INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Utterance invariant training for hybrid two-pass end-to-end speech recognition

Dhananjaya Gowda, Ankur Kumar*, Kwangyoun Kim, Hejung Yang, Abhinav Garg, Sachin Singh, Jiyeon Kim, Mehul Kumar, Sichen Jin, Shatrughan Singh, Chanwoo Kim*

Samsung Research

{d.gowda, ankur.k, ky85.kim, hejung.yang, abhinav.garg, singh.sachin, jstacey7.kim, sc.ehkim.jin, shathrughan.s, chanw.com}@samsung.com

Abstract

In this paper, we propose an utterance invariant training (UIT) specifically designed to improve the performance of a two-pass end-to-end hybrid ASR. Our proposed hybrid ASR solution uses a shared encoder with a monotonic chunkwise attention (MoChA) decoder for streaming capabilities, while using a low-latency bidirectional full-attention (BFA) decoder for enhancing the overall ASR accuracy. A modified sequence summary network (SSN) based utterance invariant training is used to suit the two-pass model architecture. The input feature stream self-conditioned by scaling and shifting with its own sequence summary is used as a concatenative conditioning on the bidirectional encoder layers sitting on top of the shared encoder. In effect, the proposed utterance invariant training combines three different types of conditioning namely, concatenative, multiplicative and additive. Experimental results show that the proposed approach shows reduction in word error rates up to 7% relative on Librispeech, and 10-15% on a large scale Korean end-to-end two-pass hybrid ASR model.

Index Terms: speech recognition, ASR, utterance invariant training, sequence summary network, two-pass ASR, streaming ASR

1. Introduction

End-to-end automatic speech recognition (ASR) models have gained popularity due to their simplicity and ease of training [1, 2, 3, 4, 5, 6, 7]. Such end-to-end (e2e) systems combine all different components of conventional hybrid DNN-HMM model into a single network, which can be jointly trained with multiple losses. These end-to-end systems perform at par or even better than conventional ASR systems with sufficiently large amount of data. However, most of the well performing e2e systems use bi-directional long short-term memory units. Hence they cannot be used for online streaming applications.

Recent efforts have focused on building e2e ASR systems with streaming capability [1, 7, 8, 9]. Recurrent neural network transducer (RNN-T) [10] and monotonic chunkwise attention (MoChA) [11, 12], in particular, have become popular and are being deployed in production setup [1, 2]. While both streaming and non-streaming e2e models perform well on clean test sets, their performance on noisy test sets still lags significantly [13, 14, 15, 16, 17]. Speaker variance is one major reason for this along with factors such as background noise.

Several speaker and/or utterance adaptive techniques have been proposed to improve conventional ASR model performance [18, 19, 20, 21, 22, 23, 24, 25]. These techniques broadly fall under two categories - (1) training on speaker specific data [26, 27, 28, 18], and (2) adaptive training on speaker agnostic input features [24, 25, 19, 29, 30, 31]. Adaptive training transforms the speaker-specific input features to a speaker-independent feature space using techniques like feature space maximum likelihood linear regression (fMLLR) [32], adversarial training [33], or with the help of auxiliary input features such as i-vectors [20]. Since the model is trained with speaker independent features, it doesn't need any post training for a new test speaker.

Many research works have tried to adapt these conventional speaker adaptive training techniques for deep neural network-based hybrid ASR systems [23, 24, 25]. In [24] authors introduced an alternate approach of sequence summary to avoid any use of external model or speaker-specific data. Specifically, an auxiliary neural network is used to extract various acoustic characteristics from input audio. Sequence summary, which is average of the output of the auxiliary network, is used to bias the input audio features and the entire network is trained jointly. Interestingly, this simple approach was able to give performance gains similar to the widely used i-vector-based techniques. However, unlike earlier methods, this technique can be integrated with neural network-based ASR components in an end-to-end fashion. In subsequent work, [25] extended this approach to attention-based end-to-end ASR systems and showed consistent performance gains. However, these experiments were done over small datasets like WSJ [34], TEDLIUM [35], containing only few hundred hours of speech.

In this work, we propose to extend the sequence summary based utterance invariant training (UIT) to the recently proposed two-pass e2e hybrid architecture [3]. The two-pass hybrid decoding strategy was proposed to enhance the overall performance of a streaming ASR system. In the first pass, an RNN-T model is used to decode the input audio for providing streaming or real-time speech-to-text capabilities. In the second pass, the embeddings from the shared encoder is passed to a listen attend and spell (LAS) [36] style bidirectional full-attention (BFA) decoder [15, 37] for improved accuracy. We also investigate a more general recipe for sequence summarization where we learn not only a biasing feature [25] but also an additional scaling factor. We propose to combine three different types of conditioning namely, concatenative, additive and multiplicative, explored in [38] for feature-wise transformation over a wide variety of application domains including speech recognition. We evaluate the performance of this utterance invariant two-pass architecture on LibriSpeech[39] corpus containing 1K hours of speech data, as well as a large scale Korean corpus with $\sim$10K hours of speech data.

The remainder of this paper is organized as follows. In Section 2, we introduce our proposed sequence summary based utterance invariant training for hybrid e2e system. In Section 3, we present our experimental setup, followed by the results on LibriSpeech [39] and a large scale Korean database along with

*Equal contribution.

http://dx.doi.org/10.21437/Interspeech.2020-3230

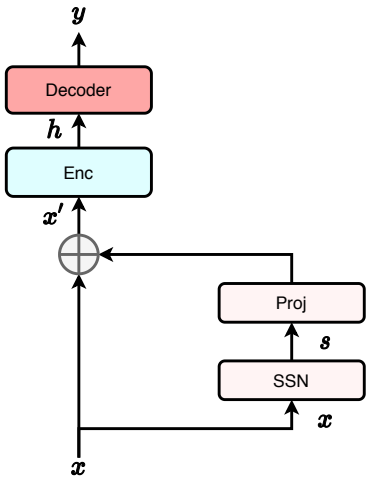

Figure 1: *Vanilla sequence summary network based utterance invariant training of end-to-end ASR models.*

their detailed analysis. Finally, we conclude in Section 4.

2. Utterance invariant two-pass hybrid e2e ASR model

In this section, we present the proposed utterance invariantly trained two-pass hybrid end-to-end ASR. It combines the idea of two-pass decoding with a shared encoder to improve the performance of a streaming e2e ASR, and utterance invariant training using a sequence summary network to improve the overall performance of the two-pass model.

2.1. Vanilla utterance invariant training

Sequence summary networks (SSN) provide an alternative to the popular speaker or utterance invariant training using i-vectors. A sequence summary network computes the length normalized summary of a sequence x non-linearly transformed by neural network model given by

$$s = 1/T \sum_{t=1}^{T} g(x_t), \tag{1}$$

where T denotes the length of the input sequence, and $g(.)$ denotes the non-linear neural transform. A block schematic of our vanilla sequence summary network based encoder-decoder architecture is shown in Fig. 1.

In the conventional utterance invariant training proposed and used in [24, 25], the SSN output is used to conditionally bias the input to the encoder after a linear transformation, given by

$$x' = x + Ps \tag{2}$$

where P denotes the linear projection layer. This approach can be referred to as additive conditioning. However, our experiments with this vanilla conditional biasing using SSN based utterance invariant training did not show any improvement over a baseline bi-directional encoder based full-attention model.

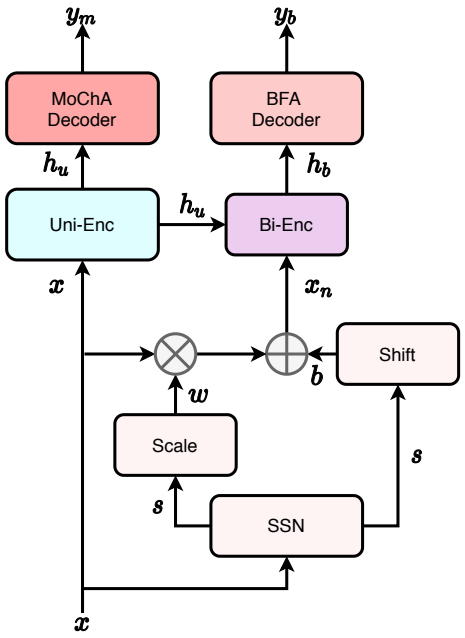

Figure 2: *Block schematic of the proposed utterance invariant training for a two-pass hybrid MoChA-BFA e2e ASR architecture.*

2.2. Conditionally scaled and biased utterance invariant training

Another popular way of conditioning the input or any layer is multiplicative conditioning. Multiplicative conditioning is considered to be useful in learning the inter-relationships between the conditioned and conditioning information. However, in the current case, the conditioned sequence is the input features and the conditioning information is also derived from the same sequence using a SSN, which can be viewed as self-conditioning. Multiplicative or scaled conditioning can also be viewed as a soft-gating mechanism which adaptively emphasizes or deemphasizes different components of an input vector that is being conditioned. In this paper, we propose to combine multiplicative and additive conditioning using the sequence summary network output to conditionally scale as well as bias the input given by

$$x_n = w \odot x + b \tag{3}$$

where $w = Ps$ and $b = Bs$ are the conditional scale and bias, with P and B denoting linear transforms or projections. $\odot$ denotes element-wise multiplication.

In this paper, however, the motivation is primarily to train a two-pass hybrid ASR with shared encoder in an utterance invariant manner. Our proposed two-pass hybrid ASR model architecture and a modified utterance invariant training to suit this architecture are discussed in the next two sections.

2.3. Two-pass hybrid MoChA-BFA e2e ASR

The block schematic of the two-pass end-to-end ASR is shown in Fig. 2. It consists of a shared encoder with several unidirectional long short-term memory (LSTM) layers stacked on top of each other. Outputs of first three layers of the encoder are temporally subsampled by a factor of two, yielding an overall subsampling factor of 8 at the shared encoder output. This

output feeds into two different decoders, one a streaming monotonic chunkwise attention (MoChA) decoder and another a bidirectional full-attention (BFA) decoder. A bidirectional encoder stack with at least one backward LSTM layer processes the output of the shared unidirectional encoder before it is decoded using the full-attention decoder. The first backward LSTM layer takes its input from the last but one forward layer in the shared encoder. The output of the top-most forward layer of the shared encoder and the first backward layer of the bidirectional encoder are concatenated before they are fed into the next bidirectional LSTM (BLSTM) layer in the bidirectional encoder stack.

2.4. Utterance invariant training for two-pass architecture

The shared encoder in our two-pass architecture is unidirectional to enable streaming capabilities. Owing to this requirement we cannot condition the input features or any of the layers higher up in the shared encoder or the MoChA decoder using a complete input sequence summary. One option is to condition with the partial sequence summary so that the entire shared encoder is conditioned to be utterance invariant. However, our experiments on conditioning the input features with partial sequence summary did not show any improvement over our baseline hybrid MoChA-BFA models.

In order to address this issue, we propose to condition the bidirectional encoder layers by feeding the self-conditioned or normalized input sequence, given by Eq. (3), into the bidirectional encoder, as shown in Fig. 2. This self-conditioned input feature sequence summary is concatenated with the unidirectional shared encoder embedding h_u and the output of the first backward layer b_1 in the bidirectional encoder stack. The concatenated input $[h_u, b_1, x_n]$ is fed into the next bidirectional encoder layer. We also concatenatively condition every layer in the bidirectional encoder stack with the same normalized input x_n in our experiments.

The proposed approach can be viewed as feeding the bidirectional decoder with a stream of utterance normalized input features, as well as conditioning the bidirectional encoder layers with the sequence summary using the bias term in Eq. (3). We also tried applying SSN on the output of the shared encoder. However, the results were not encouraging and several experiments had model convergence issues. There could be several ways of combining or conditioning, and also the choice of input sequences for computing the sequence summary. A more exhaustive exploration of all these possibilities would make an interesting piece of research, but is beyond the scope of this paper.

3. Experiments and results

In this section, we present our experimental setup, datasets, and results on the popularly used open-source Librispeech corpus [39]. We also present results on a large-scale in-house Korean corpus with large variability in terms of speakers, recording devices, environments, and domains.

3.1. Experiments on librispeech

3.1.1. Effect of vanilla and modified SSN on BFA models

Experimental results from a simple utterance invariant training on a bi-directional full-attention (BFA) model are shown in Table 1. The initial set of experiments was to validate the usefulness of the vanilla SSN (vSSN) based utterance invariant training on a simple bidirectional encoder based full-attention e2e ASR model. The model architecture and size of the BFA model is same as that used in [15]. The vanilla SSN is similar to [25] with three fully connected feedforward layers. The first two layers have tansig activation with 256 units, and the last layer is linear with 64 units. In these initial experiments, we do not use any data augmentation techniques. The models are trained for around 25 full epochs using a training recipe built *in-house* using the Tensorflow 2 Keras APIs. It can be seen that the vanilla SSN did not show any significant improvement over the baseline results, while the scaled-shifted modified SSN (mSSN) shows good improvement. It is difficult to find a reason as to why the vanilla SSN does not show any improvement since both additive and multiplicative conditioning have shown improvements in literature [38]. However, instead of using a fixed projection on the sequence summary, scaling and shifting it with a data-dependent weight and bias seem to be better suited for conditioning the encoders in the current case. One argument in favor of multiplicative conditioning is that it can better capture the correlations between the conditioning and conditioned data. It can also be interpreted as a soft-gating mechanism to relatively emphasize or deemphasize different dimensions of the input. More detailed comparative experiments of different conditioning and adaptive training variants need to be explored to find a more definitive answer.

Table 1: *Effect of vanilla SSN (vSSN) and modified SSN (mSSN) based utterance invariant training on word error rates (WERs) of a bi-directional full-attention (BFA) Librispeech model.*

Model	test-clean	test-other
BFA	4.85	15.39
BFA + vSSN	4.82	15.01
BFA + mSSN	4.55	13.93

Table 2: *Effect of adding more bidirectional layers on top of the shared unidirectional encoder in a hybrid MoChA-BFA model.*

Model	test-clean	test-other
BFA_1B	5.10	15.83
BFA_2B	**4.97**	**15.35**
BFA_3B	5.13	15.49
BFA_4B	5.15	15.51

3.1.2. Effect of deeper bidirectional encoder stack in hybrid MoChA-BFA ASR

In the proposed two-pass MoChA-BFA model architecture, increasing the depth of the bidirectional encoder could be one way of improving the overall accuracies of the hybrid ASR. Table 2 shows the effect of varying the bidirectional encoder depth from 1 to 4 layers. It can be seen that increasing the depth does not seem to have a significant impact on the performance of the BFA decoder, except for some marginal improvements for a depth of two layers. In these experiments, the models are trained from scratch by omitting the MoChA decoder and the SSN part completely from Fig. 2.

3.1.3. Effect of modified SSN based UIT on hybrid MoChA-BFA ASR

In this section, we study the effect of utterance invariant training on the proposed two-pass hybrid model architecture. Table 3 shows the results for various depths of the bidirectional encoder and for the cases with and without UIT. As compared to the previous two sections, spectral augmentation based regularization is enabled for experiments in this section. This is reflected in the improved baseline results of the two-pass model

Table 3: *Effect of modified SSN based UIT on hybrid MoChA-BFA model with varying number of bidirectional layers.*

Model	w/o UIT		with UIT	
	cln	oth	cln	oth
MoCHA_BFA_1B	4.81	12.71	4.57	12.49
MoCHA_BFA_2B	4.83	12.56	4.49	12.45
MoCHA_BFA_3B	4.77	12.75	4.45	12.50

as compared to the earlier BFA based experiments, especially for test-other. It can be seen that increasing the depth of the bidirectional encoder does not seem to have much effect in the case of Librispeech models. However, utterance invariant training provides consistent $\sim$2-7% relative improvement in WERs over the baseline training for all depths.

3.2. Experiments on large scale Korean ASR system

In this section, we present experimental results for the proposed utterance invariant training on a large scale internal Korean corpus of around 10K hours [1]. The data consists of mostly command and search queries recorded on various mobile and television devices. Multi-condition training is simulated by generating another 10K hours of data using a combination of far-field or acoustic room simulation and additive noise [40, 5, 15]. We hypothesize that this large-scale dataset with large variability in terms of speakers, devices, environments, and domains may be more suited for the utterance invariant training than the more homogeneous Librispeech dataset. A randomized set of around 1 hour is held out as validation data.

The baseline two-pass hybrid MoChA-BFA model was trained in two steps. The MoChA model, with only the shared encoder, was trained using our inhouse Tensorflow 2 Keras APIs based tool for around 10 full epochs till the performance of this model saturated. Standard layerwise pretraining [41] of the encoder along with spectral augmentation [14] were used. The shared encoder has six unidirectional LSTM layers with 1536 units each, and have an overall temporal subsampling of 8 with respect to the input sequence. In the second step, one backward layer with 1536 units was added on top of the shared encoder. This backward layer combined with the top-most forward layer of the shared encoder form the bidirectional encoder that feeds the full-attention decoder. The addition of only one backward layer was chosen to keep the overall 2nd pass latency low while trying to improve upon the MoChA decoder hypothesis during the 1st pass. After training the above model and SVD compression of this layer, it was seen that this 2nd pass decoding added around 80ms additional latency to the original approximately 150ms latency of the MoChA decoder, which is very much within the requirements of a streaming ASR solution.

This baseline model was now converted into an utterance invariant model by adding a sequence summary network, as shown in Fig. 2. Also, at this stage, the depth of the bidirectional encoder is increased from one to three, while trying to keep the overall number of parameters in the bidirectional encoder roughly the same as that in the previous stage. All layers in the bidirectional encoder are reduced to 512 units instead of the earlier single backward layer with 1536 units. The scaled-shifted input feature sequence from the SSN is concatenated with each bidirectional layer and fed to the next layer. Since the decoder from the previous stage received an encoder embedding vector of size $1536 * 2$, we use a simple projection layer to map the modified encoder output from $512 * 2 + 40$ to $1536 * 2$, where 40 is the dimension of the input features.

The results of our proposed utterance invariant training on the large scale Korean e2e ASR model are shown in Table 4. The results are presented for three different test conditions, namely typical personal assistant command and search queries (CSQ), open or general domain dictation (Dict), and far-field mixed test sets. It can be seen that the performance of the hybrid ASR improves by approximately 10.2%, 14.1%, and 15.5% relative to the base model (Base_1B) for the CSQ, Dict, and Far test sets, respectively. The base model was also allowed to train for an additional 7 full epochs to see if there are any long-train effects. It can be seen that long-train has around 2% and 5% relative improvement for CSQ and Dict cases, respectively.

Table 4: *Effect of utterance invariant training on the performance (WER) of a large scale Korean two-pass ASR system.*

Model	CSQ	Dict	Far
MoChA_BFA_1B (Base_1B)	7.05	23.43	29.98
Base_1B + long-train	6.88	22.18	31.28
Base_3B + UIT	6.33	20.12	25.34

Table 5: *Effect of UIT on the performance (WER) of MoChA model part of the two-pass ASR.*

Model	CSQ	Dict	Far
MoChA	8.47	27.58	36.44
MoChA + long-train	8.85	27.03	38.90
MoChA + UIT	7.59	23.37	35.47

3.2.1. Effect of UIT on MoChA model

It is generally known that multi-task training with multiple decoders and loss functions and with a shared encoder can have a mutually beneficial effect on both the decoders. In order to verify this hypothesis, we present the results of the MoChA decoder before and after utterance invariant training. It can be seen from the results in Table 5 that the improvements in the BFA decoder has a positive effect on the MoChA decoder. The MoChA decoder performance improves by around 10.4%, 15.2% and 2.7% for the CSQ, Dict and Far test sets. This is possibly because the shared encoder is trained better due to the UIT training, and has a multi-task regularization effect on the performance of the MoChA decoder. However, unlike the BFA model long training did not have any positive impact on the MoChA model whose performance rather fluctuated and marginally deteriorated.

4. Conclusions

In this paper, we presented a new utterance invariant training strategy using sequence summary networks for training a two-pass MoChA-BFA hybrid ASR. We proposed a scaled and shifted sequence summary network which combines both multiplicative and additive conditioning specifically designed for a two-pass ASR model architecture. In order to retain the unidirectional streaming capability of the MoChA decoder, the input sequence summary is applied directly to the bidirectional encoder by skipping the shared encoder. The proposed utterance invariant training shows up to $\sim$7% relative improvement on Librispeech models, and $\sim$10-15% for different test sets on a large scale Korean two-pass model. The much bigger improvement in the case of Korean model is probably due to the larger variability in data compared to Librispeech data. UIT training of the two-pass ASR architecture also had a positive impact on the streaming MoChA model performance as well.

5. References

[1] K. Kim, K. Lee, D. Gowda, J. Park, S. Kim, S. Jin, Y.-Y. Lee, J. Yeo, D. Kim, S. Jung *et al.*, "Attention based on-device streaming speech recognition with large speech corpus," in *Proc. ASRU*. IEEE, 2019, pp. 956–963.

[2] 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 *Proc. ICASSP*, 2019, pp. 6381–6385.

[3] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. Mcgraw, and C.-C. Chiu, "Two-pass end-to-end speech recognition," in *Proc. INTERSPEECH*, 2019.

[4] A. Garg, D. Gowda, A. Kumar, K. Kim, M. Kumar, and C. Kim, "Improved Multi-Stage Training of Online Attention-based Encoder-Decoder Models," in *Proc. ASRU*, 2019.

[5] C. Kim, S. Kim, K. Kim, M. Kumar, J. Kim, K. Lee, C. Han, A. Garg, E. Kim, M. Shin, S. Singh, L. Heck, and D. Gowda, "End-to-end training of a large vocabulary end-to-end speech recognition system," in *Proc. ASRU* , 2019.

[6] D. Gowda, A. Garg, K. Kim, M. Kumar, and C. Kim, "Multi-task multi-resolution char-to-bpe cross-attention decoder for end-toend speech recognition," in *Proc. Interspeech*, 2019.

[7] A. Garg, G. Vadisetti, D. Gowda, S. Jin, A. Jayasimha, Y. Han, J. Kim, J. Park, K. Kim, S. Kim, Y. Lee, K. Min, and C. Kim, "Streaming on-device end-to-end asr system for privacy-sensitive voicetyping," in *Proc. Interspeech*, 2020.

[8] A. Kumar, S. Singh, D. Gowda, A. Garg, S. Singh, and C. Kim, "Utterance confidence measure for end-to-end speech recognition with applications to distributed speech recognition scenarios," in *Proc. Interspeech*, 2020.

[9] A. Garg, A. Gupta, D. Gowda, S. Singh, and C. Kim, "Hierarchical multi-stage word-to-grapheme named entity corrector for automatic speech recognition," in *Proc. Interspeech*, 2020.

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

[11] C. C. Chiu and C. Raffel, "Monotonic chunkwise attention," in *International Conference on Learning Representations*, 2018.

[12] H. Miao, G. Cheng, P. Zhang, T. Li, and Y. Yan, "Online hybrid ctc/attention architecture for end-to-end speech recognition," *Proc. Interspeech 2019*, pp. 2623–2627, 2019.

[13] 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 *Proc. ICASSP*, 2018, pp. 4774–4778.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.

[15] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," *Proc. Interspeech 2019*, pp. 739–743, 2019.

[16] G. Synnaeve, Q. Xu, J. Kahn, E. Grave, T. Likhomanenko, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[17] C. Kim, K. Kim, and S. R. Indurthi, "Small energy masking for improved neural network training for end-to-end speech recognition," in *Proc. ICASSP*, 2020.

[18] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," *Proc. ICASSP*, pp. 7893–7897, 2013.

[19] S. H. K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fMLLR based feature-space speaker adaptation of dnn acoustic models," in *Proc. Interspeech*, 2015.

[20] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Proc. ICASSP*, 2014, pp. 6334–6338.

[21] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014, pp. 225–229.

[22] P. Cardinal, N. Dehak, Y. Zhang, and J. R. Glass, "Speaker adaptation using the i-vector technique for bottleneck features," in *Proc. INTERSPEECH*, 2015.

[23] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[24] K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *Proc. ICASSP*. IEEE, 2016, pp. 5315–5319.

[25] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary feature based adaptation of end-to-end asr systems." in *Proc. Interspeech*, 2018, pp. 2444–2448.

[26] F. Weninger, J. Andrés-Ferrer, X. Li, and P. Zhan, "Listen, attend, spell and adapt: Speaker adapted sequence-to-sequence asr," in *Proc. INTERSPEECH*, 2019.

[27] K. Li, J. Li, Y. Zhao, K. Kumar, and Y. Gong, "Speaker adaptation for end-to-end ctc models," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 542–549, 2018.

[28] H. Liao, "Speaker adaptation of context dependent deep neural networks," *Proc. ICASSP*, pp. 7947–7951, 2013.

[29] L. Sarı, N. Moritz, T. Hori, and J. Le Roux, "Unsupervised speaker adaptation using attention-based speaker memory for end-to-end asr," in *Proc. ICASSP*. IEEE, 2020, pp. 7384–7388.

[30] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, 1996.

[31] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*. IEEE, 2011, pp. 24–29.

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

[33] Z. Meng, Y. Gaur, J. Li, and Y. Gong, "Speaker adaptation for attention-based end-to-end speech recognition," in *Proc. Interspeech*, 2019, pp. 241–245.

[34] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proc. Workshop on Speech and Natural Language*. ACL, 1992, pp. 357–362.

[35] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks." in *LREC*, 2014, pp. 3935–3939.

[36] 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*, 2016, pp. 4960–4964.

[37] C. Kim, M. Kumar, K. Kim, and D. Gowda, "Power-law nonlinearity with maximally uniform distribution criterion for improved neural network training in automatic speech recognition," in *Proc. ASRU*, Dec. 2019, pp. 988–995.

[38] Vincent Dumoulin, et.al., "Feature-wise transformations: A simple and surprisingly effective family of conditioning mechanisms," 2018. [Online]. Available: https://distill.pub/2018/feature-wise-transformations/

[39] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[40] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Interspeech*, 2017.

[41] A. Zeyer, K. Irie, R. Schl?ter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018.