-end speech recognition models," *arXiv preprint arXiv:1904.02147*, 2019.

[21] D. Arpit, S. Jastrzkebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 233–242.

[22] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi." in *Interspeech*, 2018, pp. 12–16.

[23] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv preprint arXiv:1710.10686*, 2017.

[24] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, vol. 2013, 2013, pp. 2345–2349.

[25] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4839–4843.

[26] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. 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.

[27] N. Moritz, T. Hori, and J. Le Roux, "Streaming end-to-end speech recognition with joint ctc-attention based models," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[28] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[29] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, 2019, pp. 4696–4705.

[30] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[31] T. Sainath and B. Li, "Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks," in *INTERSPEECH*, San Francisco, USA, 2016.