Deferred NAM: Low-latency Top-K Context Injection via Deferred Context Encoding for Non-Streaming ASR

Zelin Wu **Gan Song** **Christopher Li** **Pat Rondon** **Zhong Meng**
Xavier Velez **Weiran Wang** **Diamantino Caseiro** **Golan Pundak**
Tsendsuren Munkhdalai **Angad Chandorkar** **Rohit Prabhavalkar**
Google LLC, USA
{zelinwu, gansong, chriswli, rondon, zhongmeng, xavvelez}@google.com

Abstract

Contextual biasing enables speech recognizers to transcribe important phrases in the speaker's context, such as contact names, even if they are rare in, or absent from, the training data. Attention-based biasing is a leading approach which allows for full end-to-end cotraining of the recognizer and biasing system and requires no separate inference-time components. Such biasers typically consist of a context encoder; followed by a context filter which narrows down the context to apply, improving per-step inference time; and, finally, context application via cross attention. Though much work has gone into optimizing per-frame performance, the context encoder is at least as important: recognition cannot begin before context encoding ends. Here, we show the lightweight phrase selection pass can be moved before context encoding, resulting in a speedup of up to 16.1 times and enabling biasing to scale to 20K phrases with a maximum pre-decoding delay under 33ms. With the addition of phrase- and wordpiece-level cross-entropy losses, our technique also achieves up to a 37.5% relative WER reduction over the baseline without the losses and lightweight phrase selection pass.

1 Introduction

Automatic speech recognition (ASR) applications often succeed or fail based on their ability to recognize words that are relevant in context, but may not be common, or even present, in the training data. For example, an assistant user may speak contact names from another language or titles of media entities which were released after the ASR system was trained, or the speaker may use domain-specific jargon, like legalese or medical terms, which are not common in the more-typical speech used for training. ASR contextual biasing (Hall et al., 2015; Aleksic et al., 2015) aims to account for this domain shift between training and inference.

Attention-based biasing (Pundak et al., 2018), in which the context is encoded into dense embeddings attended to during recognition, is one of the leading approaches for contextualizing end-to-end (E2E) ASR systems. As a fully-end-to-end method, it does not require separate biasing components which must be separately trained and whose integration with the core ASR system must be optimized. However, recognition cannot begin until the inference-time context has been encoded, and this context may consist of tens of thousands of items which may not be cached, for privacy or system design reasons or, more simply, because the context may change at any point up until the beginning of recognition. Thus, the context encoder must be able to efficiently handle very large contexts, as delays in context encoding translate directly into user-visible delays in ASR transcription.

In this work, we optimize context encoding by splitting it into two passes. We make the simplifying assumption that our ASR system is non-causal and can access the entire audio input; we argue that this choice is not overly restrictive, as two-pass ASR systems combining a causal ASR system to produce streaming results with a non-causal ASR system for producing the final result have become common, e.g., as in (Narayanan et al., 2021). With the speaker's entire audio available before context encoding begins, we can split context encoding into two phases. First, we encode all contextual phrases with an extremely lightweight encoder, and use the resulting encodings to determine the k phrases that are most likely to occur in the audio. We embed only the k most-likely phrases with a more powerful — but more expensive — encoder, and use the output of this "deferred" encoder for biasing.

2 Related Work

Conventional ASR contextualization relies on discrete contextualization components like (class-based) language models (LMs) (Vasserman et al.,

2016), combined with the base ASR system through on-the-fly LM rescoring (Aleksic et al., 2015; Hall et al., 2015; McGraw et al., 2016), shallow fusion (Williams et al., 2018; Zhao et al., 2019), or lattice rewriting (Serrino et al., 2019) and contextual spelling correctors (Wang et al., 2021; Antonova et al., 2023). The use of discrete contextualization components requires the implementor of an ASR system to separately train the ASR and contextualization components, to separately optimize their combination, and to take care that all relevant signals, like the input audio, are forwarded from ASR to the contextualization system.

Attention-based biasing (Pundak et al., 2018; Chang et al., 2021; Munkhdalai et al., 2022), in which the ASR network learns to use inference-time context through attention, and thus requires no separate contextualization components and can be optimized end-to-end by standard backpropagation, has become a popular method for contextualization of end-to-end ASR models. The core technique has spawned several distinct lines of complementary research. There are threads of work on improving data efficiency by adapter-style training (Sathyendra et al., 2022) or improving training on synthetic audio derived from text-only data (Naowarat et al., 2023). Work on precision improvements seeks to lower the rate of over-biasing through hierarchical or gated attention (Han et al., 2022; Munkhdalai et al., 2023; Wu et al., 2023; Tong et al., 2023a; Xu et al., 2023a; Alexandridis et al., 2023) or through slot triggering (Lu et al., 2023; Tong et al., 2023b); often, such techniques can improve not only quality but also per-step inference run time (Munkhdalai et al., 2023; Yang et al., 2023). Notably, Tong et al (Tong et al., 2023a) improve WER through an auxiliary slot-level cross-entropy loss. We do not use slots for selecting context to bias, as we have found it possible to get high performance with hierarchical attention alone; however, in this work, we do apply a phrase-level, rather than slot-level, cross-entropy loss during training to improve WER. Further we extend the technique to a novel wordpiece-level cross-entropy loss. Further work aims to improve quality on biased utterances by providing the context encoder with more information than just the graphemic biasing phrases, including phoneme-level features (Bruguier et al., 2019; Hu et al., 2019; Chen et al., 2019; Pandey et al., 2023), and augmenting the context encoder with semantics-aware embeddings from a pretrained BERT model (Fu et al., 2023). Attention-based biasing can also be complemented by shallow fusion with (contextualized) language models (Xu et al., 2023b) or combinations of language models and contextualized rescoring (Dingliwal et al., 2023) for further quality improvements.

We distinguish the current work from most of the above by focusing on the inference run time of the *context encoder* which, as noted in Section 1, is critical to the usability of contextualized ASR.

2.1 Dual-mode NAM

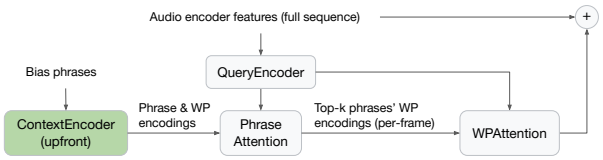

Figure 1: Inference with Dual-mode NAM.

Our work builds on Dual-mode NAM (Neural Associative Memory) (Wu et al., 2023) and its shared query encoder variant from Section 3.4.2 of (Song et al., 2023), which we use as a baseline. Dual-mode NAM computes both phrase encodings E^p and wordpiece (WP) encodings E^w through a fine-grained context encoder, where E^p corresponds to the cls embedding and E^w corresponds to the bias phrase WP embeddings of Z; $Z = \{cls; w_{n,1}, ..., w_{n,L}\}_{n=1}^{N}$, N is the number of bias phrases associated with the utterance and L is the number of WPs per bias phrase.

$$E^p, E^w = ContextEncoder(Z) \quad (1)$$

The phrase and WP attentions are trained via sampling: The phrase and WP attention contexts (c^p, c^w) are added to the audio encoder features x with a probability of p and $1 - p$, respectively.

$$x^{biased} = x + BernoulliTrial(c^p, c^w, p) \quad (2)$$

During inference (Figure 1), the model leverages the phrase-level attention logits to select per-frame top-k (k^p) phrases and feed their WP encodings to the WP attention, where $q^p_{t,h}$ and $K^p_{t,h}$ correspond to the projected audio query and E^p encodings.

$$I^p_t = TopK(\frac{1}{H}\sum_{h=1}^{H} q^p_{t,h} K^p_{t,h}, k^p) \quad (3)$$

3 Methods

Unlike conventional approaches that uniformly encode all bias phrases beforehand, Deferred NAM

(Figure 2) utilizes a lightweight phrase encoder and retrieval process to select the top-k relevant phrases, before invoking the fine-grained context encoder and WP attention at inference. Additionally, Deferred NAM employs cross-entropy losses with its phrase and WP attentions, further boosting WER performance. This design achieves both minimal latency as well as excellent recognition quality.

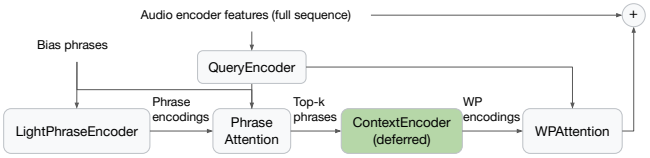

Figure 2: Inference with Deferred NAM.

3.1 Lightweight phrase encoder

We adopted Iyyer et al. (2015)'s Deep Averaging Network (DAN) as a lightweight encoder to produce phrase encodings $E^p \in \mathbb{R}^{N\times d}$, where the WP embeddings $W = \{w_{n,1}, ..., w_{n,L}\}_{n=1}^{N}$ are averaged over the L axis and then encoded by a feed-forward network with TANH activation, and d is the WP embedding/encoding dimension. We applied stop gradient ($\perp$) to prevent $LightPhraseEncoder$ from interfering with $ContextEncoder$ learning of the WP embeddings, which resulted in faster biasing WER convergence.

$$E^p = LightPhraseEncoder(\perp(W)) \tag{4}$$

3.2 Cross-entropy guided phrase attention

Algorithm 1 NO_BIAS-augmented multi-head attention logits with mean-max pooling: $f(q, k)$

Inputs: Query $q \in \mathbb{R}^{T\times d_q}$, key $k \in \mathbb{R}^{S\times d_k}$
Outputs: Mean-max pooled logits $z \in \mathbb{R}^{(1+S)}$

$$q'_{t,h} = q_t\Theta_h^Q, k'_h = [\Theta_h^{NB}; k\Theta_h^K] \tag{5}$$

$$z_t = \frac{1}{H}\sum_{h=1}^{H}\frac{q'_{t,h}(k'_h)^\top}{\sqrt{d_h}} \tag{6}$$

$$z = \max_{t=1}^{T} z_t \tag{7}$$

The trainable parameters are $\Theta_h^Q \in \mathbb{R}^{d_q\times d_h}$, $\Theta_h^K \in \mathbb{R}^{d_k\times d_h}$, $\Theta_h^{NB} \in \mathbb{R}^{d_h}$ (NO_BIAS token), where $t \in [1..T]$ denotes the time index and $h \in [1..H]$ denotes the attention head index.

As discussed in Section 5 of (Wu et al., 2023), one limitation is that the phrase/WP attentions are trained on fewer examples due to sampling (equation 2). Another limitation is that the retrieval capability (equation 3) is indirectly learned by the ASR loss. We address such limitations by learning the retrieval capability with an explicit loss.

In phrase attention, the relevance between audio query x^q and bias phrases E^p is computed via Algorithm 1. The mean-max pooled logits are then used to compute the softmax cross-entropy loss:

$$z^p = f(x^q, E^p) \tag{8}$$

$$\mathcal{L}_p = L\_SCE(z^p, labels) \tag{9}$$

where $labels \in \mathbb{R}^{1+N}$ corresponds to a probability distribution of the bias labels, with the leading "1" being the NO_BIAS token. During training, a bias phrase is marked as a correct label if it's a longest substring of the transcript truth; the NO_BIAS token is marked as the correct label if none of the bias phrases is a substring of the transcript truth.

At inference, the global top k^p phrases are used to invoke the context encoder and WP attention.

$$I^p_{global} = TopK(z^p_{[2:]}, k^p) \tag{10}$$

3.3 Deferred context encoder

By offloading phrase encoding to a dedicated lightweight encoder, the context encoder can exclusively focus on generating fine-grained WP encodings for utilization by the WP attention. While the context encoder is trained on the same set of bias phrases W as the phrase encoder, during inference, only the top-k phrases identified through the phrase attention mechanism require encoding.

$$E^w = ContextEncoder(W) \tag{11}$$

Where $E^w = \{e_{n,1}, ..., e_{n,L}\}_{n=1}^{N} \in \mathbb{R}^{N\times L\times d}$ represents WP encodings, such that $e_{i,j} \in \mathbb{R}^d$ represents the i^{th} phrase candidate's wordpiece encoding at position $j \in \{1, ..., L\}$.

3.4 Cross-entropy guided WP attention

After that, the standard NAM WP attention biasing context c^w is computed and added to the acoustic encoder feature for contextualization.

$$c^w = WPAttention(x^q, E^w) \tag{12}$$

$$x^{biased} = x + c^w \tag{13}$$

We further augment the $WPAttention$ with a cross-entropy training loss to boost the likelihood scores of the relevant WPs. Similar to the phrase attention,

we first compute the mean-max pooled logits for the WP tokens using Algorithm 1:

$$z^w = f(x^q, E^w) \in \mathbb{R}^{1+NL} \tag{14}$$

Secondly, the per-phrase average WP logits $\overline{z^w_{[2:]}}$ are computed, i.e., by summing the logits of each phrase and then dividing by the phrase's effective sequence length, ignoring padding tokens.

$$\overline{z^w_{[2:]}} = PerPhraseAvg(z^w_{[2:]}) \in \mathbb{R}^N \tag{15}$$

Thirdly, the NO_BIAS logit $z^w_{[1:2]}$ is concatenated with the per-phrase average WP logits $\overline{z^w_{[2:]}}$ to form $\overline{z^w}$ for calculating the cross-entropy loss.

$$\overline{z^w} = [z^w_{[1:2]}; \overline{z^w_{[2:]}}] \in \mathbb{R}^{1+N} \tag{16}$$

$$\mathcal{L}_w = L\_SCE(\overline{z^w}, labels) \tag{17}$$

Finally, the total loss is a weighted sum of the ASR, phrase- and WP-level cross-entropy losses.

$$\mathcal{L}_{total} = \mathcal{L}_{asr} + \lambda_p \mathcal{L}_p + \lambda_w \mathcal{L}_w \tag{18}$$

4 Experiment setup

4.1 Data sets

All data sets used aligned with the Privacy Principles and AI Principles in (Google, 2010, 2023). The ASR training data contains 520M anonymized English voice search utterances, totaling 490K hours of speech with an average of 3.4 seconds per utterance. A small percentage of the utterances are human-transcribed and the rest are machine-transcribed by a teacher ASR model (Hwang et al., 2022). We evaluate our system on the same multi-context biasing corpora described in Section 3.2.2 of (Munkhdalai et al., 2023). The corpora consist of three sets: **WO_PREFIX**: 1.3K utterances matching prefix-less patterns from $APPS, $CONTACTS, and $SONGS categories (denoted ACS). **W_PREFIX**: 2.6K utterances matching prefixed patterns such as "open $APPS", "call $CONTACTS", "play $SONGS". **ANTI**: 1K utterances simulating general voice assistant queries. Each utterance is assigned up to 3K ACS bias entities. The WO_PREFIX and W_PREFIX sets measure in-context performance, where one of the entities appears in the transcript truth; the ANTI set measures anti-context performance, where the utterances are assigned distractor entities only.

4.2 Model architecture

Our RNN-T ASR encoder architecture mimics that of Google's Universal Speech Model (USM) (Zhang et al., 2023). We use the 128-dimensional log Mel-filterbank energies (extracted from 32ms window and 10ms shift) as the frontend features, which are fed to two 2D-convolution layers, each with strides $(2, 2)$; the resulting feature sequence becomes the input to a stack of 16 Conformer layers (Gulati et al., 2020). Each conformer layer has 8 attention heads with a total dimension of 1536, and the intermediate dimension of the FFNs is 4 times the attention dimension, yielding a total of 870M parameters in the encoder. The Conformer blocks use local self-attention with a large attention span, and the encoder output has a large enough receptive field to cover the entire utterance. We apply funnel pooling (Dai et al., 2020) at the 5th to 7th conformer layers, each with a reduction rate of 2. As a result, the encoder output sequence has a low frame rate of 320ms. The model uses a $|V|^2$ embedding decoder (Botros et al., 2021), i.e., the prediction network computes LM features based on two previous non-blank tokens. The output vocabulary size consists of 4096 lowercase wordpieces.

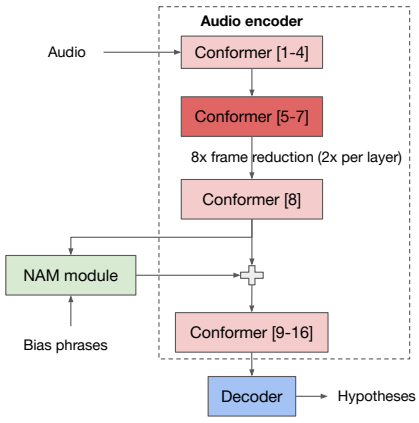

Figure 3: Non-streaming RNN-T + NAM module.

Both baseline and experiment NAM modules are placed in between 8th and 9th Conformer layer as shown in Fig. 3. In accordance with (Wu et al., 2023), bias phrases are sampled from reference transcripts during training; the bias strength $\lambda = 0.6$ is introduced at inference: $x^{biased} = x + \lambda c^w$.

The models are developed using the open-source Lingvo toolkit (Shen et al., 2019), trained on 16x16 cloud TPU v3 (Jouppi et al., 2020) with a global batch size of 4096 for 240K steps (3 days). All parameters are randomly initialized and optimized by the Adafactor optimizer (Shazeer and Stern, 2018).

4.3 Baseline: Dual-mode NAM

We configured dual-mode NAM B1, B2 modules to be the WER, latency baseline of Deferred NAM.

The B1 module has 100.8M parameters, with *QueryEncoder* (89.8M): a 2L Conformer with a model dimension of 1536, hidden dimensions of [6144, 3072] for the internal feed-forward layers; *ContextEncoder* (4M): 3L Conformer (3M) with a model, hidden dimension of 256, 512 and the WP embedding table (1M) with a vocab size of 4096. *PhraseAttention*/*WPAttention* (5.5M each): 8 heads with a per-head dimension of 192.

The B2 module is similar to B1 except that *ContextEncoder* contains 1L Conformer (1M) instead of 3L. In Table 1, we show that reducing the context encoder from 3L to 1L for a premature latency optimization would negatively impact the in-context WERs by up to 26.8% relative (3.0 → 4.1).

Expt	**B1**	B2
ANTI	2.3	1.9
WO_PREFIX	**3.0**	4.1
W_PREFIX	**2.4**	2.9

Table 1: Dual-mode NAM average WERs at k^p=32, computed by averaging over scenarios where (150, 300, 600, 1.5K, 3K) bias entities are provided per utterance.

4.4 Experiment: Deferred NAM

We explore variants D1–D3 to study the impact of each proposed training loss. When compared to B1, Deferred NAM adds a *LightPhraseEncoder*: a 4L DAN phrase encoder (788K parameters) with TANH activation, a model, hidden dimension of 256; the *PhraseAttention* (2.8M parameters) contains only parameters as described in Algorithm 1; the *ContextEncoder* consists of only a 1L Conformer (1M) identical to B2, which we found is sufficient to outperform the WERs of B1 (3L).

D1 The model learns retrieval capability via equation 2, where $p = 0.3$ and $\mathcal{L}_{total} = \mathcal{L}_{asr}$.

D2 The model learns retrieval capability through cross-entropy guided phrase attention (CE-PA): $\mathcal{L}_{total} = \mathcal{L}_{asr} + 0.1\mathcal{L}_p$.

D3 D2 with cross-entropy guided WP attention (CE-WA): $\mathcal{L}_{total} = \mathcal{L}_{asr} + 0.1\mathcal{L}_p + 0.1\mathcal{L}_w$.

5 Results

5.1 Quality

As shown in Table 2, the base Deferred NAM (D1) already outperforms the best Dual-mode NAM (B1)'s average WERs by up to 20.8% relative (2.4 → 1.9), while using fewer parameters in total. By augmenting the model with the CE-PA loss, D2 improves over D1 by up to 11.5% relative (2.6 → 2.3). With addition of CE-WA loss, D3 improves over D2 by 16.7% relative (1.8 → 1.5). As a result, the best Deferred NAM (D3) improves over the best Dual-mode NAM (B1) by up to 37.5% relative (2.4 → 1.5) on in-context recognition, and 21.7% on anti-context recognition (2.3 → 1.8).

Expt	B1	D1	D2	D3
ANTI	2.3	1.9	**1.8**	**1.8**
WO_PREFIX	3.0	2.6	2.3	**2.0**
W_PREFIX	2.4	1.9	1.8	**1.5**

Table 2: Average WERs (# entities = 0 excluded) at k^p=32 comparing Dual-mode (B1) and Deferred NAM (D1–D3). WER breakdown is shown in Table 3.

Expt	#	B1	D1	D2	D3
ANTI	0	**1.4**	**1.4**	1.5	1.6
	150	1.7	1.6	**1.5**	1.7
	300	2.0	**1.6**	**1.6**	1.8
	600	2.4	2.0	**1.7**	1.9
	1.5K	2.6	**1.9**	**1.9**	**1.9**
	3K	2.9	2.2	2.3	**1.9**
WO_ PREFIX	0	**21.1**	21.6	**21.1**	21.3
	150	1.8	2.0	1.8	**1.6**
	300	2.0	2.0	2.0	**1.7**
	600	2.7	2.3	2.3	**1.8**
	1.5K	3.4	3.1	2.4	**2.3**
	3K	4.9	3.5	3.1	**2.7**
W_ PREFIX	0	9.8	9.9	10.0	9.9
	150	1.5	1.6	1.5	**1.3**
	300	1.8	1.6	1.6	**1.4**
	600	2.1	1.8	1.7	**1.5**
	1.5K	2.8	2.1	1.8	**1.6**
	3K	3.6	2.3	2.2	**1.9**

Table 3: Detailed WER breakdown at k^p=32 on the best Dual-mode (B1) and Deferred NAM (D1–D3), where each utterance is assigned up to 3K bias entities.

We also show Deferred NAM's 1st pass retrieval recall performance in Table 4. Interestingly, although there are sizable retrieval performance in-

creases at smaller $k^p \leq 5$ from D1 to D2 (up to 25.6% relative, e.g., 69.2 $\rightarrow$ 86.9), D1's recall performance at $k^p = 32$ (where the WERs are evaluated at) is already quite high and left little room for further improvement (up to 2.3% relative, e.g., 97.4 $\rightarrow$ 99.6). On the other hand, D3's retrieval performance is more in line with D2, as expected.

Given the slightly-improved or similar recall performance at $k^p = 32$, we attribute D2's WER improvement over D1 to better-regulated audio/text embeddings (due to CE-PA) and increased data exposure of the 2nd-pass contextualization, i.e., D2 has full training data exposure while D1's phrase and WP attentions are only trained 30% and 70% of the time, respectively, due to sampling. D3's WER improvement over D2 is more obvious as the CE-WA loss is directly applied to the WP attention.

Expt	k^p	D1	D2	D3
WO_PREFIX	1	73.0	91.5	**93.4**
	5	92.8	98.7	**98.9**
	32	98.9	**99.7**	**99.7**
W_PREFIX	1	69.2	86.9	**87.9**
	5	89.8	**98.0**	97.7
	32	97.4	**99.6**	99.5

Table 4: Retrieval recall performance of D1–D3 by k^p for in-context test-sets at 3K bias entities per utterance.

5.2 Inference latency

Deferred NAM latency (ms)	# phrases 3K	20K
QueryEncoder	2.3	2.3
LightPhraseEncoder	3.5	22.8
PhraseAttention	0.9	5.2
ContextEncoder	1.3	1.3
WPAttention	0.7	0.7
Total	8.7	32.3

Table 5: Latency of Deferred NAM (D3), at k^p=32.

The ASR is benchmarked on 1x1 cloud TPU V3 at bfloat16, with batch size (number of utterances) 8; each utterance has 512 time steps (15.36s of audio), and each bias phrase has a length of 16 WPs. As shown in Table 5, the total latency of Deferred NAM at processing 3K and 20K bias phrases is only 8.7ms and 32.3ms, respectively. On the other hand, encoding all bias phrases up front (Dual-mode NAM) is significantly slower, i.e., B1's 3L Conformer context encoder latency alone is 214ms and 1549ms; B2's 1L Conformer context encoder latency alone is at 72ms and 520ms. Overall, Deferred NAM provides a speedup of at least 8.3X and 16.1X over Dual-mode NAM's best-case latency scenario (B2), and surpasses the best-case quality scenario (B1), as discussed in Section 5.1.

5.3 Exploration: NO_BIAS filter

We investigated gating the WP attention with the phrase-level NO_BIAS token, where a bias phrase is deemed "active" if its per-frame logit is higher than that of the NO_BIAS token ($z_t^p[1:2]$) at any frame. Method 1 (M1) filters inactive phrases before WP attention (equation 20). However, with fewer WPs to attend to, the remaining ones receive higher attention probabilities. This could degrade anti-context WERs if the accuracy of the NO_BIAS token is not sufficiently high, i.e., under condition m, only 43.6% of utterances are deemed inactive for D3 at ANTI (3K). Therefore, we explored Method 2 (M2), which zeroes out the WP attention *values* of inactive phrases, leaving the probabilities of the WP *keys* unaffected (equation 21).

$$m = \vee_{t=1}^{T}(z_t^p[2:] > z_t^p[1:2]) \in \mathbb{R}^N \quad (19)$$

$$Method\ 1 : (I^p_{global})' = Filter(I^p_{global}, m) \quad (20)$$

$$Method\ 2 : (v^w)' = v^w_{i,j}\Theta^V \circ m_i \quad (21)$$

$v^w_{i,j}$: The attention value of i-th phrase at j-th WP; Θ^V: The value projection matrix of WP attention.

Expt	D1		D3		
	N/A	M1	N/A	M1	M2
ANTI	1.9	**1.4**	1.8	1.9	1.7
WO_PREFIX	2.6	12.9	2.0	**1.8**	2.1
W_PREFIX	1.9	8.3	**1.5**	**1.5**	1.6

Table 6: Average WERs (# entities = 0 excluded) for Deferred NAM (D1 & D3) with NO_BIAS filtering.

Table 6 shows Deferred NAM's (D3) ability to use the phrase-level NO_BIAS token in inference. D3-M1 shows a relative improvement of up to 10% (2 $\rightarrow$ 1.8) in in-context WERs, with a relative anti-context decline of 5.6% (1.8 $\rightarrow$ 1.9). Conversely, D3-M2 shows a relative improvement of 5.6% (1.8 $\rightarrow$ 1.7) in anti-context WERs with a relative increase of up to 6.7% (1.5 $\rightarrow$ 1.6) in in-context WERs. Notably, the NO_BIAS token's logit dominates in D1 (similar to Dual-mode NAM (Wu et al., 2023)) due to the lack of supervised training (i.e., CE-PA), leading to a substantial relative increase of up to 396.2% (2.6 $\rightarrow$ 12.9) in in-context WERs.

6 Conclusion

We proposed a low-latency attention-based contextual ASR system, augmented with phrase- and WP-level cross-entropy losses, which can handle thousands of bias phrases within milliseconds while achieving up to 37.5% relative average WER reduction. This demonstrates the potential for enhancing ASR in real-world applications requiring fast and accurate contextual speech recognition.

References

Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno. 2015. Bringing contextual information to google speech recognition. In *Proc. Interspeech 2015*, pages 468–472.

Anastasios Alexandridis, Kanthashree Mysore Sathyendra, Grant P. Strimel, Feng-Ju Chang, Ariya Rastrow, Nathan Susanj, and Athanasios Mouchtaris. 2023. Gated Contextual Adapters For Selective Contextual Biasing In Neural Transducers. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alexandra Antonova, Evelina Bakhturina, and Boris Ginsburg. 2023. SpellMapper: A non-autoregressive neural spellchecker for ASR customization with candidate retrieval based on n-gram mappings. In *Proc. INTERSPEECH 2023*, pages 1404–1408.

Rami Botros, Tara N. Sainath, Robert David, Emmanuel Guzman, Wei Li, and Yanzhang He. 2021. Tied & Reduced RNN-T Decoder. In *Proc. Interspeech 2021*, pages 4563–4567.

Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N. Sainath. 2019. Phoebe: Pronunciation-aware Contextualization for End-to-end Speech Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6171–6175.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-Aware Transformer Transducer for Speech Recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 503–510.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L. Seltzer, and Christian Fuegen. 2019. Joint Grapheme and Phoneme Embeddings for Contextual End-to-End ASR. In *Proc. Interspeech 2019*, pages 3490–3494.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. 2020. Funnel-transformer: filtering out sequential redundancy for efficient language processing. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Saket Dingliwal, Monica Sunkara, Srikanth Ronanki, Jeff Farris, Katrin Kirchhoff, and Sravan Bodapati. 2023. Personalization of CTC speech recognition models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 302–309.

Xuandi Fu, Kanthashree Mysore Sathyendra, Ankur Gandhe, Jing Liu, Grant P. Strimel, Ross McGowan, and Athanasios Mouchtaris. 2023. Robust Acoustic And Semantic Contextual Biasing In Neural Transducers For Speech Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Google. 2010. Google's Privacy Principles.

Google. 2023. Artificial Intelligence at Google: Our Principles.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Keith Hall, Eunjoon Cho, Cyril Allauzen, Françoise Beaufays, Noah Coccaro, Kaisuke Nakajima, Michael Riley, Brian Roark, David Rybach, and Linda Zhang. 2015. Composition-based on-the-fly rescoring for salient n-gram biasing. In *Proc. Interspeech 2015*, pages 1418–1422.

Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. 2022. Improving End-to-End Contextual Speech Recognition with Fine-Grained Contextual Knowledge Selection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8532–8536.

Ke Hu, Antoine Bruguier, Tara N. Sainath, Rohit Prabhavalkar, and Golan Pundak. 2019. Phoneme-Based Contextualization for Cross-Lingual Speech Recognition in End-to-End Models. In *Proc. Interspeech 2019*, pages 2155–2159.

Dongseong Hwang, Khe Chai Sim, Zhouyuan Huo, and Trevor Strohman. 2022. Pseudo Label Is Better Than Human Label. In *Proc. Interspeech 2022*, pages 1421–1425.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Norman P. Jouppi, Doe Hyun Yoon, George Kurian, Sheng Li, Nishant Patil, James Laudon, Cliff Young, and David Patterson. 2020. A Domain-Specific Supercomputer for Training Deep Neural Networks. *Commun. ACM*, 63(7):67–78.

Yiting Lu, Philip Harding, Kanthashree Mysore Sathyendra, Sibo Tong, Xuandi Fu, Jing Liu, Feng-Ju Chang, Simon Wiesler, and Grant P. Strimel. 2023. Model-Internal Slot-triggered Biasing for Domain Expansion in Neural Transducer ASR Models. In *Proc. INTERSPEECH 2023*, pages 1324–1328.

Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, and Carolina Parada. 2016. Personalized speech recognition on mobile devices. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5955–5959.

Tsendsuren Munkhdalai, Khe Chai Sim, Angad Chandorkar, Fan Gao, Mason Chua, Trevor Strohman, and Françoise Beaufays. 2022. Fast Contextual Adaptation with Neural Associative Memory for On-Device Personalized Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6632–6636.

Tsendsuren Munkhdalai, Zelin Wu, Golan Pundak, Khe Chai Sim, Jiayang Li, Pat Rondon, and Tara N. Sainath. 2023. NAM+: Towards Scalable End-to-End Contextual Biasing for Adaptive ASR. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 190–196.

Burin Naowarat, Philip Harding, Pasquale D'Alterio, Sibo Tong, and Bashar Awwad Shiekh Hasan. 2023. Effective Training of Attention-based Contextual Biasing Adapters with Synthetic Audio for Personalised ASR. In *Proc. INTERSPEECH 2023*, pages 1264–1268.

Arun Narayanan, Tara N. Sainath, Ruoming Pang, Jiahui Yu, Chung-Cheng Chiu, Rohit Prabhavalkar, Ehsan Variani, and Trevor Strohman. 2021. Cascaded Encoders for Unifying Streaming and Non-Streaming ASR. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5629–5633.

Rahul Pandey, Roger Ren, Qi Luo, Jing Liu, Ariya Rastrow, Ankur Gandhe, Denis Filimonov, Grant Strimel, Andreas Stolcke, and Ivan Bulyko. 2023. Procter: Pronunciation-Aware Contextual Adapter For Personalized Speech Recognition In Neural Transducers. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep Context: End-to-end Contextual Speech Recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 418–425.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8537–8541.

Jack Serrino, Leonid Velikovich, Petar Aleksic, and Cyril Allauzen. 2019. Contextual Recovery of Out-of-Lattice Named Entities in Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 3830–3834.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, et al. 2019. Lingvo: a Modular and Scalable Framework for Sequence-to-Sequence Modeling.

Gan Song, Zelin Wu, Golan Pundak, Angad Chandorkar, Kandarp Joshi, Xavier Velez, Diamantino Caseiro, Ben Haynor, Weiran Wang, Nikhil Siddhartha, Pat Rondon, and Khe Chai Sim. 2023. Contextual Spelling Correction with Large Language Models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Sibo Tong, Philip Harding, and Simon Wiesler. 2023a. Hierarchical Attention-Based Contextual Biasing For Personalized Speech Recognition Using Neural Transducers. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Sibo Tong, Philip Harding, and Simon Wiesler. 2023b. Slot-Triggered Contextual Biasing For Personalized Speech Recognition Using Neural Transducers. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Lucy Vasserman, Ben Haynor, and Petar Aleksic. 2016. Contextual language model adaptation using dynamic classes. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 441–446.

Xiaoqiang Wang, Yanqing Liu, Sheng Zhao, and Jinyu Li. 2021. A Light-Weight Contextual Spelling Correction Model for Customizing Transducer-Based Speech Recognition Systems. In *Proc. Interspeech 2021*, pages 1982–1986.

Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach, and Tara Sainath. 2018. Contextual Speech Recognition in End-to-end Neural Network Systems Using Beam Search. In *Proc. Interspeech 2018*, pages 2227–2231.

Zelin Wu, Tsendsuren Munkhdalai, Pat Rondon, Golan Pundak, Khe Chai Sim, and Christopher Li. 2023. Dual-Mode NAM: Effective Top-K Context Injection for End-to-End ASR. In *Proc. INTERSPEECH 2023*, pages 221–225.

Tianyi Xu, Zhanheng Yang, Kaixun Huang, Pengcheng Guo, Ao Zhang, Biao Li, Changru Chen, Chao Li, and Lei Xie. 2023a. Adaptive Contextual Biasing for Transducer Based Streaming Speech Recognition. In *Proc. INTERSPEECH 2023*, pages 1668–1672.

Yaoxun Xu, Baiji Liu, Qiaochu Huang, Xingchen Song, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2023b. CB-Conformer: Contextual Biasing Conformer for Biased Word Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhanheng Yang, Sining Sun, Xiong Wang, Yike Zhang, Long Ma, and Lei Xie. 2023. Two Stage Contextual Word Filtering for Context Bias in Unified Streaming and Non-streaming Transducer. In *Proc. INTERSPEECH 2023*, pages 3257–3261.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-Fusion End-to-End Contextual Biasing. In *Proc. Interspeech 2019*, pages 1418–1422.