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, ...).

**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.

**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:
   - 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.

**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**