Introducing Speech Recognition in Non-live Subtitling to Enhance the Subtitler Experience

Željko Radić [0009-0003-9451-5557], Sabine Braun [0000-0002-6187-3812] and Elena Davitti [0000-0002-7156-9275]
University of Surrey, Guildford, Surrey GU2 7XH, UK
z.radic@surrey.ac.uk

Abstract. Interlingual Subtitle Voicing (ISV) is a new technique that focuses on using speech recognition (SR), rather than traditional keyboard-based techniques for the creation of non-live subtitles. SR has successfully been incorporated into intralingual live subtitling environments for the purposes of accessibility in major languages (real-time subtitles for the deaf and hard of hearing). However, it has not yet been integrated as a helpful tool for the translation of non-live subtitles to any great and meaningful extent, especially for lower resourced languages like Croatian. This paper presents selected results from a larger PhD study entitled ‘Interlingual Subtitle Voicing: A New Technique for the Creation of Interlingual Subtitles, A Case Study in Croatian’. More specifically, the paper focuses on the second supporting research question that explores participants feedback about the ISV technique, as a novel workflow element, and the accompanying technology. To explore this technique, purpose-made subtitling software was created, namely SpeakSubz. The constant enhancements of the tool akin to software updates are informed by participants’ empirical results and qualitative feedback and shaped by subtitlers’ needs. Some of the feedback from the main ISV study is presented in this paper.

Keywords: ISV, Interlingual Subtitles, Non-live Subtitles, Subtitler Experience, SUBX, SpeakSubz, Human-Machine Interaction, Action Research.

1 Introduction

All too often subtitling tools and, consequently, workflows are determined by clients, be they translation agencies as an intermediate link in the workflow or end clients. Subtitlers themselves always need to adapt to the existing workflows and use tools mandated by clients. Nowadays these are mostly online proprietary subtitling tools all of which differ from one another and offer various functionalities. As a result, subtitlers are expected to learn how to use multiple tools to remain competitive in the market. Unfortunately, when developing such software some of the most important agents in the subtitling workflow, subtitlers themselves, are rarely being consulted. The Interlingual Subtitle Voicing (henceforth ISV) study aims to amend this by adopting a subtitler-centric approach to the investigation of a new subtitling workflow.
The topic presented in this paper is a part of a larger doctoral study which explores the integration of speech recognition (SR) into a non-live subtitling workflow using a newly introduced technique. The ISV study generated empirical data on efficiency, measured as a combination of translation quality (from English into Croatian) and SR accuracy in Croatian, as well as the duration of various workflow elements, such as typing, voicing and editing. However, this paper will focus on the qualitative part of the study that strives to capture the subtitler experience (SUBX) with this new form of human-machine interaction through the feedback provided by participants. This feedback will inform subsequent updates of the ISV technique and the specialised software (SpeakSubz), created specifically to implement it. The ultimate goal is creating a customisable tool with a plethora of functionalities that can enhance subtitlers’ workflows and work environments based on their needs and preferences.

2 The Interlingual Subtitle Voicing Technique

ISV is a newly devised hybrid technique drawing from three established modes of language transfer: subtitling, interpreting and respeaking. The technique is meant for the translation of pre-recorded or non-live subtitles and leads to a workflow that differs from traditional subtitling workflows. It does so by introducing a new variable into the process: voicing, known as respeaking in live subtitling. This type of human-machine interaction includes a subtitler, or a voicer, translating subtitles from a source language (SL) by speaking into a microphone connected to SR software. This specialised software turns vocal input in a target language (TL) into textual output in TL in the form of subtitles. The process is similar to interlingual respeaking, without the element of immediacy since the ISV technique occurs in non-live environments.

Subtitling is a crucial component of the ISV technique. The traditional subtitling workflow typically involves a fairly linear sequence of activities. These are often determined by the client or software used to produce subtitles, especially in cloud-based environments. Within that workflow, subtitlers have some leeway in how they handle different workflow components. For the purposes of this paper, it is relevant to distinguish these two levels of workflows, which can be referred to as macro-level and micro-level workflows, respectively. The term is drawing on the distinction made by Alves et al. [1] who distinguish between macro and micro translation units.

In this paper we consider workflows coming from an end client or an intermediary translation agency as a macro-workflow that subtitlers normally cannot influence, such as the use of proprietary subtitling tools, linguistic and technical guidelines, deadlines etc. On the other hand, each subtitler has his or her own individual translation strategies, which we label as micro-workflows: these mostly have to do with the order in which a subtitler performs certain tasks and strategies s/he uses within the established macro-workflow.

In this paper, we consider workflows coming from an end client or an intermediary translation agency as a macro-workflow that subtitlers normally cannot influence, such as the use of proprietary subtitling tools, linguistic and technical guidelines, deadlines etc. On the other hand, each subtitler has his or her own individual translation strategies, which we label as micro-workflows: these mostly have to do with the order in which a subtitler performs certain tasks and strategies s/he uses within the established macro-workflow.

Figure 1 below shows the basic workflow configuration devised for the purposes of the ISV study that can, as needed, be adapted to individual subtitlers’ micro-workflows in the future. This workflow was also devised with the ISV study in mind in order to facilitate empirical measurements related to duration of various stages of a subtitling process and make participants aware of these various workflow elements.
Fig. 1. The ISV workflow consists of three stages: a pre-voicing, a voicing and a post-voicing stage.

The pre-voicing stage of the ISV workflow includes preparatory activities that occur before the translation/voicing process. The central voicing stage includes reading the template (subtitles in a source language), mentally translating the source text into a target language and speaking it out into SR software, including punctuation, while monitoring the textual output in a target language and reacting as needed. The post-voicing stage includes various edits and checks that need to be conducted to ensure the desired maximum accuracy of 100%, i.e., broadcast-ready subtitles. However, these are a part of a macro-workflow and as such not relevant for the ISV technique.

3 Related Work

The ISV technique was developed with subtitlers in mind. Such subtitler-centric research in academia is extremely rare and current studies have focused either on the final product, (i.e., subtitles), the way they are created, (i.e., the process) or the target audience. The existing scholarship ranges from research about traditional subtitling [2–4], quality of subtitles through practical experiments [5], as well as quality of templates [6] or theoretical quality assessment models [7]. More recent research has delved into media accessibility [8–10], new subtitling workflows, especially those that involve machine translation [8] and post-editing [9] and the translation process itself, be it traditional translation process research or TPR [10] or on online platforms [11]. Finally,
there are also audience reception studies [12, 13] but almost no studies that focus on subtitlers’ various micro-workflows and their evaluation of traditional or new hybrid workflows and, most importantly, their agency in creating the tools they might use in the future. The ISV technique aims to address this gap and contribute to the existing literature.

As already mentioned, the ISV research focuses on a new hybrid workflow implemented via a custom made proprietary online platform with an emphasis on the impact of integrating SR as a new major variable in the subtitler experience. To explore this workflow, subtitlers’ workspaces and needs, the current study relies on two main frameworks: action research (AR) and translator experience (TX). On one hand, AR [14–16] allows for the study of a new hybrid workflow and observes how it affects efficiency in terms of both speed and accuracy. On the other hand, the notion of TX [17], as well as studies about multimodality in translator-computer interaction [18–20] have informed the development and testing of ISV. This paved the way for the notion of Subtitler Experience (SUBX), which not only examines subtitlers themselves, but also their work processes, work environments and their evaluation of suggested techniques and workflows.

4 Methodology

This paper focuses on the second supporting research question of the ISV study: how does the ISV technique influence SUBX in a traditional subtitling environment, primarily in terms of ergonomics? We present selected findings from the qualitative portion of the study that explored the SUBX. This paper presents how participants responded and adapted to this new hybrid workflow and how they evaluated this novel form of human-machine interaction.

15 participants, Croatian native speakers, predominantly experienced subtitlers (13), took part in the ISV experiment, which included a 9-hour online guided self-training to be completed within a one-month window. All the necessary materials (SL subtitles, videos, the technique presentations, video tutorials) and tools needed for the experiment (integrated in the specialised tool SpeakSubz) were made available to participants on a website created for this study. The experiment had to be carried out entirely online because of the Covid-19 pandemic. During the ISV training, participants gradually learned voicing as a new skill and they progressively developed their abilities through a series of exercises, leading up to the final two exercises (7 & 8) that were quantitatively analysed.

To evaluate efficiency of the ISV technique and its underlying technologies, which was the focus of the first supporting research question, two variables were deemed relevant. Firstly, the speed of traditional typing compared to voicing, with and without editing. Secondly, Croatian SR accuracy was measured to gauge how close it was to the industry standard of 98% for live intralingual subtitling, while keeping in mind that ISV is meant for non-live subtitling. Moreover, there is ample time to correct SR errors in the final version of the translation. In addition, there is currently no established accuracy benchmark for interlingual respeaking, and this metric has yet to be explored.
and consolidated. This quantitative data related to accuracy and speed will be published in future papers upon the completion of the PhD thesis.

The most relevant datasets for this paper were the post-testing questionnaires and optional follow-up interviews. The questionnaires consisting of 30 questions were delivered through a Google form embedded on the ISV website\(^1\) and were completed by 15 participants. The questionnaire was divided into 5 sections: ISV Presentation, Croatian Speech Recognition and SpeakSubz, Translation Workflow and Skills, Translation Hardware and Future Translation Work.

Moreover, participants had the opportunity to apply for optional follow-up interviews (online, under 45 minutes) to provide a more in-depth evaluation of the technique upon completing their ISV training and testing. Six participants chose to do so, and despite the limited number, a wealth of information was gathered about five different topics: the ISV website, the ISV technique, specialized software SpeakSubz, participants’ workflows and workspaces and, finally, future translation practices.

Additionally, voice and screen recordings of participants’ performance in the two final exercises, initially set up for tracking durations of various processes, were used as a secondary source of data. This data gave a glimpse into participants’ behaviour during the voicing, typing and editing processes. These evaluations and general feedback, as well as voice and screen recordings, enabled exploration of human-machine interaction in non-live subtitling and how it can be further improved. Due to space limitations of this paper, the following section will only present selected findings about subtitlers’ experience with the ISV technique, while forthcoming subsequent qualitative and quantitative data will give a more complete picture of ISV and especially the SUBX.

5 Selected Findings

The selected results presented in this section come from the post-testing questionnaires and optional interviews conducted over Skype, transcribed and thematically analysed. For the participants in the ISV experiment, that was the first time they had encountered voicing professionally. Although the empirical data of the ISV study showed that Croatian SR is not yet on par with major languages such as English, Spanish, German or Italian, the results from the post-testing questionnaires and optional interviews indicate that participants were intrigued and often pleasantly surprised by the new technique.

For example, in Question 8 of the post-testing questionnaire, participants’ subjective evaluation of SR accuracy showed that they believe that the underlying technology is not yet perfect, which was also to some extent confirmed by the empirical data of the study. Most participants rated Croatian SR accuracy with a 3, with the median mark being 3.2 out of 5 (Figure 2).

\(^1\) https://isvresearch.eu/
Fig. 2. Question 8: Participants’ subjective evaluation of the accuracy of Croatian speech recognition.

Nevertheless, the rating was more positive when it came to their experience of working with Croatian SR (Question 7). The majority of participants rated it positively, with 10 participants rating it either 4 or 5, while only one participant rated it with a 2 (Figure 3). The median value was 4.06 and much more favourable than in Question 8. This suggests that participants mostly enjoyed working with Croatian SR, despite the imperfections of the underlying technology and relatively short training.

Fig. 3. Question 7: Participants’ experience of using Croatian speech recognition.

The participants’ feedback regarding the specialised software developed for this study (Question 9) was even more positive. As shown in Figure 4, the mean value amounted to 4.13 out of 5. This finding was also confirmed in Question 11 of the questionnaire and optional follow-up interviews, where participants described the software to be “user-friendly”, “easy to use”, “visually appealing” and even “game-like”. They also enjoyed its various functionalities such as reading speed markers, the comment section, pre-loaded tasks and voicing itself. Participants also suggested what other functionalities could be added in the future (Question 12) such as: automatic measurement units converter, an undo button and machine translation (MT). The latter was in the meantime added to the software and is currently in the beta phase ready for wider testing.
When asked whether participants had difficulties and what they were (Questions 5 and 6), some participants found certain exercises had too many steps in the workflow and because of that they needed to consult reminders. These reminders were in the form of “Read Me” textual files within each exercise. Other participants, on the other hand, were not always sure what to record and when. However, some participants in the follow-up interviews stated that they got used to the workflow by the final two exercises and the process became easier for them as the training progressed. Some participants also experienced some slight technical difficulties, with their voice not being captured or the recording functionality not working properly. Technical issues have been either resolved and/or put under the FAQ section of the website to clarify how to solve these issues in the future.

Regarding the comparison between voicing and typing, some participants did find voicing to be faster whereas others believed they were still faster when typing but nevertheless see the potential of the ISV technique. One participant, a highly experienced subtitler, also suggested that the technique might be better suited for interpreters since subtitlers are not used to voicing their translations and might need longer to become accustomed to it.

However, despite the potential shortcomings of the technique described above, in Question 17, 60% of participants stated they would like to continue developing their voicing skills in the future (Figure 5) and 33.3% indicated they might consider it. Taking into account the imperfections of the SR technology, the complexity of the technique, the short duration of the guided self-training and some minor bugs in the software that need to be corrected in the next version, this is a very positive result. Besides the need for training “voicers” in the future, it is encouraging to see that most participants enjoyed working with the technique, had a positive SUBX and would like to continue to use the technique in the future.
6 Future Research

After establishing the accuracy of Croatian SR within overall quality evaluation and durations of various subtitling processes to determine efficiency, it is promising to note positive results in terms of participants’ experience and reception. The future success of the ISV technique and specialised software as well as their usability in education and the audiovisual translation (AVT) industry will depend on software upgrades and technique adaptations. Experience from the AVT industry shows that machine translation seems to be increasingly used in subtitling workflows. Additionally, based on suggestions from some of the participants already working with it on a daily basis, adding MT functionality within the ISV technique would be the next logical step. This hybrid approach could populate subtitle boxes with MT (subtitle by subtitle or all subtitles at once) and subtitlers could use voicing to correct errors, thus reducing the need for typing with physical keyboards. That workflow configuration will be tested next at Croatian universities as of 2024/2025.

The technique and the software itself could be further developed in two different ways: one for academia and one for the AVT industry. SpeakSubz is not only a practical tool for professional subtitlers, but also a training and a methodological tool for students. As already mentioned, Croatian universities have already shown interest in using this tool and the technique in the training of subtitlers. The tool can be used, for example, to teach AVT students the basics of subtitling, text reduction (reading speed markers), line breaks (“new line” virtual button), voicing (respeaking) and MT post-editing.

When it comes to SUBX, a key element that needs to be studied in the next cycle of ISV research is the usability of the technique and software in real-life professional situations, especially on handheld and touchscreen devices to widen and enhance the SUBX in terms of ergonomics. While most AVT-related research nowadays is conducted in controlled media lab conditions, it is important to capture the use of the technique in subtitlers’ real workspaces which nowadays, more often than not, are their own homes. The ISV methodology allows for anonymous research in real-life professional
situations, without intruding on participants’ privacy. This methodology could be developed further to include a more meaningful utilisation of mobile and handheld devices in the next phase of the study, with a possible longitudinal study to track participants’ progress.

7 Conclusion

The ultimate goal of this study is to develop training in a hybrid translation mode (ISV) and offer users, intended as language professionals at large, customisable tools with multiple functionalities to optimise their everyday work. Unlike existing subtitling tools, users can directly influence the functionality of SpeakSubz by suggesting features that would help them work more efficiently and can be implemented in a timely manner. In addition, subtitlers can also evaluate these novel and hybrid workflows, thus giving them agency that is rarely present in subtitling software development, whereby future researchers could rapidly replicate or adapt them for their own research purposes.

As underlying technologies advance, ISV could potentially be used for live subtitling into Croatian as well as to enhance accessibility. The technique could also be applied and/or adapted to other lower-resourced languages which are lagging behind the major languages simply because the technology is not yet available to them. Rather than waiting for technology to be fully ready, proactive measures can be taken by researchers to improve existing ones through human input, as is the case in ISV. ISV can hopefully lead the way in this type of subtitler-centric research and ensure professionals are trained to live up to the challenges of new SR and MT reality both in AVT studies as well as in the AVT industry. Participants’ positive experience with the ISV technique and positive evaluation of SUBX certainly encourages us to continue with this strand of research in the future.

References


