IMPROVING CONTEXTUAL SPELLING CORRECTION BY EXTERNAL ACOUSTICS ATTENTION AND SEMANTIC AWARE DATA AUGMENTATION

Xiaoqiang Wang, Yanqing Liu, Jinyu Li, Sheng Zhao

Microsoft Corporation, Redmond, WA, US

ABSTRACT

We previously proposed contextual spelling correction (CSC) to correct the output of end-to-end (E2E) automatic speech recognition (ASR) models with contextual information such as name, place, etc. Although CSC has achieved reasonable improvement in the biasing problem, there are still two drawbacks for further accuracy improvement. First, due to information limitation in text only hypothesis or weak performance of ASR model on rare domains, the CSC model may fail to correct phrases with similar pronunciation or anti-context cases where all biasing phrases are not present in the utterance. Second, there is a discrepancy between the training and inference of CSC. The bias list in training is randomly selected but in inference there may be more similarity between ground truth phrase and other phrases. To solve above limitations, in this paper we propose an improved non-autoregressive (NAR) spelling correction model for contextual biasing in E2E neural transducer-based ASR systems to improve the previous CSC model from two perspectives: Firstly, we incorporate acoustics information with an external attention as well as text hypotheses into CSC to better distinguish target phrase from dissimilar or irrelevant phrases. Secondly, we design a semantic aware data augmentation schema in training phrase to reduce the mismatch between training and inference to further boost the biasing accuracy. Experiments show that the improved method outperforms the baseline ASR+Biasing system by as much as 20.3% relative name recall gain and achieves stable improvement compared to the previous CSC method over different bias list name coverage ratio.

Index Terms— speech recognition, contextual spelling correction, contextual biasing, external attention

1. INTRODUCTION

Contextual biasing is a challenging task for end-to-end (E2E) automatic speech recognition (ASR) systems [1], which improves the recognition performance by biasing the model to a specific domain phrase list including a user's contact names, songs, location, and other contextual information. Prior works for contextual biasing of E2E ASR system can be classified into three categories. The first method is to represent the biasing phrases as a finite state transducer (FST) and incorporate it into the beam-search decoding framework of E2E model [2, 3, 4, 5, 6]. The second method is to directly incorporate the contextual information into the E2E model with a bias encoder [4, 7, 8, 9, 10, 11]. To deal with the scalability issues [4] when with large biasing phrase list and further improve the biasing performance, the contextual spelling correction method is also proposed by biasing the recognition hypothesis with an efficient and small contextual spelling correction (CSC) model [12, 13], which acts as a post processing module. Compared to the first two categories, due to the post processing nature, CSC can pre-select biasing phrases with a filter mechanism, which reduces the effective number of biasing phrases to avoid attention diffusion.

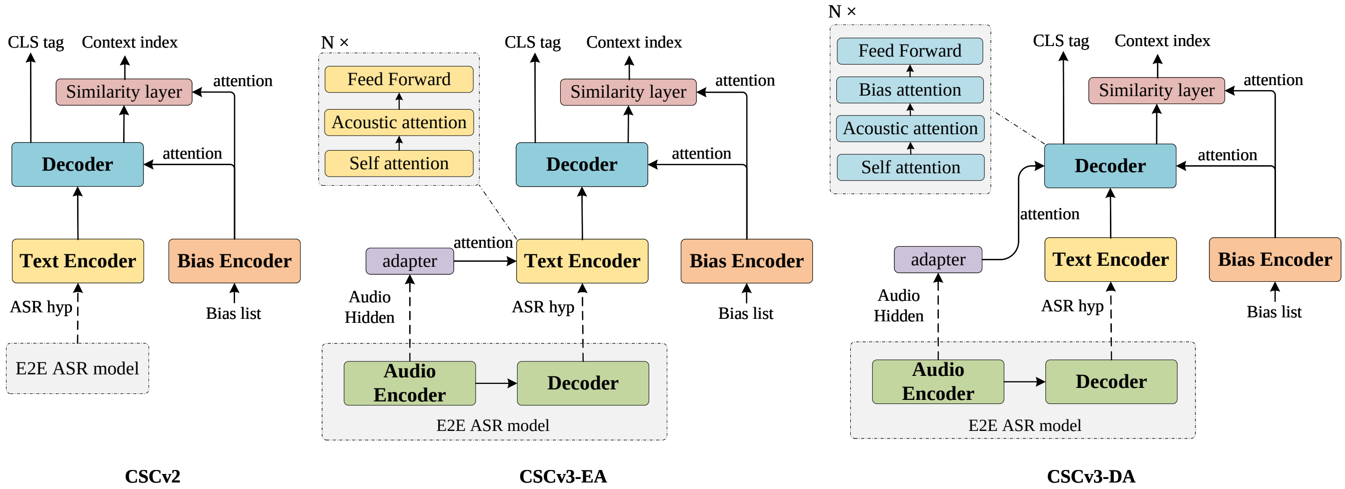
Autoregressive contextual spelling correction (CSCv1) [12] on top of ASR model as a post processing method has shown improvement on phrase biasing problems, which incorporates the contextual information into an autoregressive (AR) spelling correction model [14, 15], but the efficiency is poor due to its autoregressive nature. [13] (CSCv2) introduces a new non-autoregressive (NAR) contextual spelling correction model and incorporates context information into the decoder by attending to the contextual hidden representations from the bias encoder with an attention mechanism [16], as shown in Figure 1. In CSCv2, the decoder directly takes hidden states from text encoder as input, attends to the bias phrase hidden representations, and outputs a position-wise classification (CLS) tag cls and context index $cind$ for each input token. The CLS tag uses ”BILO” representation where ”B”, ”I”, and ”L” represent the beginning, inside and last position of a context phrase, ”O” represents a general position outside of a context phrase; $cind$ is the expected index of the ground-truth context phrase in the bias list for each position. The final correction output can then be determined by cls and $cind$. CSCv2 greatly improves the inference efficiency especially for low-end devices or resource limited systems but has similar biasing performance like its AR counterparts.

However, both CSCv1 and CSCv2 may fail on the cases that hypotheses are totally irrelevant to the ground truth context phrase or on the cases that have more biasing phrases with similar pronunciation but dissimilar written format. On the other hand, although [12, 13] use filter mechanisms for large context list to improve inference efficiency, its training hypotheses-reference pairs prepared with synthesized[17] or human speech still have similarity gap in real scenario and it's not easy for a CSC model to distinguish similar phrases with similar pronunciation or written format from limited hypothesis information only.

Acoustics information has played an important role for ASR post processing besides text hypotheses in recent research [18, 19, 20, 21, 22]. [23, 24] combine both acoustics and first-pass text hypotheses for second-pass decoding, with an RNN-T or transformer model generating the first-pass hypotheses, then a deliberation model attending to both acoustics and first-pass hypotheses for a second-pass decoding. This shows improvement over text hypotheses only post processing model and inspires us to take acoustic information to improve CSC. The intuition is that if the text hypotheses can't provide useful information for bias correction, acoustic information will help to complement the missing context.

In this paper, we proposed a new CSCv3 model, which combines both acoustic and text hypothesis for contextual spelling correction. Specifically, we introduce the designs as follows.

- we propose to combine acoustics and first-pass text hypotheses for second-pass contextual spelling correction with biasing phrases as input. The proposed CSCv3 model has a similar structure as



CSCv2 **CSCv3-EA** **CSCv3-DA**

Fig. 1. Model structure of CSCv2 and CSCv3. CSCv3 model leverages acoustic information with external acoustic attention. We propose two architectures: CSCv3-EA incorporates acoustic information in text encoder, while CSCv3-DA incorporates acoustic information in decoder.

CSCv2 [13] except for the additional acoustic attention: a text encoder generates hidden vectors of the first-pass hypothesis from ASR model conditioned on acoustics hidden from ASR audio encoder output by an acoustic attention, a bias encoder generates contextual embeddings from biasing phrases, then a transformer decoder attends to both hypothesis and contexts for a second-pass decoding to generate phrase similarity and index for each encoder input position.

- we use data augmentation for the context list construction during training with more similar pronunciation phrases or irrelevant phrases to improve the robustness of inference.

Our experiments show that the proposed method achieves 20.3% relative name recall gain improvement compared to the baseline end-to-end ASR+Biasing system, and significantly outperforms the previous CSCv2 model.

2. METHODOLOGY

2.1. Model Architecture

As shown in Figure 1, the proposed CSCv3 model consists of 4 components: text encoder, bias encoder, acoustics adapter, and decoder. The text encoder takes ASR hypothesis as input and encodes text information into hidden states. The bias encoder converts the biasing phrases into phrase level embeddings, which adopts a multi-layer transformer encoder structure as in CSCv2. The decoder uses the NAR (non-autoregressive) infrastructure which directly takes the encoder outputs as input and outputs a position-wise classification (CLS) tag cls and context index tag $cind$ for each input token with the same definition described in Section 1

2.1.1. External acoustic attention

The acoustic encoder hidden of ASR model is adapted by the acoustic adapter in CSCv3 and then fed into CSC model with external attention. The acoustic adapter consists of two linear layers with ReLU activation function and dropout in between.

We explore two different structures to leverage the acoustic information, as shown in Figure 1. In CSCv3-EA, the acoustic information is incorporated into the text encoder by a cross attention which is added between the self-attention and feedforward module in each encoder layer, the decoder follows the same transformer decoder structure as in CSCv2, with self-attention, bias cross attention, and feedforward modules in each layer. For CSCv3-DA, the acoustic information is incorporated into the decoder by cross attention which is added between the self-attention and bias attention module in each decoder layer, while the text encoder keeps the same structure as in CSCv2. The comparison results between the two structures is shown in Section 4.

Since the input audio feature sequence is typically long, we define an audio feature mask which masks audio features that are far from the position that corresponds to the current word piece token. For each token, we only attend to audio features corresponding to the surrounding S_k words of this token. S_k is randomly sampled from uniform distribution $[1, S_{kmax}]$, where S_{kmax} is a pre-defined parameter.

2.1.2. Semantic aware data augmentation

The training pairs of CSCv3 are constructed randomly with the offline prepared data during training. To generate the biasing phrase list for each utterance, CSCv2 randomly samples N_b biasing phrases from an existing large biasing phrase list besides the reference context phrase. N_b is randomly sampled from uniform distribution $[1, N_{bmax}]$, where N_{bmax} is the pre-defined max biasing phrase list size, which doesn't consider irrelevant hypotheses or pronunciation similar phrases. As shown in the following table, CSCv3 improves the sampling strategy with the prepared reference-hypotheses pairs by two ways:

(1) Except for the raw ASR hypothesis of each utterance, we also randomly replace hypothesis of the context phrases with the prepared reference-hypotheses pairs, which improves the data varieties.

(2) To deal with anti-context cases where all biasing phrases are not present in the utterance, we randomly add two types of training data with probability P_{anti}: for the first type, the ground-truth context phrase is simply removed from the bias list, and the corresponding model output is also modified as non-context case; for the second type, we not only remove the ground-truth context phrase,

but also add similar phrases into the biasing phrase list. These similar phrases come from the hypotheses in the prepared reference-hypotheses pairs.

Reference	*Call John at ten a.m.*
ASR hypothesis	*Call Joe at ten a.m.*
bias list	{*Sam*, *John*, *Dong*, *...*}
Context ref-hyp pair	*John* – {*Jane*, *Jon*, *June*, *Joe*}
(1) Replace hypothesis with ref-hyp pair:	
Hyp x	*Call Jane at ten a.m.*
Ref y	*Call John at ten a.m.*
(2.1) Remove ground-truth context phrase:	
Hyp x	*Call Joe at ten a.m.*
bias list	{*Sam*, ~~*John*~~, *Dong*, *...*}
Ref y	*Call Joe at ten a.m.*
(2.2) Add similar phrases into bias list:	
Hyp x	*Call Joe at ten a.m.*
bias list	{*Sam*, ~~*John*~~, *Dong*, *Jane*, *Jon*, *...*}
Ref y	*Call Joe at ten a.m.*

2.1.3. *Fast partial adaptation*

CSCv3 leverages both acoustics and text hypotheses information for better context biasing. Training from scratch is time consuming, for quick adaptation, we train the CSCv3 model based on a baseline CSCv2 model and only update new components in CSCv3, which includes audio adapter and acoustic attention layers. This strategy "inserts" acoustic information into the raw CSCv2 model and we will show its effectiveness in Section 4. We also use a parameter r to incorporate the acoustic information into the model. In each encoder layer of CSCv3-EA and decoder layer of CSCv3-DA, the data flow of acoustic attention layer can be expressed as:

$$x = x_0 + r \cdot \text{dropout}(\text{AcousticAtt}(\text{norm}(x_0))), \qquad (1)$$

where x_0 and x are the input and output of acoustic attention layer. r is randomly sampled from a uniform distribution $[0.0, 1.0]$, which represents the incorporation ratio of acoustic information in the model.

2.2. Training Optimization

2.2.1. *Loss Objectives*

Like CSCv2, the loss function is the sum of CLS tag loss and context index loss:

$$L = H(\widehat{y_{cls}}, y_{cls}) + H(\widehat{y_{cind}}, y_{cind}). \qquad (2)$$

We also use teacher-student learning [25, 26] to make the model smaller and more efficient. The loss function of the student model contains a hard loss L_{hard} which is the loss of student model output y_S to reference y, and a soft loss which is KL-divergence of y_S to teacher model output y_T:

$$L = \alpha L_{soft} + (1 - \alpha) L_{hard} \qquad (3)$$

$$L_{hard} = H(y_S, y) \qquad (4)$$

$$L_{soft} = D_{\text{KL}}\left(\text{softmax}(\frac{y_S}{T}), \text{softmax}(\frac{y_T}{T})\right) \cdot T^2 \qquad (5)$$

where T is a temperature hyper-parameter to adjust the smoothness of output probabilities, α determines the proportion of hard loss and soft loss.

2.2.2. *Data processing*

To generate the training data for CSCv3, we first decode the E2E ASR model for utterances with person names which are extracted from ASR model training set. The top-one hypothesis, audio encoder outputs, and forced alignment of the hypothesis and audio are needed for training. Then we locate and tag the positions of person names in each transcript, which is used to construct reference outputs during training.

Despite the raw ASR hypothesis, we also used a text to speech (TTS) system to generate synthetic audios for the person names. These synthetic audios are then fed into the ASR model to get hypotheses with more varieties. In this way we construct a set of reference-hypotheses pairs for the person names.

2.3. Inference

Like CSCv2, we use an edit distance-based relevance ranker (rRanker) to pre-select biasing phrases from the raw biasing phrase list and deal with the possible scalability issue:

$$W_r^j = -\frac{\min_i(\text{edit\_distance}(c_j, e_i))}{\text{len}(c_j)}, \qquad (6)$$

where e_i is the segment cut off from input ASR hypothesis with the same length of the context phrase c_j from the i-th word. The final relevance ranker weight is the minimum value of these edit distance normalized by the length of c_j.

The E2E ASR model decodes in a streaming way, we use the intermediate results and their corresponding decoding positions to estimate the rough alignment between audio and hypothesis. This alignment is then converted to the audio feature mask as model input.

3. EXPERIMENT

3.1. Data sets

Training set We use a small set and a large set as the training data, which include 0.2 thousand (K) hours and 17K hours of Microsoft in-house en-US data respectively. We do a full decoding of the training data with the baseline E2E ASR model to get hypothesis and audio encoder hidden for CSCv3 training.

Test set The test set consists of 12 Microsoft Teams meetings. Each meeting corresponds to a name list which consists of 600 person names, this list is expanded to a larger bias list with around 1500 phrases during inference. To evaluate the model performance on anti-context cases where the ground-truth name does not appear in the bias list, we also prepared 4 sets of bias lists with 25%, 50%, 75%, and 100% name coverage for each meeting. All the training and test data is anonymized with personally identifiable information removed.

3.2. Model settings

ASR model The baseline ASR model is a Conformer-Transducer (C-T) [27] model with the efficient low-latency implementation [28], trained with 64K hours Microsoft anonymized data. The dimension of audio encoder output is 512 and we only use top-1 text hypothesis for CSCv3 input.

Teacher model For the teacher model, each transformer layer contains a multihead-attention with 8 heads, and a 2048-dim feedforward layer. The text encoder, bias encoder and decoder all consist of 6 transformer blocks. The acoustic adapter consists of a 2048-dim

Table 1. Model performance with different bias list name coverage

Model	25% Coverage		50% Coverage		75% Coverage		100% Coverage	
	Recall	WER	Recall	WER	Recall	WER	Recall	WER
C-T	50.2	12.5	50.2	12.5	50.2	12.5	50.2	12.5
C-T+Biasing	58.0	12.6	59.0	12.6	61.4	12.6	64.1	12.6
+CSCv2	60.4	12.7	63.7	12.6	70.3	12.6	75.1	12.6
+CSCv3-EA-S0-nAnti-r1.0	58.0	12.8	60.4	12.7	70.1	12.7	74.3	12.7
+CSCv3-EA-S0-nAnti-r0.1	61.0	12.6	64.1	12.6	70.3	12.6	75.1	12.6
+CSCv3-EA-S0	61.8	12.7	64.7	12.7	71.3	12.6	75.3	12.6
+CSCv3-EA-full	61.8	12.8	64.9	12.8	72.3	12.8	76.9	12.7
+CSCv3-EA	**62.7**	12.7	**65.9**	12.7	**72.7**	12.7	**77.1**	12.6
+CSCv3-DA	62.7	12.7	64.1	12.7	71.3	12.7	75.9	12.7
+CSCv3-EA-student	62.7	12.7	65.3	12.7	72.3	12.6	77.7	12.6

feedforward layer followed by layer normalization. For text encoder of CSCv3-EA, an acoustic attention layer with 8 heads is inserted after the self-attention layer. While for the decoder of CSCv3-DA, an acoustic attention layer with 8 heads is sandwiched between the self-attention layer and biasing cross attention layer in each decoder block.

Student model For the student model, the text encoder, bias encoder, and decoder all consist of 3 transformer blocks. The embedding dimension is set to be 192, all the multi-head attentions have 4 heads, and dimension of feedforward layers is 768. The feedforward layer in the audio feature adapter is composed of a 512-dimension and a 192-dimension linear layer.

4. RESULTS

Baseline We use a C-T model as the baseline, and the C-T model with FST biasing [5] which uses the same biasing phrase list as a strong baseline (C-T+Biasing). In Table 1, we compare the name recall and WER of the models with different bias list name coverage. Where $s\%$ name coverage means there are $s\%$ of the ground truth names appear in the bias list while the rest are missing. We can see FST biasing has already achieved large name recall improvement compared to the C-T model.

Data augmentation It should be noted that CSCv3-EA-S0-nAnti-r0.1 and CSCv3-EA-S0-nAnti-r1.0 are the same model decoded with different parameters $r = 0.1$ and $r = 1.0$. It's trained with the small training set and without anti-context cases mentioned in Section 2.1.2. It shows that a small incorporation ratio of acoustic information ($r = 0.1$) leads to better performance. When the acoustic information is fully incorporated ($r = 1.0$), the model performance becomes worse. We have investigated the decoding results and found that when r is large, the model becomes more "biasing" and anti-context related errors are more likely to appear. This condition is not preferred because we hope the model be more stable on such cases. However, when anti-context data augmentation is added, as shown in CSCv3-EA-S0 which is trained with the same small set, this problem is gone and we find whether to add incorporation ratio r during training or inference does not influence the decoding results much, and the name recall is also improved. Which indicates that data augmentation leads to more stable results and better performance.

External acoustic attention CSCv3-EA-S0 shows that the model achieves name recall gain even with a small training set. CSCv3-EA is trained with the large training set, which shows significant performance improvement when the training set becomes larger compared to CSCv3-EA-S0. The comparison of CSCv3-EA and CSCv3-DA indicates that incorporating the acoustic information into the text encoder achieves larger performance improvement. We also fully trained a CSCv3-EA model with all parameters updated with the large training set, which is called CSCv3-EA-full model. We observe that CSCv3-EA-full does not perform as well as CSCv3-EA which is partially trained. One of the reasons is that CSCv3-EA can still benefit from the baseline CSCv2 model which was trained with richer text-based data; another reason is that the current training set lacks general utterances without person names, which makes the model more biasing. It should be noted that CSCv2 shows limited improvement when the bias list name coverage is small (e.g., 25% and 50%), one of the reasons is that it wrongly corrects some anti-context cases where the bias list does not contain the ground truth name but with some other phrases with similar pronunciation. With the external acoustic information, CSCv3 can deal with such issues and achieve stable improvement among different name coverage.

Model size and latency In Table 1, CSCv3-EA-student is the student model of CSCv3-EA, which shows similar performance compared to CSCv3-EA but with smaller model size. We also tested the latency of different models on the test set on a machine with 2.60GHz CPU using single thread regardless of baseline ASR model. The quantized onnx student model of CSCv2 is 5.4MB with 45.0ms latency per utterance, while CSCv3 is 6.2MB with 51.5ms per utterance, which indicates slight increase of model size and latency due to the external acoustic attention.

5. CONCLUSION

In this work, we propose an improved non-autoregressive (NAR) spelling correction model for contextual biasing in end-to-end transducer-based ASR systems with external acoustic attention and semantic aware data augmentation. The proposed model is proved to outperform the baseline ASR+Biasing system by as much as 20.3% relative name recall gain and achieves stable improvement compared to the traditional CSC method over different bias list coverage ratio.

6. REFERENCES

[1] Jinyu Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] Ian Williams, Anjuli Kannan, Petar S Aleksic, David Rybach, and Tara N Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search.," in *Proc. Interspeech*, 2018, pp. 2227–2231.

[3] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al., "Streaming end-to-end speech recognition for mobile devices," in *Proc. ICASSP*. IEEE, 2019, pp. 6381–6385.

[4] Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao, "Deep context: end-to-end contextual speech recognition," in *Proc. SLT*. IEEE, 2018, pp. 418–425.

[5] Ding Zhao, Tara N Sainath, David Rybach, D Bhatia, B Li, and R Pang, "Shallow-fusion end-to-end contextual biasing," in *Proc. Interspeech*, 2019, pp. 1418–1422.

[6] Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L Seltzer, "Deep shallow fusion for RNN-T personalization," in *Proc. SLT*. IEEE, 2021, pp. 251–257.

[7] Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf, "Contextual RNN-T for open domain ASR," in *Proc. Interspeech*, 2020, pp. 11–15.

[8] Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N Sainath, "Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition," in *Proc. ICASSP*. IEEE, 2019, pp. 6171–6175.

[9] Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer, "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," in *Proc. Interspeech*, 2021, pp. 1772–1776.

[10] Christian Huber, Juan Hussain, Sebastian Stüker, and Alexander Waibel, "Instant one-shot word-learning for context-specific neural sequence-to-sequence speech recognition," in *Proc. ASRU*. IEEE, 2021, pp. 1–7.

[11] Guangzhi Sun, Chao Zhang, and Philip C Woodland, "Tree-constrained pointer generator for end-to-end contextual speech recognition," in *Proc. ASRU*. IEEE, 2021, pp. 780–787.

[12] Xiaoqiang Wang, Yanqing Liu, Sheng Zhao, and Jinyu Li, "A light-weight contextual spelling correction model for customizing transducer-based speech recognition systems," in *Proc. Interspeech*, 2021, pp. 1982–1986.

[13] Xiaoqiang Wang, Yanqing Liu, Jinyu Li, Veljko Miljanic, Sheng Zhao, and Hosam Khalil, "Towards contextual spelling correction for customization of end-to-end speech recognition systems," *arXiv preprint arXiv:2203.00888*, 2022.

[14] Jinxi Guo, Tara N Sainath, and Ron J Weiss, "A spelling correction model for end-to-end speech recognition," in *Proc. ICASSP*, 2019, pp. 5651–5655.

[15] Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, and Yifan Gong, "Developing RNN-T models surpassing high-performance hybrid models with customization capability," in *Proc. Interspeech*, 2020, pp. 3590–3594.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[17] Yanqing Liu, Zhihang Xu, Gang Wang, Kuan Chen, Bohan Li, Xu Tan, Jinzhu Li, Lei He, and Sheng Zhao, "Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021," *arXiv preprint arXiv:2110.12612*, 2021.

[18] Jinyu Li et al., "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[19] Guoli Ye, Vadim Mazalov, Jinyu Li, and Yifan Gong, "Have best of both worlds: two-pass hybrid and e2e cascading framework for speech recognition," in *Proc. ICASSP*. IEEE, 2022, pp. 7432–7436.

[20] Ke Hu, Tara N Sainath, Arun Narayanan, Ruoming Pang, and Trevor Strohman, "Transducer-based streaming deliberation for cascaded encoders," in *Proc. ICASSP*. IEEE, 2022, pp. 8107–8111.

[21] Duc Le, Akshat Shrivastava, Paden Tomasello, Suyoun Kim, Aleksandr Livshits, Ozlem Kalinli, and Michael L Seltzer, "Deliberation model for on-device spoken language understanding," *arXiv preprint arXiv:2204.01893*, 2022.

[22] Cal Peyser, Sepand Mavandadi, Tara N Sainath, James Apfel, Ruoming Pang, and Shankar Kumar, "Improving tail performance of a deliberation e2e asr model using a large text corpus," *arXiv preprint arXiv:2008.10491*, 2020.

[23] Ke Hu, Tara N Sainath, Ruoming Pang, and Rohit Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *Proc. ICASSP*. IEEE, 2020, pp. 7799–7803.

[24] Ke Hu, Ruoming Pang, Tara N Sainath, and Trevor Strohman, "Transformer based deliberation for two-pass speech recognition," in *Proc. SLT*. IEEE, 2021, pp. 68–74.

[25] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size DNN with output-distribution-based criteria.," in *Proc. Interspeech*, 2014, pp. 1910–1914.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[27] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2019, pp. 5036–5040.

[28] Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," in *Proc. ICASSP*. IEEE, 2021, pp. 5904–5908.