Review of Thesis Proposal

Faculty of Mathematics and Physics, Charles University

Reviewer:	Jindřich Libovický, ÚFAL MFF CUNI
Date:	August 18, 2023
Candidate	Peter Polák
Thesis title:	Simultaneous and Long-Form Speech Translation
Supervisor:	Ondřej Bojar, ÚFAL MFF CUNI

The thesis proposal of Peter Polák deals with the problem of simultaneous speech translation (SST), i.e., real-time speech-to-text translation from a stream of speech in a realistic setup where sentence boundaries are not known. It is an interesting research area: it has a direct practical application, and it is hard to anticipate the dominant approach in several years. Several competing approaches exist (cascade vs. end-to-end systems, autoregressive vs. non-autoregressive generation) and competing neural architectures. This leaves a large space for empirical research and makes a very good topic for Ph.D. research.

The presented Thesis Proposal has 12 pages of content and 5 pages of references. The text is split into eight sections (including Introduction and Conclusions), five figures, and no tables. The proposal is well structured: it starts by introducing the state of the art in SST, continues by defining the candidate's research goals, and concludes with the research plan for the upcoming years. The proposal is written in good English without obvious spelling or grammatical mistakes, with only occasional stylistic flaws. The proposal comprehensively cites relevant literature and distinguishes between existing work and the candidate's original contribution to the field.

Content

The introduction briefly highlights the problem the candidate is working on and previews the proposal's content.

Section 2 introduces the problem of simultaneous speech recognition. Unfortunately, although this section explains the basic concepts and challenges of the problem (difference between translation and interpreting), at the same time, it also assumes a lot of previous knowledge (e.g., this section never says that SST is speech-to-text). When advocating for end-to-end approaches (which the candidate plans to work on), two exaggerated claims would require better factual support to be taken seriously. The first claim regards the use of prosody – without any evidence that the system can use the speech melody to distinguish between grammatical moods, this claim seems unrealistic. The second unrealistic claim is the advantage of languages without writing systems. This seems like a rather theoretical advantage because such languages will likely not have

enough data for any computational processing.

Section 3 briefly introduces the main mathematical and computer science concepts important for modeling output in speech processing. Given how complex and often unintuitive the concepts are (they combine probability with dynamic programming), there are very well explained.

Section 4 is, compared to Sections 2 and 3, very dense and hard to read for someone who does not closely follow the development in the respective subfield (e.g., MWER training is only cited without at least briefly explaining what that is). The last subsection discusses the evaluation of SST well describes the difficulties with SST evaluation. Evaluation is an important issue that should have got more attention in the proposal.

Section 5 summarizes the three main candidate's research goals. They are formulated in a rather abstract way. Except for the first one, they are not related to evaluation, making it unclear how the candidate plans to evaluate if the goal was reached (besides publishing on that topic at a respected venue). Nevertheless, if the candidate succeeded in reaching the goals, it would significantly contribute to the field.

Section 6 presents the research the author has already done. The candidate has a relatively long publication record, featuring four system description papers, three conference papers, and one pre-print. The findings of the research done so far are interesting. However, only one candidate's paper was published at a top-tier venue (Interspeech, CORE rank A).

Section 7 presents the candidate's future research plans. The plans mostly apply methods previously developed for non-autoregressive text generation and methods used in document-level machine translation. Although both might be a meaningful technical step forward in the problem of SST, their novelty is rather limited because it mostly applies existing approaches to a related problem. Given the candidate's previous work, it is the reviewer's opinion, that the plans might be more ambitious.

The Conclusions summarize and conclude the proposal.

Questions to the Candidate

The plans mentioned in the proposal rely on methods that require monotonic alignment of the input and the output. Reordering is only possible in the underlying architecture. Although Transformers, in theory, can do arbitrary reordering, literature on non-autoregressive machine translation shows it is still a problem.

1. How do you plan to tackle the problem in the model architecture?

2. Can data augmentation, e.g., modifying the training data, ensure the output is better aligned with the input?

The recent development in text-based machine translation suggests that massively multilingually approaches (such as Meta's No Language Left Behind) can efficiently leverage similarities between languages and reach good translation quality even for otherwise low-resource languages.

3. Do you think this approach is also suitable for SST? Can we expect a similar synergy of related languages?

4. Can massively multilingual systems be a solution to handle code-switching?

Compared to written text, speech is more situated. Speakers can typically assume that their communication partners see and hear the same things as they do and refer to them.

5. Do you see multimodality (especially referring to the outside world) as an important problem that should be tackled soon?

Conclusion

Sections 2–4, despite all the reviewer's criticism, clearly show that the author closely follows the recent development and is already an expert in the area. The candidate's work so far (participation in several shared tasks, co-authoring three conference papers) is of good quality and, if continued, will lead to enough material for a doctoral thesis. The future plan is realistic, with a high chance of making a good technical contribution to the problem of SST.

Prague, August 18, 2023

Jindřich Libovický