CLIN 35
Leuven 2025

KU LEUVEN

Department Electrical Engineering ESAT-PSI
KU Leuven University, Belgium

Explicit Pronunciation Modelling in LLM-based Speech Recognition

Jakob Poncelet, Hugo Van hamme

Topic

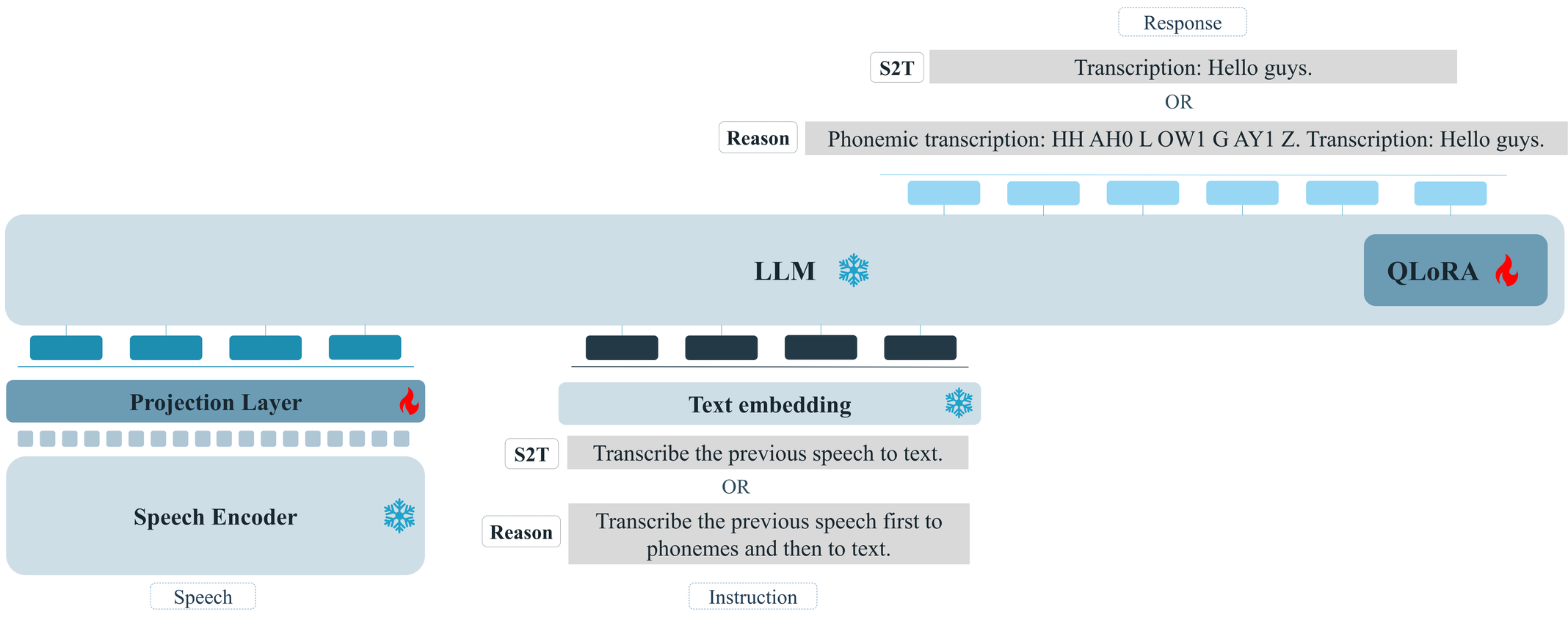
Multimodal LLMs are emerging that can process non-text inputs, such as images or audio, by incorporating modality-specific encoders. LLMs equipped with audio encoders provide **speech-aware LLMs** that can process both spoken and written inputs (Voxtral, Phi4-MM, Gemma-3n, …). Such systems can harness the contextual language abilities of the LLM to bridge the gap between speech recognition, spoken language understanding and textual reasoning.

However, finding a good **alignment** between pre-trained speech models and LLMs is difficult and requires many examples of speech inputs:
- speech encoders provide acoustic representations related to pronunciation, e.g. *bank – rank – dank*
- LLMs expect embeddings that are more related to concepts and semantics, e.g. *bank – economy – money*

Typically, alignment consists of training an **adapter** to map the speech features of a pre-trained speech encoder into the embedding space of the LLM. This mapping is learned through training with a **speech-to-text task**, where the LLM is prompted to transcribe the projected speech features to text.

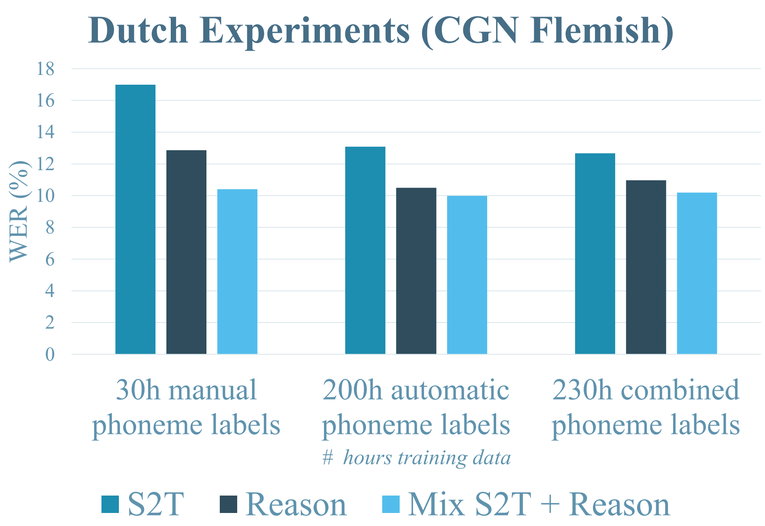
Method

We propose an approach that promotes **pronunciation modelling** in the LLM to improve speech recognition, by teaching the LLM to predict **phonemes** in addition to words. These phonemes can be automatically generated from labeled speech with a forced aligner and a pronunciation dictionary.
→ By **first** predicting the phonemes and then using them to ***reason*** over which words were actually spoken, we reduce the acoustic confusion in the LLM and attain a more fine-grained interpretation of acoustic features. Additionally, the predicted phonemes can be used to improve the **explainability** of the system and provide pronunciation feedback.

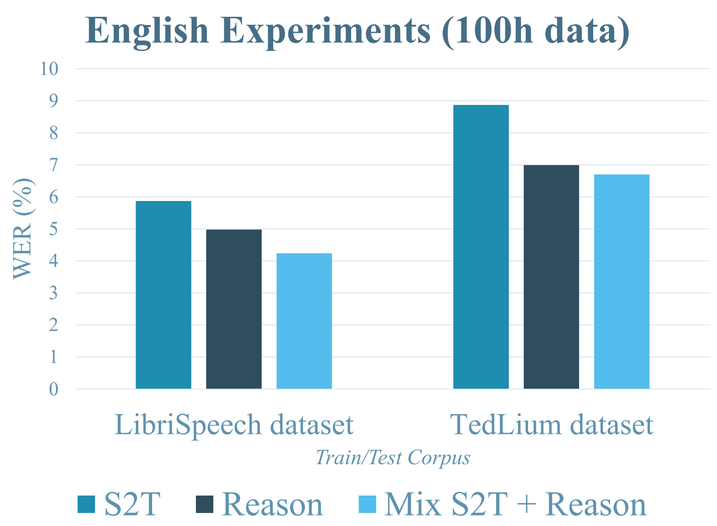


Results

Finetuning the LLM and adapter with the ***reasoning*** objective improves Word Error Rates of the predicted transcripts. Randomly **mixing S2T and reasoning** prompts helps against overfitting (i.e. overfitting on the reference phoneme labels during text prediction).



Encoder = NeLF (KUL) ASR model
LLM = 4-bit Tweety-7B



Encoder = Whisper-medium
LLM = 4-bit Llama-3.1-8B

The predicted transcripts make are **acoustically closer** to the actual transcripts, when evaluating G2P-phonemized versions of the transcripts.

Objective	WER reduction	Phonemized ER reduction
S2T	(5.9)	(4.3)
Reason	**- 15 %**	**- 37 %**
Mix S2T + Reason	**- 27 %**	**- 44 %**

Example Transcripts	
Reference	**Rhein** and **Moselle** eighteen ninety-three
Hyp. S2T	**Romeo** and **Juliet** eighteen ninety-three
Hyp. Reason	**R AY1 AH0 N** AE1 N D **M OW0 S EH1 L** ... **Ryan** and **Mozel** eighteen ninety-three
Reference	Look at their **bellies** pink they're feasting
Hyp. S2T	Look at their **ballots** pink they're feasting
Hyp. Reason	L UH K AE T DH EH R **B AE L IY Z** ... Look at their **bellies** pink they're feasting