Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation

Dávid Javorský **Dominik Macháček** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{surname}@ufal.mff.cuni.cz

Abstract

Simultaneous speech translation (SST) can be evaluated on simulated online events where human evaluators watch subtitled videos and continuously express their satisfaction by pressing buttons (so called Continuous Rating). Continuous Rating is easy to collect, but little is known about its reliability, or relation to comprehension of foreign language document by SST users. In this paper, we contrast Continuous Rating with factual questionnaires on judges with different levels of source language knowledge. Our results show that Continuous Rating is easy and reliable SST quality assessment if the judges have at least limited knowledge of the source language. Our study indicates users' preferences on subtitle layout and presentation style and, most importantly, provides a significant evidence that users with advanced source language knowledge prefer low latency over fewer re-translations.

1 Introduction

Simultaneous speech translation (SST) is a technology that assists users to understand and follow a speech in a foreign language in real-time. The users may need such an assistance because of limited knowledge of the source language, the speaker's non-native accent, or the topic and vocabulary. The technology can be used for the target languages, for which human interpretation is unavailable, e.g. due to capacity reasons.

Candidate systems for simultaneous speech translation differ in quality of translation, latency and the approach to stability. Some are streaming, only adding more words (Grissom II et al., 2014; Gu et al., 2017; Arivazhagan et al., 2019; Press and Smith, 2018; Xiong et al., 2019; Ma et al., 2019; Zheng et al., 2019; Iranzo Sanchez et al., 2022), some allow re-translation as more input arrives (Müller et al., 2016b; Niehues et al., 2016; Dessloch et al., 2018; Niehues et al., 2018; Arivazhagan et al., 2020). Finally, subtitle presentation options

Figure 1: A detail of the default layout with the video document "Dinge Erklärt: Impfen...".[1] The video is at the top, overlaid by two lines of subtitles in Czech, followed by buttons for Continuous Rating. The button labels are: 1: Worse; 2: Average; 3: Good; 0: I do not understand at all.

(size of subtitling window, layout, allowed reading time, font size, etc.) also affect users' impression. The combination of the re-translating approach and limited space for subtitles is challenging because of "flicker", i.e. the updates to the text that the user is reading at the moment, has already read, or that has been scrolled away. The subtitling options impact the amount of flicker, reading comfort and delay and may affect the general usability.

The evaluation of the traditional, text-to-text machine translation (MT) has been researched for many years (see e.g. Han, 2018 or developments and discussion within the series of WMT, Akhbardeh et al., 2021). It targets only the translation quality. SST evaluation faces new challenges: simultaneity, latency, and readability to humans. Evaluating only selected aspects in isolation is reasonable (as MT quality in Elbayad et al., 2020, latency in Ma et al., 2018; Cherry and Foster, 2019), however, a complete evaluation must be end-to-end, from sound acquisition to subtitling, and take into

[1] https://youtu.be/4E0dwFS72gk

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 154 - 164

account the intent of communication. We generalize the intent to passing pieces of information from the speaker (sender) to a participant in an online session (receiver).

Our Contributions In this paper, we run an experimental evaluation campaign on 2 hours of documents with German-Czech SST using 32 judges with different levels of source language proficiency. (i) We contrast two methods of SST evaluation: Continuous Rating and factual questionnaires. We find out that Continuous Rating by bilinguals is easy and reliable for assessing the comprehension. (ii) We measure how much comprehension is lost by simultaneity, flicker and presentation options. (iii) We evaluate different presentation options and layouts and find the most preferred one. (iv) We find a statistically significant evidence that the users with an advanced, but limited knowledge of the source language reach higher comprehension with low latency subtitles than with large latency and low flicker. (v) We publish our implementation of the subtitling tool, web application for simulating live events with SST subtitling, and SST human evaluation framework.

Since Continuous Rating is easily applicable to any speech documents, even to those without transcripts and reference translations, and requires minimal time overhead for both preparation and user evaluation, we believe it is suitable to become a standardized way for human manual evaluation of SST.

2 Related Work

Hamon et al. (2009) propose user evaluation of speech-to-speech simultaneous translation. To test the adequacy and intelligibility, they prepared questionnaires with factual questions from the source speech. The judges listened either to the interpreter, or the machine, and answered the questions. They evaluated the offline mode, the judges were allowed to stop and replay the audio while answering. This way the authors measured the comprehension loss caused by the automatic translation or interpretation. Each sample was processed by multiple judges, to eliminate human errors. Fluency was assessed by the judges on a scale.

Macháček and Bojar (2020) propose a technique for collecting continuous user rating while the user watches video and simultaneous subtitles. The user is asked to express the satisfaction with the subtitles at any moment by pressing one of four buttons as the rating changes.

Müller et al. (2016a) analyzed the feedback from foreign students using KIT Lecture Translator within two semesters. Such a long-term and informal evaluation differs considerably from judging in controlled conditions. On one hand, it summarizes the real-life situation with all the variables and corner cases that a lab test could only approximate or omit. On the other hand, the users may not be motivated to give the feedback, and can give only personal opinions that may be biased. This way it is also difficult to compare multiple system candidates.

3 Evaluation Campaign

In our evaluation, we simulate live events on which participants need assistance with understanding the spoken language. The source and target languages in our study are German and Czech, respectively. This is an interesting example of two neighbouring countries, distinct language families and yet a relatively well studied pair with sufficient direct training data.

3.1 Translation System

We use the ASR system originally prepared for German lectures (Cho et al., 2013). It is a hybrid HMM-DNN model emitting partial hypotheses in real time and correcting them as more context is becoming available. The same system was used also by KIT Lecture Translator (Müller et al., 2016b).

The system is connected in a cascade with a tool for removing disfluencies and inserting punctuations (Cho et al., 2012), and with a German–Czech NMT system.

The machine translation is trained on 8M sentence pairs from Europarl and Open Subtitles (Koehn, 2005; Lison and Tiedemann, 2016), and validated on newstest. The Transformer-based (Vaswani et al., 2017) system runs in Marian (Junczys-Dowmunt et al., 2018) and reaches 18.8 cased BLEU on WMT newstest-2019.

Despite the translations are pre-recorded and only played back in our simulated setup, we ensured we keep the original timing as emitted by the online speech translation system.

3.2 Selection of Documents

We selected German videos or audio resources that fulfilled the following four conditions: 1) Length 5 to 10 minutes (with some exceptions). 2) The

Type	#	Length	Description
TP	3	18:08	European Parliament
TP	3	17:34	DG SCIC, Repository for interpretation training
A	3	27:52	A mock interpreted conference at interpretation school
V	2	14:43	Maus, Educative videos for children
A	2	18:48	DW, For learners of German
V	2	16:09	Dinge, Educative videos for teens
All	15	114:52	

Table 1: Summary of domains of selected documents. Type distinguishes audio only (A), talking person only (TP) and video (V) with illustrative or informative content. Length is reported in minutes and seconds.

translations had to be of a sufficient quality. Based on a manual check, we discarded several candidate documents: a math lecture and broadcast news due to many mistranslated technical terms and named entities. Another group of documents was mistranslated and discarded because they were not long-form speeches, but isolated utterances with long pauses. 3) Informative content. We intend to measure adequacy and comprehension by asking the judges complementary questions. We thus excluded the documents where the speaker is not giving information by speech, but uses mostly paralinguistic means, e.g. singing, poetry, or non-verbal communication. 4) Non-technicality. We expect the judges answer in several plain words in their mother tongue. They may lack knowledge of any specialized vocabulary.

We selected audios, videos with informative or illustrative content, and videos of talking persons, to compare user feedback for these types of documents. Table 1 summarizes the selected documents.

3.3 Subtitler: Subtitle Presentation

Subtitler is our implementation of the algorithm by Macháček and Bojar (2020) extended by automatic adaptive reading speed in addition to the "flicker" parameter as defined in Macháček and Bojar (2020). The speed varies between 10 and 25 characters per second depending on the current size of the incoming buffer. The default font size is 4.8 mm. The default subtitling window is 2 lines high and 163 mm wide.[2] By default, we use the maximum flicker and the lowest delay (presenting all translation hypotheses, not filtering out the partial and possibly unstable ones), no colour highlighting, and smooth slide-up animation while scrolling. The example of the setup can be seen in Figure 1.

With the default subtitling window, 90% of the words in the test documents are finalized in subtitles at most 3 seconds after translation. In 99%, it is at most 7 seconds. More details and the comparison to fixed reading speed are provided in Appendix A.1.[3]

3.4 Web Application as Simulation Environment

We implemented a web application for presenting video and audio documents with embedded Subtitler. We use it for simulation of live subtitled events. The application is equipped with a tool for collecting users' feedback. It also allows administrators to design experiments with different variables (document, subtitling layout, subtitling option) and distribute them to individual judges.[4]

3.5 Types of Feedback

Continuous Rating Inspired by Macháček and Bojar (2020), we add 4 buttons below the audio/video document. While watching, the participants are asked to press the buttons to indicate their current satisfaction with the subtitles. We let participants decide the frequency of rating but we suggest clicking each 5-10 seconds or when their assessment has changed. We encourage them to provide feedback as often as possible even if their assessment has not changed. The scores of the rating range between 0 (the worst) and 3 (the best). The order 1, 2, 3, 0 matches the keyboard layout; participants are encouraged to use keyboard shortcuts. The layout is illustrated in Figure 1.

Questionnaires Answering questions as an evaluation approach has been already used (Hamon et al., 2009; Berka et al., 2011). Our questionnaires were composed of two parts: factual questions and general questions.

For **factual questions** we used the open style, i.e. asking for a short response, instead of yes/no or multiple choice to exclude guessing. We asked a Czech teacher of German to prepare the questions and an answer key from the original German documents, regardless of the machine translation. The teacher wrote the questions in Czech, and was instructed to prepare one question from every 30

[2]All typographical properties follow https://bbc.github.io/subtitle-guidelines/

[3]The source code of Subtitler is available at https://github.com/ufal/subtitler

[4]The source code of the application is available at https://github.com/ufal/continuous-rating

	Layout Experiments							
CEFR	0	A1	A2	B1	B2	C1	C2	all
Count	5	5	1	2	1	-	-	14

	Flicker Experiments							
	Z	**Begin.**		**Advanced**				
CEFR	0	A1	A2	B1	B2	C1	C2	all
Count	3	1	3	-	2	8	1	18
All	8	6	4	2	3	8	1	32

Table 2: The judges by their German proficiency levels on CEFR scale and their assignment to experiments. In Flicker experiment, the distribution to groups: Zero level, Beginners, Advanced.

seconds of the stream and distribute them evenly, if possible. The questions had to be answerable only after listening to the document, and not from the general knowledge. The complexity of the questions was targeted on the level that an ordinary high-school student could answer after listening to the source document once, if the student would not have any obstacles in understanding German. To reduce the effect of limited memory, the judges had an option in the questionnaire to indicate they knew the answer but forgot it. Furthermore, they had to fill, from which source they knew the answer: from the subtitles, from the speech, from an image on the video, or from their previous knowledge.

Finally, we evaluated the factual questions manually against the key, rating them at three levels: correct, incorrect, and partially correct.

After the factual questions, all the questionnaires had a common part with **general questions** where we asked the judges on their impression of translation fluency, adequacy, stability and latency, overall quality, video watching comfort, and a summary comment.

3.6 Judges

We have conducted two groups of experiments, each with different and distinct groups of judges.

In Comprehension and Layout experiments (Sections 4.1 and 4.2), we examined distinct subtitling features. We selected 14 native Czech speakers as judges. Their self-reported knowledge of German had to be between zero and B2 on the CEFR[5] scale, to ensure they need some level of assistance with understanding German. We also ensured they do not have knowledge of any other language which could help them understanding German.

[5]Common European Framework of Reference for Languages

Type	**w. avg±std**	***t*-test**
Offline+voting	0.81±0.11	
Offline	0.59±0.16	***
Online, without flicker	0.36±0.16	***
Online, flicker, top layout	0.33±0.13	
Online, flicker, least preferred	0.31±0.16	

Table 3: Comprehension scores on all documents and judges. The average weighted by number of questions in document. *** denote the statistically significant difference (p-value$<$ 0.01) between the current and previous line.

For Flicker experiments (Section 4.3), we found other 18 native Czech speakers with an unrestricted German proficiency, to contrast their feedback and level of German. For further analyses, we divided them into three groups. For brevity further in the paper, we denote the judges with no proficiency of German as "Zero" level group, with proficiency between A1 and A2 as "Beginners", and the others as "Advanced". See summary of the judges in Table 2.

The judges were paid for participation in the study. Each judge spent in total 2 hours on watching and 3 hours on the questionnaires. They watched the videos at their homes on their own devices. They were asked to customize their screen resolution and eye-screen distance to suit their comfort.

4 Results

First, we analyzed the comprehension levels (Section 4.1) and presentation layouts (Section 4.2). Then, we selected the most preferred layout and used it for examining the impact of flicker on comprehension in Flicker experiments (Section 4.3).[6]

4.1 Comprehension Levels

In our study, we assume comprehension can be assessed as a proportion of correctly answered questions. We assume the following model: A person without any language barrier and with non-restricted access to the document during answering the questionnaire can answer all questions correctly. With a language barrier and offline MT (unlimited perusal of the document while answering), some information may be lost in machine translation. More information is lost with one-shot access to online machine translation because of forgetting and temporal inattention. Some more information may be

[6]The collected data are available at http://hdl.handle.net/11234/1-4913

lost because of flicker, and some more because of suboptimal subtitling layout.

Our results confirm the assumed hierarchy of comprehension levels. Moreover, we notice that even the judges with offline MT give inconsistent answers. Combining them and counting answers as correct if at least one judge is correct leads to higher scores. We explain it by insufficient attention.

Table 3 summarizes the results on all documents. We measured that on average, 81% of information was preserved by machine translation (Offline+voting, i.e. one of two judges answered correctly). A single judge could find 59% of information (Offline). In an oracle experiment without flicker, when the machine translation gives the final hypotheses with the timing of the partial ones (i.e. as if it knew the best translation of the upcoming sentence), a single judge could answer 36%. In real setup with flicker and the most preferred subtitling layout (Online, flicker, top layout), 33% information was found, and 31% with less preferred. The standard deviation is between 11 and 16%.

We found statistically significant difference (two-sided t-test) between offline MT with voting and without it, and between offline and online MT.

4.2 Layout Preference

We analyzed effects of distinct subtitling features by contrastive experiments differing only at one feature, see the paragraphs in this section. We distributed them randomly among the judges, regardless of their German skills. After watching each document, the judge fills the questionnaire.

In all cases, the results show a slight insignificant preference towards one variant of the feature in all three types of feedback that we collect: "Comprehension" is the proportion of correctly answered factual questions, "Averaged Continuous Rating" is an averaged feedback from button clicks, and "Final rating" summarizes the responses in the general section of questionnaires.

For visually informative videos, we separately report the scores of "Watching comfort" which we collected in the general section of questionnaires. Some judges provided also textual feedback, examples are in Appendix B.2.

Side vs Below For videos and videos with a talking person, we consider two locations for the subtitle window: on the left side of the video, or below. The side window can be high but narrow (17 lines of 60 mm width, to match the height of the video), while the window underneath is short and wide (2 lines of 163 mm width). The first is more comfortable for reading, the latter for watching the video.

The results are in Table 4 on the left. There is a preference for the layout "below" when the video is informative, and for "side" otherwise.

Below vs Overlay The subtitling window can be placed over the video, as in films, or below. In the first case, the subtitles possibly hide an informative image content, in the latter case, there is a larger distance between the image and the subtitles. The results on non-German speaking judges are insignificantly in favor of overlay, see the middle of Table 4.

Highlighting Flicker Status The underlying rewriting speech translation system distinguishes three levels of status for segments (automatically identified sentences): "Finalized" segments no longer change. "Completed" segments are sentences which received a punctuation mark. They can be changed by a new update and the prediction of the punctuation may also change or disappear. They usually flicker once in several seconds. "Expected" segments are incomplete sentences, to which new translated words are still appended. They flicker several times per second.

It is a user interface question if the status of the segments should be indicated by highlighting, or if this piece of information would be rather disturbing. We experimented only with colouring text background in large and medium subtitling window for audio-only documents.

Our experiments show that the judges prefer highlighting flicker status in the large window. For the medium window, this inclination is less clear, see Table 5.

Size of Subtitling Window The subtitling window can be of any size. If the window is short and narrow, there is a short gap between an image and subtitles, which simplifies focus switching. On the other hand, a small window contains short history, so the user can miss translation content if it disappears while paying attention to the video. A small window may also accidentally cause a long subtitling delay if the translation was updated in the scrolled-away part of text. In this situation, Subtitler has to "reset" the subtitles and repeat the part. With a large window, the distance between the growing end of the subtitles and the image is

		Side vs Below		Below vs Overlay		Size of subtitling window	
		Side	Below	Below	Overlay	2 l.×163mm	5 l.×200mm
Final rating	audio					10 1.80 ±0.87	8 **2.75 ±0.97**
	talking	5 **2.80 ±1.33**	7 2.43 ±1.05	9 2.33 ±1.05	9 **2.78 ±1.13**	9 2.33 ±1.05	5 **2.80 ±1.60**
	video	1 1.00 ±0.00	3 **1.67 ±0.94**	5 1.40 ±0.80	8 **2.38 ±0.86**	5 1.40 ±0.80	3 **2.33 ±0.47**
	sum, avg	6 **2.50 ±1.38**	10 2.20 ±1.08	14 2.00 ±1.07	17 **2.59 ±1.03**	24 1.92 ±1.00	16 **2.69 ±1.16**
Comprehension	audio					10 0.25 ±0.15	8 **0.31 ±0.15**
	talking	5 **0.34 ±0.25**	7 0.28 ±0.27	9 0.29 ±0.25	9 **0.39 ±0.20**	9 0.29 ±0.25	5 **0.40 ±0.21**
	video	1 0.18 ±0.00	3 **0.36 ±0.04**	5 0.26 ±0.14	8 **0.37 ±0.11**	5 0.26 ±0.14	3 **0.28 ±0.05**
	sum, avg	6 **0.31 ±0.24**	10 0.30 ±0.23	14 0.28 ±0.21	17 **0.38 ±0.17**	24 0.26 ±0.19	16 **0.33 ±0.16**
Avg. Cont. Rating	audio					10 0.90 ±0.71	8 **1.66 ±0.95**
	talking	5 1.56 ±1.00	7 **1.78 ±0.35**	9 1.65 ±0.52	9 1.65 ±0.99	9 **1.65 ±0.52**	5 1.09 ±0.78
	video	1 0.23 ±0.00	3 **1.21 ±0.45**	5 1.11 ±0.50	8 **1.15 ±0.77**	5 1.11 ±0.50	3 **1.35 ±0.31**
	sum, avg	6 1.33 ±1.04	10 **1.64 ±0.45**	14 **1.47 ±0.57**	17 1.42 ±0.93	22 1.21 ±0.70	16 **1.42 ±0.85**
Watching comfort	talking	5 2.80 ±0.75	7 **3.33 ±0.75**	9 3.43 ±0.73	9 **4.11 ±0.74**	7 **3.43 ±0.73**	5 2.80 ±0.98
	video	1 2.00 ±0.00	3 **3.00 ±1.63**	5 2.20 ±1.60	8 **3.00 ±1.00**	5 2.20 ±1.60	3 **2.33 ±1.25**
	sum, avg	6 2.67 ±0.75	10 **3.22 ±1.13**	14 2.92 ±1.32	17 **3.59 ±1.03**	12 **2.92 ±1.32**	8 2.62 ±1.11

Table 4: Results of the contrastive experiments for Side vs Below, Below vs Overlay and Subtitling window size: 2 lines height × 163 mm width vs 5 lines height × 200 mm width. The three numbers in each row and cell are the number of experiments, average and standard deviation. The higher score, the better. Comprehension rate is between 0 and 1, average continuous rating is between 0 and 3, the others on a discrete scale 1 to 5. Higher score in each experiment is bolded. The last row of each section summarizes the scores across document types.

Highlighting	No	Yes	No	Yes	No	No
Size [lines,mm width]	18×250 ("Large")		5×200 ("Medium")		18×250	5×200
Final rating	14 2.93 ±0.80	13 **3.31 ±1.14**	2 2.50 ±0.50	1 **4.00 ±0.00**	11 **2.91 ±0.79**	8 2.75 ±0.97
Comprehension	14 0.25 ±0.15	13 **0.30 ±0.12**	2 **0.44 ±0.18**	1 0.39 ±0.00	11 0.23 ±0.14	8 **0.31 ±0.15**
Avg. Cont. Rating	14 1.32 ±0.82	13 **1.42 ±0.74**	2 **2.19 ±0.50**	1 2.12 ±0.00	11 1.50 ±0.79	8 **1.66 ±0.95**

Table 5: Results of highlighting experiments on audio documents and subtitling window size 5 lines × 200 mm vs 18 lines × 250 mm. Description of numbers as in Table 4.

larger. The content stays longer, but it is more complicated to find a place where the user stopped reading before the last focus switch.

Depending on spatial constraints, it is always recommended to use as large window as possible, especially for documents without visual information, where focus switching between an image and subtitles is not expected. We tested two pairs of sizes on the same documents. The results are in Table 4 on the right. As we expected, the window with 5 lines was rated insignificantly better than with 2 lines in most scales and setups, but the 2-line reached a higher average watching comfort (2.92) that the 5-line setup (2.62).

For an audio-only document, we also tested the large (18 lines) vs. medium (5 lines) window, observing users' reported preference for the large one but slightly higher comprehension and continuous feedback for the medium one, see the right part of Table 4.

4.3 Flicker Experiments

We assume that the user behaviour differs by knowledge of the source language. We hypothesize that the Zero group of users and Beginners read all the subtitles all the time and do not pay attention to the speech. They do not mind large latency, but demand high quality translation, and comfortable reading without flicker. On the other hand, the users with an advanced knowledge of the source language may listen to the speech, try to understand on their own, and look at the subtitles only occasionally, when they are temporarily uncertain or need assistance with an unfamiliar word. They need low latency, and do not mind slightly lower quality.

To empirically test our hypothesis, we prepared two realistic setups: With flicker, the subtitles are presented immediately as available, but with frequent rewriting which discomforts the reader. Without flicker, the translations are delayed until the SST system confirms they will not change, and that usually happens during uttering the next sentence. We selected two videos for this experiment and distributed these setups uniformly between all groups of judges.

The results of comprehension are in Table 6. It shows that Advanced users achieve higher compre-

	Zero level	Beginners	Advanced
flic.	27 **0.34 ±0.16**	33 **0.33 ±0.16**	91 **0.58 ±0.19**
no f.	29 0.30 ±0.15	38 0.31 ±0.12	81 0.49 ±0.20
	insignificant	insignificant	$p < 0.01$

Table 6: Comprehension scores on a setup with flicker and no flicker, as rated by judges with different source language proficiency. The three numbers in each row and cell are the number of samples, average and standard deviation. Higher scores bolded. The difference between setups within Advanced group is statistically significant with $p < 0.01$.

	χ^2-test p-values		
	Zero level	Beginners	Advanced
OK/OK-	0.24	$\mathbf{1.8 \cdot 10^{-5}}$	$\mathbf{5.6 \cdot 10^{-5}}$
unknown	0.033	$\mathbf{1.7 \cdot 10^{-4}}$	$\mathbf{9.1 \cdot 10^{-4}}$
wrong	0.59	0.45	$\mathbf{2.9 \cdot 10^{-3}}$
forgot	0.9	0.48	0.019

Table 7: The results of χ^2-test for independence of Continuous Rating and answer correctness. Bolded values are where the two variables are **dependent** with statistical significance $p < 0.01$.

hension with flicker (58%) than without (49%). We found the difference statistically significant, which confirms the second part of our hypothesis.

The Zero level speakers and Beginners also report higher comprehension with flicker (Zero: 30% vs 34% and Beginners: 31% vs 33%), but this difference is statistically insignificant. Even though the preference inclines towards flicker, it is less noticeable compared to the Advanced group, and we consider this difference negligible. The other types of feedback (Average Continuous Rating and Overall rating from the end of questionnaire; not shown) confirm the trend of Comprehension for all groups.

4.4 Comprehension vs Continuous Rating

We collected Continuous Rating of the overall quality of subtitles at given times. For every comprehension question, we know the time span when the answer appears in the source speech document. Based on this timing information, we can relate comprehension and Continuous Rating. For a given time span answering a particular question, we find the most frequent Continuous Rating (button clicked most often) for every annotator. This gives us a histogram of Continuous Rating scores reported by different judges. In Figure 2 top, we show the correct ("OK") or partially correct answers ("OK-") and the histogram of Continuous Ratings by judges of distinct German proficiency levels. For a more detailed plot including all evaluation classes see Appendix B.1. This data aggregates observations for all documents and all setups excluding the offline SST and the oracle online SST without flicker.

For the judges with zero knowledge of German, we can not see any dependency of their comprehension to their Continuous Rating. On the other hand, the more the judges are proficient in German, the more their Continuous Rating reflects their comprehension. For example, for the C1 judges (Advanced) we can estimate their comprehension (and thus subtitle quality) from their clicking well: When they understand the content, the most probable given rating is 3 or 2. A less probable rating is 1, and they almost never rate 0 when they understand the content.

Listening while Rating In Figure 2 bottom, we show, from which source the judges knew the correct answer, either from the subtitles, or from sound. We can observe that indeed, the judges with German proficiency level B1 and higher listen to the source sound and understand, while the Zero level judges and Beginners rely only on subtitles.

Statistical Test To test the relation rigorously, we divide the judges into three groups by proficiency levels, their counts (see Table 2), their relation of Continuous Rating to correct answers and approach to listening versus reading (Figure 2). We run χ^2-test for statistical independence of Continuous Rating and answer results on the three groups. Test results are in Table 7. It shows that for the judges with Zero level of German, their Continuous Rating is independent on answer results. They do not follow the sound at all because they do not understand it, and rate only the readability and flicker. In case of the Beginners (A1 and A2, recall Table 2), we observe the dependency of their Continuous Rating on correct answers ("OK/OK-") and on cases when they did not answer ("unknown"). Their wrong answers and forgetting is independent of Continuous Rating, they probably make random mistakes uniformly. The Advanced group of judges give their correct, unknown or wrong answers consistently with their Continuous Rating. We therefore assume that they follow and understand the source speech and include the adequacy in their Continuous Rating.

We can also see that in all the three groups, the forgotten answers are independent on Continuous Rating. We assume that random and uniform out-

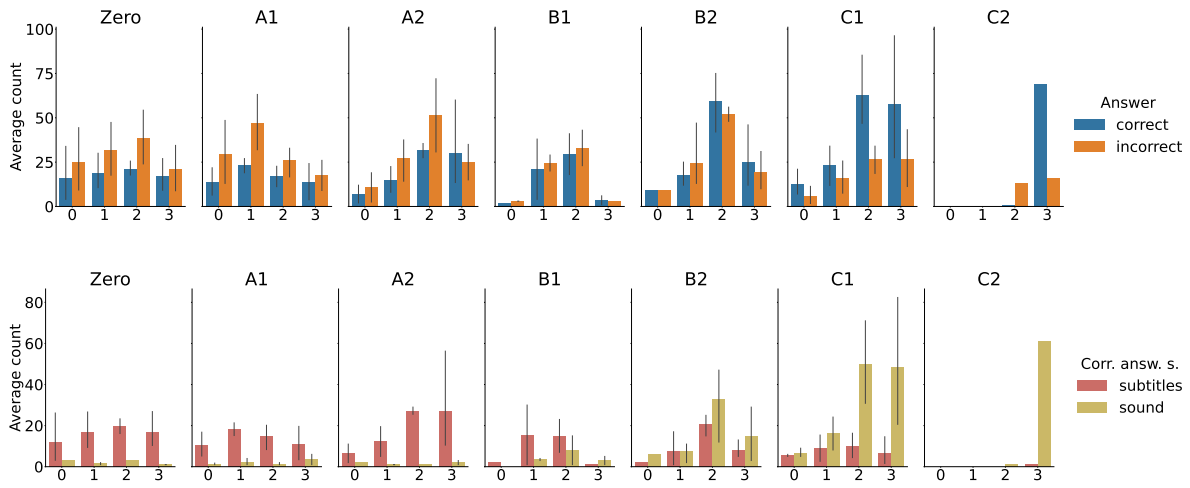

Figure 2: The average count of answers per judge for each proficiency level. Top: Correct (OK/OK-, blue bars) and incorrect (wrong/unknown, orange bars) answers vs Continuous Rating at the time when the answer was disclosed in the original document (x-axis, 0 means worst, 3 the best), distributed by source language proficiency level of the judges. Bottom: From which source the judges learned the correct or partially correct answer; subtitles in red, sound in yellow.

ages may be characteristic for human memory.

Practical Conclusions We conclude that Continuous Rating is a suitable for manual evaluation of simultaneous machine translation. The judges who speak the source language on at least B2 level on CEFR scale have an ability to assess SST quality reliably only by Continuous Rating, without the need for questionnaires which are laborious to prepare, answer and evaluate.

5 Conclusion

We proposed a novel and effective method for end-to-end user evaluation of simultaneous speech translation SST called Continuous Rating, publishing an open source evaluation tool for the future use. We showed that this method can be used for measuring comprehension and evaluating subtitling parameters. We demonstrated how user comprehension differs from offline MT to online MT. We showed that the users with a knowledge of the source language prefer low latency despite higher instability. We demonstrated that Continuous Rating can be used as a time-efficient human evaluation metric when employing judges with at least B2 (or, preferrably, C1) level of source language proficiency.

Limitations

This work is limited to only one direction of SST and lacks the comparison of multiple SST variants. Additionally, due to the number of investigated subtitling features and the smaller sample of judges, the results of layout experiments show only statistically insignificant preference towards one variant.

Acknowledgments

The research was partially supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, "Grant Schemes at CU" (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935) and SVV project number 260 575.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation

versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Jan Berka, Ondrej Bojar, et al. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Eunah Cho, C. Fügen, T. Hermann, K. Kilgour, Mohammed Mediani, C. Mohr, J. Niehues, Kay Rottmann, C. Saam, Sebastian Stüker, and A. Waibel. 2013. A real-world system for simultaneous translation of german lectures. pages 3473–3477.

Eunah Cho, J. Niehues, and Alexander H. Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *IWSLT*.

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, and Alexander Waibel. 2018. KIT lecture translator: Multilingual speech translation with one-shot learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 89–93, Santa Fe, New Mexico. Association for Computational Linguistics.

Maha Elbayad, Michael Ustaszewski, Emmanuelle Esperança-Rodier, Francis Brunet-Manquat, Jakob Verbeek, and Laurent Besacier. 2020. Online versus offline NMT quality: An in-depth analysis on English-German and German-English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5047–5058, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Olivier Hamon, Christian Fügen, Djamel Mostefa, Victoria Arranz, Muntsin Kolss, Alex Waibel, and Khalid Choukri. 2009. End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 345–353, Athens, Greece. Association for Computational Linguistics.

Lifeng Han. 2018. Machine translation evaluation resources and methods: a survey. In *IPRC – Irish Postgraduate Research Conference*, Dublin, Ireland.

Javier Iranzo Sanchez, Jorge Civera, and Alfons Juan-Císcar. 2022. From simultaneous to streaming machine translation by leveraging streaming history. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6972–6985, Dublin, Ireland. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 2018. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. *arXiv preprint arXiv:1810.08398*.

Dominik Macháček and Ondřej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies – Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org.

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016a. Evaluation of the KIT lecture translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1856–1861, Portorož, Slovenia. European Language Resources Association (ELRA).

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016b. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, volume 08-12-September-2016 of *Proceedings of the Annual Conference of the International Speech Communication Association. Ed. : N. Morgan*, pages 2513–2517. International Speech and Communication Association, Baixas.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Interspeech 2018*, Hyderabad, India.

Ofir Press and Noah A. Smith. 2018. You may not need attention. *CoRR*, abs/1810.13409.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun Hea, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354, Hong Kong, China. Association for Computational Linguistics.

	Delay						
	70%	80%	90%	95%	99%	max	resets
ARS	**0.01**	**1.44**	**3.06**	**4.51**	**7.05**	**12.06**	8.80
FRS	1.74	3.54	5.18	7.52	10.65	16.78	**5.47**

Table 8: The adaptive reading speed (ARS) in comparison to the fixed reading speed (FRS), set to 18 char/sec. Percentages denote the proportion of words that have a delay less than the given number. The delay is in seconds, resets in the average count per document.

A Subtitler

A.1 Adaptive Reading Speed: Delay

We compared adaptive to fixed reading speed, averaging over all documents. We set the value of fixed reading speed to 18 characters per seconds, which we obtained by averaging all delays in the setting without adaptive reading speed.

The comparison is in Table 8. The delay was measured for all presented words. We used a subtitling window of 2 lines $\times$163 mm because it represents an upper bound for the delay of bigger subtitling windows.

B Results

B.1 Comprehensions vs Continuous Rating

In Figure 3, we show the average count of answers per judge for each proficiency level. Note two observations: 1) The number of already known answers is negligible, which proves that the questions were selected based on the content of documents. 2) The number of answers whose source was not given is high for all answers (Figure 3, right column), whereas it is low when correct and partially correct answers were selected (Figure 3, middle column). It means that judges provided the source when they answered a question.

B.2 Textual Feedback

In Table 9, we depict several textual ratings from Flicker Experiment. We select judges with C1 source language proficiency and contrast their feedback for flicker and no flicker.

The judges report higher satisfaction with flicker. They notice increased latency when the presentation mitigate flicker. This is consistent with our findings in Flicker experiment for Advanced group.

	Feedback
Setting	**C1 proficiency, Overlay layout**
Flicker	The subtitles weren't so bad in terms of content or latency.
	The subtitles were very good, they just got stuck in the middle of the video, but after a short pause they worked again without any problems.
	The subtitles were relatively good, but despite their intelligibility and relative linguistic accuracy, they seemed very chaotic and very uncomfortable to read.
No flicker	A big delay of subtitles was sometimes inconvenient. If the subtitles are very delayed, it is almost impossible to follow them.
	The subtitles were small and dense, it was hard to orientate, especially when they were even delayed.
	At first, the delay was small. Then, at one point the subtitles got stuck and there was a lot of delay behind the sound.

Table 9: The selection of textual feedback from judges.

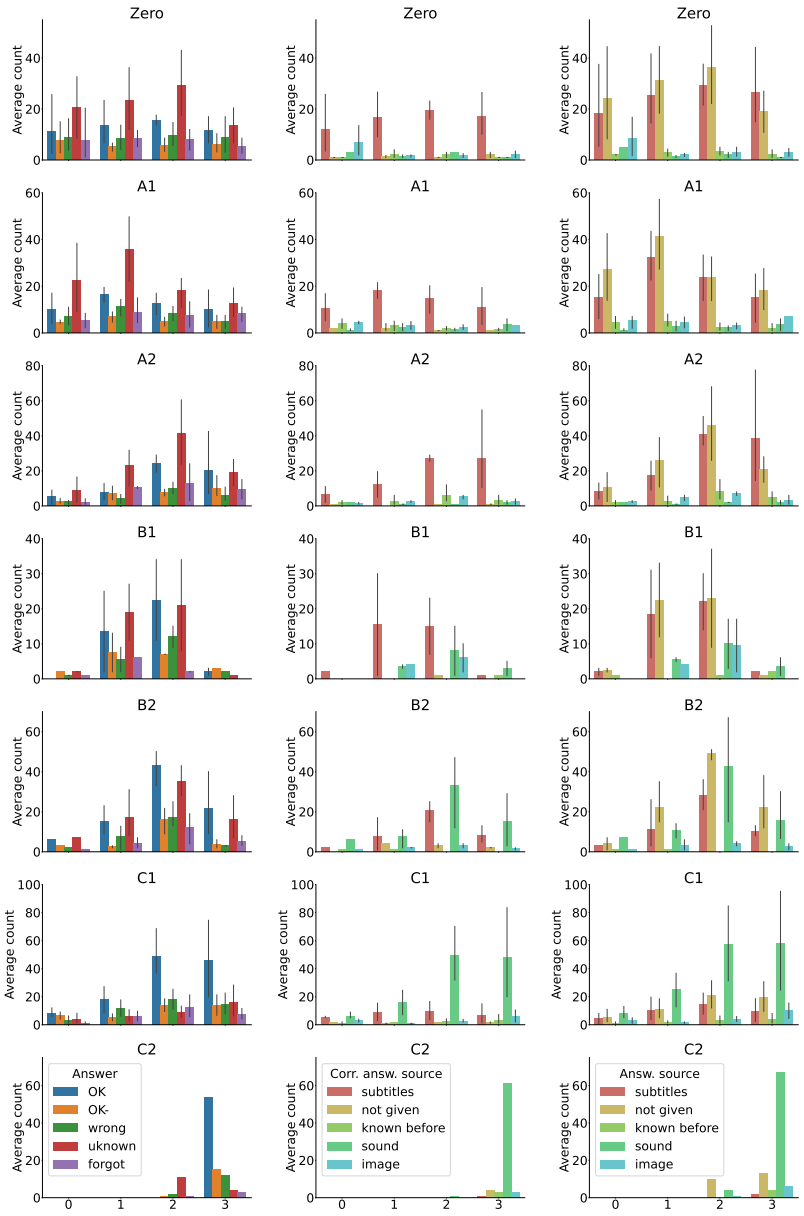

Figure 3: The average count of answers per judge for each proficiency level. Left: OK, OK-, wrong, unknown and forgotten answers vs Continuous Rating at the time when the answer was disclosed in the original document (x-axis, 0 means worst, 3 the best), distributed by source language proficiency level of the judges: from zero through beginners (A1, A2) and intermediate (B1, B2) to advanced (C1, C2). Middle: From which source the judges learned the correct (OK) or partially correct (OK-) answer. Right: From which source the judges learned all answers, regardless of their evaluation.