Direct Speech Translation for Automatic Subtitling

Sara Papi[1,2], Marco Gaido[1,2], Alina Karakanta[3]*, Mauro Cettolo[1], Matteo Negri[1], Marco Turchi[4]*

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy
[3]Leiden University Centre for Linguistics, Leiden, The Netherlands
[4]Zoom Video Communications, Karlsruhe, Germany
{spapi,mgaido,cettolo,negri}@fbk.eu

Abstract

Automatic subtitling is the task of automatically translating the speech of audiovisual content into short pieces of timed text, i.e., subtitles and their corresponding timestamps. The generated subtitles need to conform to space and time requirements, while being synchronized with the speech and segmented in a way that facilitates comprehension. Given its considerable complexity, the task has so far been addressed through a pipeline of components that separately deal with transcribing, translating, and segmenting text into subtitles, as well as predicting timestamps. In this paper, we propose the first direct speech translation model for automatic subtitling that generates subtitles in the target language along with their timestamps with a single model. Our experiments on 7 language pairs show that our approach outperforms a cascade system in the same data condition, also being competitive with production tools on both in-domain and newly released out-domain benchmarks covering new scenarios.

1 Introduction

With the growth of websites and streaming platforms such as YouTube and Netflix,[1] the amount of audiovisual content available online has dramatically increased. Suffice to say that the number of hours of Netflix original content has increased by 2,400% from 2014 to 2019.[2] This phenomenon has led to a huge demand for subtitles, which is becoming more and more difficult to satisfy only with human resources. Consequently, automatic subtitling tools are spreading to reduce subtitlers' workload by providing them with suggested subtitles to be post-edited (Álvarez et al., 2015; Vitikainen and Koponen, 2021). In general, subtitles can be either *intralingual* (hereinafter *captions*), if source audio and subtitle text are in the same language, or *interlingual* (hereinafter *subtitles*), if the text is in a different language. In this paper, we focus on automatizing interlingual subtitling, framing it as a speech translation (ST) for subtitling problem.

Differently from ST, in automatic subtitling the generated text has to comply with multiple requirements related to its length, format, and the time it should be displayed on the screen (Cintas and Remael, 2021). These requirements, which depend on the type of video content and target language, are dictated by the need to keep users' cognitive effort as low as possible while maximizing comprehension and engagement (Perego, 2008; Szarkowska and Gerber-Morón, 2018). This often leads to a condensation of the original spoken content, aimed at reducing the time required for reading subtitles while increasing that of watching the video (Burnham et al., 2008; Szarkowska et al., 2016).

Being such a complex task, automatic subtitling has so far been addressed by dividing the process into different steps (Piperidis et al., 2004; Melero et al., 2006; Matusov et al., 2019; Koponen et al., 2020; Bojar et al., 2021): automatic speech recognition (ASR), timestamp extraction from audio, segmentation into captions, and their machine translation (MT) into the final subtitles. More recently, drawing from the evidence that direct models achieve competitive quality with cascade architectures (Ansari et al., 2020), Karakanta et al.

* Work done while at FBK.

[1]https://www.insiderintelligence.com/insights/ott-video-streaming-services/.

[2]https://www.statista.com/statistics/882490/netflix-original-content-hours/.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1355–1376, 2023. https://doi.org/10.1162/tacl_a_00607
Action Editor: George Foster. Submission batch: 3/2023; Revision batch: 7/2023; Published 11/2023.

(2020a) proposed an ST system that jointly translates and segments into subtitles, arguing that direct models are able to better exploit speech cues and prosody in subtitle segmentation. However, their system does not generate timestamps, hence missing a critical aspect to reach the goal of fully automatic subtitling. Furthermore, the current lack of benchmarks hinders a thorough evaluation of the technologies developed for automatic subtitling. In fact, the only corpus publicly available to date is MuST-Cinema (Karakanta et al., 2020b), which contains only single-speaker audio in the TED-talks domain with verbatim translations.

To fill these gaps, this paper presents the first automatic subtitling system that performs the whole task with a single direct ST model, and introduces two new benchmarks. Our contributions can be summarized as follows:

- We propose the first direct ST model for automatic subtitling able to produce both subtitles and timestamps. Code and pre-trained models are released under the Apache License 2.0 at: `https://github.com/hlt-mt/FBK-fairseq/`;
- We introduce two (en→{de, es}) benchmarks for automatic subtitling, covering new domains, news/documentaries and interviews, with the presence of background noise and multiple speakers. We release them under the CC BY-NC 4.0 license at: `https://mt.fbk.eu/ec-short-clips/` and `https://mt.fbk.eu/europarl-interviews/`;
- We conduct the first extensive comparison between automatic subtitling systems based on cascade and direct ST models on all 7 language pairs of MuST-Cinema (en→{de, es, fr, it, nl, pt, ro}), showing the superiority of our direct solution, while also demonstrating its competitiveness with production systems on both MuST-Cinema and out-of-domain benchmarks.

2 Background

2.1 Direct Speech Translation

While the first cascaded approach to ST was proposed decades ago (Stentiford and Steer, 1988; Waibel et al., 1991), direct models[3] have recently become increasingly popular (Bérard et al., 2016; Weiss et al., 2017) due to their ability to avoid error propagation (Sperber and Paulik, 2020), their superior exploitation of prosody and better audio comprehension (Bentivogli et al., 2021), and their lower computational cost (Weller et al., 2021). Motivated by these advantages, direct models are rapidly evolving and their initial performance gap with cascade architectures (Niehues et al., 2019) has been significantly reduced, leading to a substantial parity in the latest IWSLT campaigns (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022). Such improvements can be partly attributed to the development of specialized architectures for speech processing (Chang et al., 2020; Papi et al., 2021; Burchi and Vielzeuf, 2021; Kim et al., 2022; Andrusenko et al., 2022), which are all variants of a Transformer model (Vaswani et al., 2017) preceded by convolutional layers that reduce the length of the input sequence (Bérard et al., 2018; Di Gangi et al., 2019). Among them, Conformer (Gulati et al., 2020) is currently the best-performing model in ST (Inaguma et al., 2021). For this reason, we build our systems with this architecture and test, for the first time, its effectiveness in the challenging task of fully automatic subtitling.

2.2 Subtitling Requirements

Subtitles are short pieces of timed text, generally displayed at the bottom of the screen, which describe, transcribe, or translate the dialogue or narrative. A subtitle is composed of two elements: the text, shown into ‘‘blocks’’, and the corresponding start and end display time (or timestamps).[4]

Depending on the subtitle provider and the audiovisual content, different requirements have to be respected concerning both the text space and its timing. These constraints typically consist in: *i)* using at most two lines per block; *ii)* keeping linguistic units (e.g., noun and verb phrases) in the same line; *iii)* not exceeding a pre-defined number of characters per line (CPL), spaces included;

[3]According to the official IWSLT definition (`https://iwslt.org/2023/offline`), a direct model is a system that does not use intermediate discrete representations to generate the outputs from audio segments and whose parameters used during decoding are all trained altogether on the ST task, while it does not consider the audio segmentation.

[4]The most widespread subtitle format is SubRip or srt.

iv) not exceeding a pre-defined reading speed, measured in number of characters per second (CPS). While a typical value used as maximum CPL threshold is 42 for most Latin languages,[5] there is no agreement on the maximum CPS allowed. For instance, Netflix guidelines[6] allow up to 17 CPS for adult and 15 for children programs, TED guidelines[7] up to 21 CPS, and Amara guidelines[8] up to 25 CPS.

To convey the meaning of the audiovisual product while adhering to time and space constraints, in some domains and scenarios subtitles require compression or condensation (Kruger, 2001; Gottlieb, 2004; Aziz et al., 2012; Liu et al., 2020a; Buet and Yvon, 2021). Due to the rehearsed nature of TED talks, the subtitles in MuST-Cinema have a limited degree of condensation, and the translation is mostly verbatim. In addition, the audio conditions (no background noise and a single speaker) are not representative of all the diverse contexts where subtitling is applied, such as news and movies. To fill this gap, we introduce two new benchmarks that feature different domains, scenarios (e.g., multiple speakers), and levels of subtitle condensation.

2.3 Automatic Subtitling

Attempts to (semi-)automatize the subtitling process have been done with cascade systems made of an ASR, a segmenter, and an MT model. Most work focused on adapting the MT module to subtitling with the goal of producing shorter and compressed texts. This has been performed either using statistical approaches trained on subtitling corpora (Volk et al., 2010; Etchegoyhen et al., 2014; Bywood et al., 2013) or by developing specifically tailored decoding solutions on statistical (Aziz et al., 2012) and neural models (Matusov et al., 2019). In particular, recent research efforts focused on controlling the MT output length so as to satisfy isometric requirements between source transcripts and target translations (Lakew et al., 2019; Matusov et al., 2020; Lakew et al., 2021, 2022). In addition, several research groups (Öktem et al., 2019; Federico et al., 2020; Virkar et al., 2021; Tam et al., 2022; Effendi et al., 2022) proved the usefulness of injecting prosody information about speech cues, such as pauses, in determining subtitle boundaries. Given the possibility for direct ST systems to access this information and their advantages mentioned in §2.1, Karakanta et al. (2020a, 2021) built the only (to the best of our knowledge) automatic subtitling system using a direct ST model, confirming with their results that the ability of direct ST systems to leverage prosody has particular importance for subtitle segmentation. However, their solution only covers the translation and segmentation into subtitles, neglecting the timestamp generation. Our study is hence the first to complete the entire subtitling process with a direct ST model and to evaluate its performance on all aspects of the subtitling task.

3 Direct Speech Translation for Subtitling

Motivated by all the advantages discussed in §2.1 and §2.3, we build the first automatic subtitling system solely based on a direct ST model (Figure 1). Our system works as follows: *i)* the audio is fed to a *Subtitle Generator* (§3.1) that produces the (untimed) subtitle blocks; *ii)* the computed encoder representations are passed to the *Source Timestamp Generator* (§3.2) to obtain the caption blocks and their corresponding timestamps; *iii)* the subtitle timestamps are estimated by the *Source-to-Target Timestamp Projection* (§3.3) from the generated subtitles, captions, and source timestamps. These modules are described in the rest of this section.

3.1 Subtitle Generation

We train a direct ST Conformer-based model that jointly performs the ST task and the segmentation of the generated translation into (untimed) subtitle blocks and lines. To this end, we add two special tokens to the vocabulary of our system, `<eob>` and `<eol>`, which respectively represent the end of a subtitle block and the end of a line within a block. Both at training and inference time, `<eob>` and `<eol>` are treated as any other token, without giving them different weights, or adding specific loss. Additionally, we do not incorporate losses aimed at minimizing the number

[5] https://www.ted.com/participate/translate/subtitling-tips.

[6] https://partnerhelp.netflixstudios.com/en-us/articles/219375728-Timed-Text-Style-Guide-Subtitle-Templates.

[7] https://www.ted.com/participate/translate/subtitling-tips.

[8] https://blog.amara.org/2020/10/22/create-quality-subtitles-in-a-few-simple-steps/.

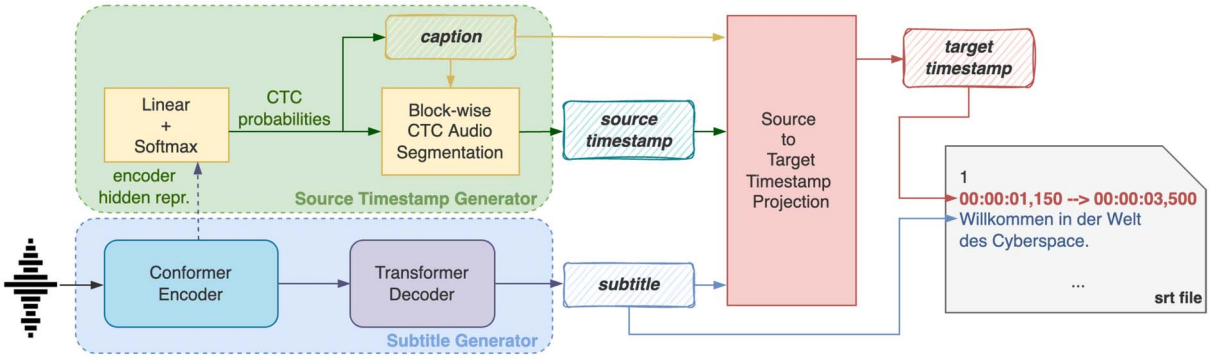

Figure 1: Architecture of the direct ST system for automatic subtitling.

of generated characters or explicitly optimizing for CPL and CPS compliance.

3.2 Source Timestamp Generation

Estimating timestamps for the generated subtitle blocks from source audio is a challenging task. Current sequence-to-sequence models, in fact, generate target sequences that are decoupled from the input and, therefore, their tokens do not have a clear relationship with the frames they correspond to. To recover this relationship, we start from the observation that direct ST models are often trained with an auxiliary Connectionist Temporal Classification or CTC loss (Graves et al., 2006) in the encoder to improve model convergence (Kim et al., 2017; Bahar et al., 2019). The CTC maps the input frames to the transcripts—in our use case, captions—and we propose to leverage this CTC module at inference time to estimate the block timestamps.

In particular, the encoder representations computed during the forward pass are fed to the CTC module that provides the frame-level probability distribution over the source vocabulary tokens (including `<eob>`, `<eol>`, and the additional CTC *blank* token). This sequence of CTC probabilities over the source vocabulary serves two purposes. First, it is used to predict the caption with the CTC beam search algorithm (Graves and Jaitly, 2014).[9] Second, it is fed, together with the generated caption, to the CTC-based segmentation algorithm (Kürzinger et al., 2020), whose task is to find the most likely alignment between caption tokens and audio frames. The algorithm builds a trellis over the time steps for the generated tokens and, at each time step, only three paths are possible: *i)* staying at the same token (self-loop); *ii)* moving to the *blank* token; *iii)* moving to the next token. To avoid forcing the caption to start at the beginning of the audio, the transition cost for staying at the first token is set to 0. Otherwise, the transition cost is the CTC-predicted probability for a given token in that time step. The trellis is then backtracked from the time step with the highest probability in the last token of the generated caption, until the first token is reached. In our case, since we are interested in the timestamps of the subtitle blocks, we extract block-wise alignments that correspond to the start and the end time of each block. This means finding the time in which the first word of each subtitle is pronounced and the time in which the corresponding `<eob>` symbol is emitted by using the aforementioned algorithm.

3.3 Source-to-Target Timestamp Projection

After generating the untimed subtitles (§3.1), and captions with their timestamps (§3.2), the next step is to obtain the timestamps for subtitle blocks on the target side. In general, caption and subtitle segmentations may differ for many reasons (e.g., due to different syntactic patterns between languages) and imposing the caption segmentation on the subtitle side—as done in most cascade approaches (Georgakopoulou, 2019; Koponen et al., 2020)—could be a sub-optimal solution. For this reason, we introduce a caption-subtitle alignment module that projects the source timestamps to the target blocks. To perform this task, we tested the three alternative methods described below.

Block-Wise Projection (BWP) This method operates at character level to project the predicted

[9]We also tested a greedy decoding, in which the most likely label for each time step is chosen to obtain the output sequence. However, this approach did not prove effective.

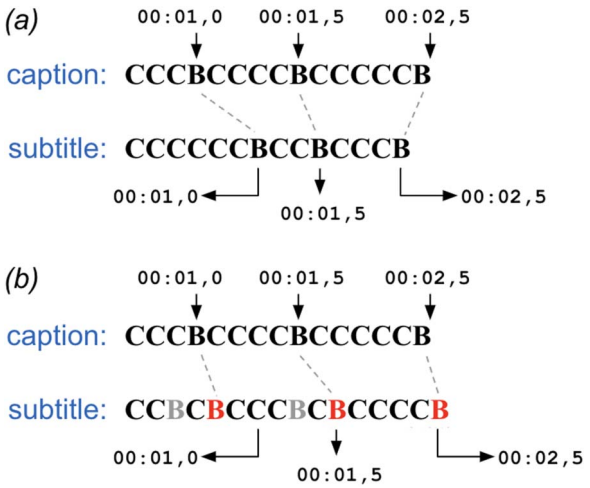

Figure 2: Example of BWP projection with *(a)* same number of blocks and *(b)* different number of blocks between caption and subtitle.

source-side (captions) timestamps on the target side (subtitles) without alterations. When the number of caption and subtitle blocks is equal, a condition that occurs in ∼80% of the cases, the timestamps of each caption block are directly assigned to the corresponding subtitle block.[10] This process is depicted in Figure 2.a, in which ''C'' and ''B'' respectively stand for characters and blocks in the caption and subtitle. When the number of caption and subtitle blocks is different (Figure 2.b), the target segmentation is discarded and replaced with the caption segmentation. In this case, line and block boundaries (<eol>/<eob>) are inserted in the target side by matching the number of characters each line/block has in the caption. If the insertion falls in the middle of a word, the <eol>/<eob> is appended to the word. This approach has two main weaknesses. First, it assumes that, when captions and subtitles have the same number of blocks, these blocks contain the same linguistic content, although this is not guaranteed. Second, it ignores the subtitle segmentation in ∼20% of the cases.

Levenshtein-based Projection (LEV) To overcome the above limitations, our second method exploits the Levenshtein distance-based alignment (Levenshtein, 1966) between captions and subtitles. This method estimates the target-side

[10]Selecting the candidates with the closest number of blocks among the source and target n-best lists had negligible effects.

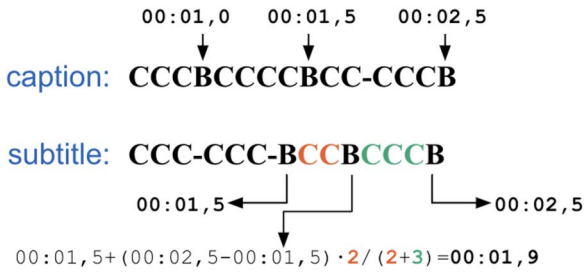

Figure 3: Example of Levenshtein-based projection.

timestamps from the source-side timestamps without ever altering the original target-side segmentation. First, all the non-block characters are masked with a single symbol (''C''). For instance, ''*This is a block*'' <eob> is converted into ''CC-CCCCCCCCCCCCCB'', where ''B'' stands for <eob>. Then, the masked caption and subtitle are aligned with the weighted version of Levenshtein distance, in which the substitution operation is forbidden so as to avoid the replacement of a character with a block and vice versa. If the positions of a block in the aligned caption and subtitle match, its caption timestamp is directly assigned to the subtitle block. If they do not match, the timestamps of the subtitle blocks are estimated from the caption timestamps based on the alignment of ''B''s and the number of characters. For instance, given the caption ''CCCBCCCCBCCCCCB'' and the subtitle ''CCCCCCBCCBCCCB'', the optimal source-targ et alignment with the corresponding timestamp calculation is shown in Figure 3. In detail, the first subtitle block (CCC-CCC-B) is matched with the first two caption blocks (CC-CBCCCCB) and the corresponding timestamp (00:01,5) is directly mapped. This also happens with the timestamp 00:02,5 of the last caption (BCC-CCCB) and subtitle block (CCCB). For the second subtitle block (CCB), the timestamp (00:01,9) is estimated proportionally from the caption (BCC-CCCB) using the character ratio between the orange block and the orange + green blocks.

Semantic-based Projection (SEM) The third method projects the predicted source-side timestamps on target blocks by looking at the semantic content of the generated captions and subtitles. The method is based on SimAlign (Jalili Sabet et al., 2020), which combines semantic embeddings from fastText (Bojanowski et al., 2017),

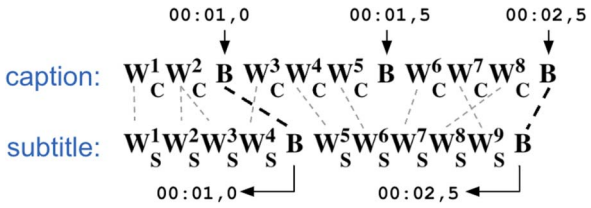

Figure 4: Example of Semantic-based projection.

Dataset	de	es	fr	it	nl	pt	ro
MuST-Cinema	388	479	469	441	421	364	410
Europarl-ST	75	74	–	–	–	–	–
CoVoST2	412	412	–	–	–	–	–
CommonVoice	885	885	–	–	–	–	–
TEDlium	444	444	–	–	–	–	–
VoxPopuli	519	519	–	–	–	–	–

Table 1: Number of hours of the training sets.

VecMap (Artetxe et al., 2018), mBERT,[11] and XLM-RoBERTa (Conneau et al., 2020) to align source and target texts at the word level. Specifically, we first align captions and subtitles word by word (<eol>/<eob> included) with SimAlign. Then, when all <eob>s of a subtitle are aligned with <eob>s in the caption (66% of the cases), we assign the corresponding timestamp (Figure 4). Otherwise, i.e., when at least one <eob> in the subtitle is aligned with a caption word or <eol> or is not aligned at all, one of the two previous methods is applied as a fallback solution.

4 Data

4.1 Training Data

For the comparison between cascade and direct architectures (§6.2), we train the models in a controlled and easily reproducible data setting by using MuST-Cinema v1.1, the only publicly available subtitling corpus also containing the source speech. It covers one general domain (TED talks), and 7 language pairs, namely, en→{de, es, fr, it, nl, pt, ro}. The number of hours in the training set of each language pair is shown in the first row of Table 1.

For the comparison with production tools (§6.3), we experiment in a more realistic unconstrained data scenario and we focus on en→de, and en→es.[12] For training, we use MuST-Cinema, two ST datasets (Europarl-ST [Iranzo-Sánchez et al., 2020] and CoVoST2 [Wang et al., 2020b]) and three ASR datasets (CommonVoice [Ardila et al., 2020], TEDlium [Hernandez et al., 2018], and VoxPopuli [Wang et al., 2021]). We translate the ASR corpora with the Helsinki-NLP MT models (Tiedemann and Thottingal, 2020) and filter out data with a very high or low transcript/translation character ratio, as per Gaido et al. (2022). The use of automatic translations as targets, also known as sequence-level knowledge distillation (Kim and Rush, 2016), is a popular data augmentation method used in the most recent IWSLT evaluation campaigns (Anastasopoulos et al., 2021, 2022) to enhance the performance of ST systems. Since none of the training sets, except for MuST-Cinema, includes the subtitle boundaries (<eob> and <eol>) in the target translation, we automatically insert them by employing the publicly released multimodal and multilingual segmenter by Papi et al. (2022). The segmenter takes the source audio and the unsegmented text as input and outputs the segmented text, i.e., containing <eob> and <eol>. By doing this, we can train our system to jointly translate from speech and segment into subtitles without the need for manually-curated subtitle targets, which are hard to find and costly to create. The number of training hours is reported in Table 1.

4.2 Test Data

The models are tested in both in-domain and out-of-domain conditions. For in-domain experiments, we use the MuST-Cinema test set, for which we adopt both the original audio segmentation (for reproducibility and for the sake of comparison with previous and future work) and more realistic automatic segmentation obtained with SHAS (Tsiamas et al., 2022). Notice that this audio segmentation is a completely different task from determining subtitle boundaries. Its only goal is splitting long audio files into smaller chunks (or utterances) that can be processed by ST systems, limiting performance degradation due to information loss caused by sub-optimal splits (e.g., in the middle of a sentence). In general, each resulting utterance contains multiple subtitle blocks. For instance, in the MuST-Cinema

[11] https://github.com/google-research/bert/blob/master/multilingual.md.

[12] We select these two language pairs due to, respectively, a different and similar word ordering with respect to the source.

training set there are ∼2.5 blocks per utterance, even though utterances are quite short (6.4s on average). When automatic segmentation methods like SHAS are applied, this ratio significantly increases, as audio segments are typically much longer, with many segments lasting between 14 and 20 seconds (Gaido et al., 2021b; Tsiamas et al., 2022).

For out-of-domain evaluations, we introduce the two new (en→{de, es}) test sets described below, which we also segment with SHAS.

EC Short Clips The first test set is composed of short videos from the Audiovisual Service of the European Commission (EC)[13] recorded between 2016 and 2022. These informative clips have an average duration of 2 minutes and cover various topics discussed in EC debates such as economy, environment, and international rights. This benchmark presents several additional difficulties compared to TED talks since the videos often contain multiple speakers, and background music is sometimes present during the speech. We selected the videos with the highest subtitle conformity (at least 80% of the subtitles conforming to 42 CPL, and 75% conforming to 21 CPS), and removed subtitles describing on-screen text. This resulted in 27 videos having a total duration of 1 hour. The target srt files contain ∼5,000 words per language.

EuroParl Interviews The second test set is compiled from publicly available video interviews from the European Parliament TV[14] (2009–2015). We selected 12 videos of 1 hour total duration, amounting to ∼6,500 words per target language. The videos present multiple speakers and sometimes contain short interposed clips with news or narratives. Apart from the more challenging source audio properties compared to the clean single-speaker TED talks, here the target subtitles are not verbatim and demonstrate a high degree of compression and reduction. As a consequence, the CPL and CPS conformity is very high (∼100%) but this comes at the cost of being more difficult for automatic systems to perfectly match the non-verbatim translations. Nonetheless, to achieve real progress in automatic subtitling, it is particularly relevant to evaluate automatic systems on realistic and challenging benchmarks like the ones we provide.

[13] https://audiovisual.ec.europa.eu/.

[14] https://www.europarltv.europa.eu/.

5 Experimental Settings

5.1 Training Settings

Our systems are implemented on Fairseq-ST (Wang et al., 2020a), following the default settings unless stated otherwise. The input is represented by 80 audio features extracted every 10ms with sample window of 25 and pre-processed by two 1D convolutional layers with stride 2 to reduce the input length by a factor of 4. All segments longer than $30s$ in the training set are filtered out to speed up training. The models are based on encoder-decoder architectures and composed by a stack of 12 Conformer encoder layers and 8 Transformer decoder layers. We apply CTC loss to the 8^{th} encoder layer and use its predictions to compress the input sequences to reduce RAM consumption (Liu et al., 2020b; Gaido et al., 2021a). Both the Conformer and Transformer layers have a 512 embedding dimension and 2,048 hidden units in the linear layer. We set dropout to 0.1 in the linear, attention, and convolutional modules. In the convolutional modules, we also set a kernel size of 31 for the point- and depth-wise convolutions.

For the comparison between cascade and direct architectures, we train a one-to-many multilingual ST model that prepends a token representing the selected target language for decoding (Inaguma et al., 2019) on all the 7 languages of MuST-Cinema. Conversely, for the comparison with production tools, we develop a dedicated ST model for each target language (de, es). For inference, we set the beam size to 5 for both subtitles and captions.

We train with the Adam optimizer (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$) for 100,000 steps. The learning rate increases linearly up to 0.002 for the first 25,000 warm-up steps and then decays with an inverse square root policy, apart from fine-tunings, where it is the fixed value 0.001. Utterance-level Cepstral Mean and Variance Normalization (CMVN) and SpecAugment (Park et al., 2019) are applied during training, as per Fairseq-ST default settings. The vocabularies are based on SentencePiece models (Sennrich et al., 2016) with size 8,000 for the source language. For the multilingual model trained on MuST-Cinema, a shared vocabulary is built with

System	Num. params
Direct	124.6M
Cascade	341.9M
- ASR	116.4M
- Audio forced aligner	9.7M
- Segmenter	40.6M
- Multilingual MT	175.2M

Table 2: Number of parameters for the direct (both multilingual and monolingual) and cascade systems.

a size of 16,000; for the two models developed to compare with production tools, we build German and Spanish vocabularies with a size of 16,000 subwords each. The ASR of our cascade model is trained using the same source language vocabulary of size 8,000 used in the translation setting. The MT model is trained using the standard hyper-parameters of the Fairseq multilingual MT task (Ott et al., 2019), with the same source and target vocabularies of the ST task.

For all models, we stop the training when the validation loss does not improve for 10 epochs and the final models are obtained by averaging 7 checkpoints (the best, the 3 preceding, and the 3 succeeding). Training is performed on 4 NVIDIA A100 (40GB RAM), with 40k max tokens per mini-batch and an update frequency of 2, except for the MT models for which 8 NVIDIA K80 (12GB RAM) are used with 4k max tokens and an update frequency of 1. Table 2 lists the total number of parameters of our direct models, showing that it is ∼1/3 of the cascade system used as a term of comparison.

5.2 Terms of Comparison

We compare our direct ST system both with a cascade pipeline trained under the same data conditions and with production tools.

Cascade We build an in-domain cascade composed of an ASR, an audio forced aligner, a segmenter, and an MT system. The ASR has the same architecture as our ST system (Conformer encoder + Transformer decoder), and it is trained on MuST-Cinema transcripts without <eob> and <eol>. The audio forced aligner used to estimate the timestamps (Gretter et al., 2021) is based on the Kaldi[15] acoustic model. The subtitle segmenter is the same multimodal segmenter we used to segment the training data for the direct system (§4.1). The MT is a multilingual model trained on the MuST-Cinema (*transcript*, *translation*) pairs without <eob> and <eol>. The pipeline works as follows. The audio is first transcribed by the ASR and word-level timestamps are estimated with the forced aligner. Then, the transcript is segmented into captions with the segmenter and each block timestamp is obtained by averaging the end time of the word before an <eob> and the start time of the word after it. The segmented text is then split into sentences according to the <eob> and, finally, these sentences are translated by the MT. The <eob>s are automatically re-inserted at the end of each sentence while <eol>s are added to the subtitle translation using the same segmenter.

Production Tools As a term of comparison for the unconstrained data condition, we use production tools for automatic subtitling. These tools take audio or video content as input and return the subtitles in various formats, including srt. We test three online tools,[16] namely: MateSub,[17] Sonix,[18] and Zeemo.[19] We also compare with the AppTek subtitling system,[20] a cascade architecture whose ASR component is equipped with a neural model that predicts the subtitle boundaries before feeding the transcripts to the MT component (Matusov et al., 2019). For this system, two variants of the MT model are evaluated: a standard model and a model specifically trained to obtain shorter translations in order to better conform to length requirements (Matusov et al., 2020). Since we are not interested in comparing the tools with each other, all system scores are anonymized.

5.3 Evaluation

Translation quality, timing, and segmentation of subtitles are measured with multiple metrics. First, we compute SubER (Wilken et al., 2022),[21] a tailored TER-based metric (the lower, the better) that scores the overall subtitle quality by

[15] https://github.com/kaldi-asr/kaldi.
[16] All outputs were collected in August 2022.
[17] https://matesub.com/.
[18] https://sonix.ai.
[19] https://zeemo.ai/.
[20] https://www.apptek.com/.
[21] Version 0.2.0.

	en-de				en-es			
Model	SubER (↓)	Sigma (↑)	CPL (↑)	CPS (↑)	SubER (↓)	Sigma (↑)	CPL (↑)	CPS (↑)
Gold audio segmentation								
Baseline	63.5	65.6	77.7	64.4	52.0	70.4	80.7	68.0
BWP	60.8	75.6	86.1	64.0	48.6	78.5	90.9	66.9
LEV	**58.7**	**78.8**	**88.8**	**65.4**	**46.7**	**81.1**	93.9	**68.4**
SEM	60.7	75.5	88.6	63.7	48.6	78.8	**94.0**	65.5
Automatic audio segmentation								
Baseline	66.9	62.0	78.2	70.5	55.7	66.0	79.9	75.1
BWP	62.8	73.3	86.2	70.3	51.8	75.9	89.6	73.5
LEV	**60.3**	**78.5**	**88.9**	**72.1**	**48.5**	**80.6**	**94.2**	**76.1**
SEM	62.8	75.8	**88.9**	69.7	51.4	78.3	**94.2**	72.9

Table 3: Comparison of timestamp projection methods on the MuST-Cinema en→{de, es} test set.

considering translation, segmentation, and timing altogether. We adopt the cased and with punctuation version of the metric since these aspects are crucial for the quality and comprehension of the subtitles. Next, specifically for translation quality, we use SacreBLEU (Post, 2018)[22] on texts from which <eol> and <eob> have been removed. The quality of segmentation into subtitles is evaluated with Sigma from the EvalSub toolkit (Karakanta et al., 2022). Since BLEU and Sigma require the same audio segmentation between reference and predicted subtitles, we re-align the predictions in case of non-perfect alignment with the mWERSegmenter (Matusov et al., 2005). Lastly, to check the spatio-temporal compliance described in §2.2, we compute CPL conformity as the percentage of lines not exceeding 42 characters, and CPS conformity as the percentage of subtitle blocks having a maximum reading speed of 21 characters per second.[23] Confidence intervals (CIs) are computed with bootstrap resampling (Koehn, 2004).

6 Results

In this section, we first (§6.1) choose the best timestamp projection method among those introduced in §3.3. Then (§6.2), we compare the cascade and direct approaches trained in the same data conditions. Lastly (§6.3), we show that our direct model, even though trained in laboratory settings, is competitive with production tools. In addition, in Appendix A, we analyze the performance of the CTC-segmentation algorithm for timestamp estimation compared to forced aligner tools.

6.1 Timestamp Projection

The quality of source-to-target timestamp projection (§3) is crucial to correctly estimate the target-side timestamps and, in turn, to produce good subtitles. To select the best strategy, we compare the methods in §3.3 using the constrained model on the MuST-Cinema test sets for en→{de, es}. To test the robustness of the various methods when gold-segmented audio is not available, we also report the results using the automatic audio segmentation in addition to that obtained using the gold one.

Results are shown in Table 3. BLEU is not reported because the translated text is always the same, regardless of the timestamp projection method. We also report, as a baseline, a method that completely ignores the target segmentation and always maps the caption segmentation onto the subtitle as in BWP when the number of caption and subtitle blocks is different (§3.3). For the SEM method, if the source-targ et alignment is not found by SimAlign, the LEV method is applied instead.[24]

The results highlight the superiority of the LEV method, which outperforms the others on almost

[22] case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1.

[23] We used version 1.1 of the script adopted for the IWSLT subtitling task (https://iwslt.org/2023/subtitling): https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/subtitle_compliance.py.

[24] We also applied the baseline and the BWP method as a fallback method for SEM but it led to worse results.

all metrics, with similar trends for both language pairs. The gap is more marked in the realistic scenario of automatically-segmented audio, likely due to the fact that the audio segments produced by SHAS are longer than the manually annotated ones (8.6s vs 5.5s). As such, each audio segment contains more blocks to align, so the difference between the methods emerges more clearly. The low scores obtained by the baseline confirm that the caption segmentation is not optimal for the target language. Furthermore, SEM yields results that are either comparable to or slightly better than those obtained by BWP, especially in terms of Sigma and CPL, while being always worse than LEV. In addition, SEM exhibits lower CPS conformity even compared to the baseline. Consequently, its performance suggests that semantically-motivated approaches are not the best solution for timestamp projection.

Focusing on the LEV method, we observe that segmentation quality (higher Sigma) and overall subtitle quality (lower SubER) are slightly better when the gold segmentation is used, as expected. Conversely, CPS conformity is higher with the automatic audio segmentation. This counter-intuitive result can be explained as follows: audio segmentation not only splits but sometimes also cuts the audio according to speakers' pauses, while the manual segmentation delimits speech boundaries more aggressively than the automatic one. In our case, manual segmentation results in audio segments that are about 2% shorter than those obtained with the automatic segmentation, thus ''forcing'' the generated subtitles to appear on screen for a shorter time, which in turn leads to a higher reading speed.

6.2 Cascade vs. Direct

After selecting LEV as our best timestamp projection method, we evaluate cascade and direct ST systems trained in the same data condition. Before this, to ensure the competitiveness of our cascade baseline, we compare it with the results obtained on the MuST-Cinema test set by the other cascade systems presented in literature, namely: en→{de, fr} by Karakanta et al. (2021), and en→fr by Xu et al. (2022). As these works report only BLEU with breaks, i.e., BLEU computed including also <eob> and <eol>, we compare our cascade baseline with them on that metric.[25] Although these studies leverage large additional training corpora for both ASR (e.g., LibriSpeech [Panayotov et al., 2015]) and MT (e.g., OPUS [Tiedemann 2016] and WMT-14 [Bojar et al. 2014]), our cascade trained only on MuST-Cinema performs on par with them. It scores 20.2 on German and 26.2 on French, which are similar, or even better than, respectively, the results reported in Karakanta et al. (2021) (19.9 and 26.9, respectively), and the result of 25.8 on French from Xu et al. (2022). These results confirm the strength of our baselines, and the soundness of our experimental settings.

Table 4 reports the scores of the constrained direct and cascade models. The overall subtitle quality of the direct solution is significantly higher compared to that of the cascade on all language pairs, with a SubER decrease of 3.8–5.5 points, corresponding to an ∼8% improvement on average. Since SubER measures translation, segmentation, and timestamp quality altogether, to disentangle the contribution of each of these aspects we leverage the other metrics. The higher Sigma of our system (+1.2 average improvement) demonstrates that the joint generation of subtitle content and boundaries results in superior segmentation. This finding corroborates previous research on the value of prosody (see §2.3), and the ineffectiveness of projecting caption segmentation onto subtitles, as done by cascade approaches (Georgakopoulou, 2019; Koponen et al., 2020). The sub-optimal placement of block boundaries in the cascade system can also account for the superior translation quality of our method (+3.9 BLEU average improvement): As the MT component translates the caption block-by-block, inaccurate boundaries can impede access to information required for proper translation.

Looking at the conformity metrics, the direct system complies with the length requirement of 42 characters (CPL) in almost 90% of cases while the cascade system does so in only 78.1% of cases. This difference is explained by the higher number of <eol> generated by the direct model (10–15% more than the cascade), although being still lower than that of the reference (8–10% less). According to the statistics computed on the outputs of the two systems, the cascade does not

[25] <eob> and <eol> are considered as a single token and replaced, respectively, with § and μ as in the EvalSub toolkit.

Sys.	en-de	en-es	en-fr	en-it	en-nl	en-pt	en-ro	Avg.
	SubER (↓)							
Casc.	64.2 (64.2±2.4)	50.5 (50.5±2.2)	57.0 (57.0±1.8)	54.2 (54.2±1.7)	52.8 (52.8±1.8)	49.7 (49.7±1.7)	52.7 (52.7±2.0)	54.4
Dir.	**58.7** (58.7±2.3)	**46.7** (46.7±2.1)	**52.9** (52.9±1.7)	**50.4** (50.4±1.7)	**47.4** (47.4±1.9)	**44.6** (44.6±1.7)	**48.5** (48.5±2.1)	**49.9**
	BLEU (↑)							
Casc.	18.9 (18.9±1.4)	32.4 (32.4±1.8)	25.1 (25.1±1.5)	26.0 (26.0±1.6)	25.8 (25.8±1.5)	31.4 (31.4±1.7)	28.4 (28.3±1.6)	26.9
Dir.	**22.1** (22.1±1.6)	**35.9** (35.8±1.9)	**28.0** (28.0±1.6)	**29.6** (29.6±1.8)	**31.6** (31.6±1.8)	**36.8** (36.7±1.7)	**31.9** (31.8±1.8)	**30.8**
	Sigma (↑)							
Casc.	**79.5** (79.5±2.0)	80.9 (80.9±1.5)	84.0 (84.0±1.7)	83.8 (83.8±1.6)	77.5 (77.4±1.8)	81.2 (81.2±1.7)	**86.4** (86.4±1.5)	81.9
Dir.	78.8 (78.8±2.0)	**81.1** (81.1±1.5)	**84.1** (84.1±1.7)	**85.1** (85.1±1.5)	**83.1** (83.1±1.6)	**84.5** (84.4±1.4)	85.3 (85.3±1.4)	**83.1**
	CPL (↑)							
Casc.	81.8 (81.8±1.9)	83.4 (83.3±1.8)	85.2 (85.2±1.7)	81.4 (81.4±1.9)	83.3 (83.2±1.9)	78.1 (78.1±2.0)	53.3 (53.3±3.0)	78.1
Dir.	**88.9** (88.9±1.5)	**94.0** (94.0±1.1)	**91.9** (91.9±1.2)	**89.3** (89.2±1.5)	**84.0** (84.0±1.8)	**88.2** (88.2±1.5)	**92.1** (92.1±1.2)	**89.8**
	CPS (↑)							
Casc.	**69.1** (69.1±2.6)	**74.0** (73.9±2.7)	**64.3** (64.3±2.9)	**71.2** (71.2±2.8)	**74.4** (74.4±2.5)	**74.7** (74.7±2.6)	**76.2** (76.2±2.4)	**72.0**
Dir.	65.4 (65.4±2.7)	68.4 (68.3±2.7)	60.7 (60.8±2.8)	67.9 (67.9±2.6)	72.2 (72.2±2.6)	71.9 (71.8±2.7)	76.0 (75.9±2.4)	68.9

Table 4: Cascade (Casc.) and direct (Dir.) results on all MuST-Cinema language pairs with 95% CI in parentheses.

only have a higher average number of characters per line (32 vs. 29), but its variance is 1.5–2 times greater, with lines sometimes close to or even longer than 100 characters on all language pairs. In contrast, most of the CPL violations of the direct system are caused by lines shorter than 60 characters, and lines never exceed 70 characters. The trend for CPS is instead different, since the cascade generates subtitles with a higher conformity to the 21-CPS reading speed (72.0 vs 68.9). This can be partially explained by looking at the generated timestamps: Upon a manual inspection of 100 subtitles, we noticed that the direct model tends to assign the start times of the subtitles slightly after those of the cascade (within 100ms of difference), and end times slightly before those of the cascade (mostly within 200ms). Overall, on the MuST-Cinema test sets, this leads to a total of ∼2,940s with subtitles on screen for the cascade, and ∼2,850s for the direct (∼3% lower).

To sum up, our direct system proves to be the best choice to address the automatic subtitling task in the constrained data condition, reaching better translation quality and more well-formed subtitles. Our results also indicate that improving the reading speed of the generated subtitles is one of the main aspects on which to focus future works.

6.3 Comparison with Production Tools

To test our approach in more realistic conditions, we train our models on several openly available corpora (unconstrained condition) and compare them with production tools, which represent very challenging competitors as they can leverage large proprietary datasets. We focus on two language pairs (en→{de, es}) for both the in-domain MuST-Cinema, and on the two out-of-domain EC Short Clips and EuroParl Interviews test sets. We feed all systems with the full test audio clips, so each system has to segment its audio. Only in the case of EC Short Clips and EuroParl Interviews do we clean the audio using Veed[26] before processing it, for the sake of a fair comparison with production tools that have similar procedures.[27] The impact of audio cleaning is analyzed in Appendix B.

MuST-Cinema The results of the unconstrained models on the in-domain MuST-Cinema test set are shown in Table 5. Compared to production tools, our system shows better translation and segmentation quality as well as a significantly

[26] https://www.veed.io/.

[27] For example, see https://www.apptek.com/post/asr-in-captions-accessibility-series-article-7 and https://sonix.ai/articles/how-to-remove-background-audio-noise.

en-de					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	66.9 (66.9±2.8)	20.1 (20.2±1.5)	71.7 (71.6±2.4)	**100** (100±0.0)	58.7 (58.6±3.1)
System 2	61.5 (61.5±2.4)	22.3 (22.2±1.6)	71.8 (71.8±2.3)	**100** (100±0.0)	76.2 (76.2±2.7)
System 3	68.1 (68.1±1.5)	13.5 (13.5±1.2)	62.1 (62.0±2.6)	91.6 (91.7±1.4)	**89.3** (89.3±1.8)
System 4	67.5 (67.1±7.3)	23.3 (23.2±1.7)	57.9 (57.9±2.2)	96.4 (96.4±0.9)	83.7 (83.7±2.3)
System 5	66.8 (66.8±2.9)	19.5 (19.5±1.5)	74.0 (74.0±2.0)	44.1 (42.8±3.0)	50.2 (50.2±3.1)
Ours	**59.9** (59.9±3.2)	**23.4** (23.4±1.6)	**77.9** (78.0±2.1)	86.9 (86.9±1.6)	68.6 (68.6±2.7)
en-es					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	52.2 (52.2±2.7)	33.4 (33.3±1.8)	76.9 (76.9±1.9)	**100** (100±0.0)	64.6 (64.6±2.9)
System 2	51.3 (51.2±2.4)	32.7 (32.6±1.8)	77.1 (77.0±2.0)	**100** (100±0.0)	77.6 (77.6±2.5)
System 3	58.3 (58.3±1.7)	23.3 (23.2±1.4)	66.1 (66.0±2.3)	94.1 (94.1±1.2)	**87.1** (87.1±2.0)
System 4	53.8 (53.8±4.7)	35.3 (35.3±2.0)	65.7 (65.7±1.8)	81.3 (81.3±2.2)	86.2 (86.1±2.2)
System 5	64.6 (64.6±2.0)	18.6 (18.6±1.3)	79.3 (79.3±1.9)	48.5 (48.5±3.0)	63.0 (62.9±2.8)
Ours	**46.8** (46.7±2.2)	**37.4** (37.5±2.0)	**81.6** (81.7±1.5)	93.2 (93.3±1.1)	74.6 (74.6±2.5)

Table 5: Unconstrained results on MuST-Cinema with 95% CI in parentheses.

better overall quality on both languages. Gains in BLEU are more evident in Spanish, where we obtain a ∼6% improvement compared to the second-best model (System 4). Also, considerable Sigma improvements are observed with gains of 5.3–34.5% for German and 2.9-24.2% for Spanish, which are in line with SubER improvements of, respectively, 2.6–12.0% and 8.8–27.6%. A perfect CPL conformity is reached by Systems 1 and 2 for both languages, while our system is on par with System 3 on en-es and falls slightly behind Systems 3 and 4 on en-de, with a ∼90% average conformity for the two language pairs. System 5 is by far the worst, as it violates the 42 CPL constraint in more than 50% of the lines. As for CPS conformity, we observe that our system achieves better scores compared to Systems 1 and 5 but it is worse than Systems 2, 3, and 4 on both language directions, highlighting again the need to improve this aspect in future work.

EC Short Clips This out-of-domain test set presents additional difficulties compared to TED talks, namely, the presence of multiple speakers and background music during speech. It is worth mentioning that our direct ST models have not been trained to be robust to these phenomena, as they are not present in the training data, whereas production tools are designed to deal with any condition, and may have dedicated modules to handle them.

Nevertheless, the results in Table 6 show that, even in these challenging conditions, our direct ST models are competitive with production tools on BLEU, Sigma, and SubER. Indeed, there is no clear winner between the systems as the best score for each metric is obtained by a different model, which also varies across languages. Looking at the conformity constraints, Systems 1, 2, and 4 achieve a perfect CPL conformity (100%), while ours is comparable with System 3 and better than System 5. This difference is likely motivated by the number of `<eol>` inserted by our system, which is considerably lower than that of System 4 (368 vs. 635 for German and 451 vs. 594 for Spanish). Instead, the results for CPS conformity follow the same trend observed in the constrained data condition (§6.2).

Even though this scenario features completely different domain and audio characteristics, some trends are in line with the results shown in Table 5. System 3 always achieves the best CPS conformity, while Systems 1, 2, and 4 achieve perfect CPL conformity on both languages. Moreover, although System 4 achieves the best translation quality (and it is the second best on MuST-Cinema, after our system), its segmentation quality (Sigma) is always the worst, indicating that its subtitles are not segmented in an optimal way to facilitate comprehension. All in all, these results suggest that each production tool has been optimized on a different aspect of automatic subtitling

en-de					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	63.0 (63.0±2.4)	23.8 (23.8±1.9)	**71.6** (71.5±2.7)	**100** (100±0.0)	76.1 (76.1±2.8)
System 2	60.8 (60.8±1.8)	22.1 (22.1±1.9)	67.2 (67.1±2.9)	**100** (100±0.0)	91.1 (91.1±1.9)
System 3	**59.0** (58.9±1.9)	25.0 (25.0±1.9)	70.4 (70.4±2.8)	84.6 (84.6±1.9)	**95.4** (95.4±1.4)
System 4	61.5 (61.5±3.3)	**28.2** (28.3±2.0)	59.4 (59.4±2.2)	**100** (100±0.0)	94.9 (95.0±1.5)
System 5	62.4 (62.4±2.2)	24.2 (24.2±1.8)	71.3 (71.2±2.2)	39.8 (39.7±3.4)	71.3 (71.3±3.3)
Ours	59.9 (59.9±2.2)	25.3 (25.3±1.9)	70.8 (70.7±2.4)	81.3 (81.3±2.2)	79.9 (80.0±2.7)
en-es					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	52.9 (52.9±1.8)	33.7 (33.7±1.8)	76.0 (75.9±2.2)	**100** (100±0.0)	80.4 (80.3±2.8)
System 2	51.7 (51.6±1.6)	32.2 (32.3±1.9)	75.6 (75.6±2.2)	**100** (100±0.0)	93.5 (93.5±1.7)
System 3	**49.7** (49.7±1.8)	35.5 (35.5±1.8)	74.9 (74.9±1.9)	87.3 (87.4±1.8)	**95.3** (95.3±1.4)
System 4	50.2 (50.2±2.2)	**39.6** (39.6±1.9)	61.9 (61.9±1.8)	**100** (100±0.0)	93.4 (93.4±1.4)
System 5	64.9 (64.9±1.6)	21.9 (21.9±1.5)	**79.7** (79.6±2.0)	41.7 (41.6±3.3)	73.1 (73.0±3.2)
Ours	52.7 (52.7±2.0)	34.8 (34.9±2.0)	72.6 (72.7±2.0)	88.6 (88.5±1.6)	79.1 (79.0±2.6)

Table 6: Unconstrained results on EC Short Clips with 95% CI in parentheses.

(e.g., System 3 has been optimized to achieve high CPS conformity). In contrast, our direct model, which has been trained without prioritizing any specific aspect, performs on average, also achieving competitive results in out-of-domain scenarios.

EuroParl Interviews The EuroParl Interviews set represents the most difficult of the three test sets: It contains multiple speakers, and the target translations are not verbatim since they are compressed to perfectly fit the subtitling constraints (§2.2). This characteristic is very challenging for current automatic subtitling tools, especially for our direct model since it has not been trained on similar data.

The results are shown in Table 7. As on the EC test set, our system performs competitively with production tools, even achieving the best Sigma for German. For CPL, instead, most systems have high length conformity, even reaching 100%. As already noticed on the other test sets, the CPL conformity is strongly correlated with the number of <eol> inserted by a system: Our model has an average conformity of 85.5% with only 451 <eol> inserted, nearly half of those inserted by System 1 (864), System 2 (711), and System 4 (774) that always comply with the CPL constraint. CPS conformity shows the same trend as with the other test sets.

Compared to the results in Tables 5 and 6, we can see that all systems struggle in achieving a comparable overall subtitle quality (SubER), high-quality segmentations (Sigma), and, above all, high translation quality (BLEU). The translation quality of all systems degrades by at least 10 BLEU compared to the values observed on the MuST-Cinema and EC test sets. However, as previously mentioned, these results are expected since the EuroParl Interviews test set contains condensed translations of the source speech.

All in all, we can conclude that our direct ST model, even though not developed as a production-ready system (it is not trained on huge amounts of data and different domains), is competitive with production tools. Indeed, considering the SubER metric computed over the three test sets (Table 8), our direct ST approach is the best on both German (67.0) and Spanish (57.2). As only the scores of System 2 fall within the confidence interval of our direct model in both cases, we can conclude that our model is on par with the best production system and outperforms the others in terms of SubER.

en-de					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	84.9 (85.0±2.4)	12.3 (12.3±1.1)	64.8 (64.8±2.8)	**100** (100±0.0)	67.6 (67.7±2.8)
System 2	78.4 (78.4±2.0)	13.2 (13.2±1.1)	63.9 (63.9±2.9)	**100** (100±0.0)	79.8 (79.8±2.3)
System 3	**78.1** (78.1±1.9)	13.6 (13.6±1.1)	69.6 (69.6±2.8)	86.9 (86.9±1.6)	**93.2** (93.3±1.4)
System 4	80.1 (80.1±2.7)	**15.8** (15.8±1.3)	56.9 (56.9±2.8)	**100** (100±0.0)	83.8 (83.9±2.2)
System 5	85.1 (85.1±1.9)	11.4 (11.4±1.1)	69.8 (69.8±2.5)	44.4 (44.4±2.8)	59.2 (59.3±2.7)
Ours	80.3 (80.3±2.4)	12.5 (12.5±1.1)	**70.0** (70.0±2.8)	80.9 (81.0±1.9)	68.8 (68.8±2.5)
en-es					
Model	SubER (↓)	BLEU (↑)	Sigma (↑)	CPL (↑)	CPS (↑)
System 1	75.5 (75.5±2.3)	19.8 (19.8±1.3)	72.7 (72.7±2.2)	**100** (100±0.0)	72.7 (72.8±2.5)
System 2	71.4 (71.4±2.1)	20.9 (20.9±1.4)	73.8 (73.8±2.0)	**100** (100±0.0)	81.4 (81.5±2.3)
System 3	70.0 (70.1±2.2)	20.8 (20.8±1.4)	72.8 (72.8±2.0)	90.5 (90.5±1.4)	**93.7** (93.7±1.3)
System 4	**68.6** (68.5±2.5)	**25.4** (25.4±1.4)	61.6 (61.6±2.0)	**100** (100±0.0)	91.5 (91.5±1.8)
System 5	80.8 (80.8±1.7)	13.0 (12.9±1.1)	**77.3** (77.3±2.4)	52.1 (52.1±2.8)	67.4 (67.5±2.7)
Ours	72.3 (72.3±2.2)	20.8 (20.9±1.4)	70.4 (70.4±2.0)	90.1 (90.1±1.3)	76.9 (76.9±2.4)

Table 7: Unconstrained results on EuroParl Interviews with 95% CI in parentheses.

	System 1	System 2	System 3	System 4	System 5	Ours
en-de	72.0 (72.0±1.6)	67.2 (67.1±1.3)	69.0 (69.0±1.2)	70.1 (70.1±3.3)	71.9 (71.9±1.7)	**67.0** (67.0±1.7)
en-es	60.3 (60.3±1.5)	58.2 (58.2±1.3)	59.8 (59.8±1.2)	57.8 (57.8±2.4)	70.2 (70.2±1.1)	**57.2** (57.1±1.5)

Table 8: SubER (↓) over the three test sets with 95% CI in parentheses.

7 Conclusions

In this paper, we proposed the first approach based on direct speech-to-text translation models to fully automatize the subtitling process, including translation, segmentation into subtitles, and timestamp estimation. Experiments in constrained data conditions on 7 language pairs demonstrated the potential of our approach, which outperformed the current cascade architectures with a ∼7% improvement in terms of SubER. In addition, to test the generalizability of our findings across subtitling genres, we extended our evaluation setting by collecting two new test sets for en→{de, es} covering different domains, degrees of subtitle condensation, and audio conditions. Finally, we compared our models with production tools in unconstrained data conditions on both existing benchmarks and the newly collected test sets. This comparison further highlighted that our approach represents a promising direction: Although trained on a relatively limited amount of data, our systems achieved comparable quality with production tools, with improvements in SubER ranging from 0.2 to 5.0 on en→de and from 0.6 to 13.0 on en→es over the three test sets.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The authors thank Marco Matassoni (and the SpeechTek unit at FBK) for their help in providing forced aligners, and Evgeny Matusov (and AppTek) for sharing their production system outputs, as well as the anonymous reviewers for their insightful comments that improved the manuscript.

References

Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, Tiago Luís, Sérgio Paulo, Nicola Piccinini, Haritz Arzelus, João Neto, Carlo Aliprandi, and Arantza Pozo. 2015. Automating live and batch subtitling of multimedia contents for several european languages. *Multimedia Tools and Applications*, 75:1–31. https://doi.org/10.1007/s11042-015-2794-z

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). https://doi.org/10.18653/v1/2022.iwslt-1.10

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. Findings of the IWSLT 2021 evaluation campaign. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). https://doi.org/10.18653/v1/2021.iwslt-1.1

Andrei Andrusenko, Rauf Nasretdinov, and Aleksei Romanenko. 2022. Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition. *arXiv preprint arXiv:2208.07657*. https://doi.org/10.1109/ICASSP49357.2023.10095430

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. Findings of the IWSLT 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. https://doi.org/10.18653/v1/2020.iwslt-1.1

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1073

Wilker Aziz, Sheila C. M. de Sousa, and Lucia Specia. 2012. Cross-lingual sentence compression for subtitles. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 103–110, Trento, Italy.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *Proceedings of International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore. https://doi.org/10.1109/ASRU46091.2019.9003774

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade

versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. https://doi.org/10.18653/v1/2021.acl-long.224

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *Proceedings of ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada. https://doi.org/10.1109/ICASSP.2018.8461690

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. https://doi.org/10.3115/v1/W14-3302

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. https://doi.org/10.18653/v1/2021.eacl-demos.32

François Buet and François Yvon. 2021. Toward genre adapted closed captioning. In *Proceedings of Interspeech 2021*, pages 4403–4407. https://doi.org/10.21437/Interspeech.2021-1762

Maxime Burchi and Valentin Vielzeuf. 2021. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. https://doi.org/10.1109/ASRU51503.2021.9687874

Denis Burnham, Greg Leigh, William Noble, Caroline Jones, Michael Tyler, Leonid Grebennikov, and Alex Varley. 2008. Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension. *The Journal of Deaf Studies and Deaf Education*, 13(3):391–404. https://doi.org/10.1093/deafed/enn003, PubMed: 18372297

Lindsay Bywood, Martin Volk, Mark Fishel, and Panayota Georgakopoulou. 2013. Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*, 21(4):595–610. https://doi.org/10.1080/0907676X.2013.831920

Xuankai Chang, Aswin Shanmugam Subramanian, Pengcheng Guo, Shinji Watanabe, Yuya Fujita, and Motoi Omachi. 2020. End-to-end asr with adaptive span self-attention. In *INTERSPEECH*. https://doi.org/10.21437/Interspeech.2020-2816

Jorge Díaz Cintas and Aline Remael. 2021. *Subtitling: Concepts and Practices*. Translation practices explained. Routledge. https://doi.org/10.4324/9781315674278

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451,

Online. https://doi.org/10.18653/v1/2020.acl-main.747

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting transformer to end-to-end spoken language translation. In *Proceedings of Interspeech 2019*, pages 1133–1137. https://doi.org/10.21437/Interspeech.2019-3045

Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico. 2022. Duration modeling of neural tts for automatic dubbing. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8037–8041. https://doi.org/10.1109/ICASSP43922.2022.9747158

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland.

Marcello Federico, Yogesh Virkar, Robert Enyedi, and Roberto Barra-Chicote. 2020. Evaluating and optimizing prosodic alignment for automatic dubbing. In *Proceedings of Interspeech 2020*, pages 1481–1485. https://doi.org/10.21437/Interspeech.2020-2983

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021a. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. https://doi.org/10.18653/v1/2021.eacl-main.57

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021b. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 55–62, Trento, Italy. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022. Efficient yet competitive speech translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online). https://doi.org/10.18653/v1/2022.iwslt-1.13

Panayota Georgakopoulou. 2019. Template files: The holy grail of subtitling. *Journal of Audiovisual Translation*, 2(2):137–160. https://doi.org/10.47476/jat.v2i2.84

Henrik Gottlieb. 2004. Subtitles and international anglification. *Nordic Journal of English Studies*, 3:219. https://doi.org/10.35360/njes.32

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA. https://doi.org/10.1145/1143844.1143891

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1764–II–1772.

Roberto Gretter, Marco Matassoni, and Daniele Falavigna. 2021. Seed words based data selection for language model adaptation. *arXiv preprint arXiv:2107.09433*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pages 198–208, Cham. https://doi.org/10.1007/978-3-319-99579-3_21

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. https://doi.org/10.1109/ASRU46091.2019.9003832

Hirofumi Inaguma, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2021. Non-autoregressive end-to-end speech translation with parallel autoregressive rescoring. *arXiv preprint arXiv:2109.04411*. https://doi.org/10.1109/ICASSP39728.2021.9415093

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. https://doi.org/10.1109/ICASSP40776.2020.9054626

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.147

Alina Karakanta, Franćois Buet, Mauro Cettolo, and Franćois Yvon. 2022. Evaluating subtitle segmentation for end-to-end generation systems. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 3069–3078, Marseilles, France.

Alina Karakanta, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Between flexibility and consistency: Joint generation of captions and subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 215–225, Bangkok, Thailand (online). https://doi.org/10.18653/v1/2021.iwslt-1.26

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. https://doi.org/10.18653/v1/2020.iwslt-1.26

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020b. MuST-cinema: A speech-to-subtitles corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. *arxiv:2206.00888*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839, New Orleans, Louisiana. https://doi.org/10.1109/ICASSP.2017.7953075

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. https://doi.org/10.18653/v1/D16-1139

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124. Lisboa, Portugal.

Helena Kruger. 2001. The creation of interlingual subtitles: Semiotics, equivalence and condensation. *Perspectives*, 9(3):177–196. https://doi.org/10.1080/0907676X.2001.9961416

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. CTC-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer*, pages 267–278, Cham. https://doi.org/10.1007/978-3-030-60276-5_27

Surafel M. Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*. Hong Kong.

Surafel M. Lakew, Marcello Federico, Yue Wang, Cuong Hoang, Yogesh Virkar, Roberto Barra-Chicote, and Robert Enyedi. 2021. Machine translation verbosity control for automatic dubbing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7538–7542. https://doi.org/10.1109/ICASSP39728.2021.9414411

Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isometric mt: Neural machine translation for automatic dubbing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6242–6246. https://doi.org/10.1109/ICASSP43922.2022.9747023

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777. https://doi.org/10.1109/TASLP.2014.2304637

Danni Liu, Jan Niehues, and Gerasimos Spanakis. 2020a. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256, Online. https://doi.org/10.18653/v1/2020.iwslt-1.30

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020b. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. https://doi.org/10.18653/v1/W19-5209

Evgeny Matusov, Patrick Wilken, and Christian Herold. 2020. Flexible customization of a single neural machine translation system with multi-dimensional metadata inputs. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 204–216.

Maite Melero, Antoni Oliver, and Toni Badia. 2006. Automatic multilingual subtitling in the eTITLE project. In *Proceedings of ASLIB Translating and the Computer 28*. https://doi.org/10.1.1.107.6011&rep=rep1

Vikramjit Mitra, Horacio Franco, Richard M. Stern, Julien van Hout, Luciana Ferrer, Martin Graciarena, Wen Wang, Dimitra Vergyri, Abeer Alwan, and John H. L. Hansen. 2017. Robust features in deep-learning-based speech recognition. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-64680-0_8

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2019. Prosodic phrase alignment for machine dubbing. *ArXiv*, abs/1908.07226. https://doi.org/10.21437/Interspeech.2019-1621

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-4009

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Speechformer: Reducing information loss in direct speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1706, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.127

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 480–487, Online only.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Elisa Perego. 2008. Subtitles and line-breaks: Towards improved readability. *Between Text and Image: Updating Research in Screen Translation*, 78(1):211–223. https://doi.org/10.1075/btl.78.21per

Stelios Piperidis, Iason Demiros, Prokopis Prokopidis, Peter Vanroose, Anja Hoethker, Walter Daelemans, Elsa Sklavounou, Manos Konstantinou, and Yannis Karavidas. 2004. Multimodal, multilingual resources in the subtitling process. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. https://doi.org/10.18653/v1/W18-6319

Michael L. Seltzer, Dong Yu, and Yongqiang Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7398–7402. https://doi.org/10.1109/ICASSP.2013.6639100

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. https://doi.org/10.18653/v1/P16-1162

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. https://doi.org/10.18653/v1/2020.acl-main.661

Fred Stentiford and M. G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6:116–123.

Agnieszka Szarkowska and Olivia Gerber-Morón. 2018. Viewers can keep up with fast subtitles: Evidence from eye movements. *PLOS ONE*, 13(6):1–30. https://doi.org/10

.1371/journal.pone.0199331, PubMed: 29920538

Agnieszka Szarkowska, Izabela Krejtz, Olga Pilipczuk, Lukasz Dutka, and Jan-Louis Kruger. 2016. The effects of text editing and subtitle presentation rate on the comprehension and reading patterns of interlingual and intralingual subtitles among deaf, hard of hearing and hearing viewers. *Across Languages and Cultures*, 17(2):183–204. https://doi.org/10.1556/084.2016.17.2.3

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isochrony-aware neural machine translation for automatic dubbing. In *Proceedings of Interspeech 2022*, pages 1776–1780. https://doi.org/10.21437/Interspeech.2022-11136

Jörg Tiedemann. 2016. OPUS–parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT—Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. https://doi.org/10.21437/Interspeech.2022-59

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 *Advances in Neural Information Processing Systems*, volume 30.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. 2021. Improvements to prosodic alignment for automatic dubbing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7543–7574. https://doi.org/10.1109/ICASSP39728.2021.9414966

Kaisa Vitikainen and Maarit Koponen. 2021. Automation in the intralingual subtitling process: Exploring productivity and user experience. *Journal of Audiovisual Translation*, 4(3):44–65. https://doi.org/10.47476/jat.v4i3.2021.197

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 53–62, Denver, Colorado, USA.

Alexander Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 793–796 vol. 2. https://doi.org/10.1109/ICASSP.1991.150456

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. https://doi.org/10.18653/v1/2021.acl-long.80

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2: A massively multilingual speech-to-text translation corpus. https://doi.org/10.21437/Interspeech.2021-2027

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017.

Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-503

Orion Weller, Matthias Sperber, Christian Gollan, and Joris Kluivers. 2021. Streaming models for joint speech recognition and translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2533–2539, Online. https://doi.org/10.18653/v1/2021.eacl-main.216

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - A metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). https://doi.org/10.18653/v1/2022.iwslt-1.1

Jitao Xu, François Buet, Josep Crego, Elise Bertin-Lemée, and François Yvon. 2022. Joint generation of captions and subtitles with dual decoding. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 74–82, Dublin, Ireland (in-person and online). https://doi.org/10.18653/v1/2022.iwslt-1.7

A Timestamp Extraction Method

To validate the effectiveness of extracting source-side timestamps with the CTC-based segmentation algorithm, we conduct an ablation study where we replace it with the forced aligner tool of the Cascade architecture (§5.2). Table 9 reports the scores. The forced aligner tool (FA) achieves similar results compared to the CTC-based segmentation algorithm (CTC), with a slightly worse SubER (+0.1) on average on the three test sets. Moreover, it is important to highlight that our method does not require an external model. These findings support our choice and align with previous research by Kürzinger et al. (2020), which highlighted the competitiveness of the CTC-based segmentation approach compared to widely used forced aligners (in their case, Gentle[28]).

[28]https://github.com/lowerquality/gentle.

Method	en-de			en-es			Avg.
	MC	ECSC	EPI	MC	ECSC	EPI	
CTC	59.9	**59.9**	**80.3**	46.8	**52.7**	72.3	**62.0**
FA	**59.7**	60.3	80.7	**46.7**	**52.7**	**72.2**	62.1

Table 9: SubER scores (↓) on MuST-Cinema test set (MC), EC Short Clips (ECSC), and EuroParl Interviews (EPI) when the CTC-based audio segmentation (CTC) or the forced aligner (FA) method is used to extract the source-side timestamps.

Noise Removed	en-de		en-es		Avg.
	ECSC	EPI	ECSC	EPI	
1. Yes	59.9	80.3	66.3	72.3	52.7
2. Only Segm.	61.4	82.0	68.4	73.9	56.4
3. No	63.1	81.7	69.5	75.3	58.1

Table 10: SubER scores (↓) on EC Short Clips (ECSC) and EuroParl Interviews (EPI) with background noise removal for: both the audio segmentation with SHAS and the prediction of the direct ST system (*1.*); only the audio segmentation, but the noisy audio is fed as input to the direct ST model (*2.*); no noise removal (*3.*).

B Effect of Background Noise

The presence of background noise in the test sets complicates both the audio segmentation (performed with SHAS) and the generation with the direct ST model. For this reason, for the sake of a fair comparison with production tools, we used Veed to remove the background noise from EC Short Clips and EuroParl Interviews, as mentioned in §6.3. Table 10 shows the impact of background noise on the resulting subtitling quality. By comparing *1.* and *3.*, we notice that the presence of background noise causes an overall relative error increase of ∼5% on average over the two test sets and two language pairs. The degradation is caused both by the lower quality of the audio segmentation of SHAS and by worse outputs produced by the direct ST system, as the absence of noise during segmentation (*2.*) improves by an average of 1.7 SubER the results obtained without noise removal (*3.*). Creating models robust to background noise, though, is a task *per se* (Seltzer et al., 2013; Li et al., 2014; Mitra et al., 2017) and goes beyond the scope of this work.