

# PROCTER: PRONUNCIATION-AWARE CONTEXTUAL ADAPTER FOR PERSONALIZED SPEECH RECOGNITION IN NEURAL TRANSDUCERS

Rahul Pandey<sup>1,2\*</sup> Roger Ren<sup>1†</sup> Qi Luo<sup>1†</sup> Jing Liu<sup>1†</sup> Ariya Rastrow<sup>1</sup>  
 Ankur Gandhe<sup>1</sup> Denis Filimonov<sup>1</sup> Grant Strimel<sup>1</sup> Andreas Stolcke<sup>1</sup> Ivan Bulyko<sup>1</sup>

<sup>1</sup>Amazon Alexa AI, USA  
<sup>2</sup>George Mason University, USA

## ABSTRACT

End-to-End (E2E) automatic speech recognition (ASR) systems used in voice assistants often have difficulties recognizing infrequent words personalized to the user, such as names and places. Rare words often have non-trivial pronunciations, and in such cases, human knowledge in the form of a pronunciation lexicon can be useful. We propose a **PRONunCiation-aware conTextual adaptER** (PROCTER) that dynamically injects lexicon knowledge into an RNN-T model by adding a phonemic embedding along with a textual embedding. The experimental results show that the proposed PROCTER architecture outperforms the baseline RNN-T model by improving the word error rate (WER) by 44% and 57% when measured on personalized entities and personalized rare entities, respectively, while increasing the model size (number of trainable parameters) by only 1%. Furthermore, when evaluated in a zero-shot setting to recognize personalized device names, we observe 7% WER improvement with PROCTER, as compared to only 1% WER improvement with text-only contextual attention.

**Index Terms**— speech recognition, personalization, pronunciation, neural transducer, RNN-T, contextual biasing, attention

## 1. INTRODUCTION

End-to-end (E2E) ASR systems [1–4] have achieved significant WER improvements over traditional hybrid systems. However, they often have difficulty in correctly recognizing words that appear infrequently in the training data. Entity lists personalized to specific users (e.g., personalized entity lists or user-defined personalized device names) can help improve ASR accuracy for virtual voice assistants.

There has been much recent prior work to improve recognition of rare words for E2E ASR systems by incorporating additional contextual information [5–17]. The contextual information is ingested either post-training (e.g., by shallow fusion [15, 18] and on-the-fly rescore [19–21]) or during training (e.g., with contextual adapters [6]). However, the efficacy of the post-training method is limited due to lack of joint training with the core recognition components, making these approaches sensitive to heuristics used for tuning their rescore weights. In contextual adapter approaches, however, context is directly and jointly used as part of the recognition (e.g., RNN-T) model to improve the E2E ASR loss.

Most prior work leverages only the textual representation when incorporating contextual information. However, given that contexts

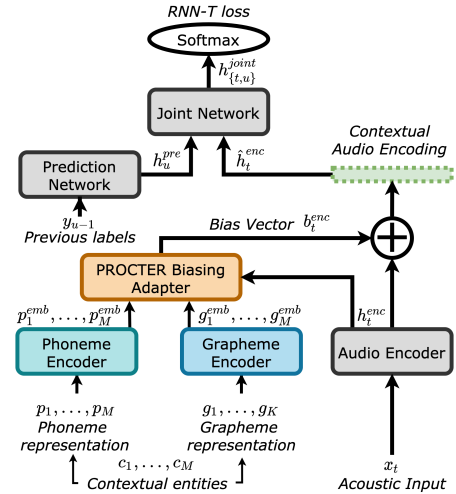


Fig. 1. Architecture of PROCTER.

like personalized entities are infrequent during training and often have multiple pronunciations, the E2E ASR model can get confused by similar textual contexts. For example, “Jana” has a pronunciation “Yana” that is not evident from the graphemes (in English). Thus, adding pronunciation information would make it easier for the model to align the entity “Jana” to the input audio, and hence infer the correct output. There is little prior work that utilizes phonemic representations of context, via a simple additive attention mechanism [7, 22] that is specific to the LAS [1] E2E architecture.

In this paper, we propose a jointly-trained contextualization adapter for the RNN-T model by incorporating both textual (graphemic) and phonemic representations of context to create a pronunciation-aware contextual adapter (**PROCTER**). First, given the contextual entities, we encode its grapheme and phoneme representation to get the grapheme and phoneme embeddings, respectively. Then, they are used to bias the intermediate outputs from the audio encoder of the RNN-T model to fuse the contextual information using the PROCTER biasing adapter before passing it to the joint network of RNN-T. In contrast to previous approaches that select one pronunciation per contextual entity [7, 22], we preserve all pronunciations by considering them as separate entities and duplicate the corresponding textual representation in multiple grapheme-phoneme pairs. Additionally, unlike the attention mechanism used in prior work [7], our adapter biases the audio encoder outputs with contextual information using a more optimized scaled dot-product attention [23]. Prior work has shown that the output corresponding to the last layer of the audio encoder in E2E ASR models has less

\*Work done as an applied scientist intern at Amazon Alexa.

†Contributed equally.

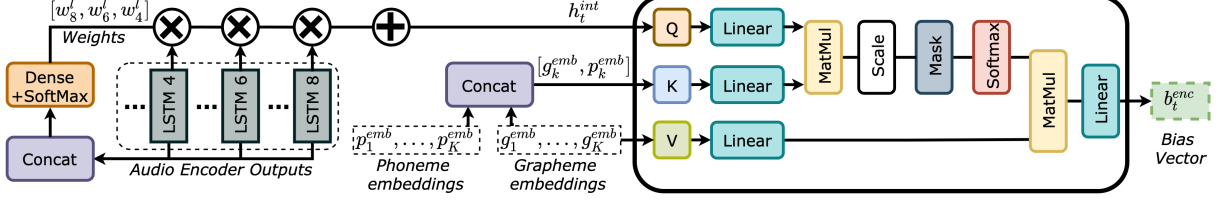


Fig. 2. Proposed PROCTER biasing adapter

phonemic information since ASR is optimized to output the correct graphemes irrespective of phonemic variations [24, 25]. Thus, instead of letting only the last layer attend to contextual entities [6, 7], we also bias earlier layers with our phonemic embeddings. Providing phonemic context allows the biasing network to pay attention to idiosyncratic pronunciation variants. Moreover, allowing multiple grapheme-phoneme pairs lets the model attend to the best-matching pronunciation of a context item. We train the model in adapter style, freezing the rest of the RNN-T model with weights from pretraining. Thus, it requires a small fraction of the data and time needed for full training. To demonstrate the performance of our proposed model, we compared PROCTER with a state-of-art model (i.e., a text-only) contextual adapter model [5, 6]) on three in-house far-field datasets.

## 2. PRIOR WORK

### 2.1. Neural Transducer

The recurrent neural transducer (RNN-T) is widely used as an E2E ASR streaming model. The grey blocks in Figure 1 represent the streaming RNN-T ASR system, which has three main components: an audio encoder, a prediction network, and a joint network.

The audio encoder typically consists of stacked LSTM layers [3]. It generates the high-level audio encoding representations  $h_t^{enc}$  given the input audio frames  $x_{0,t} = (x_0, \dots, x_t)$ . The prediction network also uses stacked LSTM layers to encode the last nonblank word pieces  $y_{0,u-1} = (y_0, \dots, y_{u-1})$ , and generate  $h_u^{pre}$ . The joint network takes the output from both audio encoder  $h_t^{enc}$  and prediction network  $h_u^{pre}$ . It fuses the two outputs before passing them to a series of dense layers (denoted by  $\phi$ ). Finally, a softmax operation is applied to obtain the probability distribution over word pieces. Equation 1 shows that RNN-T outputs probability at time  $t$  given a sequence  $u$ .

$$\begin{aligned} h_{t,u}^{joint} &= \phi(\text{JoinOp}(h_t^{enc}, h_u^{pre})) \\ P(y_u | t, u) &= \text{softmax}(h_{t,u}^{joint}) \end{aligned} \quad (1)$$

### 2.2. Contextual adapters

Contextual adapters have been shown effective in adapting pre-trained RNN-T model for the recognition of contextual entities, achieving up to 31% relative word error rate reduction (WERR) compared to the vanilla RNN-T model [5, 6]. The biasing network has two main components—a catalog encoder and a biasing adapter.

The catalog encoder embeds a catalog of contextual entities  $C = c_1, c_2, \dots, c_K$ . It takes the graphemic (textual) representation of catalog entities as input, passing them to a subword tokenizer [26], followed by a stack of BiLSTM layers. The last state of the BiLSTM is used as the embedding of a contextual entity, which is an encoded representation  $c_e$ . Given a catalog with  $K$  entities, and the generated entity embedding size of  $D$ , the catalog encoder outputs  $C^e \in \mathbb{R}^{K \times D}$  as follows:

$$c_i^e = \text{BiLSTM}(\text{Embedding}(c_i)) \quad (2)$$

The biasing adapter transforms an intermediate representation computed by the RNN-T, such as the audio encoder output ( $h_t^{enc}$ ), using the graphemic entity embeddings from the personalized catalog. It uses a cross-attention-based biasing adapter to attend over all entity embeddings  $C^e$  (key and value) based on  $h_t^{enc}$  as the input query. The attention scores  $\alpha_i$  for each catalog entity are calculated using the scaled dot-product attention mechanism [23]. The biased intermediate representation (bias vector) is calculated as  $b_t^{enc} = \sum_i^K \alpha_i W^v c_i^e$ . Finally, the intermediate representations are updated with the bias vector using element-wise addition. Thus, the contextual audio encoding ( $\hat{h}_t^{enc}$ ) that contains the biased representation is calculated as  $\hat{h}_t^{enc} = h_t^{enc} \oplus b_t^{enc}$ .

One major shortcoming in designing the contextual adapters is its heavy reliance on the textual representation of the catalog entities. While the contextual adapter can learn to effectively represent rare catalog entities, it struggles to encode entities with uncommon pronunciation. Prior work [7, 27] showed that the phonemic representation of such rare entities can be compared more effectively to grapheme-only representations in the case of a contextual LAS model [8]. These findings motivate combining phonemic and graphemic representations within contextual adapters to improve accuracy on rare entities.

## 3. PROCTER

The proposed PROCTER model takes as input the phonemic representation of catalog entities in addition to its textual grapheme representation. Figure 1 shows the overview of the architecture. It has three components: phoneme encoder, grapheme encoder, and PROCTER biasing adapter. The objective of PROCTER is to learn the phonemic variations of catalog entity representations such that it improves the biasing of intermediate audio encodings for the contextual catalog entities with nonstandard pronunciations.

### 3.1. Phoneme Encoder

The phoneme encoder embeds the pronunciation representation of entities in the catalog. The pronunciation information is retrieved from the dictionary. Same-text entities can have multiple pronunciations, which are all preserved as separate entities for biasing the intermediate output representation. Let  $P = [p_1, p_2, \dots, p_M]$  denote all pronunciations (phoneme sequences) of catalog entities  $C = [c_1, c_2, \dots, c_K]$  where  $M \geq K$ . The phoneme encoder embeds the phoneme sequences using an embedding lookup, followed by a stack of BiLSTM layers. With  $M$  pronunciations and an embedding size of  $D_p$ , the phoneme encoder outputs  $P^{emb} \in \mathbb{R}^{M \times D_p}$  where  $p_m^{emb} \in P^{emb}$  is the phoneme embedding of pronunciation  $p_m$ , calculated as

$$p_j^{emb} = \text{BiLSTM}(\text{Embedding}(p_j))$$

We also add a *no\_bias* token to the pronunciation catalog for disabling adapter biasing, similar to prior work [5, 7].

### 3.2. Grapheme Encoder

Our grapheme encoder is similar to the catalog encoder described in Section 2.2. It encodes the grapheme representation  $G = [g_1, g_2, \dots, g_K]$  of catalog entities  $C = [c_1, c_2, \dots, c_K]$  using Equation 2. The grapheme encoder also duplicates the output grapheme embeddings for the entities with more than one pronunciation to match up with the phoneme embeddings for the various pronunciations of a given entity. Hence, given a total of  $M$  pronunciations and grapheme embedding size  $D_g$ , the grapheme encoder outputs  $G^{emb} \in \mathbb{R}^{M \times D_g}$  where  $g_m^{emb} \subset G^{emb}$  is the grapheme embedding of the textual representation in the catalog with  $p_m$  as one of its pronunciations.

### 3.3. PROCTER Biasing Adapter

The proposed PROCTER biasing adapter adapts an intermediate representation from the audio encoder with both textual and pronunciation representations of contextual catalog entities, as shown in Figure 2. It is based on cross-attention using the scaled dot-product [23]. The cross-attention module of the proposed adapter takes three inputs: query, key, and value.

*Query:* The query comes from the intermediate layer output of the pretrained audio encoder. Instead of using only the final layer output, we also include intermediate layer outputs earlier in the pretrained RNN-T audio encoder. More specifically, we utilize the outputs corresponding to the last, third-last, and fifth-last LSTM layers. We compute weighted addition of these encoder outputs as a gating mechanism. Given the intermediate layer outputs  $\{h_t^{enc^l}, h_t^{enc^{l-2}}, h_t^{enc^{l-4}}\}$ , we use a dense layer to project the concatenated intermediate layer outputs to a 3 dimensional vector for weights  $[w_t^l, w_t^{l-2}, w_t^{l-4}]$ . And these weights are used to calculate the final weighted sum of encoder output  $h_t^{qry}$ , which is used as the query to the scaled dot-product attention as shown in Figure 2.

$$\begin{aligned} h_t^{enc^{cat}} &= \text{Concat}(\{h_t^{enc^l}, h_t^{enc^{l-2}}, h_t^{enc^{l-4}}\}) \\ [w_t^l, w_t^{l-2}, w_t^{l-4}] &= \text{Softmax}(\text{Dense}(h_t^{enc^{cat}})) \end{aligned} \quad (3)$$

$$h_t^{qry} = \sum_{i \in \{l, l-2, l-4\}} w_t^i h_t^{enc^i}$$

*Key:* The key is used to compute the attention score based on the provided query. We use both textual and pronunciation representations coming from the grapheme and phoneme encoders, respectively, as the key. Since we have a one-to-one mapping of corresponding phoneme and grapheme embeddings, we concatenate them to obtain the final embedding for the key. For each catalog grapheme-phoneme embedding pair  $g_m^{emb}$  and  $p_m^{emb}$  where  $m \in 1, \dots, M$ , we compute the input to key  $c_m^{key}$  as

$$c_m^{key} = \text{Concat}(g_m^{emb}, p_m^{emb}) \quad (4)$$

Thus, the key for all grapheme-phoneme embedding pairs becomes  $C^{key} = [c_1^{key}, c_2^{key}, \dots, c_M^{key}]$ .

*Value:* The value is used to bias the query with the computed attention weights. Since the biased vector passed to the joint network of RNN-T is expected to contain only textual information, we use the textual representation  $g_m^{emb}$  as the value.

We compute an attention score  $\alpha_i$  based on the weighted sum of intermediate encoder outputs (Eqn. 3) as query and concatenated grapheme-phoneme representations of all catalog entities as the key:

$$\alpha_i = \text{Softmax}_i \left( \frac{W^q h_t^{qry} \cdot (W^k C^{key})^T}{\sqrt{d}} \right) \quad (5)$$

The final bias vector  $b_t^{enc}$  is calculated as  $b_t^{enc} = \sum_i^K \alpha_i W^v g_i^{emb}$ . Finally, similar to [5], the intermediate representations are fused with the contextual bias vector using element-wise addition. Thus, the contextual audio encoding ( $\hat{h}_t^{enc}$ ) that contains the biasing contextual representation is calculated as  $\hat{h}_t^{enc} = h_t^{enc} \oplus b_t^{enc}$ .

The intuition behind the proposed PROCTER strategy is to learn better attention weights by including the phonemic representation of the catalog entities. Moreover, to better index into the phonemic representation, we also include the intermediate layer outputs of the pretrained audio encoder to capture acoustic-phonetic variations in the audio inputs.

## 4. EXPERIMENTS

### 4.1. Dataset and Evaluation

We use in-house de-identified far-field datasets coming from interactions with a virtual voice assistant. The training data consists of text-audio pairs of utterances randomly sampled from more than 20 domains such as Communications, Weather, SmartHome, and Music. The baseline RNN-T model was trained using 114k hours of data. However, following the work in [5, 6], the attention components in PROCTER are trained using a fine-tuning step where the core components of RNN-T (encoder, prediction) are kept frozen. For training the adapter, we use much less,  $\sim 290$  hours of data. Personalized entities from real users are used as context information. We test the model on three test sets: 1) 16k utterances of far-field English data, 2) 29k utterances of far-field English data, which includes mentions of the personalized entities, 3) 3558 utterances of data containing mentions of personalized device names. We report the WERR on all three test sets, as well as the relative named-entity WER reduction (NE-WERR) of personalized entities and personalized device names, respectively, on the second and third test sets. For the second test set, we further compute the NE-WERR of rare personalized entities (those appearing only once in the test set), to specifically measure performance on infrequent words.

### 4.2. Experimental Setup

**Pretrained RNN-T model.** We use the RNN-T model described in Section 2.1. The input is a 64-dim LFBF feature for every 10 ms of audio, with a window size of 25 ms, with three frames stacked together resulting in 192 features per frame. Each ground truth text token is passed through a 4000 word-piece tokenizer [26]. The RNN-T audio encoder consists of 8 LSTM layers with 1280 units per layer. The prediction network consists of 2 LSTM layers with 1280 units per layer. The joint network consists of a dense layer of 512 units, followed by a softmax activation over RNN-T output tokens. The decoding is performed using the standard beam search with a beam size of 8. The output vocabulary consists of 4000 word pieces.

**PROCTER configuration.** The phoneme sequences of contextual entities are generated from a lexicon. A text can have multiple pronunciations, and we found it important to keep all phoneme sequences. The grapheme and phoneme encoders consist of BiLSTMs with 64 and 128 units, respectively. The PROCTER biasing adapter projects the query, key, and value to 128 dimensions with attention weights. The maximum number of contextual entities is set to 300. For the PROCTER-based experiments, we further divide the full personalized entities or personalized device names into individual words. The words are duplicated based on their number of pronunciations in the lexicon file for a one-to-one correspondence during

**Table 1. Results.** Relative change in WER (WERR), and NE-WER (NE-WERR) over vanilla RNN-T models for various models. The rare personalized entities appear only once, and personalized device name test set shows the zero-shot capability of the model.

Model	Phoneme		Int-Layers	General WERR	Personalized Entity			Personalized Device Name	
	key	value			Overall	All Entities	Rare Entities	Overall	Personalized Device Names
					WERR	NE-WERR		WERR	NE-WERR
<i>Text-only adapter</i>	N	N	N	0.0%	32.5%	37.1%	50.0%	1.1%	0.9%
PROCTER	Y	N	Y	0.2%	<b>38.2%</b>	<b>43.9%</b>	<b>57.2%</b>	<b>2.6%</b>	<b>6.9%</b>
<i>PROCTER+Ph-in-value</i>	Y	Y	Y	0.2%	<b>38.2%</b>	43.2%	55.1%	1.3%	5.3%
<i>PROCTER+No-Int-Layers</i>	Y	N	N	<b>1.0%</b>	<b>38.2%</b>	43.6%	57.1%	0.2%	3.9%
<i>PROCTER+Ph-in-value+No-Int-Layers</i>	Y	Y	N	0.2%	37.6%	43.0%	54.6%	0.0%	4.8%

concatenation. We keep a maximum size of 600 for these phoneme-grapheme pairs. For training the PROCTER, we use the Adam optimizer with a learning rate of  $5 \times 10^{-4}$  configured to converge with early stopping. The number of trainable adapter parameters is 1.5M, or about 1% of the RNN-T.

**Experiment setup.** The primary baseline model is a vanilla RNN-T model without personalization described in Section 2.1. We compare the proposed PROCTER adapter to four additional model configurations: 1) *Text-only adapter*: Contextual adapter proposed in [5] using only text representations of context entities; 2) *PROCTER + Ph-in-value*: PROCTER with the value embeddings created the same way as the keys (Equation 4). This experiment assesses the effect of using phonemic information in the final contextual embedding that is passed into the joint network; 3) *PROCTER + No-Int-Layers*: PROCTER with the queries defined using the final layer output of the pretrained audio encoder (as opposed to using the weighted sum of intermediate outputs). This ablation study examines the importance of information from intermediate audio encoder layers when attending to the contextual embeddings; 4) *PROCTER + Ph-in-value + No-Int-Layers*: Combination of variants (2) and (3).

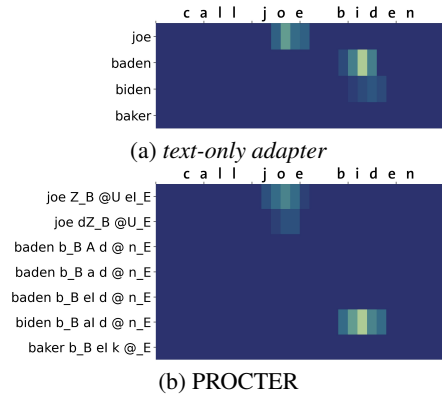
## 5. RESULTS

Table 1 shows the results relative to the vanilla RNN-T model on all three test sets. Given the general test set with no personalized context, we observe that the previous *text-only adapter*, proposed PROCTER, and all ablation experiments have no degradation over the baseline RNN-T model. It shows that the contextual adapters are immune to the performance of general ASR. For the personalized entity test set, in terms of overall WERR, we observe that all the experiments significantly improved over the vanilla RNN-T, which shows the effectiveness of contextual adapters. However, we see a gain when we introduce the phonemic context representation compared to the *text-only adapter* (38.2% vs. 32.5%). Furthermore, the performance gain increases compared to the *text-only adapter* as the sparsity of context increases, as seen in the NE-WERR of all personalized entities and rare personalized entities (43.9% & 57.2% vs. 37.1% & 50%). Although all the ablation experiments perform similarly to the PROCTER model in overall WERR and NE-WERR for all personalized entities, we see a performance gain for PROCTER on rare personalized entities over the first and third ablation experiments (57.2% vs. 55.1% and 54.6%). This shows the importance of not passing the phoneme information to the joint network of the RNN-T model.

Additionally, we test the zero-shot capability of PROCTER. We test the personalized device name test set with personalized device names as context, with all the experiments trained with personalized entities for contextualization. Since the context domain differed, we did not see much improvement over the vanilla RNN-T model.

However, we observe the highest performance gain by PROCTER in terms of overall WERR and personalized device name NE-WERR. In particular, we achieved 6.9% NE-WERR of the personalized device name context compared to 1.0% with *text-only adapter*. Moreover, we also observe better performance of PROCTER with the second ablation experiment. Therefore, when the sparsity of context is at the maximum, it is essential to include the intermediate layer audio encoder outputs for better contextualization.

Finally, to demonstrate the learning of phonemic context representation, Figure 3 shows an example (the personalized entity was altered to protect privacy) where the PROCTER model identifies the correct entity while the *text-only adapter* fails. *Text-only adapter* attends to another textually similar personalized entity, ‘baden’, for the audio frames corresponding to ‘biden’, leading to an incorrect hypothesis, “call joe baden”. However, the PROCTER model, with access to the pronunciations of personalized entities ‘biden’ and ‘baden’, identifies the context ‘biden’ to arrive at the correct ASR hypothesis.



**Fig. 3.** Attention visualization

## 6. CONCLUSION

We proposed PROCTER, a pronunciation-aware contextual adapter for a vanilla RNN-T model for improving the recognition of rare words. It fuses the pronunciation representations with the textual representations of context words via scaled dot-product attention. PROCTER is trained in adapter style with only 1% of the trainable parameters of a vanilla RNN-T model. As a result, we observe over 57% improvement compared to the baseline RNN-T model on rare personalized entities and up to 7% improvement in zero-shot testing on a personalized device name test set. Although PROCTER relies on a dictionary for pronunciations, we found less than 0.5% of contexts lacking pronunciation information. In the future, we plan to use a neural grapheme-to-phoneme model for those rare cases. We also plan to try applying PROCTER-style biasing on decoder outputs in addition to audio encoder outputs.

## 7. REFERENCES

- [1] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2016, pp. 4960–4964, IEEE.
- [2] Linhao Dong, Shuang Xu, and Bo Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” in *Proc. IEEE ICASSP*, 2018, pp. 5884–5888.
- [3] Alex Graves, “Sequence transduction with recurrent neural networks,” *Proc. ICML Workshop on Representation Learning*, 2012.
- [4] Ching feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, and Michael L. Seltzer, “Transformer-transducer: End-to-end speech recognition with self-attention,” *ArXiv*, vol. abs/1910.12977, 2019.
- [5] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann, “Contextual adapters for personalized speech recognition in neural transducers,” in *Proc. IEEE ICASSP*, 2022, pp. 8537–8541.
- [6] Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann, “Context-aware transformer transducer for speech recognition,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 503–510.
- [7] Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N. Sainath, “Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition,” in *Proc. IEEE ICASSP*, 2019, pp. 6171–6175.
- [8] Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjali Kannan, and Ding Zhao, “Deep context: end-to-end contextual speech recognition,” in *Proc. IEEE Spoken Language Technology Workshop*, 2018, pp. 418–425.
- [9] Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf, “Contextual RNN-T for open domain ASR,” *Proc. Interspeech*, pp. 11–15, 2020.
- [10] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates, “Cold fusion: Training SeqSeq Models Together with Language Models,” in *Proc. Interspeech*. Sept. 2018, pp. 387–391, ISCA.
- [11] Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe, “Multi-modal data augmentation for end-to-end ASR,” in *Proc. Interspeech*. Sept. 2018, pp. 2394–2398, ISCA.
- [12] Anjali Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in *Proc. IEEE ICASSP*, 2018, pp. 5824–5828.
- [13] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Atsunori Ogawa, and Marc Delcroix, “Semi-supervised end-to-end speech recognition,” in *Proc. Interspeech*. Sept. 2018, pp. 2–6, ISCA.
- [14] Antoine Bruguier, Fuchun Peng, and Françoise Beaufays, “Learning personalized pronunciations for contact name recognition,” in *Proc. Interspeech*. Sept. 2016, pp. 3096–3100, ISCA.
- [15] Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xi-angyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko, “Personalization strategies for end-to-end speech recognition systems,” in *Proc. IEEE ICASSP*, 2021, pp. 7348–7352.
- [16] Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shanguan, Christian Fuegen, Ozlem Kalinli, and others, “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion,” *arXiv:2104.02194*, 2021.
- [17] Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L Seltzer, “Deep shallow fusion for RNN-T personalization,” in *Proc. IEEE Spoken Language Technology Workshop*, 2021, pp. 251–257.
- [18] Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang, “Shallow-fusion end-to-end contextual biasing,” in *Proc. Interspeech*, Sept. 2019, pp. 1418–1422.
- [19] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Razi-ziel Alvarez, Ding Zhao, David Rybach, Anjali Kannan, Yonghui Wu, Ruoming Pang, et al., “Streaming end-to-end speech recognition for mobile devices,” in *Proc. IEEE ICASSP*, 2019, pp. 6381–6385.
- [20] Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, Anirudh Raju, Denis Filimonov, and Ivan Bulyko, “Multi-task language modeling for improving speech recognition of rare words,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 1087–1093.
- [21] Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko, “RescoreBERT: Discriminative speech recognition rescoring with bert,” in *Proc. IEEE ICASSP*, 2022, pp. 6117–6121.
- [22] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L. Seltzer, and Christian Fuegen, “Joint grapheme and phoneme embeddings for contextual end-to-end ASR,” in *Proc. Interspeech*, Sept. 2019, pp. 3490–3494.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [24] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 914–921.
- [25] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen, “Voice2series: Reprogramming acoustic models for time series classification,” in *Proc. ICML*. 18–24 Jul 2021, vol. 139, pp. 11808–11819, PMLR.
- [26] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.
- [27] Vasileios Papadourakis, Markus Müller, Jing Liu, Athanasios Mouchtaris, and Maurizio Omologo, “Phonetically induced subwords for end-to-end speech recognition,” in *Proc. Interspeech*, 2021, pp. 1992–1996.