DIALOG ACT GUIDED CONTEXTUAL ADAPTER FOR PERSONALIZED SPEECH RECOGNITION

Feng-Ju Chang, Thejaswi Muniyappa, Kanthashree Mysore Sathyendra, Kai Wei, Grant P. Strimel, Ross McGowan

Alexa, Amazon, USA

ABSTRACT

Personalization in multi-turn dialogs has been a long standing challenge for end-to-end automatic speech recognition (E2E ASR) models. Recent work on contextual adapters has tackled rare word recognition using user catalogs. This adaptation, however, does not incorporate an important cue, the dialog act, which is available in a multi-turn dialog scenario. In this work, we propose a dialog act guided contextual adapter network. Specifically, it leverages dialog acts to select the most relevant user catalogs and creates queries based on both – the audio as well as the semantic relationship between the carrier phrase and user catalogs to better guide the contextual biasing. On industrial voice assistant datasets, our model outperforms both the baselines - dialog act encoder-only model, and the contextual adaptation, leading to the most improvement over the no-context model: 58% average relative word error rate reduction (WERR) in the multi-turn dialog scenario, in comparison to the prior-art contextual adapter, which has achieved 39% WERR over the no-context model.

Index Terms— Contextual adapter, personalized speech recognition, RNN-T, dialog act, early-late fusion

1. INTRODUCTION

Personalized speech recognition has gained considerable attention in recent years as industrial voice assistants (IVAs) rise in popularity. To provide the best customer experience, an IVA should be able to recognize personalized requests correctly. For example, *"drop in on [Proper Name]"*, where the name could be user defined words.

End-to-end deep neural networks have become the mainstream approach for ASR in IVAs, due to their great capacity to learn the audio-to-text mapping without the alignment information between the audio and text transcript [1]. These systems include the Connectionist Temporal Classification (CTC) [2], Listen-Attend-Spell (LAS) [3], Recurrent Neural Network Transducer (RNN-T) [4], and the transformer network [5]. However, E2E ASR models face challenges in generalizing to recognize user-specific rare words, such as contact names, proper nouns, and named entities due to the scarcity of paired audio-to-text training data [6–10].

Prior works have explored various types of contextual information to improve personalized speech recognition [8, 11–18]: The weighted finite state transducers (WFSTs) [11] on the rare words [12]; Domain-aware neural language models [13]; Text metadata of video [14, 15], and the catalogs [16, 17] provided by speakers such as contact names and device names. The contexts are incorporated to the model either with the shallow fusion [18, 19] or the neural biasing approaches [8, 10, 16, 17, 20]. The neural solutions have been shown outperforming the use of grammars or dynamic WFSTs in the shallow fusion. In [16, 17], the catalog entities are encoded by BiLSTM or BERT based encoders, and then integrated with the audio encoder via a multi-head attention based contextual biasing network. The model is taught to leverage user-specific contexts to improve the recognition of rare words. The performance of this biasing approach, however, may be hindered when irrelevant user catalogs or/and wrongly predicted carrier phrases bias the next predictions towards a wrong token.

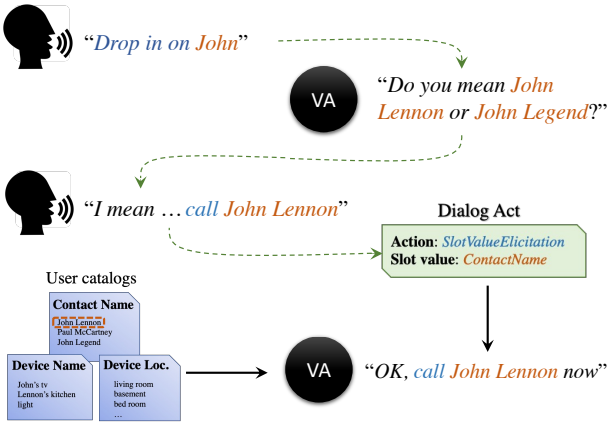

Fig. 1. A high-level diagram illustrating our idea to leverage both dialog act and user catalogs to help a voice assistant (VA) recognizing proper nouns in a multi-turn dialog. Note that prior contextual adapters [16, 17] consider only user catalogs for contextual biasing.

To address the above issues, we propose to leverage both dialog act and user catalogs for the contextual biasing (Fig. 1) and introduce a dialog act guided contextual adapter network (Fig. 2). The dialog acts (DA) are used to identify the action and the slot value of spoken utterances in multi-turn dialogues (Fig. 1). An example of DA is *SlotValueElicitation(ProperName)*, which is associated with a dialog when a user tries to call someone. The DA can help the VA to elicit a contact name from the user catalog for the next turn, usually to clarify or confirm the entity. Given the semantic relationship between "*drop in on*" or "*call*" versus "*SlotValueElicitation*", DAs can help improve the prediction of the carrier phrases. We then use an attention based contextual adapter to bias the predictions towards the tokens in the user's *[Contact Name]* catalog. Note that prior work with DAs [21–24] primarily focus on the generic multi-turn utterance improvements, while the contextual biasing approaches [16, 17] were mainly dedicated to the single-turn named entity improvement. With the proposed network, we can bridge the gap between a single turn and multi-turn personalization in ASR.

Our contributions are summarized as follows. First, we present the dialog act (DA) encoder (Section 3.1) and the fusion networks (Section 3.2) to integrate the dialog act context with both the low level and high level acoustic representations, and improve the queries for the biasing network. Second, we use dialog acts to select the most rel-

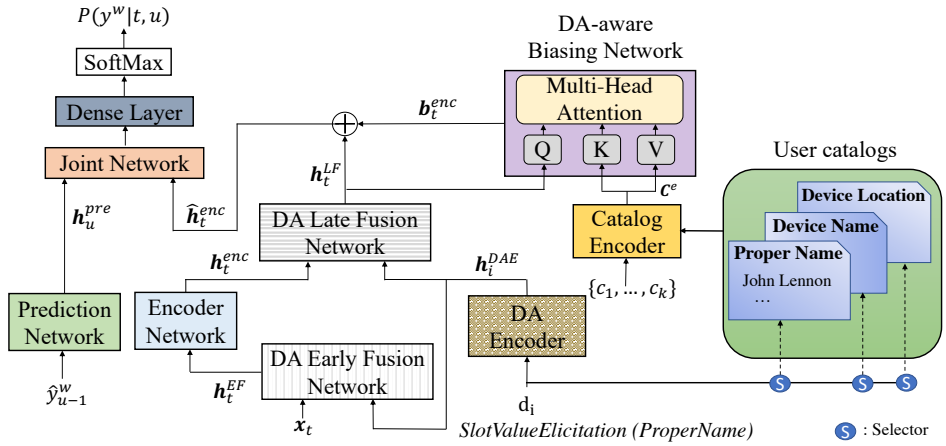

Fig. 2. **The proposed DA guided contextual adapter.** The DAs are used for catalog selection and are fused with different levels of audio representations to build better queries for the biasing network.

evant catalogs (Section 3.3), which directly influence the keys/values used in the biasing network. Third, we introduce a DA-aware biasing network, which takes the improved queries and keys/values above to bias toward the right catalog entities (Section 3.4). Finally, we propose a two-stage adapter training and investigate the impacts of freezing and unfreezing DA encoder and fusion networks during the contextual biasing stage (Section 3.5).

2. CONTEXTUAL ADAPTER

The contextual adapter used in this work, [17], is built upon a type of streaming E2E ASR, RNN-T [4], which consists of an RNN based encoder, an RNN based prediction network, and a joint network. The encoder network produces high-level representations $\mathbf{h}_t^{enc}$ for the audio frames, while the prediction network encodes the previously predicted word-pieces and produces the output $\mathbf{h}_u^{pre}$. The joint network fuses $\mathbf{h}_t^{enc}$ and $\mathbf{h}_u^{pre}$ via the join operation followed by a series of dense layers with activations and a softmax function to obtain the probability distribution over word-pieces plus a *blank* symbol. In [17], a catalog encoder is introduced to embed the user catalogs and types. In addition, a multi-head attention based biasing network is introduced to measure the relevance of the user catalog context using the encoder network outputs as queries.

3. DIALOG ACT GUIDED CONTEXTUAL ADAPTER

In order to leverage DAs to guide the contextual adapter learning, we introduce four new components as shown in Fig. 2: (1) DA Encoder, (2) DA Fusion Networks, (3) Catalog selection with DAs, and (4) DA aware Biasing Network, along with the two stage adapter training.

3.1. DA Encoder

Each DA string, d_i, contains the action string a_i and the slot string s_i, in the form of $\mathrm{a}_i(\mathrm{s}_i)$, e.g. *SlotValueElicitation(ProperName)*, as shown in Figure 3(a). Motivated by [21], we convert a_i and s_i by two embedding matrices into the action embeddings $\mathbf{h}_i^a$ and slot embeddings $\mathbf{h}_i^s$, with the same embedding size. Both embeddings are then fused via an element-wise addition followed by a Feed Forward Network (FFN) and the ReLU activation, denoted as σ, to produce the fixed dimensional DA embedding, $\mathbf{h}_i^{DAE} \in \mathbb{R}^{1\times d_{da}}$:

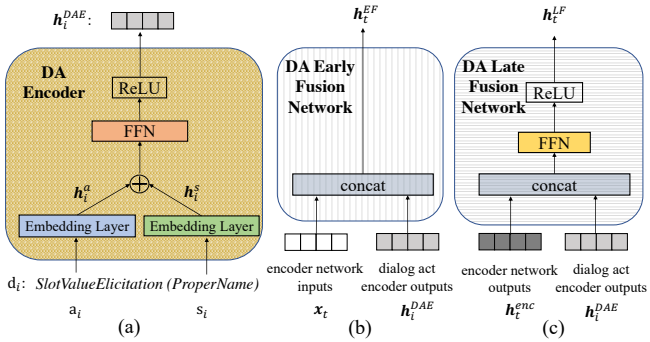

Fig. 3. **The DA encoder and fusion networks**. (a) DA encoder, (b) DA early fusion network, and (c) DA late fusion network

$$\mathbf{h}_i^{DAE} = \sigma(\mathrm{FFN}(\mathbf{h}_i^a + \mathbf{h}_i^s)) \tag{1}$$

3.2. Fusion of DAs and Multi-level Audio Embeddings

After we obtain the DA embedding, $\mathbf{h}_i^{DAE}$, it is combined with two levels of the audio representations using one of the three approaches presented below:

Early Fusion: We simply concatenate $\mathbf{h}_i^{DAE}$ with the input audio features per frame, $\mathbf{x}_t$, see Figure 3(b).

$$\mathbf{h}_t^{EF} = \mathrm{Concat}(\mathbf{x}_t, \mathbf{h}_i^{DAE}) \tag{2}$$

Late Fusion: The encoder network output, $\mathbf{h}_t^{enc}$ is first concatenated with $\mathbf{h}_i^{DAE}$. The concatenated output is then transformed by a FFN with the ReLU activation, denoted as σ. See Figure 3(c).

$$\mathbf{h}_t^{LF} = \sigma(\mathrm{FFN}(\mathrm{Concat}(\mathbf{h}_t^{enc}, \mathbf{h}_i^{DAE}))) \tag{3}$$

Early-Late Fusion: In this approach, we inject $\mathbf{h}_i^{DAE}$ jointly with the input audio (Equation 2) and with the encoder output (Equation 3) as shown in Figure 2. Note that the DA encoder is shared in this approach.

3.3. Catalog Selection with DAs

In addition to fusing the DA embeddings with the audio representations (Section 3.2), we also investigate the selection of catalogs based on the DA string $\mathrm{d}_i = \mathrm{a}_i(\mathrm{s}_i)$, during training. Specifically, the

slot string s_i in d_i is used for selecting the catalog for biasing. We consider three types of catalogs in this work: contact names, device names, and device locations. By matching s_i with the catalog type, we can summarize the selection rule as follows:

$$\{c_1, ...c_K\} \in \begin{cases} \text{proper names catalog,} & \text{if } s_i = \textit{ProperName} \\ \text{device names catalog,} & \text{if } s_i = \textit{DeviceName} \\ \text{device locations catalog,} & \text{if } s_i = \textit{DeviceLocation} \\ \text{all three catalogs,} & \text{otherwise} \end{cases}$$

Using this approach, we can prevent irrelevant catalogs from deteriorating biasing network training.

3.4. DA Aware Biasing Network

The biasing network learns which catalog entities to bias toward. Inspired by [16, 17], we employ cross-attention to attend over the catalog embedding matrix $\boldsymbol{C}^e \in \mathbb{R}^{K \times D_c}$ based on the input query $\mathbf{q}$, which comes from the DA late fusion network, $\mathbf{h}_t^{LF} \in \mathbb{R}^{1 \times D_a}$ (Equation 3). Here K, D_c, D_a are the total number of catalog entities, the catalog embedding size, and the audio embedding size respectively. This query intuitively encodes both the audio and the DA information – the semantic relationship between the carrier phrase and the user catalogs.

The query and the catalog entity embeddings are first projected via $\boldsymbol{W}^q \in \mathbb{R}^{D_a \times d}$ and $\boldsymbol{W}^k \in \mathbb{R}^{D_c \times d}$ to dimension d. An attention score vector $\boldsymbol{\alpha}_t$ for catalog entity embeddings is computed by the scaled dot product attention mechanism [25] as follows:

$$\boldsymbol{\alpha}_t = \text{Softmax}\left(\frac{(\boldsymbol{C}^e\boldsymbol{W}^k)(\mathbf{h}_t^{LF}\boldsymbol{W}^q)^T}{\sqrt{d}}\right) \quad (4)$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^{K \times 1}$. The biasing vector $\mathbf{b}_t^{enc}$ is then computed as $\mathbf{b}_t^{enc} = (\boldsymbol{\alpha}_t)^T(\boldsymbol{C}^e\boldsymbol{W}^v) \in \mathbb{R}^{1 \times D_a}$, where $\boldsymbol{W}^v \in \mathbb{R}^{D_c \times D_a}$. The biasing vector $\mathbf{b}_t^{enc}$ is then combined with the $\mathbf{h}_t^{LF}$: $\hat{\mathbf{h}}_t^{enc} = \mathbf{h}_t^{LF} \oplus \mathbf{b}_t^{enc}$.

While the audio representation enables the biasing network to bias towards entities based on acoustic similarity, the DA helps the biasing network attend over the right *type* of entity. For instance, if the DA is *SlotValueElicitation(DeviceName)*, the biasing network can be guided to bias towards entities that are of type *DeviceName*.

3.5. Two-Stage Training: DA-Aware RNN-T Training followed by Contextual Adaptation

While improving user-specific word recognition, a desirable E2E ASR should be able to maintain the performance on general speech recognition. To this end, we adopt the two-stage training similar to [17]. We first pre-train the DA-Aware RNN-T, i.e. RNN-T (the encoder network, prediction network, joint network, dense layer) augmented with the DA encoder and DA fusion networks on the live IVA distribution of utterances. In the second stage of contextual adaptation on user catalogs, we keep the RNN-T weights frozen while training the catalog encoder and the biasing network on upsampled personalized data mixed with the original general data under two conditions: (1) Freeze the DA encoder (DA Enc.) and fusion network (FN), (2) Unfreeze/ keep finetuning the DA Enc. and FN.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We use in-house IVA datasets where audio to transcription paired utterances are de-identified and randomly sampled from more than 20 domains such as *Global*, *Communications*, *SmartHome*, *Weather*, and *Music*. We consider the top 49 frequently occurring DAs in the IVA data sets including the *DefaultDialogAct* (usually associated with the first-turn utterances) and the non-default DAs (associated with more follow-up turns)[1]. 114k hours of data are used to pre-train the DA-Aware RNN-T. For training the catalog encoder and the biasing network, we use approximately 290 hours of data, containing a mix of *user-specific* and *general* training data with the ratio of 1.5:1 as suggested in [17]. *User-specific* datasets contain utterances with contact names, device names, and/or device locations. *General* utterances are sampled from the original training data distribution. We evaluate the models and report results on multiple test sets including a 20 hour *User-specific* dataset and a 75 hour *General* dataset. The *User-specific* test set is further split into the default DA and non-default DA set for further comparisons. To evaluate our model on individual turns of a dialog, we further created turn-wise test sets, which contain *user-specific* and non-default DA utterances from the first turn (turn 1), the second turn (turn 2) and the third turn (turn 3).

The relative word error rate reduction (WERR) is used throughout the experiments to summarize the overall, slot-type-wise, or turn-wise ASR performances. Given a model A's WER (WER_A) and a baseline B's WER (WER_B), the WERR of A over B can be computed by $(\text{WER}_B - \text{WER}_A)/\text{WER}_B$; a higher WERR indicates a better WER. We use three types of catalogs (Section 3.3). The maximum number of catalog entities fed into the catalog encoder is set to $K = 500$, and contains the correct *user-specific* entities.

4.2. Experimental Setup

4.2.1. Model Configurations

The DA guided contextual adapter has the following configuration. The encoder network has 5 LSTM layers and the prediction network is composed of 2 LSTM layers, both with 736 units followed by a feed-forward network of 512 units. The joint network is a fully-connected feed-forward component with one hidden layer followed by a $\tanh$ activation function. The DA encoder contains a 49-dim embedding layer followed by a feed-forward network of 64 units (i.e., $d_{da} = 64$). The DA late fusion network has a feed-forward network of 512 units. The catalog encoder is a BiLSTM layer with 128 units (each for forward and backward LSTMs) with input size 64. The final output from the catalog encoder is projected to 64-dim (i.e., $D_c = 64$). The biasing network projects the query, key and values to 64-dimensions (i.e., $d = 64$). The resulting biasing vector is then projected to the same size as the encoder output, 512-dim (i.e., $D_a = 512$).

We compare our model to the following baselines: (1) **No Context**: This is a vanilla RNN-T [4]. (2) **DA only**: This is an RNN-T using only DAs integrated with the input audio features [21], and (3) **CA**: This is an RNN-T based Contextual Adapter (CA) using only user catalogs [17]. All the models including the proposed one contain ~ 35 million parameters.

4.2.2. Input, Output, and Learning Rate Configurations

The input audio features consists of 64-dim LFBE features, which are extracted every 10 ms with a window size of 25 ms from audio samples. The features of each frame are then stacked with the left two frames, followed by a downsampling of factor 3 to achieve low frame rate, resulting in 192 feature dimensions. The subword tokenizer

[1]There does not exist an equivalent, publicly available contextual dataset containing both user catalogs and dialog acts in multi-turn dialogue scenario.

Table 1. WERRs (%) over No-context baseline [4]: Comparisons for DA guided CA, DA only, and CA models on IVA *User-specific* and *General* test sets. A higher number indicates a better WER.

Model	DA Fusion	DA Enc.&FN	User-Specific	General
No context [4]	N/A	N/A	baseline	baseline
DA Only [21]	N/A	N/A	1.49	-0.95
CA [17]	N/A	N/A	28.23	-0.59
DA guided CA (Ours)	Early Fusion	freeze	29.72	-1.02
		un-freeze	28.83	-0.66
	Late Fusion	freeze	28.75	-1.31
		un-freeze	30.61	-2.58
	Early-Late Fusion	freeze	30.53	-2.3
		un-freeze	**32.39**	-2.5

Table 2. WERRs (%) over No-context baseline [4] on Non-Default/Default DA splits of the *User-specific* test set. A higher number indicates a better WER.

Model	DA Fusion	DA Enc.&FN	Non-Default DA	Default DA
No context [4]	N/A	N/A	baseline	baseline
DA Only [21]	N/A	N/A	13.81	-1.27
CA [17]	N/A	N/A	29.91	27.85
DA guided CA (Ours)	Early Fusion	freeze	40.35	27.77
		un-freeze	39.71	26.84
	Late Fusion	freeze	39.01	27.00
		un-freeze	39.54	29.03
	Early-Late Fusion	freeze	40.04	28.69
		un-freeze	**41.56**	**30.46**

[26, 27] is used to create tokens from the transcriptions; we use 4000 tokens in total. We trained the RNN-T, the DA encoder, and the DA fusion network by minimizing the RNN-T loss [4] using the Adam optimizer [28], and varied the learning rate following [5, 25].with the starting LR = 1.5e-7 and ramping up for 3000 steps to reach LR = 4e-4. The LR is then held constant for $150K$ steps after which there is an exponential decay. For training the catalog encoder and biasing network in the CA stage, we use the Adam optimizer with LR = 1e-3 trained to convergence with early stopping.

4.3. Results

Table 1 presents the WERRs of the DA guided CA and baseline approaches over the no-context model for the *User-specific* and *General* test sets. The DA guided CA outperforms the no-context model by 32.39% relative improvement, while the DA only model and CA achieve 1.49%, and 28.90% WERRs respectively on the *User-specific* set. While unfreezing the DA Enc. and FN boosts the performance on the *User-specific* sets, it is more prone to overbiasing, resulting in more degradations on the *General* set compared to freezing

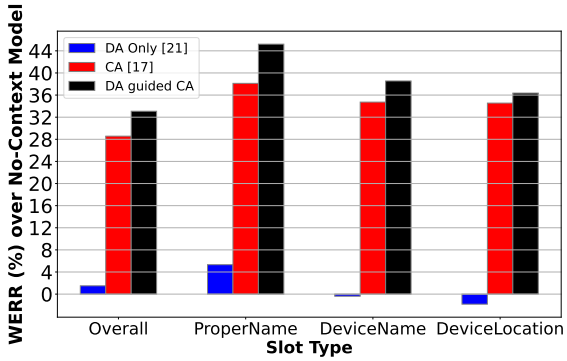

Fig. 4. Slot WERRs (%) over No-Context model [4] on the IVA *User-specific* test set.

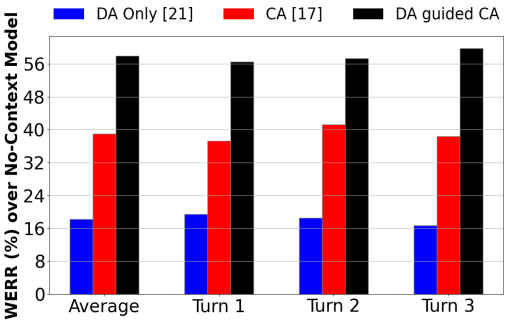

Fig. 5. Turn-wise WERRs (%) over No-context baseline [4] on the IVA *User-specific* multi-turn test set.

their weights: -1.31% WERR (freeze) vs. -2.58% WERR (unfreeze) for the late dialog act fusion, and -2.3% WERR (freeze) vs. -2.5% WERR (unfreeze) for the early-late dialog act fusion in Table 1. We also observe that bypassing the catalog selection with DAs on the best performing DA guided CA, the WERR over No-Context model drops from 32.39% to 31.43%, which indicates selecting catalogs and filtering catalog embeddings, i.e. keys/values, to the most relevant ones, leads to better biasing.

In Table 2, we show the WERRs on non-default DA and default DA splits. The DA guided CA, again, improves WER over the No-Context model the most (41.56%), compared to the WERRs achieved by the DA only (13.81%) and CA (29.91%) on the non-default DA split. The huge improvements seen in the non-default DA split can be explained by the fact that the *user-specific* utterances are usually associated with non-default DA.

We further compute the slot WERRs on three major *user-specific* slots in Fig. 4 along with the overall performances across all 131 slots. The DA guided CA model shows the most relative improvements in terms of Overall Slot WER against the No-Context model (33.08%), compared to the WERRs of DA Only (1.5%), and CA (28.57%) respectively. Here, we can see the improvements on the contact name slot are the highest across all models. The proposed model achieves the most improvement (45.21% relative). Notably, although DA Only model performs slightly worse than the No-Context model for the device name and device location slots (-0.38% and -1.82% WERRs), DA guided CA still leads to better relative improvement over No-Context model, compared to the CA's improvement for the device name (38.55% vs. 34.73%) and device location slots (36.36% vs. 34.55%) respectively.

Finally, we report the results on the turn-wise test sets in Fig. 5. The average of WERRs across turns are presented as well. While the DA Only model and CA have significant improvements, DA guided CA leads to the best improvements on average (58%), turn 1 (56.5%), turn 2 (57.3%), and turn 3 (59.8%) over the No-Context model. This shows that the DA guided CA model significantly improves performances for multi-turn personalized ASR over the baselines.

5. CONCLUSION

We proposed a dialog act guided contextual adapter approach to address personalized ASR in multi-turn dialogs. We leverage DAs to short list the most relevant catalogs and create better queries to guide the biasing network. The experimental results on IVA test sets show that the DA guided CA achieves on average 58% WER relative improvements over the no-context model on the *user-specific* multi-turn test set, in comparison to the prior-art contextual adapter model which achieved 39% over the no-context model.

6. REFERENCES

[1] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

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

[4] Alex Graves, "Sequence transduction with recurrent neural networks," *CoRR, vol. abs/1211.3711, 2*, 2012.

[5] Linhao Dong, S. Xu, and Bo Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *ICASSP*, 2018.

[6] Tony Bruguier, Fuchun Peng, and Françoise Beaufays, "Learning personalized pronunciations for contact names recognition," 2016.

[7] Tara N Sainath, Rohit Prabhavalkar, Shankar Kumar, Seungji Lee, Anjuli Kannan, David Rybach, Vlad Schogol, Patrick Nguyen, Bo Li, Yonghui Wu, et al., "No need for a lexicon? evaluating the value of the pronunciation lexica in end-to-end models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5859–5863.

[8] Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 418–425.

[9] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr.," in *Interspeech*, 2019.

[10] Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N Sainath, "Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition," in *ICASSP*, 2019.

[11] Mehryar Mohri, Fernando C Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, pp. 69–88, 2002.

[12] 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 *Interspeech*, 2018.

[13] Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko, "Domain-aware neural language models for speech recognition," in *ICASSP*, 2021.

[14] Mahaveer Jain, Gil Keren, Jay Mahadeokar, and Yatharth Saraf, "Contextual rnn-t for open domain asr," *arXiv preprint arXiv:2006.03411*, 2020.

[15] Da-Rong Liu, Chunxi Liu, Frank Zhang, Gabriel Synnaeve, Yatharth Saraf, and Geoffrey Zweig, "Contextualizing asr lattice rescoring with hybrid pointer network language model," *arXiv preprint arXiv:2005.07394*, 2020.

[16] Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann, "Context-aware transformer transducer for speech recognition," *ASRU*, 2021.

[17] 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 *ICASSP*, 2022, pp. 8537–8541.

[18] Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xiangyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, et al., "Personalization strategies for end-to-end speech recognition systems," in *ICASSP*, 2021.

[19] 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 2019*, 2019, pp. 1418–1422.

[20] Waheed Ahmed Abro, Guilin Qi, Huan Gao, Muhammad Asif Khan, and Zafar Ali, "Multi-turn intent determination for goal-oriented dialogue systems," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[21] Kai Wei, Thanh Tran, Feng-Ju Chang, Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Jing Liu, Anirudh Raju, Ross McGowan, Nathan Susanj, Ariya Rastrow, et al., "Attentive contextual carryover for multi-turn end-to-end spoken language understanding," *ASRU*, 2021.

[22] Suyoun Kim, Siddharth Dalmia, and Florian Metze, "Gated embeddings in end-to-end speech recognition for conversational-context fusion.," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1131–1141.

[23] Zelin Wu, Bo Li, Yu Zhang, Petar S Aleksic, and Tara N Sainath, "Multistate encoding with end-to-end speech rnn transducer network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7819–7823.

[24] Raghav Gupta, Abhinav Rastogi, and Dilek Hakkani-Tür, "An efficient approach to encoding context for spoken language understanding," in *Interspeech 2018*, 2018, pp. 3469–3473.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016.

[27] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, vol. 1, pp. 66–75.

[28] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.