Contextual Speech Recognition with Difficult Negative Training Examples

Uri Alon, Golan Pundak, Tara N. Sainath

TECHNION Israel Institute of Technology   Google

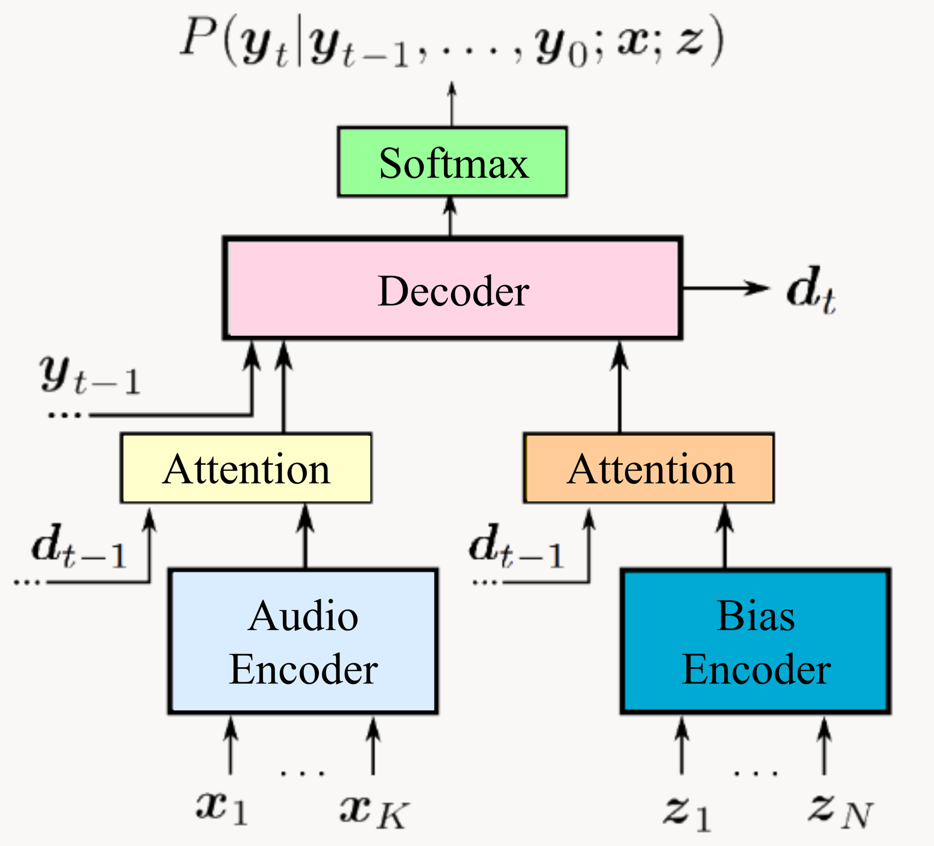
urialon@cs.technion.ac.il   {golan,tsainath}@google.com

1. Task: Contextualized ASR

- **Context** provided in addition to audio can help reduce WER significantly.
- Such user-specific contextual information can include:
  - The user's list of songs
  - The user's contact list
  - The currently installed apps
- **Proper nouns** are very frequent in various ASR tasks:
  - "Call *Joan*'s mobile"
  - "Play *Taylor Swift*"
  - "How tall is *LeBron James*?"
- But contextual ASR models usually perform poorly on **rare words** and especially on **proper nouns** (NNPs).

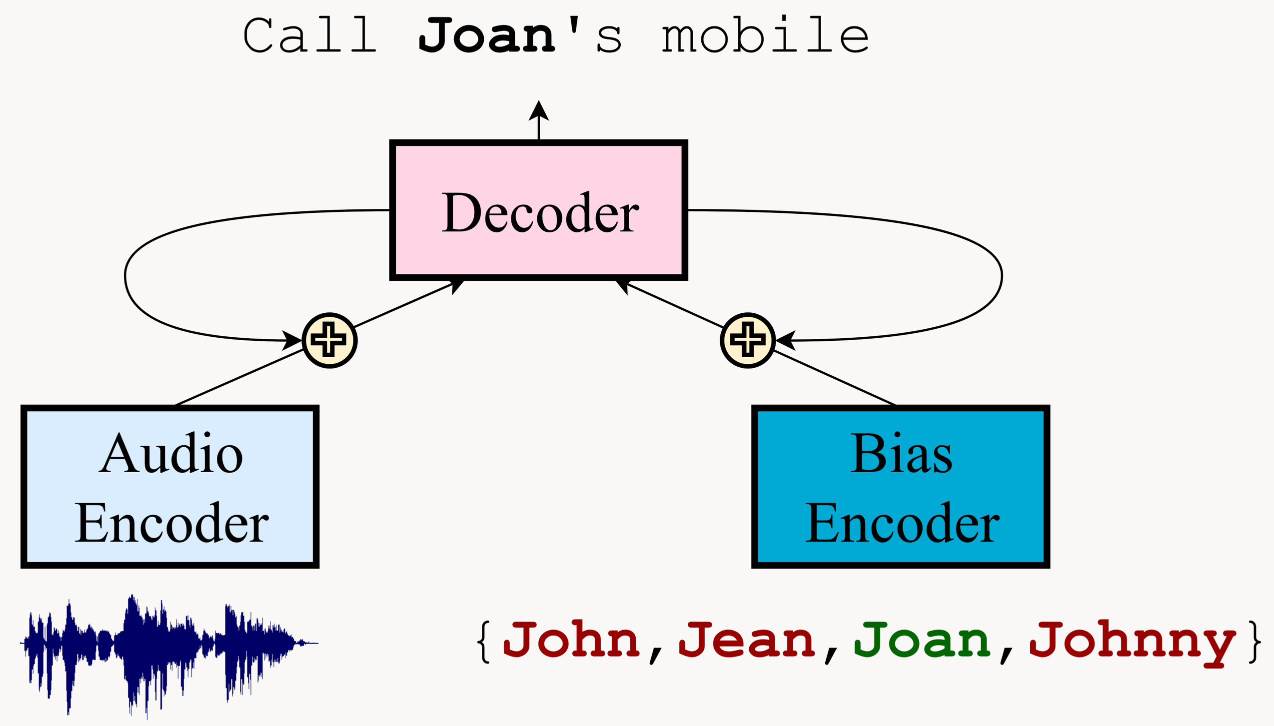
2. The Contextualized LAS (CLAS) Model (Pundak et al., SLT'2018)

- **CLAS** is an E2E ASR model based on the **Listen-Attend-and-Spell (LAS)** encoder-decoder architecture.
- The key difference from LAS: **biasing** sub-module.



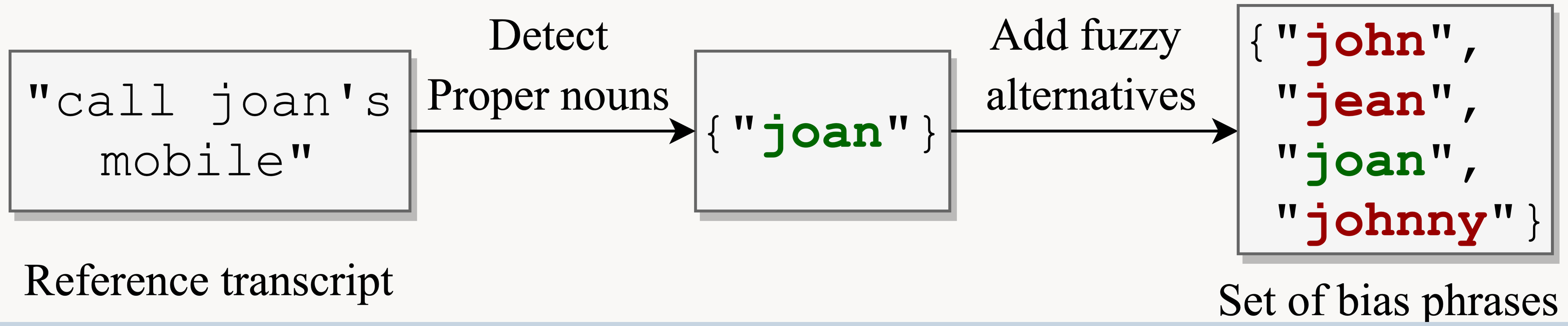
3. The Problem: The Network Fails to Distinguish Between Phonetically Similar Phrases

- Disambiguation of similarly sounding phrases is challenging.
- The network makes even more mistakes as the set of bias phrases becomes larger.



4. Training with Difficult Negative Examples

- During **training**, we provide the network with **phonetically similar proper nouns** (NNPs) as the "distractors".
- This way, we encourage the network to:
  - **Distinguish** between similarly sounding phrases
  - Learn more **discriminative representations**.



5. Evaluation

- We experimented with the following training schemes:

	Vanilla CLAS	CLAS+NNP	CLAS+fuzzy	CLAS NNP+fuzzy
Bias Phrases Selection	Random	**NNPs from reference**	Random n-grams from reference	**NNPs from reference**
Distractors Selection	Random	Random NNPs	**Fuzzy alternatives**	**Fuzzy alternatives**

- Results:

Test Set	Vanilla CLAS	CLAS+NNP	CLAS+fuzzy	CLAS NNP+fuzzy
Songs	9.8	6.7 (31.6%)	10.4	**5.4** (44.9%)
Contacts	11.3	6.1 (46.0%)	16.5	**5.3** (53.1%)
Talk-To	15.2	14.8 (2.6%)	**11.1** (27.0%)	11.3 (25.7%)

Table: WER of the compared training schemes. In parentheses: the relative improvement over Vanilla CLAS.

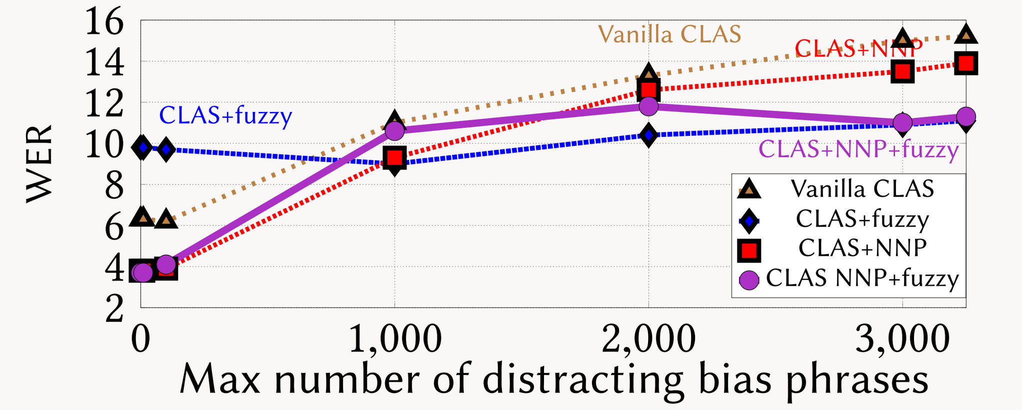


Figure: CLAS NNP+fuzzy achieves the lowest WER with a small set of bias phrases, and almost the lowest WER when presented with 3255 bias phrases.

6. Qualitative Analysis

True ref: `creepy carrots</bias>`

Fuzzy: **creepy** carrots</bias>   Non-fuzzy: **sleepy** carrots</bias>

Figure: The fuzzy model attends mostly to "creepy carrots" and makes a correct prediction, while the non-fuzzy model attends to "sleepy carrots" and predicts the wrong word "sleepy".