A Neural Model for Contextual Biasing Score Learning and Filtering

Wanting Huang *Weiran Wang* University of Iowa

ASRU 2025

Poster No. 208

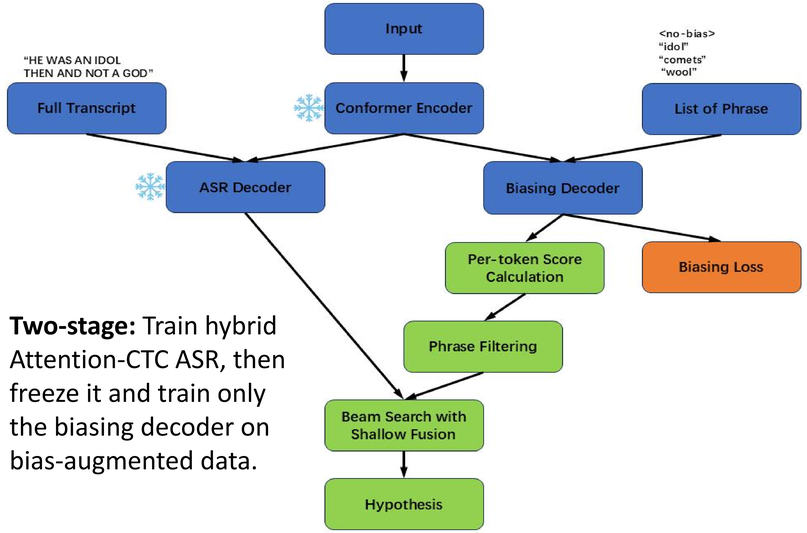
INTRODUCTION

Motivation

- ASR is still relatively weak in identifying **rare words, user-specific words, and entity names.**
- **Contextual biasing** injects a curated list of task or user-specific phrases into decoding, encouraging the ASR system to favor them.

Contributions

- **Attention-based biasing decoder** with audio-conditioned phrase scores.
- **Per-token discriminative training** over phrases.
- Learned scores drive **phrase filtering** and **adaptive shallow-fusion bonuses**, yielding better WER and efficiency.



Two-stage: Train hybrid Attention-CTC ASR, then freeze it and train only the biasing decoder on bias-augmented data.

METHOD

- **Biasing Decoder:** Same functionality as a normal ASR decoder, except that it models the (shorter) phrase sequences only.

$$P(p_i|X) = \prod_{t=1}^{L_i} P\left(p_{it}|\{<sos>, p_{i1}, ..., p_{i(t-1)}\}, X\right)$$

- **Phrase-level Log Loss:** All **positive phrases** to have high probability.

$$\mathcal{L}_{log} = -\sum_{i=1}^{M} l_i \cdot log\, P\left(p_i|X\right)$$

- **Per-token Discriminative Loss:** Ensure that **true biasing phrases** are strongly preferred **over distractors**.

$$s_i = log\, P\left(p_i|X\right)/L_i \qquad L_{disc} = -\sum_{i=0}^{M} l_i \cdot log \frac{exp\left(s_i\right)}{\sum_{j=0}^{M} exp\left(s_j\right)}$$

- **Final Biasing Loss**

$$L_{bias}(X, \{p_0, p_1, ..., p_M\}) = (1-\beta)\,\mathcal{L}_{log} + \beta\,\mathcal{L}_{disc}$$

Inference

- **Before search**
  - Perform **beam search** with search-based (shallow-fusion) biasing.
  - **Quick filtering:** Compare each phrase per-token score s_i and with the no-bias score s_0.

$$tol + s_i - s_0 \geq 0$$

- **During search**
  - Attention decoder proposes top expansions for each hypothesis, which are combined with CTC prefix scores.
  - **Dynamic per-token bonus:**

$$bonus = \max_i \{tol + s_i - s_0\}$$

RESULTS & CONCLUSION

Setup & Evaluation: Librispeech biasing setup (960 h)

- N phrases with rare words and randomly sampled distractors
- U-WER (words not in the bias list), and B-WER (words in the bias list)

tol	dev-clean WER(U-/B-WER)	phrases	dev-other WER(U-/B-WER)	phrases
0.0	1.9 (1.6/4.1)	3.2	5.2 (4.8/8.8)	2.8
1.0	**1.8** (1.6/3.7)	5.5	5.2 (4.8/7.8)	5.4
2.0	1.9 (1.7/3.4)	10.1	**5.1** (4.9/7.1)	10.7
3.0	1.9 (1.7/3.1)	17.9	5.2 (5.1/6.7)	20.2
4.0	2.0 (1.9/2.9)	30.2	5.4 (5.3/6.4)	35.1
5.0	2.1 (2.0/2.8)	47.9	5.6 (5.6/5.9)	56.2

TABLE: Sensitivity with respect to tol. (N=1000 β=0.9)

- Small positive tol (=1 or 2) B-WERR without degrading U-WER.
- Model **only** keeps **≈5 -10 active phrases per utterance,** close to the average number of ground-truth phrases (2.1 / 1.6 for dev-clean/dev-other).

Method	test-clean $N=100$	test-clean $N=500$	test-clean $N=1000$	test-clean $N=2000$	test-other $N=100$	test-other $N=500$	test-other $N=1000$	test-other $N=2000$
Attention	5.1 (3.9/14.1)				8.8 (6.6/27.9)			
Biasing with BPB	2.8 (2.3/6.0)	3.2 (2.7/7.0)	3.5 (3.0/7.7)		5.6 (4.9/12.0)	6.3 (5.5/13.5)	7.3 (6.4/15.8)	
Transducer	2.2 (1.3/9.7)				5.2 (3.3/21.8)			
Shallow Fusion Biasing	1.5 (1.2/4.0)	1.6 (1.3/4.2)	1.6 (1.3/4.3)		4.0 (3.3/10.5)	4.1 (3.3/11.1)	4.3 (3.5/11.2)	
Neural biasing, Intermediate layers	1.5 (1.1/4.7)	1.7 (1.2/5.8)	1.9 (1.3/6.6)		3.7 (3.1/9.2)	4.0 (3.2/11.4)	4.4 (3.4/13.5)	
+ text perturbation	1.2 (1.1/2.3)	1.5 (1.2/3.6)	1.7 (1.3/5.1)		3.3 (3.1/5.5)	3.7 (3.2/8.2)	4.2 (3.5/10.8)	
Ours, Attention + CTC (without biasing)	2.7 (1.7/11.1)				6.3 (4.3/23.3)			
Manual bonus (5.0) (no filtering)	**1.9** (1.6/4.4)	**2.0** (1.7/4.6)	**2.0** (1.7/4.6)	2.2 (1.8/4.9)	**4.6** (4.1/9.2)	**4.7** (4.2/9.4)	**4.8** (4.2/9.6)	4.9 (4.4/9.9)
Biasing decoder (tol 0.0)	2.0 (1.6/5.2)	2.1 (1.7/5.2)	2.1 (1.7/5.2)	**2.1** (1.7/5.3)	4.9 (4.2/11.4)	4.9 (4.2/11.4)	5.0 (4.2/11.6)	5.0 (4.3/11.5)
Biasing decoder (tol 2.0)	2.0 (1.6/4.8)	**2.0** (1.7/4.6)	**2.0** (1.7/4.6)	**2.1** (1.8/4.6)	4.7 (4.1/9.6)	**4.7** (4.2/9.4)	**4.8** (4.3/9.3)	5.0 (4.5/9.5)
Biasing decoder filtering (tol 2.0) + manual bonus (5.0)	**1.9** (1.6/4.5)	**2.0** (1.6/4.6)	**2.0** (1.7/4.6)	**2.1** (1.7/4.9)	**4.6** (4.1/9.3)	**4.7** (4.1/9.6)	**4.8** (4.2/9.7)	**4.8** (4.2/9.7)

TABLE: Biasing WERs of different models on Librispeech test sets.

- Setting tol = 2.0 achieves an effect comparable to a manual bonus of 5.0, **but only 1% of original phrases** → faster search.
- >20% relative WERR and >50% relative B-WERR
- Filtering-only (tol = 2.0) + manual bonus 5.0, best at N = 2000 → filtering + use another method for biasing; **avoid over-biasing**.