Thesis Title: "*Direct Speech Translation in Constrained Contexts: the Simultaneous and Subtitling Scenarios*"

Sara Papi
University of Trento and Fondazione Bruno Kessler
spapi@fbk.eu

ADVISORS:

- Marco Turchi (Zoom Video Communications)
- Matteo Negri (Fondazione Bruno Kessler)

EXTERNAL REVIEWERS:

- Claudio Fantinuoli (University of Mainz and KUDO Inc.)
- Juan Pino (Meta AI)

SUMMARY:

1 Motivation

The shift to online communications in various sectors like business, education, and entertainment has highlighted the need for effective language translation to enable seamless interaction among users with diverse linguistic and accessibility needs. Speech-to-text translation (ST) emerges as a core technology for overcoming language barriers and facilitating communication by converting spoken words into another language, offering a natural understanding of language. However, developing ST systems is challenging due to the inherent complexities of speech, such as variations in accents, speaking rates, and background noise. These challenges are further complicated by constraints such as time (e.g., output latency), space (e.g., characters to be displayed on the screen), computational resources (e.g., using CPUs or GPUs), or limited data availability (e.g., low resource languages).

2 Research Questions

The objective of ST is to achieve the highest quality of automatic textual translations. However, many applications require more than just high translation quality. When additional constraints are present, the challenge becomes balancing translation quality with these specific requirements. This PhD thesis focuses on two constrained scenarios: simultaneous speech translation and automatic subtitling. Both tasks are of significant scientific and industrial interest.

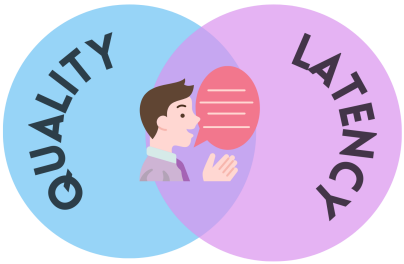

Figure 1: Simultaneous Translation Constraints.

2.1 Simultaneous Translation

Simultaneous Speech Translation (SimulST) aims to minimize latency–the delay from when an utterance is spoken in the source language to when it is translated into the target language. This requires translations to be displayed promptly and aligned with the natural pace of speech. Balancing translation quality and latency is essential for user comprehension and experience (Figure 1). Current SimulST systems face challenges in achieving this balance and often require complex training procedures with multiple training stages and sometimes the need to develop several models for different latency requirements. This PhD research explored whether existing ST systems possess intrinsic knowledge that can be leveraged for real-time applications without complex, ad-hoc training procedures. The main research questions were:

- Are complex training procedures necessary for SimulST?
- Can the knowledge acquired by standard ST models be used to guide them during simultaneous inference?

Proceedings of Machine Translation Summit XX Volume 1, pages 2–3, June 23-27, 2025

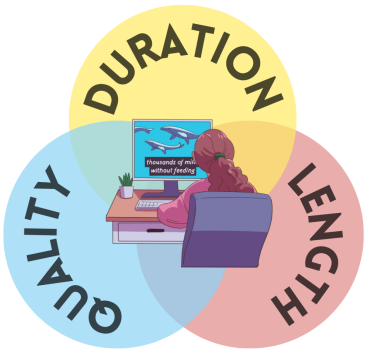

Figure 2: Automatic Subtitling Constraints.

2.2 Automatic Subtitling

Automatic Subtitling translates spoken dialogue in audiovisual media into text, which has to conform to spatial constraints (subtitle length) and temporal constraints (synchronization with audiovisual content). Long subtitles may overwhelm viewers, while short ones risk losing information; thus, proper subtitle length and synchronization ensure they remain on screen long enough to be read without disrupting the video's flow (Figure 2). In this scenario, prosody and speech cues are crucial for subtitle segmentation and timing, but current cascade architectures lose this information. Therefore, this PhD research aimed to leverage direct models that have direct access to this information, addressing two key questions:

- Is there a way to exploit prosody and speech cues to build automatic subtitling datasets starting from already existing ST corpora, overcoming data scarcity?
- Can a direct ST model produce fully segmented and timed subtitles?

3 Contibutions

3.1 Simultaneous Translation

In SimulST, the goal was to assess if standard ST systems could be repurposed for real-time use by leveraging their intrinsic knowledge, advocating a paradigm shift in model development. The contributions can be summarized in the findings below:

- Standard ST systems used for simultaneous inference achieve competitive or superior quality and latency compared to those ad-hoc trained for the tasks (Papi et al., 2022a);
- Intrinsic knowledge, particularly cross-attention, can be effectively used for SimulST, resulting in low-latency translation with minimal computational costs (Papi et al., 2023b);
- Using cross-attention for aligning speech and translation to guide simultaneous inference achieves an optimal balance between quality and latency (Papi et al., 2023c).

3.2 Automatic Subtitling

In Automatic Subtitling, the goal was to use direct systems, able to exploit speech cues, for subtitle segmentation and to generate complete subtitles. Specifically, the main findings are:

- To cope with data scarcity, direct multilingual multimodal models, which utilize both audio and textual cues to identify optimal segmentation points, revealed their effectiveness in automatic subtitle segmentation, delivering performance comparable to gold segmentation (Papi et al., 2022b);
- Direct ST models demonstrate the capability of generating full subtitles, which consist of segmented translations with corresponding timestamps, showing competitive performance against existing production tools (Papi et al., 2023a).

References

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022a. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates.

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022b. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 480–487, Online only.

Sara Papi, Matteo Negri, and Marco Turchi. 2023b. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada.

Sara Papi, Marco Turchi, and Matteo Negri. 2023c. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Interspeech 2023*, pages 3974–3978.