Thesis Review by the Supervisor

Reviewer & supervisor: doc. RNDr. Ondřej Bojar, Ph.D.; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, 181 00

Date: May 30, 2024

Thesis Title: Multi-Source Simultaneous Speech Translation

Candidate: Mgr. Dominik Macháček

The doctoral thesis submitted by Dominik Macháček studies the task of speech translation in a multilingual or highly multilingual setting. Ideally, we would like fully automatic systems to reliably complement human interpreters and benefit from their simultaneous interpreting. Given the source speech and the speech of one or more interpreters, a multi-source system could be able, in principle, to merge the information from the original speech and from the interpretation when translating into a third language. Reaching the envisioned live speech translation system in practice proved to be a hard task.

When supervising Dominik's work towards this ambitious goal, I particularly value Dominik's systematic and well organized approach. As described in the thesis, Dominik's scientific contributions start with a very carefully prepared test set, the ESIC corpus. We then move to the analysis of interpreters' outputs in ESIC which provides quantitative estimates of e.g. the latency that needs to be achieved by the system for practical usability, as well as qualitative differences that make the difference between automatic speech translation and the distant goal of machine interpreting obvious.

Working towards the goal of multi-source speech translation, Dominik documents that errors occurring in speech recognition system outputs are complementary for two sources. This positive sign is exploited in Dominik's main result in multi-sourcing: for certain levels of (simulated) speech recognition errors, Dominik achieves a significant improvement in translation quality when using two sources instead of any one of them alone. As Dominik clearly notes, this result is achieved in a simplified setting where the two sources are perfectly synchronized at the level of individual sentences, so future research is needed before this synchronization can be carried out in realistic conditions on the fly.

The thesis offers two more technical but perhaps more visible contributions: a careful evaluation of our manual quality measure for speech-to-text live translation, and a very attractive wrapper for the Whisper speech recognition model which allows Whisper deployment in low-latency live regime.

The thesis itself is written in very good English, well structured at the chapter, section, as well as paragraph levels and complemented with many illustrations, tables and charts, all clearly commented. After a brief introduction in Chapter 1, Chapter 2 provides a detailed motivation and related work. It is worth mentioning that Dominik's topic is not a particularly frequently studied one. There have been related attempts in the past but not many. Chapter 3 carefully delimits what is the focus of Dominik's work and what will be intentionally left for future. Chapter 4 describes available relevant datasets and the creation and annotation of the ESIC corpus. Analysis of interpreters' work on ESIC is

presented in Chapter 5. Chapter 6 can be seen as the peak of the thesis, describing multi-sourcing experiments and presenting empirical evidence of the improvements in a specific range of speech recognition errors. Chapter 7 deals with evaluation: describing Continuous Rating (CR), a relatively cheap manual evaluation method for which Dominik (with co-authors) carried out an in-depth study, documenting a good correlation of standard automatic MT evaluation metrics with CR. Chapter 8 returns to the multi-sourcing experiment in a slightly more realistic setting: the ASR outputs are now genuine, with errors appearing naturally in the outputs of the best available models, and the evaluation is carried out also manually. The results now indicate that multi-sourcing did not bring the expected gains and Dominik carefully lists the possible steps that would allow us to reach them in the future. One clear limitation came from the relatively simple input combination method (late averaging). The tool created for continuous stream processing of speech, Whisper-streaming, as mentioned above, in any case makes a very practical and appreciated contribution to the speech community. The thesis is concluded with Chapter 9 summarizing the results.

Throughout our collaboration, I was very content with Dominik's approach, progress, focus and technical abilities. Dominik was independent, self-managed and able to handle very well all the obstacles we ran into. Our collaboration extended beyond the thesis, Dominik was also a very reliable colleague of mine in the research projects (esp. ELITR and NEUREM3), always ready to do all what was needed for an overall success of experiments as well as demonstration sessions. I sincerely hope that this collaboration will continue. Among other things, the topic of multi-source speech translation definitely remained a valid and promising path, just too ambitious to be handled within one thesis.

In sum, I consider the doctoral thesis by Dominik Macháček as excellent work, documenting Dominik's research proficiency and expertise. The full goal of live combination of speech and interpretation was not reached but significant progress in the direction has been made and the next needed steps have been clearly listed. The scientific contributions by Dominik are solid and valuable both short-term and long-term. Therefore, I fully recommend accepting the thesis.

In Prague, May 30, 2024.

Ondřej Bojar