INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Rapid Speaker Adaptation for Conformer Transducer: Attention and Bias are All You Need

Yan Huang, Guoli Ye, Jinyu Li, and Yifan Gong

Microsoft Corporation, USA
{yanhuang, guoye, jinyli, ygong}@microsoft.com

Abstract

Conformer transducer achieves new state-of-the-art end-to-end (E2E) system performance and has become increasingly appealing for production. In this paper, we study how to effectively perform rapid speaker adaptation in a conformer transducer and how it compares with the RNN transducer. We hierarchically decompose the conformer transducer and compare adapting each component through fine-tuning. Among various interesting observations, there are three distinct findings: First, adapting the self-attention can achieve more than 80% gain of the full network adaptation. When the adaptation data is extremely scarce, attention is all you need to adapt. Second, within the self-attention, adapting the value projection outperforms adapting the key or the query projection. Lastly, bias adaptation, despite of its compact parameter space, is surprisingly effective. We conduct experiments on a state-of-the-art conformer transducer for an email dictation task. With 3 to 5 min source speech and 200 minute personalized TTS speech, the best performing encoder and joint network adaptation yields 38.37% and 19.90% relative word error rate (WER) reduction. Combining the attention and bias adaptation can achieve 90% of the gain with significantly smaller footprint. Further comparison with the RNN-T suggests the new state-of-the-art conformer transducer can benefit as much as if not more from personalization.

Index Terms: Speaker Adaptation, Personalization, Conformer Transducer, Transformer, Speech Recognition

1. Introduction

Among the various forms of E2E models [1–15], conformer transducer has achieved new state-of-the-art performance [16–18]. It models both local and global dependencies through parameter efficient combination of convolution network and transformer. Together with the recent development of effective streaming techniques [19–21], conformer transducer has become increasingly appealing and practical for production. Thus personalization, a widely practiced strategy in industry systems [22–24], naturally becomes our next problem to tackle.

Personalizing an E2E system is challenging due to its unified end-to-end structure. Rapid speaker adaptation refers to adapting a speech model to a specific speaker with limited data (e.g. less than 10 min). It has been studied substantially both in the hybrid [24–32] and E2E systems [33–38]. A comprehensive and up-to-date overview of this topic can be found in [39].

Complexity and cost are the important practical considerations for personalization. In this paper, we would like to answer the following two questions: First, how to effectively perform rapid speaker adaptation in a conformer transducer; Second, to what extent this new state-of-the-art E2E model can benefit from speaker adaptation and how it compares with the RNN transducer (RNN-T) [36].

We adopt a top-down analytic approach to decompose the conformer transducer and compare adapting each component as speaker signature. We found that adapting the self-attention can achieve more than 80% gain of the full network adaptation. When data is scarce, attention is all you need to adapt. In particular, despite with similar size, the value projection adaptation yields significantly larger gain comparing to the key and the query projection. Adapting the bias of the network is surprisingly effective, especially given the parameter space is highly compact. Furthermore, we found adapting the lower and middle conformer layers slightly outperforms the top layers. Avoiding adapting the top conformer layers may help alleviate the impact of imperfect labeling in unsupervised adaption. Lastly, at the top transducer level, the encoder adaptation is significantly more effective than adapting the prediction, joint, and softmax network, consistent with our previous study in the RNN-T [36].

We conduct experiments on a state-of-the-art conformer transducer for an email dictation task. Each speaker has 3 to 5 min source adaptation speech. The best performing encoder and joint adaptation yields 21.02% and 6.68% relative WER reduction for the supervised and unsupervised setup respectively. After adding personalized TTS speech, the gain increases to 38.37% and 19.90%. Attention and bias adaptation can achieve 90% of the above gain. Conformer transducer, although with much stronger baseline performance, can still benefit from personalization as in the RNN transducer if not even more.

To the best of our knowledge, we are not aware of any previous work on rapid speaker adaptation on conformer transducer, especially on an industry quality system. Our analytic study also provides some good insights on the conformer transducer.

The rest of this paper is organized as: Section 2 introduces the methodology; Section 3 presents the experiments and results; Section 4 concludes the paper.

2. Methodology

In this section, we first introduce the motivation and background work, then review the model architecture and describe the speaker signature for conformer transducer adaptation.

2.1. Motivation

Most speaker adaptation methodologies dedicate specific parameter space to modeling speaker trait; thus achieves more accurate modeling and better performance for individual speakers. We can think of this speaker specific parameter space as speaker signature. There are abundance of literature in formulating speaker signature using additional network structure, such as i-vector [29,40], svd-based adaptation [26], learning hidden unit contribution (LHUC) [27], factorized analysis [31], speaker-aware persistent memory [35]. Regularization is another commonly used technique to address overfitting [23,25,41].

http://dx.doi.org/10.21437/Interspeech.2021-1884

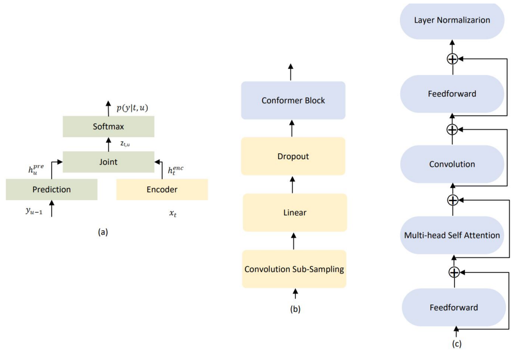

Figure 1: *Hierarchical structure breakdown of conformer transducer: (a) transducer; (b) encoder; (c) conformer block.*

Complexity and cost are the important considerations when designing personalization strategies for a practical speech recognition system with large number of users. One needs to find the best trade-off between the accuracy and the computation, storage, system maintenance cost. Therefore, we focus on searching for the salient and compact speaker signature within the conformer transducer without introducing additional model parameters or modifying the original model structure.

2.2. Architecture of Conformer Transducer

The conformer transducer was first proposed in [16, 18]. The architecture of our conformer transducer is depicted in Fig. 1. It has a similar model structure as in [16].

At the top-level, conformer transducer is a standard transducer, which consists of an encoder, a prediction, and a joint network. The topology is depicted in Fig. 1 (a) and the network is formally specified in Eq. (1). The encoder network converts the acoustic feature x_t into a high-level representation h_t^{enc}, where t is the index of time. The prediction network generates a high-level representation h_{pre}^u by conditioning on the previous non-blank target y_{u-1} predicted by the transducer, where u is the index of the label. The joint network $z_{t,u}$ is a feed-forward network that combines the encoder output h_{enc}^t and the prediction output h_u^{pre}. $z_{t,u}$ is connected to the output layer with a linear transform followed by a softmax. $P(k|t,u)$ is the posterior of each output token k.

$$\begin{aligned}
h_t^{enc} &= f^{enc}(x_t) \\
h_u^{pre} &= f^{pre}(y_{u-1}) \\
z_{t,u} &= f_{joint}(h_t^{enc}, h_u^{pre}) \\
P(k|t,u) &= \text{softmax}(W_y z_{t,u}^k + b_y)
\end{aligned} \tag{1}$$

The encoder network consists of a convolution sub-sampling layer followed by stacks of conformer blocks as depicted in Fig. 1 (b). The convolution sub-sampling layer is a stack of 2D convolution layers which sub-samples the speech frames by a certain pre-defined factor. For example, the 10 ms frame rate is converted to the 40 ms frame rate after passing through the convolution layers with sub-sampling factor of 4.

The conformer block contains the feedforward (FFN), convolution (Conv), and multi-head self-attention (MHSA) as depicted in Fig. 1 (c). The convolution layer models the local dependency. The multi-head attention, placed between two feedforward layers, models the global dependency. All layers are connected with residue connections. Formally,

$$\begin{aligned}
\tilde{x}_i &= x_i + \frac{1}{2}\text{FFN}(x_i) \\
x_i' &= \tilde{x}_i + \text{MHSA}(\tilde{x}_i) \\
x_i'' &= x_i' + \text{Conv}(x_i') \\
y_i &= \text{Layernorm}(x_i'' + \frac{1}{2}\text{FFN}(x_i''))
\end{aligned} \tag{2}$$

The self-attention is an important modeling component in a conformer. It projects the input (x_t) to the query, key, and value space through the learnable parameters W_q, W_k, and W_v. Self-attention uses dot-product to compute the attention distribution ($\alpha_{t,\tau}$) over the query ($W_q x_t$) and the key ($W_k x_t$). This distribution is then used to weight the value ($W_v x_t$) and calculate the output embedding (z_t). β is a scaling factor.

$$\begin{aligned}
\alpha_{t,\tau} &= \frac{\beta(W_q x_t)^T W_k x_\tau}{\sum_{\tau'} \beta(W_q x_{\tau'})^T W_k x_{\tau'}} \\
z_t &= \sum_{\tau} \alpha_{t,\tau} W_v x_\tau
\end{aligned} \tag{3}$$

2.3. Speaker Signature

We adopt a top-down analytic approach to search for the salient and compact speaker signature within the conformer transducer.

2.3.1. Encoder: transducer-level speaker signature

At the top transducer level, we expect the encoder to be a more salient speaker signature as the speaker voice trait modeling largely resides in the encoder. The prediction network, generally believed to carry primarily language-level information or simply handling speech alignment according to a recent study [42], is expected to be less relevant in rapid adaptation. The joint network specifies the fusion of acoustic and language information, which can also potentially be optimized per speaker basis but with limited adaptation capacity.

2.3.2. Conformer: encoder-level speaker signature

Within the encoder, the convolution sub-sampling layer contains the low-level acoustic environment and channel information. It can be relevant in some speaker adaptation scenario when such kinds of mismatch are distinct. Nevertheless, we expect that the conformer blocks modeling detailed speech pattern carries more salient speaker information.

2.3.3. Attention: conformer-level speaker signature

Within the conformer block, as described earlier, a multi-head self-attention and convolution layer are sandwiched between two feedforward layers. The convolution layer models local dependency, which is expected to be less speaker specific; while the self-attention models global dependency, which potentially carries more speaker specific information.

2.3.4. Value projection: attention-level speaker signature

The multi-head self-attention is parameterized by projections of the query (W_q), key (W_k), and value (W_v), which project the input to the query, key, and value space as in Eq. (3). In our model, they have the same number of parameters. Value projection is the basis of the embedding representation, while the query and key projection are only used to determine the combination weight. Thus the value projection has larger representation capacity, potentially a more effective speaker signature.

2.3.5. Bias and layer normalization as speaker signature

Bias specifies a systematic drift of the neuron activation, which is used in different neurons throughout the conformer transducer. For example, vector b_y in Eq. (1) is an example of bias. Layer normalization, measuring the distribution within the layer, may as well carry speaker specific information. They are both compact in size, potentially more robust to data scarcity and labeling errors. We collect all the bias or the normalization layers as potential speaker signatures to be adapted.

3. Experiment and Result

In this section, we present speaker adaptation experiment. We first compare different speaker signatures for adaptation on the PowerPoint presentation task; then further verify our selected adaptation signatures on an email dictation task.

3.1. Baseline and Setup

The baseline conformer transducer is trained from 65K hour speech with the cross-entropy criterion. The training data is anonymized with personal identifiable information removed. The encoder network follows the similar structure as [20]. It consists of two convolution layers that sub-sample the time frame by a factor of 4, followed by 18 conformer layers. Each conformer layer has a multi-head attention with 8 heads and a depth-wise convolution with kernel size 3. The multi-head attention and the depth-wise convolution are sandwiched between two 1024-dim feedforward layers. The prediction network consists of two layer-normalized LSTM layers. Each LSTM layer has 1024 hidden units. The output size is reduced to 512 using a linear projection layer. We used an 80-dim log-mel filterbank extracted every 10 ms for the input feature. The output layer models 4030 word pieces plus an additional blank label.

The PowerPoint presentation consists of six speakers, each with 10 min for training and 20 min for testing. The email dictation consists of seven speakers, each with 5 min for training and 10 min for testing. Both are internally collected data sets. We use the Adam optimizer with fixed small learning rate (0.00001) in the adaptation with no regularization performed.

3.2. Speaker Signature

The comparison of different speaker signature is presented in Table 1. We use the PowerPoint presentation task for this study. The parameter size of each signature are presented in Fig. 2.

First, we compare adapting different components at the top transducer level. Adapting the encoder yields significantly larger gain than adapting the prediction, joint, and softmax network. This confirms that speaker voice trait largely resides in the encoder, consistent with our previous study [36].

We then decompose the encoder into the convolution subsampling network and the conformer blocks. We found that adapting convolution sub-sampling can yield moderate gain, but majority of the encoder adaptation gain comes from the conformer block adaptation. As mentioned earlier, the convolution sub-sampling adaptation may help compensate the low-level acoustic and channel mismatch for individual speakers.

Next, we compare adapting the feedforward, convolution, and attention network within the conformer block. Attention adaptation achieves more than 80% of the full network adaptation. In the Section 3.5, we will show that when adaptation data is extremely scarce, attention adaptation can achieve robust accuracy gain. Nevertheless, when labeling error exists, the performance is notably affected. In comparison, the convolution layer adaptation yields much smaller gain for the supervised adaptation, while exhibiting robustness to the labeling error.

Lastly, we zoom into the self-attention and adapt the key (W_K), query (W_Q), and value (W_V) projection separately. Despite of sharing similar parameter size, the value projection adaptation results in significantly larger gain. We observe the same pattern in both supervised and unsupervised adaptation. As a matter of fact, in unsupervised adaptation, it outperforms full network adaptation. In Section 3.5, we will also show the similar pattern in extremely rapid speaker adaptation.

Table 1: *Comparison of different speaker signature on the PowerPoint presentation task. Results of 10 min supervised and unsupervised are presented. The component annotated with + is further broken down to its building blocks in the next table session. WER.R refers to the relative WER reduction.*

Model	SUP	**WER.R**	UNSUP	**WER.R**
Baseline	11.38	**NA**	11.38	**NA**
ALL^+	9.45	**16.94**	10.85	**4.67**
$Encoder^+$	9.43	**17.10**	10.78	**5.23**
$Prediction$	11.38	**-0.01**	10.34	**0.31**
$Joint$	11.20	**1.57**	11.34	**0.34**
$Softmax$	11.13	**2.14**	11.32	**0.50**
$Conformer^+$	9.51	**16.42**	10.82	**4.86**
$Conv(2D)$	10.86	**4.53**	10.86	**4.53**
$Attention^+$	9.65	**15.22**	10.87	**4.44**
$Feedforward$	9.80	**13.84**	10.65	**6.40**
$Conv$	10.71	**5.90**	10.71	**5.87**
$Attention(W_V)$	10.11	**11.16**	10.84	**4.75**
$Attention(W_K)$	11.11	**2.34**	11.32	**1.32**
$Attention(W_Q)$	11.16	**1.88**	11.27	**0.94**

Table 2: *Comparison of adapting the bias, layer normalization, and conformer layers at different network depth on the PowerPoint presentation task. Results of 10 min supervised and unsupervised are presented.* $Layer[1-6]$, $Layer[7-12]$, *and* $Layer[13-18]$ *refer to the bottom, middle, and top six conformer layers. WER.R is the relative WER reduction.*

Model	SUP	**WER.R**	UNSUP	**WER.R**
Baseline	11.38	**NA**	11.38	**NA**
ALL	9.45	**16.94**	10.85	**4.67**
$Norm$	10.81	**5.01**	10.83	**4.85**
$Bias$	9.72	**14.58**	10.80	**5.11**
$Layer[1-6]$	9.81	**13.76**	10.74	**5.58**
$Layer[7-12]$	9.65	**15.16**	10.77	**5.30**
$Layer[13-18]$	10.08	**11.38**	10.87	**4.48**

3.3. Bias and Layer Normalization

Taking a slightly different perspective, we choose to adapt alternative distinct components of the network, e.g. bias and layer normalization. Adapting the neuron bias, as in Table 2, is surprisingly effective and parameter efficient. Adapting a tiny fraction of the full network achieves more than 80% of the full network adaptation gain for the supervised setup. It even outperforms the full network adaptation for the unsupervised setup. In comparison, adapting the normalization layers is not as nearly effective in the supervised adaption, which yet exhibits good robustness to labeling errors in the unsupervised setup.

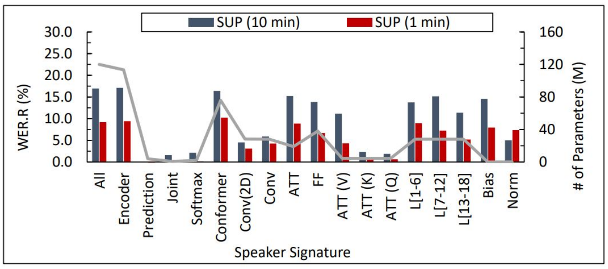

Figure 2: *Comparison of different speaker signature in 1 min and 10 min supervised adaptation on the PowerPoint presentation task. The grey curve is the number of parameters for speaker signature. WER.R refers to the relative WER reduction.*

3.4. Conformer Block at Different Network Depth

Taking yet another different perspective, we compare adapting conformer layers at different network depth. Our conformer transducer consists of 18 conformer blocks, which are grouped into three sets of layers from bottom to top, e.g. $Layer[1-6]$, $Layer[7-12]$, $Layer[13-18]$. As presented in Table 2, adapting bottom and mid conformer layers slightly outperforms the top layers, especially for the unsupervised setup. Avoiding adapting top conformer layers may slightly help alleviate the impact of the imperfect transcription.

3.5. Extremely Rapid Speaker Adaptation

To fully understand the behavior of different speaker signature in a more challenging adaptation setup, we further reduce the adaptation data from 10 min to 1 min. Fig. 2 presents the adaptation results. The parameter size for each speaker signature is also included for reference. We observe similar performance trend in the 1 min adaptation. In particular, we found that the compact signature such as attention and bias adaptation is more beneficial when data is extremely scarce.

3.6. Email Dictation Task

In this session, we conduct conformer speaker adaptation experiment on the email dictation task. We apply the data augmentation with personalized TTS speech. The details of leveraging personalized TTS for speaker adaptation, including the impact of the source data amount, TTS data amount, and the label quality, can be found in [36]. Supervised and unsupervised adaptation with or without TTS speech are presented in Table 3. Five different adaptation structures are presented: full network (ALL), encoder and joint (E + J), attention and joint (A + J), bias (B), attention and bias (A + B). The best performing encoder and joint adaptation yields 21.02% and 6.68% relative WER reduction for the supervised and unsupervised setup respectively. After adding 200 minute personalized TTS speech, the gain increases to 38.37% and 19.90%, consistent with our previous findings [36]. Combining the attention and bias adaptation can achieve 90% of the above gain. After the submission of this paper, we conduct experiments on a new data set with more speakers and obtain similar results.

3.7. Compare with RNN Transducer

We compare with the RNN-T adaptation on the email dictation task. Results of supervised and unsupervised adaptation, with/without augmented personalized TTS speech, are presented in Fig. 3. Both the conformer and the RNN transducer show substantial improvement especially after applying data augmentation. The lack of gain for some speakers when adapting only with original data is due to sever data scarcity. Unsupervised adaptation is more challenging for both models. Nevertheless, with the personalized TTS speech, the unsupervised adaption yields more than 15% relative WER reduction. Conformer transducer, as a stronger baseline (generally indicating less room to improve), can benefit from personalization as much as the RNN transducer. This is likely because the conformer structure provides more flexibility for improvement.

Table 3: *Speaker adaptation on the email dictation task. Five adaptation structures are presented: full network (ALL), encoder (E) and joint (E + J), attention and joint (A + J), bias (B), attention (A) and bias (A + B). $+T200$ refers to adding 200 minute personalized TTS speech for adaptation.*

Model	SUP	**WER.R**	UNSUP	**WER.R**
Baseline	11.72	**NA**	11.72	**NA**
ALL	9.27	**20.93**	10.90	**7.01**
ALL_{+T200}	7.48	**36.21**	9.58	**18.29**
$E+J$	9.26	**21.02**	10.94	**6.68**
$E+J_{+T200}$	7.22	**38.37**	9.39	**19.90**
$A+J$	10.15	**13.38**	11.06	**5.64**
$A+J_{+T200}$	7.56	**35.54**	9.62	**17.89**
B	10.45	**10.86**	11.14	**5.00**
B_{+T200}	9.60	**18.12**	10.49	**10.47**
$A+B$	10.19	**13.07**	10.97	**6.39**
$A+B_{+T200}$	7.99	**35.21**	9.89	**15.61**

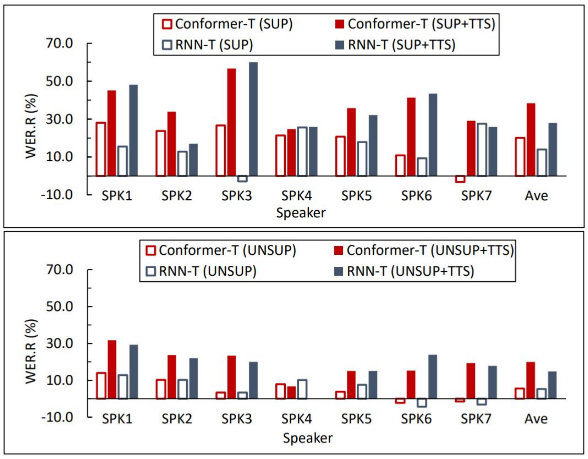

Figure 3: *RNN-T and Conformer-T speaker adaptation comparison on the email dictation task with/without augmented personalized TTS speech. All are based on the encoder adaptation.*

4. Conclusions

In summary, we studied the salient and compact speaker signature for conformer transducer personalization. We found that self-attention and bias are two most distinct speaker signatures. They provide a good balance between accuracy performance and computation/storage/maintenance cost for personalization in a practical system. We also conclude that personalization in the conformer transducer is as promising as in the RNN-T.

5. Acknowledgements

The authors would like to thank Lei He, Wenning Wei, Sheng Zhao, and Yanqing Liu for providing the personalized TTS.

6. References

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

[2] A. Kim, T. Hori, and W. S., "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *Proceedings of ICASSP*, 2017.

[3] C. C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, K. Kannan, R. J. Weiss, R. K., and K. Goninaetal, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proceedings of ICASSP*, 2018.

[4] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer," in *Proceedings of ASRU*, 2017.

[5] Y. He, T. Sainath, and et al., "Streaming end-to-end speech recognition for mobile devices," in *Proceedings of ICASSP*, 2019.

[6] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving RNN transducer modeling for end-to-end speech recognition," in *Proceedings of ASRU*, 2019.

[7] K. Hu, T. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *Proceedings of ICASSP*, 2020.

[8] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*, 2006.

[9] A. Graves, "Sequence transduction with recurrent neural networks," in *CoRR, vol. abs/1211.371*, 2012.

[10] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of ICASSP*, 2016.

[11] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proceedings of ICASSP*, 2017.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of NIPS*, 2017.

[13] C. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," in *Proceedings of ICASSP*, 2019.

[14] S. Karita, N. Chen, and et.al., "A comparative study on transformer vs RNN in speech applications," in *Proceedings of ASRU*, 2019.

[15] A. Tripathi, J. Kim, Q. Zhang, and H. Sakv, "Transformer transducer: One model unifying streaming and non-streaming speech recognition," in *Proceedings of ICASSP*, 2020.

[16] A. Gulati, J. Qin, and et.al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of Interspeech*, 2020.

[17] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," in *Proceedings of Interspeech*, 2020.

[18] W. Huang, W. Hu, Y. Yeung, and X. Chen, "Conv-transformer transducer: Low latency, low frame rate, streamable end-to-end speech recognition," in *Proceedings of Interspeech*, 2020.

[19] Y. Shi, Y. Wang, C. Wu, C. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *arXiv preprint arXiv:2010.10759*, 2020.

[20] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech," in *arXiv preprint arXiv:2010.11395*, 2020.

[21] Y. Wang, Y. Shi, F. Zhang, C. Wu, J. Chan, C. Yeh, and A. Xiao, "Transformer in action: A comparative study of transformer-based acoustic models for large scale speech recognition applications," in *arXiv preprint arXiv:2010.14665*, 2020.

[22] I. McGraw, R. Prabhavalkar, R. Alvarez, and et.al., "Personalized speech recognition on mobile devices," in *Proceedings of ICASSP*, 2016.

[23] K. C. Sim, L. Johnson, G. Motta, and H. Zhang, "Personalization of end-to-end speech recognition on mobile devices for named entities," in *Proceedings of ASRU*, 2019.

[24] Y. Huang and Y. Gong, "Acoustic model adaptation for presentation transcription and intelligent meeting assistant systems," in *Proceedings of ICASSP*, 2020.

[25] D. Yu, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proceedings of ICASSP*, 2013.

[26] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Proceedings of ICASSP*, 2014.

[27] P. Swietojanski, J. Li, and S. Renal, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 1450–1463, 2016.

[28] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proceedings of ICASSP*, 2013.

[29] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, 2013.

[30] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, p. 459–468, 2016.

[31] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," in *Proceedings of ICASSP*, 2014.

[32] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong, "Using personalized speech synthesis and neural language generator for rapid speaker adaptation," in *Proceedings of ICASSP*, 2020.

[33] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: End-to-end contextual speech recognition," in *Proceedings of SLT*, 2018.

[34] F. Weninger, J. Andres-Ferrer, X. Li, and P. Zhan, "Listen, attend, spell and adapt: Speaker adapted sequence-to-sequence ASR," in *Proceedings of Interspeech*, 2019.

[35] Y. Zhao, C. Ni, C. Leung, S. Joty, E. Chng, and B. Ma, "Speech transformer with speaker aware persistent memory," in *Proceedings of Interspeech*, 2020.

[36] Y. Huang, J. Li, L. He, W. Wei, W. Gale, and Y. Gong, "RNN-T adaptation using personalized speech synthesis and neural language generator," in *Proceedings of Interspeech*, 2020.

[37] K. Li, Z. Liu, T. He, H. Huang, F. Peng, D. Povey, and S. Khudanpur, "An empirical study of transformer-based neural language model adaptation," in *Proceedings of ICASSP*, 2020.

[38] L. Sarı, N. Moritz, T. Hori, and J. L. Roux, "Unsupervised speaker adaptation using attention-based speaker memory for end-to-end ASR," in *Proceedings of ICASSP*, 2020.

[39] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for speech recognition: An overview," in *IEEE Open Journal of Signal Processing*, February 2021.

[40] A. Senior and I. L. Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proceedings of ICASSP*, 2014.

[41] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, and et.al., "Overcoming catastrophic forgetting in neural networksms," in *arXiv:1612.00796*, 2016.

[42] M. Mohammadreza Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "RNN-transducer with stateless prediction network," in *Proceedings of ICASSP*, 2020.