ICASSP 2024 SLP-L13.6: Improving ASR Contextual Biasing using Guided Attention Loss

Jiyang Tang[1], Kwangyoun Kim[2], Suwon Shon[2], Felix Wu[2], Prashant Sridhar[2]
[1]Carnegie Mellon University, [2]ASAPP Inc.

[1] jiyangta@andrew.cmu.edu

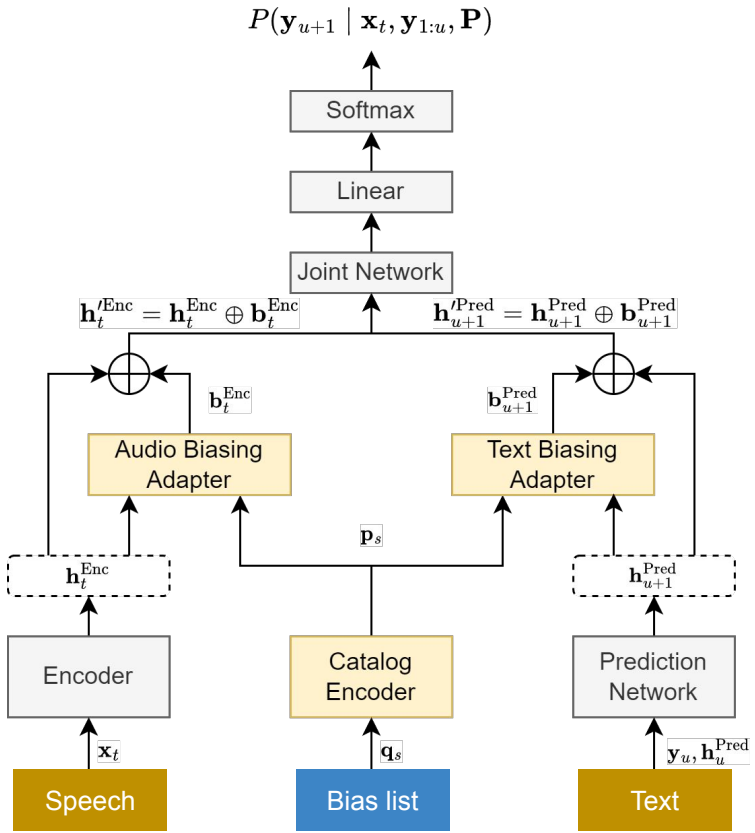
Background: Contextual Biasing (CB)

Adapting ASR systems to recognize custom entities (**bias phrases**)

- Example utterance: “I work at ASAPP WTC”
- “ASAPP” and “WTC” are uncommon/unseen in the training dataset
- Sometimes we have prior knowledge of possible vocabulary to occur
- We give the ASR system a list of **bias phrases** as a reference
  - [“ASAPP”, “WTC”, “ICASSP 2024”, “Gangnam District” …]
- Advantage: the **bias list** is dynamic, so no need to retrain/finetune ASR model for different conversation scenarios

Related Work 1: Contextual Adapter [1]

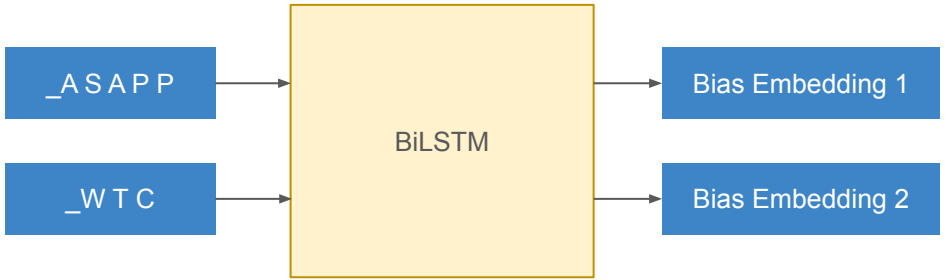
- Based on RNN Transducer (RNN-T)
- **Catalog Encoder** encodes a bias phrase token sequence into a vector representation
- **Biasing Adapter** attends to a bias phrase at each time step using cross attention
- The output of **Biasing Adapter** is added to the encoder output or prediction network output to bias ASR output towards bias phrases

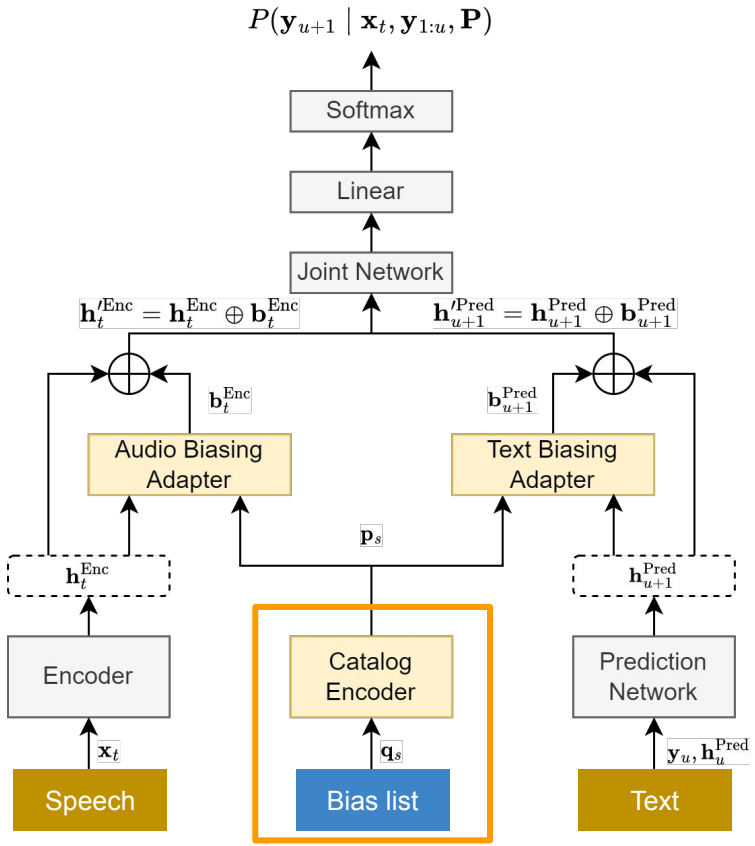


[1] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann, "Contextual adapters for personalized speech recognition in neural transducers," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech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

Related Work 1: Contextual Adapter [1]

Catalog Encoder:

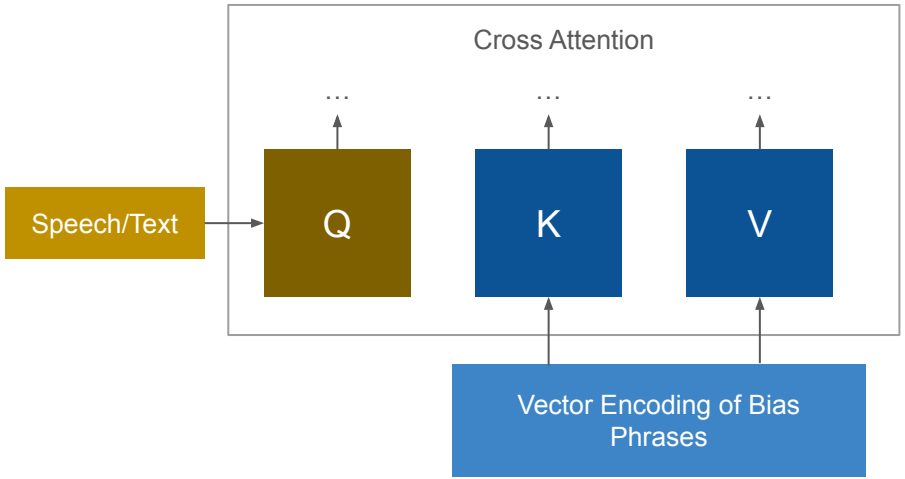


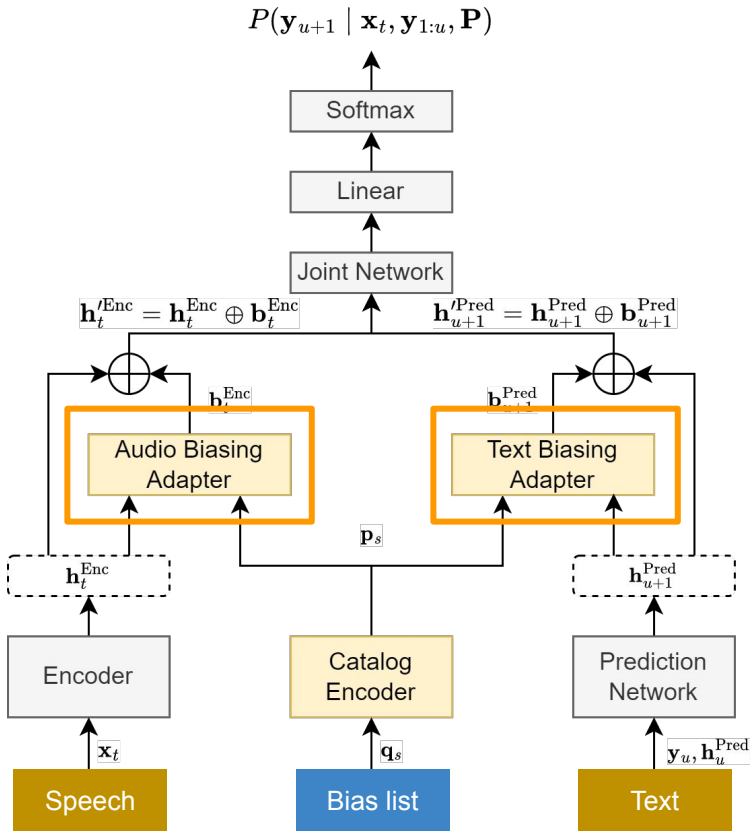


[1] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann, "Contextual adapters for personalized speech recognition in neural transducers," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech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

Related Work 1: Contextual Adapter [1]

Biasing Adapter:

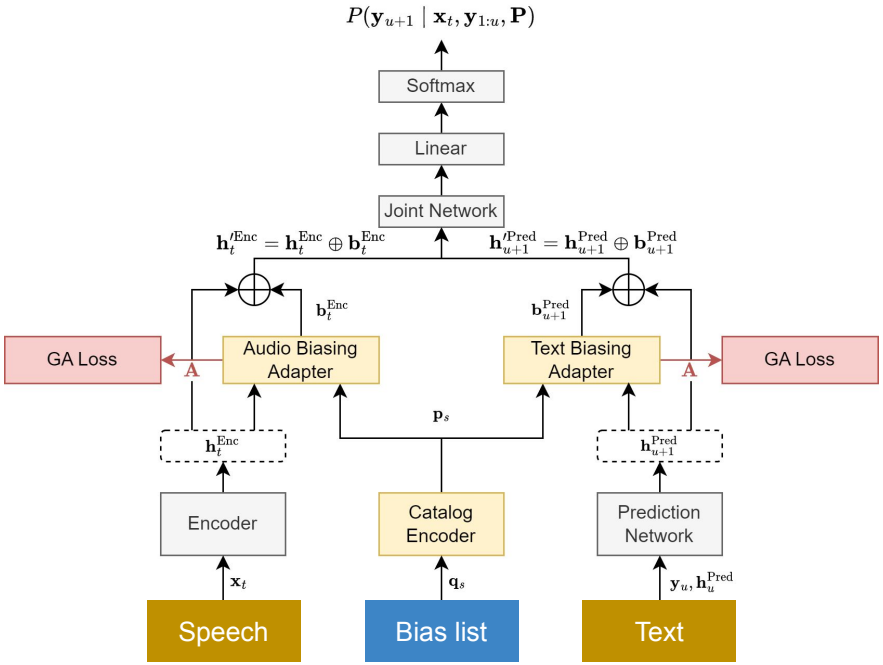




[1] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann, "Contextual adapters for personalized speech recognition in neural transducers," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech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

Related Work 2: Adding Supervision [2]

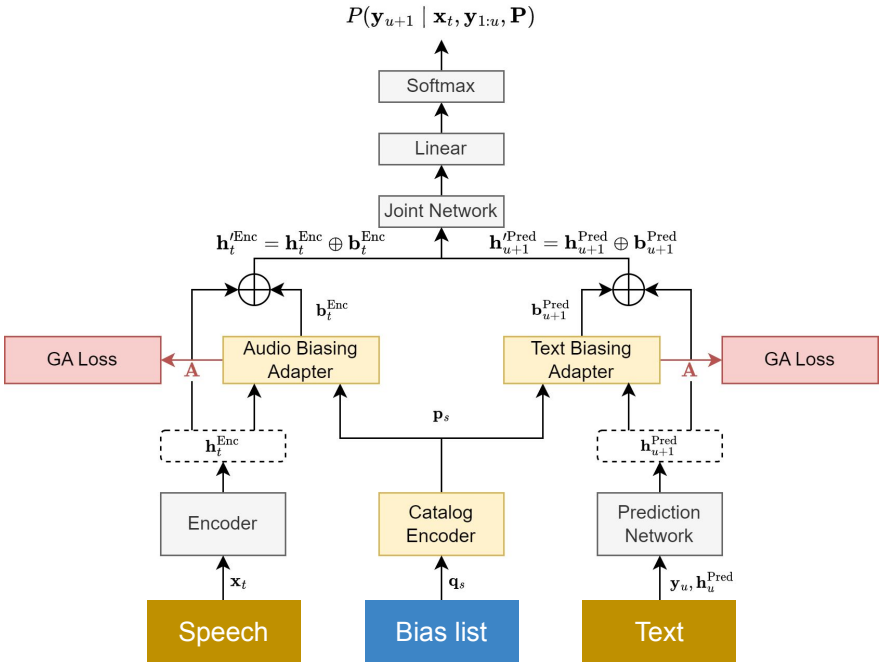
- Idea: force the **Biasing Adaper** to learn which bias phrase to attend to at each time step
- One of the contributions of this paper is using **cross entropy** as an auxiliary loss
- Treating the attention weight A_{ts} as the likelihood of observing phrase *s* at time step *t*, we calculate the cross entropy against the reference phrase sequence
- The auxiliary loss is added to the final loss



[2] Devang Kulshreshtha, Saket Dingliwal, Brady Houston, and Sravan Bodapati, "Multilingual Contextual Adapters To Improve Custom Word Recognition In Low-resource Languages," in Proc. INTERSPEECH 2023, 2023, pp. 3302–3306.

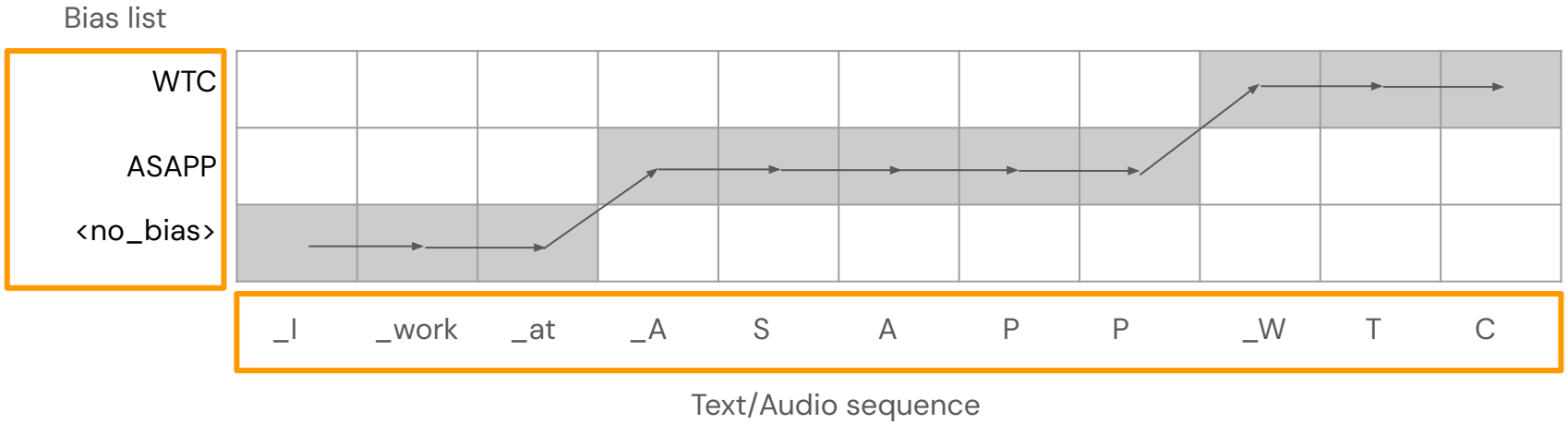
Our Contribution

- The **Guided Attention** loss using cross entropy (**GA-CE**) requires phrase index labels for every text token or audio frame
  - Extra work to generate this, for example force alignment
- We propose using **CTC loss** as the guided attention loss function (**GA-CTC**)
  - Only requires an ordered label sequence of bias phrases that occurred in the utterance



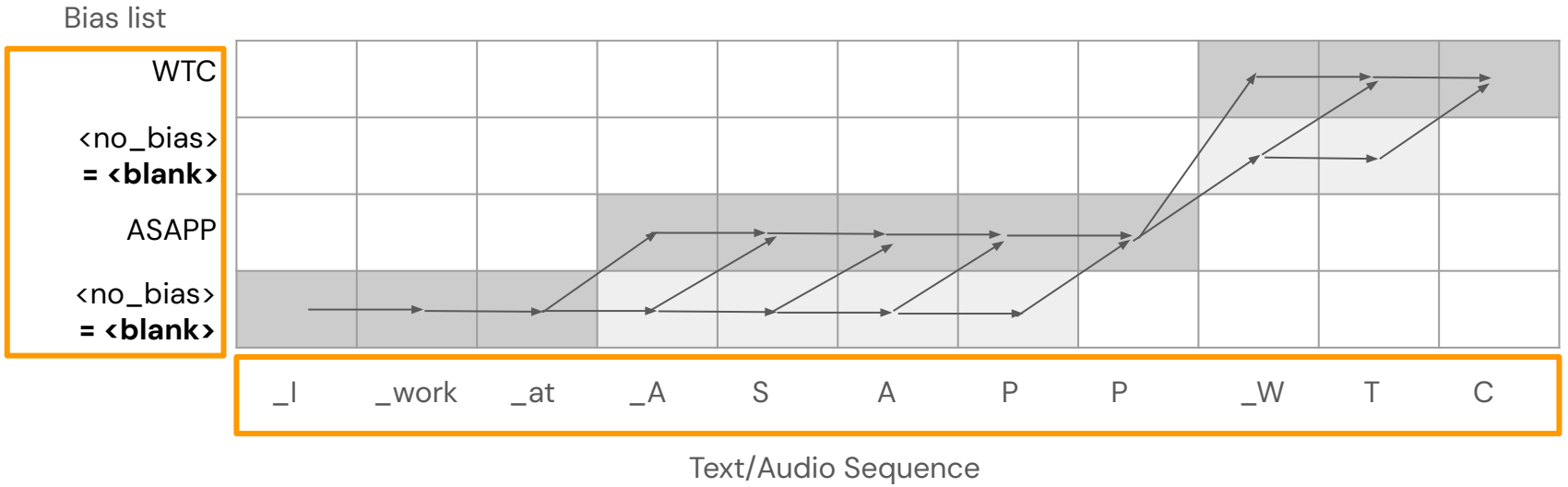
Attention weights: GA-CE

- Label: [0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2]
- Valid “paths” are marked using arrows



Attention weights: GA-CTC

- Label: [1, 2]
- Multiple valid “paths” -> more relaxed supervision



Results on LibriSpeech

B-WER = WER of biased words
Distractors = Unrelated words added to the bias list as distraction. We want to verify whether the Biasing Adapter is able to distinguish and choose the correct phrases

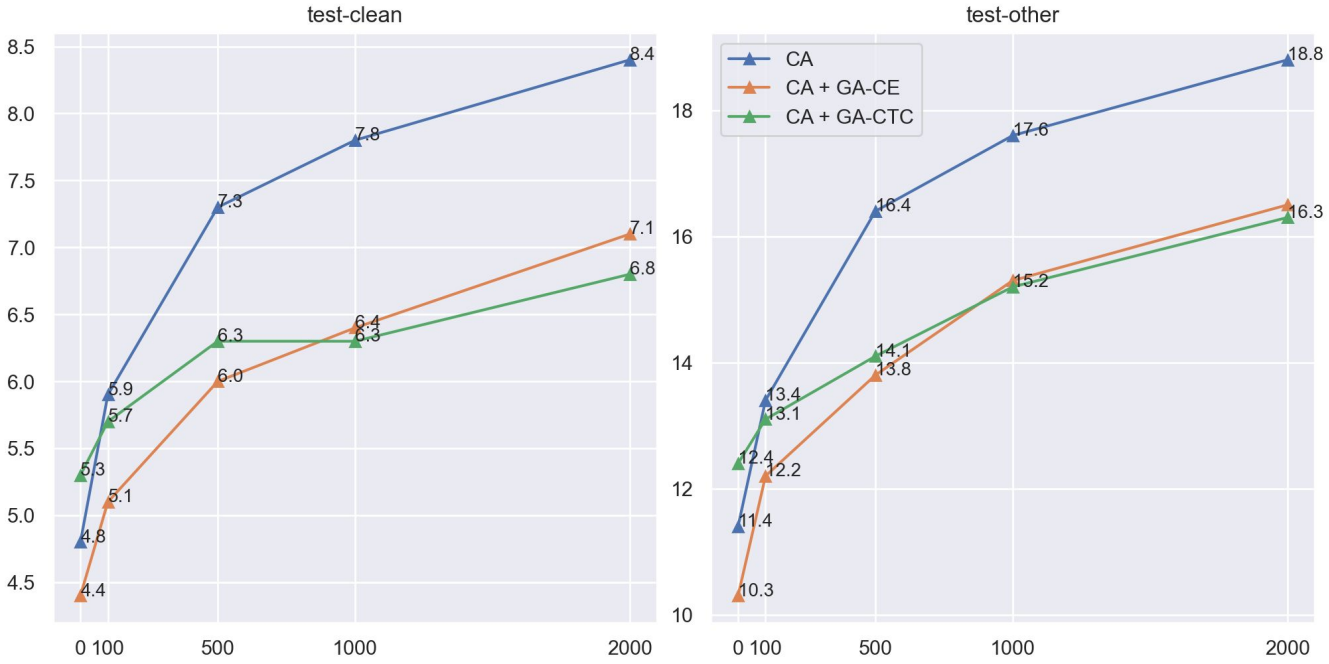


Fig. 2: B-WER of Contextual Adapter (CA) baseline, CA with GA-CE and CA with GA-CTC, using different number of distractors N ∈ {0, 100, 500, 1000, 2000}.

Conclusion

- Explicit supervision (both GA-CE and GA-CTC) can significantly improve performance of contextual biasing
- Proposed GA-CTC is easier to implement in practice while its error rate is on par with GA-CE
    - Especially when many distractors exist (common scenario in application)

Thank you!

Questions?