Direct Speech Translation
Toward High-Quality, Inclusive, and Augmented Systems

Marco Gaido
Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
University of Trento, Trento, Italy
Email: mgaido@fbk.eu, Phone: +39 3482670470
Supervisors: Marco Turchi, Matteo Negri
marco.turchi@zoom.us, negri@fbk.eu

When this PhD started, in November 2019, the translation of speech into text in a different language was mainly tackled with a cascade of automatic speech recognition (ASR) and machine translation (MT) models. However, a new paradigm was emerging, with the proposal of direct (or end-to-end) models designed to tackle the speech-to-text translation (ST) task in a single step. At that time, the main question within the ST community was: *will direct ST models be able to keep their promise and reach (or even outperform) the quality of cascade approaches?* Therefore, the initial phase of the PhD has been dedicated to building **high-quality** direct models, specifically under the practical scenario where lengthy audio files necessitate automated segmentation. The positive outcomes attained in terms of overall translation quality enabled the study of specific aspects of direct systems that are pivotal for meeting the real needs of end-users. Consequently, a significant portion of the PhD has been dedicated to analyzing and improving their behavior concerning two critical aspects: **inclusivity** (in terms of gender bias) and **augmented translation** (the integration of useful concepts and contextual information to help users' understanding). Below, I summarize the work I carried out on the above lines of research, and the related findings and achievements.

Translation Quality. Through the continuous experimentation of new techniques compared with the state of the art and evaluated in the challenging yearly international IWSLT evaluation campaign for speech translation, I contributed to closing the gap between the two paradigms, as attested by the first success of a direct system in the competition in 2020 (where the FBK model ranked 2nd, first among academic participants) and a through manual analysis carried out to compare the solutions (**ACL 2021**). Specifically, on one side I introduced training procedures and architectural solutions aimed at improving the translation quality of direct ST systems and their efficiency, reducing computational costs. On the other, I focused on how to limit the quality drops observed when the audio is not segmented according to a known reference but has to be automatically segmented into chunks processable by ST models.

As part of the first group of activities, I studied the best methods to transfer knowledge from an MT model into a direct ST system with knowledge distillation, highlighting not only the benefits but also its limitations, for which I provided an easy yet effective solution (**IWSLT 2020**). I also proposed a compression mechanism that leverages the prediction of a CTC module and dynamically reduces the length of the input sequence in the encoder of ST systems, improving both translation quality and computational efficiency (**EACL 2021**). Building on the CTC-compression module, I introduced Speechformer, the first architecture for direct ST that, enabled by an attention implementation with reduced computational complexity, avoids any fixed compression of the audio input, respecting the variability of the amount of information in speech signals and bringing significant quality gains (**EMNLP 2021**). Lastly, I showed the superfluity of the ASR pre-training when using an auxiliary CTC loss and the effectiveness of a simple data filtering procedure based on the transcript-to-translation character ratio (**IWSLT 2022**).

Moving to the second goal of coping with sub-optimal audio segmentation, I increased the robustness of direct ST models with regard to au-

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 2–3
June 24-27, 2024 ©2024 European Association for Machine Translation

tomatic segmentation of the audio by fine-tuning them on resegmented training corpora and by providing the previous audio segment as contextual information (**Interspeech 2020**). Moreover, I proposed a new hybrid segmentation method that limits the quality degradation with respect to optimal segmentation based on the transcripts, which are unknown at inference time (**ICNLSP 2021**).

Inclusivity. Reckoning that a high overall quality is not enough to consider a technology ready for the users and driven by the ethical commitment and deep belief in the importance of raising awareness of the limitations – and even potential harms – of automatically-generated text in contemporary society, I devoted part of my PhD to studying the gender bias of direct ST systems. The goal was to ensure the fairness of automatic systems and equal opportunities for different groups of users to benefit from them. In this context, I disclosed how the pursuit of higher general performance can exacerbate gender representational disparities and proposed mitigation techniques that reduce the gender bias of ST models. To this aim, I explored different solutions to control the grammatical gender of words referred to the speaker (assuming that the gender of the speaker is known in advance), investigating for the first time the case in which the speakers' gender conflicts with their vocal characteristics (**COLING 2020 Oustanding Paper**). In this context, I proposed automatic metrics tailored at disentangling the gender bias of a system from its overall quality, which has been validated through an extensive manual analysis, which also showed that ST models are nearly perfect in handling gender agreement and that the most biased part of speech is nouns (**ACL 2022**). Then, I unveiled the exacerbation of gender bias caused by a BPE segmentation of the target text in comparison with a character-based segmentation, and the proposal of a solution that goes beyond the trade-off between translation quality – BPE – and gender accuracy – char – (**ACL-Findings 2021**). Lastly, I demonstrated the increase in gender bias caused by distilling knowledge from MT and how to solve the issue with a simple fine-tuning (**CLiC-it 2020 Best Paper, IJCoL 2022**).

Augmented Translation. At last, motivated by the practical needs of interpreters and translators, my PhD evaluated the potential of direct ST systems in the "augmented translation" scenario, where the translation is enriched with contextual information that eases its fruition. In particular, within the Smarter Interpreting[1] research project – aimed at the creation of to a new generation of computer-assisted interpreting (CAI) tools – the main focus was the translation and recognition of named entities (NEs), which constitute one of the most demanding challenges for interpreters. This strand of research activities started with the creation of a new benchmark (NEurRoparl-ST), used to assess the similar weaknesses of cascade and direct ST systems when it comes to NEs (**EMNLP 2021**). Having ascertained that person names are the most complex NE type for ST systems, I isolated the factors that contribute to this difficulty of ST systems (low frequency in the training data, names associated with languages not included in the source side of the training set) and proposed the adoption of multilingual models that jointly predict the transcript and the translation (giving more weight to the transcription) to mitigate such errors (**IWSLT 2023 Best Paper**). Moreover, in cases in which a dictionary of entities likely to appear in a given domain is available (a frequent condition in the interpreting sector), I showed that the accuracy of NEs (especially of person names) can be significantly improved by means of additional modules that first recognize which of them are present and then inject the corresponding translations as suggestions while generating the output (**ICASSP 2023**). The project was concluded by the introduction of models that jointly perform ST and NER, outperforming a pipeline of ST and NER systems while keeping the computational cost as low as that of a single direct ST model (**Interspeech 2023**).

Besides automatic evaluations on the proposed benchmark, the effectiveness of our solutions has been proved in two demos, carried out in April and December 2022, in which our joint ST and NER systems have been integrated into a new CAI tool that displays the translated NEs and domain-specific terminology in real time to the interpreter. In the first demo, students and professionals of the University of La Laguna performed a human-centric evaluation to assess the usefulness of the system for interpreters. The positive feedback of this analysis led to presenting the tool at international interpreting conferences,[2] where it has been introduced as the first 4th generation CAI system.

[1] `https://smarter-interpreting.eu/` – financed by CDTI Neotec funds.

[2] `https://ctn.hkbu.edu.hk/interpreting_conf2022/`