Implementing contextual biasing in GPU decoder for online ASR

Iuliia Thorbecke (Nigmatulina)

Contextualisation (personalisation)

Goal: to improve recognition of **key entities** when **contextual information** is available.

“Call John Smith mobile.”

“Play Beatles Strawberry fields.”

“But yeah it's scheduled for friday twelve and two.”

“Guten morgen turkish seven alfa whiskey pushback is approved area two.”

Context

- *list of contacts*
- *music playlist*
- *organisation names*
- *dates*
- *street names*

etc.

Contextual information (knowledge) is typically a list of words or word sequences, which are more probable to appear in speech.

Nigmatulina, Iuliia, Srikanth Madikeri, Esaú Villatoro-Tello, Petr Motlíček, Juan Zuluaga-Gomez, Karthik Pandia, and Aravind Ganapathiraju. "Implementing contextual biasing in GPU decoder for online ASR." INTERSPEECH (2023).

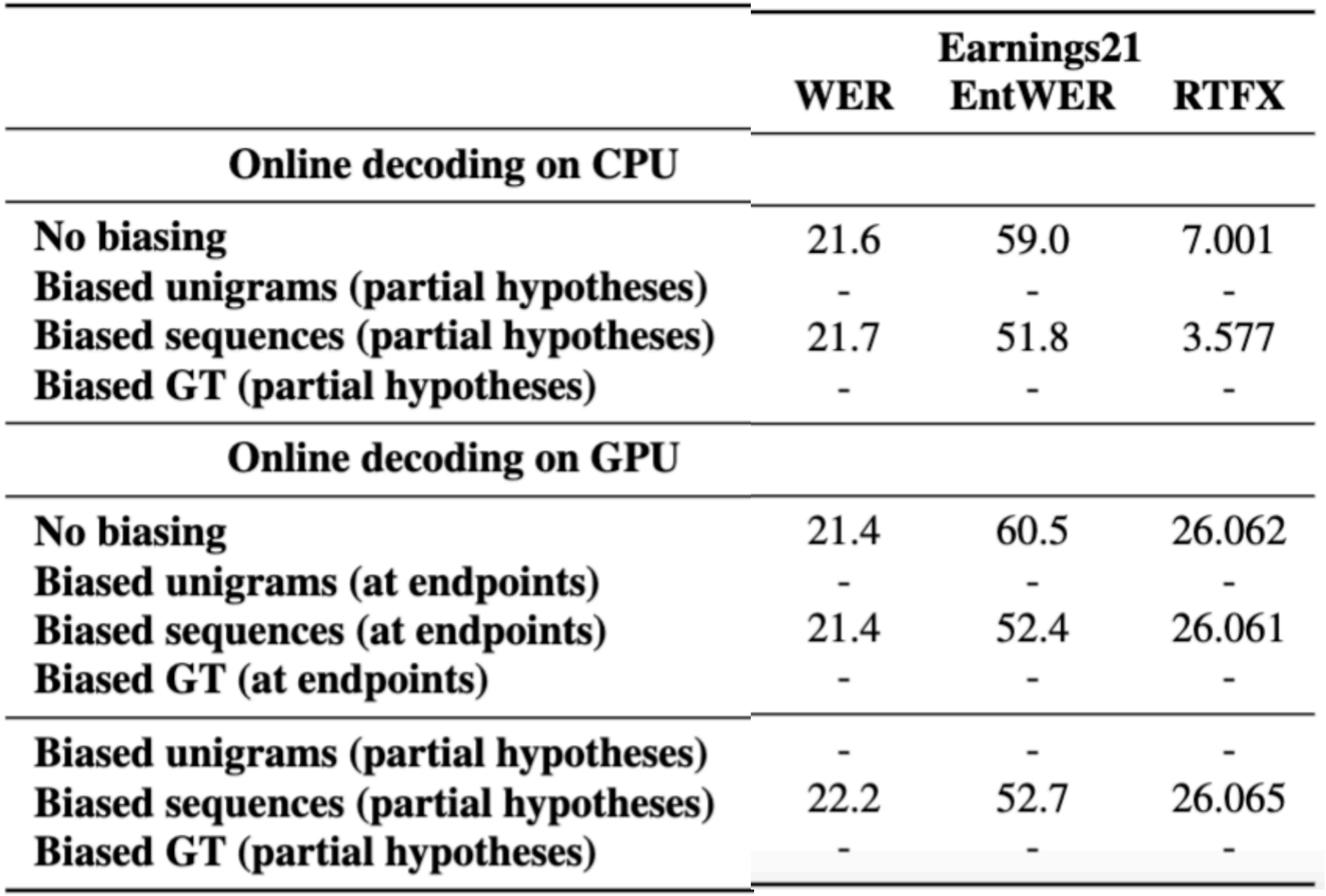
Motivation and Contribution

- **Given**: previous studies on contextualisation (rescoring) for hybrid ASR.
- **Missing**: no rescoring done directly on GPUs.
- **Problem**: rescoring is typically done with lattice composition; in online GPU, no lattices are produced.
- **Goal**: rescoring without lattices.
- **Our main contribution**: an algorithm for rescoring without lattices; the rescoring approach inside Kaldi GPU decoder which is fully integrated into the parallelized decoding process, with no need of lattices.

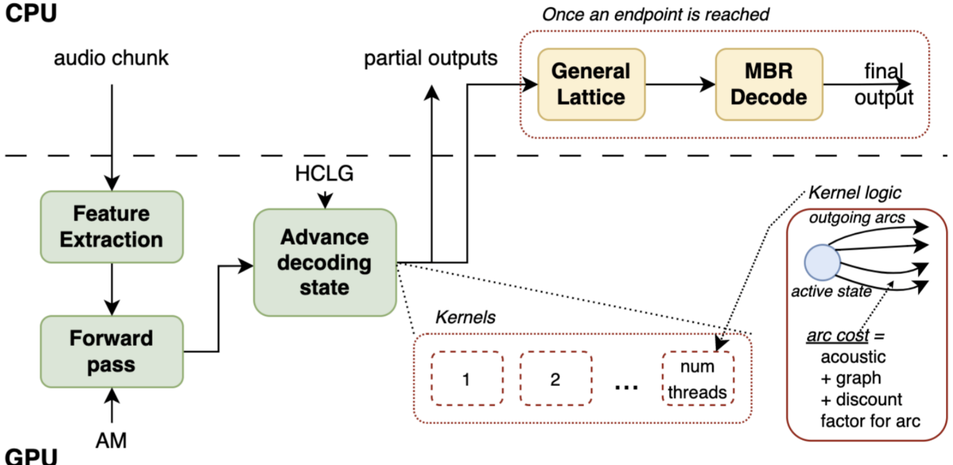
https://github.com/idiap/contextual-biasing-on-gpus

Nigmatulina, Iuliia, Srikanth Madikeri, Esaú Villatoro-Tello, Petr Motliček, Juan Zuluaga-Gomez, Karthik Pandia, and Aravind Ganapathiraju. "Implementing contextual biasing in GPU decoder for online ASR." INTERSPEECH (2023).

Rescoring on GPUs



	Earnings21		
	WER	EntWER	RTFX
Online decoding on CPU			
No biasing	21.6	59.0	7.001
Biased unigrams (partial hypotheses)	-	-	-
Biased sequences (partial hypotheses)	21.7	51.8	3.577
Biased GT (partial hypotheses)	-	-	-
Online decoding on GPU			
No biasing	21.4	60.5	26.062
Biased unigrams (at endpoints)	-	-	-
Biased sequences (at endpoints)	21.4	52.4	26.061
Biased GT (at endpoints)	-	-	-
Biased unigrams (partial hypotheses)	-	-	-
Biased sequences (partial hypotheses)	22.2	52.7	26.065
Biased GT (partial hypotheses)	-	-	-



Decoding a chunk of audio

Nigmatulina, Iuliia, Srikanth Madikeri, Esaú Villatoro-Tello, Petr Motlíček, Juan Zuluaga-Gomez, Karthik Pandia, and Aravind Ganapathiraju. "Implementing contextual biasing in GPU decoder for online ASR." INTERSPEECH (2023).