SIMULTANEOUS SPEECH TRANSLATION
Leveraging Neural Models for Efficient Cross-Language Communication

Sara Papi, MT Unit (FBK) - email: spapi@fbk.eu

ABSTRACT

Simultaneous speech translation (SimulST) is the process of real-time translation of spoken language into another language, enabling cross-language communication. This technology has become increasingly popular in recent years, with the development of advanced models which, however, still face the main challenge of providing accurate but fast translations. My research aims to fill this gap, by proposing new strategies that leverage the knowledge already acquired by neural models to reduce their latency.

PROBLEM

Neural models for SimulST usually require:

- **Specific architectures** that introduce additional modules to be optimized;

- **Long and complicated training procedures** (e.g., different optimization objectives);

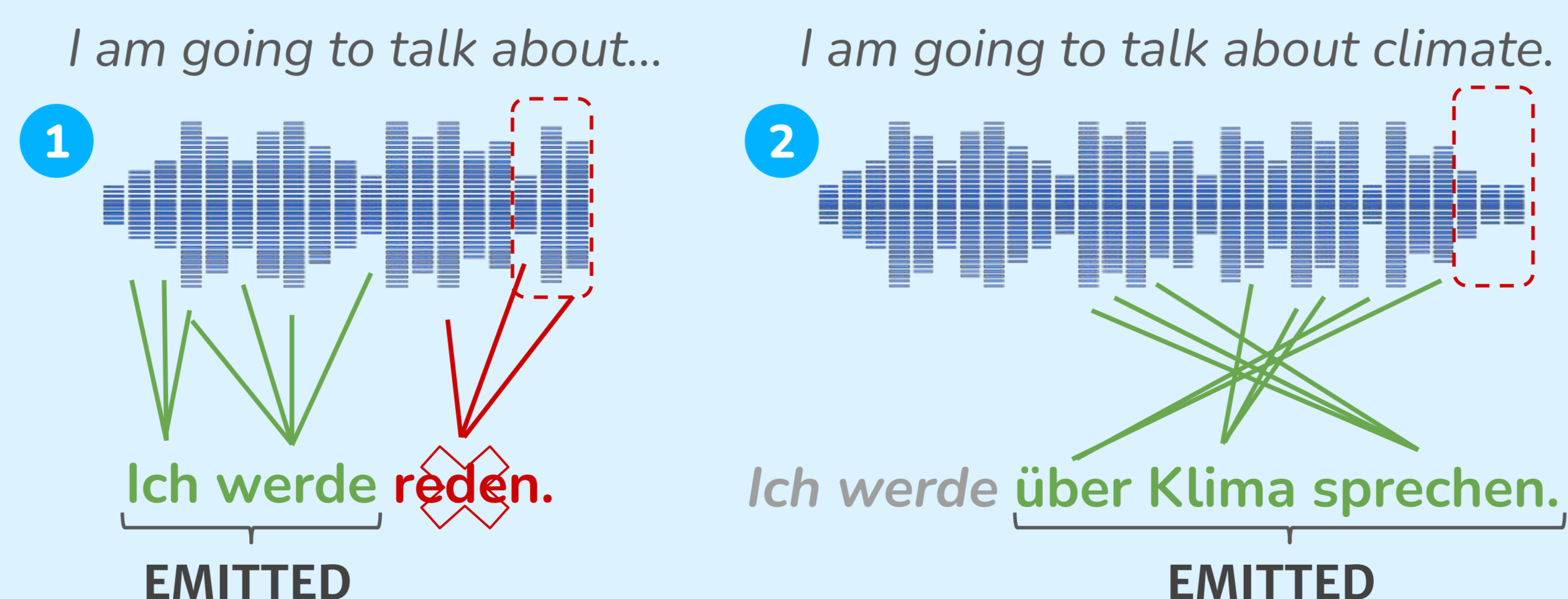- Training and maintaining several models to reach **different latency regimes** (e.g., *1s*, *2s*, ...).
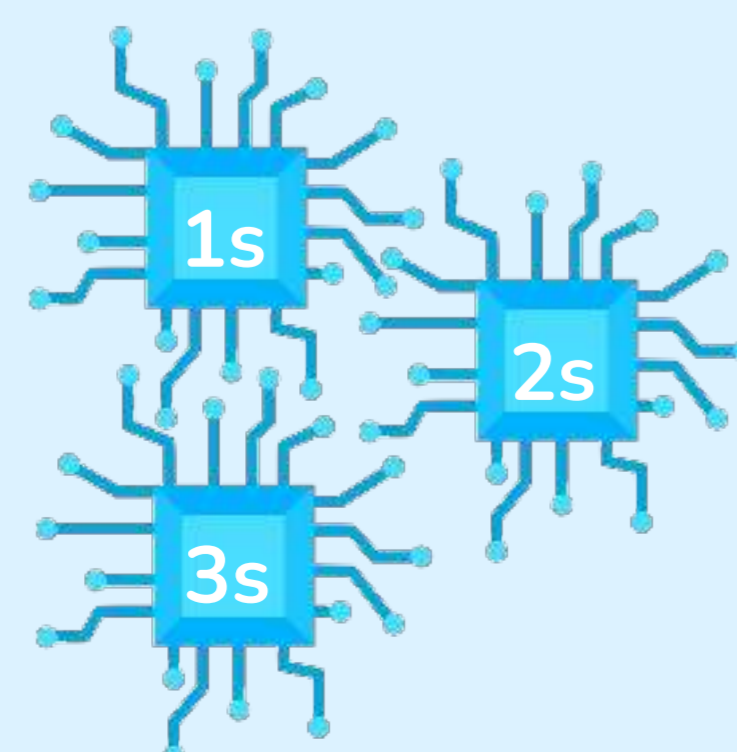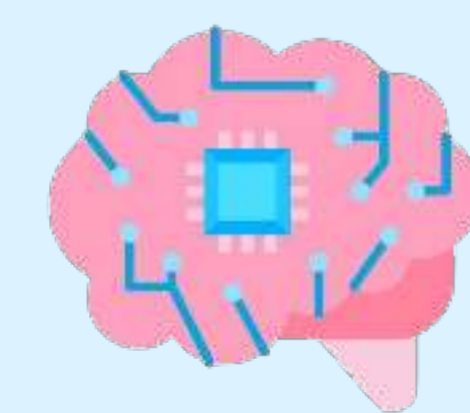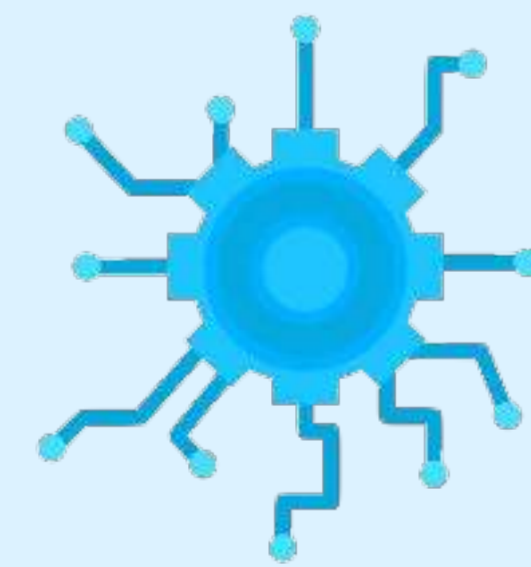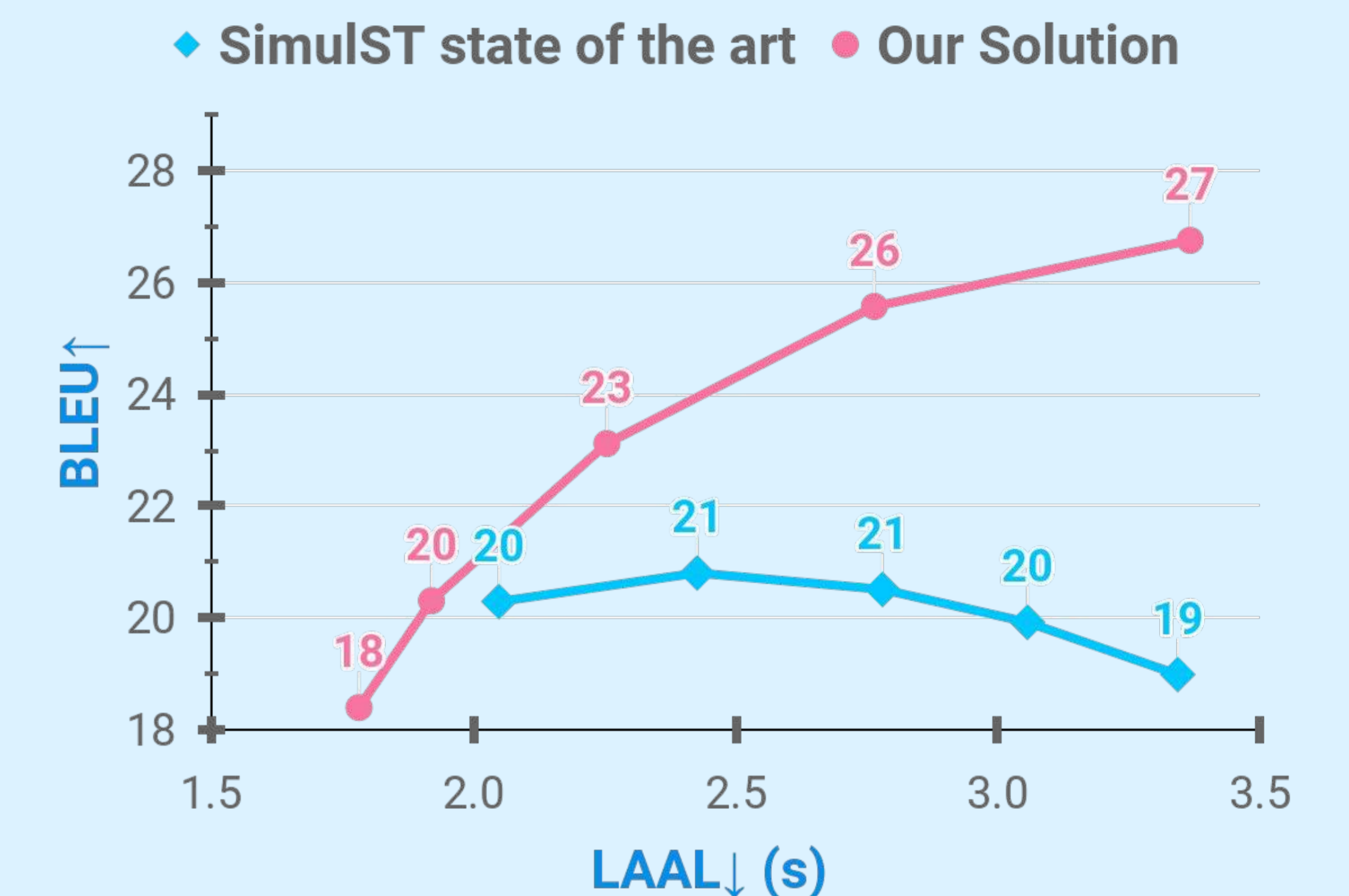
SOLUTION

1. Use **already existing neural models** trained for offline speech translation without re-training or adopting specific architecture for SimulST;

2. Use **only one model** for every latency regime and handle latency through specific parameters;

3. Leverage **translation knowledge** already acquired by the model through the **attention mechanism** between audio input and textual output:

- I
- am
- a
- student.
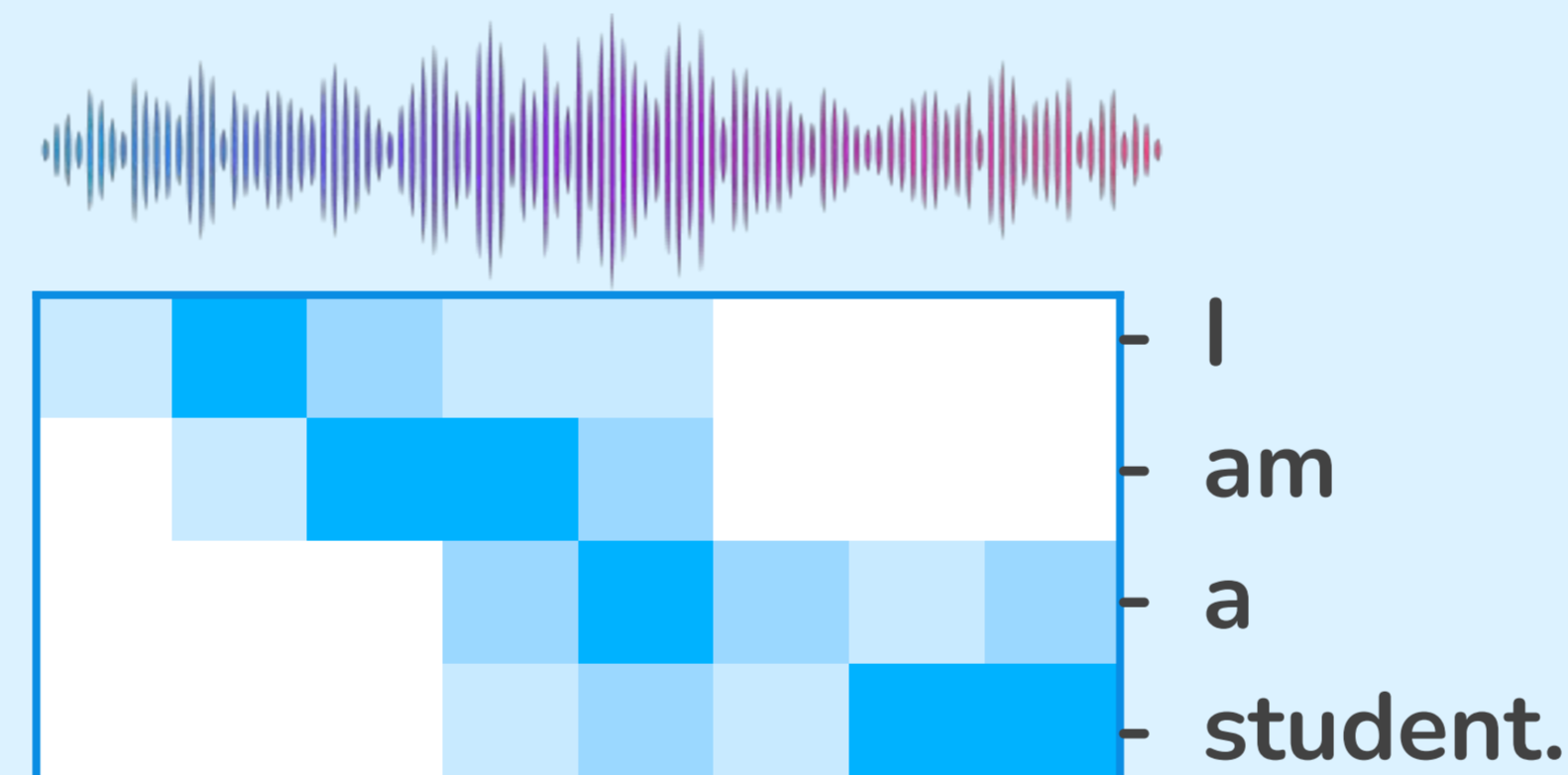
4. Decide whether to **emit or not a partial translation** based on where attention points to: a word is emitted if the attention is not concentrated towards the end, meaning that the received information is enough stable.

I am going to talk about...

Ich werde ~~reden.~~

EMITTED

I am going to talk about climate.

Ich werde über Klima sprechen.

EMITTED

PERFORMANCE

We achieve **higher quality (BLEU) with lower latency (LAAL)** compared to the state-of-the-art SimulST solution (a model specifically trained for the task).

REFERENCE PAPERS

1. Papi et al., 2022. *Does simultaneous speech translation need simultaneous models?*. In "Findings of EMNLP 2022".
2. Papi et al., 2022. *Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation*. In "Proceedings of the Third Workshop on Automatic Simultaneous Translation".
3. Papi et al., 2023. *Attention as a Guide for Simultaneous Speech Translation*. Submitted to ACL 2023.