# SBAAM! Eliminating Transcript Dependency in Automatic Subtitling

**Marco Gaido, Sara Papi, Matteo Negri, Mauro Cettolo, Luisa Bentivogli**
Fondazione Bruno Kessler
Trento, Italy
{mgaido,spapi,negri,cettolo,bentivo}fbk.eu

## Abstract

Subtitling plays a crucial role in enhancing the accessibility of audiovisual content and encompasses three primary subtasks: translating spoken dialogue, segmenting translations into concise textual units, and estimating timestamps that govern their on-screen duration. Past attempts to automate this process rely, to varying degrees, on automatic transcripts, employed diversely for the three subtasks. In response to the acknowledged limitations associated with this reliance on transcripts, recent research has shifted towards transcription-free solutions for translation and segmentation, leaving the direct generation of timestamps as uncharted territory. To fill this gap, we introduce the first direct model capable of producing automatic subtitles, entirely eliminating any dependence on intermediate transcripts also for timestamp prediction. Experimental results, backed by manual evaluation, showcase our solution's new state-of-the-art performance across multiple language pairs and diverse conditions.

## 1 Introduction

Subtitling aims to facilitate the accessibility of audiovisual media, such as movies, TV shows, and video lectures, by providing users with a textual translation of spoken content. Subtitles consist of two components: a textual block, typically encompassing one or two lines, and its corresponding time duration, indicated by start and end timestamps. To ensure effective on-screen presentation and minimize users' cognitive load, subtitles should conform to spatio-temporal constraints (Bogucki, 2004; Khalaf, 2016). These include restrictions on the maximum number of characters per line and a display duration that guarantees synchronization with the video while granting viewers sufficient time to read the entire text.

Automating the task involves addressing three main subtasks: 1 translation of the spoken content, 2 segmentation of the translated text into blocks and lines, and 3 estimation of the timestamps for each block. Early approaches to automatic subtitling (AS) adopted a cascade architecture (Piperidis et al., 2004; Oliver Gonzalez, 2006; Alvarez et al., 2017; Bojar et al., 2021), i.e. a pipeline of components, including automatic speech recognition (ASR) and machine translation (MT) models. In these systems, transcripts served as the foundational element for all three subtasks, despite the well-documented limitations of this reliance on them, such as error propagation (Sperber and Paulik, 2020), the loss of useful prosody information (Lakew et al., 2022; Tam et al., 2022), inapplicability to source languages lacking written forms (Lee et al., 2022), and higher computational and environmental cost (Strubell et al., 2019) due to the need to run multiple models.

To cope with these limitations, subsequent studies aimed to streamline the subtitling pipeline by reducing its dependency on transcripts. In this endeavor, direct speech-to-text translation (ST) systems (Bérard et al., 2016; Weiss et al., 2017), capable of translating speech without recourse to intermediate symbolic representations, were successfully used by Karakanta et al. (2020a) for the translation step (subtask 1). For subtitle segmentation (subtask 2), efforts focused on adapting MT (Etchegoyhen et al., 2014; Bywood et al., 2013; Volk et al., 2010; Matusov et al., 2019; Koponen et al., 2020b; Cherry et al., 2021) and language models (Ponce et al., 2023) to directly produce translations that incorporate block and line boundaries as specific tags, also by exploiting audio information (Papi et al., 2022).

To date, compared to subtasks 1 and 2, timestamp estimation (subtask 3) has received much less attention. This regards not only the elimination of intermediate transcription steps, the primary focus of this work, but also evaluation, as there is still no reliable and informative metric for directly assessing the quality of the predicted timestamps.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3673–3691

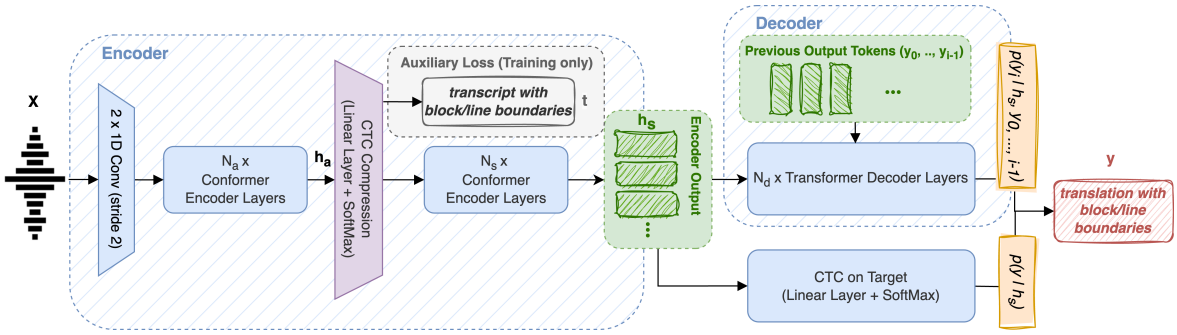


Figure 1: Architecture of our model.

As for the task itself, current approaches, including those exclusively based on direct ST models (Papi et al., 2023a,b; Bahar et al., 2023), still require transcripts for timestamp estimation, which involves *i)* generating captions (i.e., segmented transcripts), *ii)* estimating timestamps for caption blocks, typically through a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006), and *iii)* projecting them onto target subtitles. Regarding automatic evaluation, current metrics (Wilken et al., 2022) are by design holistic and therefore inadequate to precisely measure timestamp estimation quality.

To bridge these gaps, we introduce and evaluate the first fully end-to-end AS solution[1] that seamlessly produces both subtitles (i.e., segmented translations) and their timestamps without any reliance on intermediate transcripts. This innovation is realized by incorporating into our model the capability to directly determine the temporal alignment between the spoken utterances and their corresponding translation in subtitle form. Along this direction, our contributions are the following:

• **We propose two methods for timestamp estimation** (§3), respectively based on applying the CTC loss directly on translations (Zhang et al., 2022a), and on estimating the audio-text temporal alignment from the attention mechanism (Papi et al., 2023c; Alastruey et al., 2024). Both approaches are complemented by the joint CTC decoding (Watanabe et al., 2017), which yields higher translation quality and a more precise alignment between the generated text and the corresponding audio (Yan et al., 2023);

• **We introduce SubSONAR,[2] a novel metric for evaluating timestamp quality** (§4), which is based on SONAR (Duquenne et al., 2023) and designed to be sensitive to time shifts, enabling a focused evaluation of timestamps;

• **We validate our approach through comparative experiments** (§5) on 7 language pairs, 2 data conditions, and 4 domains, achieving new state-of-the-art results on different benchmarks and demonstrating better performance compared to cascade architectures for automatic subtitling;

• **We verify the efficacy of our model through manual evaluation** (§6), attesting a significant reduction in timestamp adjustments of ∼24% compared to the previous state of the art.

## 2 Direct Model for Automatic Subtitling

Following previous work on direct ST (Liu et al., 2020; Xu et al., 2021a, 2023; Wu et al., 2023), we build a direct autoregressive encoder-decoder model, where the encoder is composed of three blocks: *i)* an acoustic encoder – made of two 1D convolutions with stride 2 and $N_a$ Conformer (Gulati et al., 2020) layers, *ii)* a length adaptor – a CTC Compression (Gaido et al., 2021) module that averages the vectors corresponding to the same CTC prediction, and *iii)* a semantic encoder – made of $N_s$ Conformer layers. The encoder output is then fed to an autoregressive decoder $\mathcal{D}$ and, in parallel, to a CTC on Target (TgtCTC) module (Zhang et al., 2022a; Yan et al., 2023). The full architecture is shown in Figure 1.

We train our model with a composite loss ($\mathcal{L}$), which is a linear combination of a label smoothing cross-entropy (CE) loss (Szegedy et al., 2016) on the decoder $\mathcal{D}$, a CTC loss on the TgtCTC module, and a CTC loss on the CTC Compression module:

$$\mathcal{L} = \lambda_1 \,\mathrm{CTC}(h_a, t) + \lambda_2 \,\mathrm{CTC}(h_s, y) + \lambda_3 \,\mathrm{CE}(\mathcal{D}(h_s, y), y)$$

[1] All our code and pre-trained models are available at https://github.com/hlt-mt/FBK-fairseq/ under Apache Licence 2.0.

[2] Available at https://github.com/hlt-mt/subsonar/ under Apache License 2.0 and on PyPi (https://pypi.org/project/SubSONAR/).

where $\lambda_{1,2,3}$ are the loss weights, $h_a$ is the acoustic encoder output, $h_s$ is the encoder output, $y$ is the target subtitle, and $t$ is the caption. It is noteworthy that the caption ($t$), albeit optionally used for training, is not strictly required by our model. Additionally, **captions are neither generated nor utilized by our novel timestamp estimation methods** (§3), which exclusively rely on the generated subtitles. Indeed, the auxiliary CTC loss on captions and the CTC compression module can be entirely omitted, at the cost of a slight reduction in quality (Zhang et al., 2022a), and captions can be replaced with speech units (Hsu et al., 2021; Zhang et al., 2022b) as the module target. However, we opt to retain the CTC compression module (as well as the related auxiliary CTC loss on captions), not only for its benefits in terms of computational efficiency and translation quality but, most importantly, to enable the use of the same model with earlier timestamp estimation solutions, which require a CTC module with captions as the target (Bahar et al., 2023; Papi et al., 2023a). This allows a direct comparison between different timestamp estimation methods, employing the same model and generated subtitles.

At inference time, subtitles are generated by estimating their likelihood with the joint CTC/attention framework with CTC rescoring (Yan et al., 2023) through a linear combination of the probabilities of the TgtCTC module and the decoder:

$$p = p_{\mathcal{D}}(y_i|h_s, y_{0,\dots,i-1}) + \alpha\, p_{\text{TgtCTC}}(y_{0,\dots,i}|h_s)$$

where $\alpha$ is a hyperparameter.

**Multilinguality.** We experiment both in bilingual (English to a target language) and in multilingual (English to many languages) settings. In the latter case, we prepend a learned language embedding to the previous output tokens to be fed to the decoder (Inaguma et al., 2019; Wang et al., 2020a). In addition, we sum the same learned embedding to all the vectors of the encoder output ($h_s$) before processing them with the TgtCTC module, to inform the module about the language to generate.[3]

[3]We explored alternative solutions: adding the language embedding at the beginning of the encoder before all Conformer layers, as per (Di Gangi et al., 2019); adding it to the vectors obtained after the CTC compression module; multiplying the language embedding vector instead of summing it; using separate learned language embeddings for the decoder and the encoder. However, none of them led to better results.

## 3 Estimation of Block Timestamps

In this section, we present novel methods for estimating the block timestamps without requiring the transcripts. These solutions not only avoid error propagation and are applicable to unwritten source languages but also exhibit increased speed compared to current methods, skipping the transcript generation step, thereby minimizing significant overhead. Our methods involve either directly applying the CTC-segmentation algorithm (Kürzinger et al., 2020) to the CTC on Target module (§3.1) or leveraging the encoder-decoder attention scores to find the audio-text alignment (§3.2).

### 3.1 Subtitle CTC-based Estimation

The error propagation and approximations caused by the cross-lingual Levenshtein alignment proposed by Papi et al. (2023b) can be avoided by directly estimating the timestamps on the target, i.e., on the generated subtitles. This is realized by exploiting the predictions of the TgtCTC module placed on top of the encoder, which are used by the CTC segmentation algorithm (Kürzinger et al., 2020), together with the input audio, to retrieve the timestamp information. This subtitle CTC-based method (SubCTC) enables the direct alignment between the input audio features (representing the time over the speech sequence) and the boundaries of the generated subtitle blocks, eliminating the need for intermediate transcript alignments with the audio and their projection into the target side.

The SubCTC method builds upon the assumption that the CTC module can learn meaningful alignments between source audio and target texts, as it does when predicting transcripts (Sak et al., 2015; Sainath et al., 2020; Chen et al., 2023) without direct supervision. However, the validity of this assumption has not been verified yet. Rather, related works in non-autoregressive translation showed the inability of such models to implicitly learn complex word reordering (Chuang et al., 2021; Ran et al., 2021; Xu et al., 2021b). Motivated by this potential issue, we devise an alternative method that leverages the cross-attention matrix to infer source-target alignments.

### 3.2 Attention-based Estimation

Building on the extensive literature in text-to-text translation and language modeling discussing the quality of source-target alignments learned by the attention mechanism (Tang et al., 2018; Zenkel

**Algorithm 1** Attention-based DTW

1: $N, L \leftarrow nTokens, audioLen$
2: $B \leftarrow blockIdxs$ ▷ List of block boundary indices
3: $A \leftarrow attnMatrix$ ▷ $A \in \mathbb{R}^{N \times L}$
4: $A \leftarrow$ **StdNorm**$(A, axis = 0)$ ▷ Column-wise normalize
5: $A \leftarrow$ **MedianFilter**$(A, width = 7, axis = 1)$
6: $D \leftarrow$ **DTWDistance**$(-A)$ ▷ Negated attentions as costs
7: $n, l, P, blockTimings \leftarrow N, L, [(N, L)], [\,]$
8: **while** $P[-1] \neq (0, 0)$ **do** ▷ Backtrack best DTW path
9: **if** $n \in B$ **then**
10: $P \leftarrow$ **Append**$(P, (n-1, l-1))$
11: $blockTimings \leftarrow$ **Append**$(blockTimings, l)$
12: **else**
13: $P \leftarrow$ **Append**$(P, \underset{i \in \{n, n-1\}, j \in \{l, l-1\}}{\text{argmin}} D[i, j])$
14: **end if**
15: $n, l \leftarrow P[-1]$
16: **end while**
17: **return reverse**$(blockTimings)$

**Algorithm 2** SBAAM

1: $N, L \leftarrow nTokens, audioLen$
2: $B \leftarrow blockIdxs$ ▷ List of block boundary indices
3: $A \leftarrow attnMatrix$ ▷ $A \in \mathbb{R}^{N \times L}$
4: $A \leftarrow$ **StdNorm**$(A, axis = 0)$ ▷ Row-wise normalize
5: $A[A < 0] \leftarrow -\epsilon$ ▷ $\epsilon$ set to 0.01
6: $n, l, blockTimings \leftarrow 0, 0, [\,]$
7: **for all** $i_b \in [0, |B|)$ **do**
8: $l \leftarrow \underset{j \in (l, L-|B|+i_b)}{\text{argmax}} (\sum A[n : B[i_b], l : j] + \sum A[B[i_b] + 1 : N, j + 1 : L])$
9: $blockTimings \leftarrow$ **Append**$(blockTimings, l)$
10: $n \leftarrow B[i_b]$
11: **end for**
12: **return** $blockTimings$

et al., 2019; Garg et al., 2019; Chen et al., 2020) and recent works on speech-to-text (Papi et al., 2023c; Alastruey et al., 2024), we propose to exploit the encoder-decoder attention scores to determine subtitle block timings and devise two different modalities of leveraging this information. For the first one, we adapt the Dynamic Time Warping (DTW) algorithm (Sakoe and Chiba, 1978), which is commonly used to determine token-level timestamps in ASR (§3.2.1).[4] For the second one, recognizing that DTW relies on the assumption of monotonic source-target alignments, which does not always hold in our case, we introduce a new algorithm named SBAAM (SPEECH BLOCK ATTENTION AREA MAXIMIZATION), specifically designed for subtitles (§3.2.2).

### 3.2.1 Attention-based DTW for Subtitling

The DTW algorithm is a dynamic programming algorithm similar to Viterbi (Juang, 1984), which finds the best path (i.e., the one that minimizes a distance function) between two temporal sequences with varying speeds. This algorithm operates under the assumption that all the elements of the two sequences are aligned and that the mapping between the two sequences is monotonic. In our case, while the assumption may not be applicable at the token (or word) level, it remains valid at the block level, where the order has to be maintained; we therefore investigate its applicability to our task.

We use the additive inverse of the attention matrix as a distance function and follow the improved DTW algorithm by Yuxin and Miyanaga (2011) that applies a 1D median filter (Huang et al., 1979)[5] over the speech sequence for each text token, after a standard normalization over the text tokens (see Alg. 1). After computing the accumulated DTW distance matrix, we define the best path in the backtracking phase, with three possible moves: *i)* moving to the preceding element of the first sequence, *ii)* moving to the preceding element of the second sequence, or *iii)* moving to the preceding elements of both sequences. In this phase, we force moving across both (speech and text) sequences when the block boundary token (<eob>) is encountered. This decision is based on two observations: *i)* a manual inspection of attention matrices revealed that the attention of <eob> tokens is often unreliable (e.g., focused on the last token of the speech sequence or silence), and *ii)* we want to ensure that at least one speech segment (equivalent to 40ms) is assigned to each block by skipping the <eob> token in the DTW search (see lines 9-12 of Alg. 1). As a result, we use the timings assigned to the <eob> tokens as the temporal boundaries (timestamp) for the corresponding subtitle blocks.

### 3.2.2 SBAAM

To relax the constraint of alignment monotonicity, we devise a new algorithm to estimate the timing of block boundaries. Intuitively, our method (see Alg. 2) maximizes the attention scores within a rectangular area encompassing the tokens belonging to a block (i.e. all the tokens between two <eob>) along one axis, and the assigned span across the temporal dimension of the speech sequence along the other axis. For this reason, we named it **S**PEECH **B**LOCK **A**TTENTION **A**REA **M**AXIMIZATION or SBAAM.

[4] For example, the DTW is used in the Transformers implementation of Whisper (Radford et al., 2023): https://github.com/huggingface/transformers/blob/v4.34.0/src/transformers/models/whisper/modeling_whisper.py#L1817.

[5] A filter that takes the median value within a pre-defined window defined as the width of the filter itself.

| Model | SubER (↓) | | | | | | | | SubSONAR (↑) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | de | es | fr | it | nl | pt | ro | AVG | de | es | fr | it | nl | pt | ro | AVG |
| LEV | 60.2 | 47.9 | 53.7 | 52.0 | 49.0 | 46.1 | 49.8 | 51.2 | .677 | .692 | .670 | .689 | .688 | .691 | .688 | .685 |
| - joint CTC | 61.2 | 49.0 | 54.7 | 52.5 | 49.7 | 47.1 | 50.7 | 52.1 | .667 | .693 | .673 | .690 | .683 | .688 | .687 | .683 |
| SubCTC | 59.9 | 47.5 | 53.5 | 51.7 | 48.7 | 45.7 | 49.4 | 50.9 | .718 | .739 | .713 | .724 | .731 | .731 | .719 | .725 |
| ATTN DTW | 59.9 | **47.5** | **53.4** | **51.6** | **48.6** | **45.5** | **49.2** | **50.8** | .745 | .775 | .747 | .761 | .758 | .765 | .754 | .758 |
| SBAAM | **59.8** | **47.5** | **53.4** | **51.6** | 48.7 | **45.5** | 49.3 | **50.8** | **.749** | **.780** | **.753** | **.765** | **.767** | **.770** | **.761** | **.764** |

Table 1: Results of the timestamp estimation methods in terms of SubER (cased) and subSONAR for all the 7 languages of MuST-Cinema.

The SBAAM algorithm encompasses two steps. First, the attention matrix is normalized (following the previous procedure) and all negative values are set to a small negative value ($-\epsilon$). This is required as the attention values are typically peaky in the range $[0, 1]$ due to the `softmax`, with a few large values and random fluctuations close to 0 for all the others. After normalization, some small values may become very negative, while others may be closer to 0, even though both indicate low attention. Despite this, we maintain them as negative values to penalize the integration of areas with very low attention. Second, for each generated subtitle block (i.e., for each <eob>), we iteratively determine the timing by selecting the splitting point that maximizes the area of the first block with the audio up to that point and the rest of the text with the remaining audio. At the end of this process, we obtain the start and end timing (timestamp) for all blocks.

## 4 SubSONAR

Since it was proposed, the SubER[6] metric (Wilken et al., 2022) has been used to provide a holistic evaluation of subtitles, encompassing translation quality, block/line segmentation accuracy, and timing. Specifically, SubER computes the number of word edits and block/line edits required to match the reference, where hypothesis and reference words are allowed to match only within subtitle blocks that overlap in time. This definition highlights how SubER is sensitive only to major errors in terms of timing, as it solely checks whether two blocks overlap, even if only for a few milliseconds. This limitation motivates our proposal of SubSONAR, a new metric designed to be more sensitive to time shifts and, consequently, more suitable for evaluating the quality of subtitle timestamps.

To this aim, we leverage the multimodal and multilingual SONAR model (Duquenne et al., 2023), designed to generate sentence embeddings within a shared multimodal (text and audio) semantic space for all languages. Specifically, we calculate the cosine similarity between the SONAR embeddings of the text within a subtitle block and its corresponding audio, determined by the timestamp of the block. Subsequently, we average the similarity scores across all subtitle blocks, which results in a single score in the $[-1, 1]$ range. Being SONAR trained to capture both text and audio semantics, higher SubSONAR scores indicate better alignment between text and audio content, which is influenced by both translation quality and timestamp accuracy. Nevertheless, as revealed by the empirical validation presented in the following sections (§5,6), SubSONAR exhibits higher sensitivity to timing accuracy than to translation quality.

[6]In this work, we compute SubER-cased unless specified otherwise as per (Papi et al., 2023b).

## 5 Automatic Evaluation

In this section, we evaluate our solutions automatically through comparative experiments in different resource conditions and language settings. First, we validate the adoption of the joint CTC rescoring and compare the timestamp estimation methods introduced in §3 using a multilingual system trained and tested on all the 7 language pairs (en→{de,es,fr,it,nl,pt,ro}) of MuST-Cinema (Karakanta et al., 2020b) (§5.1). Then, we confirm their strength by comparing two bilingual systems (en-de and en-es) trained in the high resource conditions of the IWSLT 2023 subtitling track (Agarwal et al., 2023) with the results reported for the current state of the art in direct AS and production tools on several test sets (§5.2). Lastly, we demonstrate that our solutions close the gap with the cascade approach, outperforming the best IWSLT cascade models on the 4 publicly available validation sets (§5.3). Full experimental settings and details about training data are given in Appendix A to ensure the reproducibility of our work.

| decoding | de | es | fr | it | nl | pt | ro | AVG |
|---|---|---|---|---|---|---|---|---|
| standard | 20.9 | 34.3 | 27.9 | 28.6 | 30.9 | 35.1 | 30.1 | 29.7 |
| joint CTC | **21.8** | **35.2** | **28.6** | **28.9** | **31.1** | **35.7** | **30.8** | **30.3** |

Table 2: Translation quality results measured by AS-BLEU (↑) with and without joint CTC decoding for all the 7 languages of MuST-Cinema.

### 5.1 Timestamp Estimation Methods

Table 1 shows the results of our multilingual system with the proposed timestamp estimation methods, comparing them to the Levenshtein-based approach (LEV) of Papi et al. (2023a). Notice that we employ the same model for all rows, meaning that the outputs vary solely in terms of block timings, except for the ablation row (- joint CTC) where we analyze the impact of the joint CTC rescoring on the subtitle quality.

First of all, we notice a notable SubER decrease with joint CTC rescoring (-0.9 on average), while the different timing strategies have a lower impact on it (improvements remain below 0.4 on average), with the two attention-based methods (ATTN DTW and SBAAM) achieving the same scores. On the contrary, SubSONAR is not significantly affected by the decoding strategy but exhibits increasingly higher scores as block timings become more accurate. The substantial gains (>6% relative improvement) of SubCTC over LEV underscore the importance of directly estimating timestamps on the subtitle blocks, thereby avoiding inherent error propagation of mapping them from the caption blocks (as done in LEV). Consistently with the SubER scores, the attention-based methods lead to superior scores, with SBAAM emerging as the best timestamp estimation strategy, outperforming ATTN DTW by a limited, yet consistent, margin across all languages. Overall, SBAAM improves SubSONAR by 11.8% over the current state of the art (LEV) for direct subtitling.

In addition to certifying the effectiveness of our solutions, these results demonstrate that SubER is more sensitive to translation quality than timestamp accuracy. Comparing the results in Table 1 with those in Table 2, it is evident that the substantial gains achieved through joint CTC rescoring are proportional to the improvements in terms of AS-BLEU,[7] which measures the pure translation quality. Notably, the languages exhibiting the smallest decrease in SubER coincide with those where BLEU scores are closer (it and nl), while the language with the highest BLEU increase (es) also demonstrates the widest margin in terms of SubER. Conversely, timestamp accuracy shows limited effects on SubER scores, despite the important difference between methods, as attested by SubSONAR scores and validated by our manual analysis (§6).

Lastly, we verify the effect of the proposed solutions in terms of subtitle conformity, namely the percentage of blocks that respect character-per-line (CPL) and character-per-second (CPS) limits. Such limits, respectively set to 42 and 21 in TED guidelines,[8] ensure that subtitles can be understood by users without excessive cognitive effort. As it has been demonstrated that the joint CTC rescoring solves the end detection problem of purely attentional models and promotes hypotheses with correct length (Yan et al., 2023), we hypothesized that its use could improve both CPL and CPS metrics. However, inconsistent outcomes across various settings and languages (shown in Appendix B) dispute this hypothesis, whereas the attention-based timing methods consistently improve the CPS, with SBAAM being superior also in this respect.

### 5.2 High-Resource Conditions

To further validate the robustness of our findings, we experiment in high-resource conditions with bilingual systems (en-de and en-es) and compare our solutions with the state of the art. To cover different domains and data conditions besides MuST-Cinema, we leverage two additional test sets: European Commission (EC) Short Clips (Papi et al., 2023a), made of multi-speaker short informative videos about various topics with background music, and European Parliament (EP) Interviews (Papi et al., 2023a), interviews with non-verbatim subtitles. The results of our solution are reported in Table 3 and compared with the current best scores published on these test sets.

Upon comparing rows 1 and 3, we observe that our baseline model, even without joint CTC rescoring and replicating the LEV method, outperforms previous results across nearly all test sets, with the sole exception of MuST-Cinema en-es, affirming its competitiveness and reinforcing our findings. Furthermore, we observe consistent trends and similar improvements as in the previous section when applying joint CTC rescoring and employing the

[7] AS-BLEU, computed with SubER tool (AppTek, 2022), is calculated with the popular sacreBLEU metric (Post, 2018) after aligning blocks of the hypothesis and reference with the minimum Levenshtein distance method (Matusov et al., 2005).

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

| Model | align. | joint | SubER (↓) MSTCIN de | MSTCIN es | ECSC de | ECSC es | EPI de | EPI es | AVG | SubSONAR (↑) MSTCIN de | MSTCIN es | ECSC de | ECSC es | EPI de | EPI es | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Papi et al. (2023a) | LEV | ✗ | 59.9 | 46.8 | 59.9 | 52.7 | 80.3 | 72.3 | 62.0 | - | - | - | - | - | - | - |
| Best Production* | na | na | 61.5 | 51.3 | 59.0 | 49.7 | 78.1 | **68.6** | 61.4 | | | | | | | |
| This work | LEV | ✗ | 58.0 | 48.6 | 59.6 | 49.9 | 78.5 | 70.2 | 60.8 | .656 | .707 | .691 | .713 | .670 | .697 | .689 |
| | | ✓ | 56.5 | 45.1 | 58.9 | 49.0 | 77.6 | 69.7 | 59.5 | .668 | .712 | .693 | .722 | .676 | .701 | .695 |
| | DTW | ✓ | 56.3 | **44.7** | **58.5** | **48.9** | **77.3** | 69.7 | **59.2** | **.733** | .780 | **.742** | .778 | .733 | .763 | .755 |
| | SBAAM | ✓ | **56.2** | **44.7** | 58.6 | **48.9** | **77.3** | 69.6 | **59.2** | .728 | **.781** | .740 | **.785** | **.734** | **.770** | **.756** |

Table 3: SubER (cased) and SubSONAR results of our high-resource models with and without joint CTC generation (joint) and with the LEV, DTW or SBAAM timestamp alignment methods (align.), compared with previous work and production tools on MuST-Cinema (MSTCIN), EC Short Clips (ECSC) and EP Interviews (EPI) en-de and en-es. * Best results of the production tools reported by Papi et al. (2023a) for every language and test set.

**en-de**

| Model | *TED* SubER cased | *TED* SubER uncased | *EPTV* SubER cased | *EPTV* SubER uncased | *ITV* SubER cased | *ITV* SubER uncased | *PELOTON* SubER cased | *PELOTON* SubER uncased | ***AVG*** SubER cased | ***AVG*** SubER uncased |
|---|---|---|---|---|---|---|---|---|---|---|
| Best Cascade | - | 63.0 | - | 78.7 | - | 83.6 | - | 87.6 | - | 78.2 |
| Best Direct | 69.4 | - | 80.6 | - | 83.7 | - | 79.1 | - | 78.2 | - |
| This work | **61.6** | **62.1** | **78.7** | **78.3** | **80.0** | **80.7** | **75.6** | **78.2** | **74.0** | **74.8** |

**en-es**

| Model | *TED* SubER cased | *TED* SubER uncased | *EPTV* SubER cased | *EPTV* SubER uncased | *ITV* SubER cased | *ITV* SubER uncased | *PELOTON* SubER cased | *PELOTON* SubER uncased | ***AVG*** SubER cased | ***AVG*** SubER uncased |
|---|---|---|---|---|---|---|---|---|---|---|
| Best Cascade | - | 48.8 | - | **70.2** | - | 82.1 | - | **79.0** | - | 70.0 |
| Best Direct | 52.5 | - | 73.7 | - | 82.2 | - | 80.3 | - | 72.2 | - |
| This work | **49.5** | **47.5** | **73.1** | 71.0 | **79.1** | **79.5** | **79.3** | 80.8 | **70.2** | **69.7** |

Table 4: SubER (↓) comparison with the best cascade (AppTek) and direct (FBK) models trained on constrained conditions from the IWSLT 2023 Evaluation Campaign on automatic subtitling for en-de and en-es validation sets.

attention-based DTW and SBAAM methods for timestamp estimation. On average across all languages and test sets, joint CTC rescoring decreases the SubER by 1.3 with negligible effects on SubSONAR, while SBAAM and DTW yield limited improvement (0.3) in SubER but a substantial relative increase in SubSONAR, which amounts to 8.8% in the case of SBAAM that confirms to be the best method overall.

In fact, our proposed solution (last row) emerges as the best one on all test sets, except for EP Interviews en-es, where the best production system reported by Papi et al. (2023a)[9] has a better SubER (-1.0). On average, our model is 2.2 SubER better than cherry-picking the best production system for each test set and language, and 1.6 SubER better than our baseline. To further confirm its strength, in the next section, we compare it with the best systems of the last IWSLT campaign.

### 5.3 Is the Gap with Cascade Systems Closed?

In this section, we compare our high-resource models with the best direct (FBK – Papi et al. 2023b) and the best cascade (AppTek – Bahar et al. 2023) systems of the last IWSLT campaign (Agarwal et al., 2023), in the constrained data condition, meaning the models are trained on the same data as ours. As the references for the official test sets are not public, we present the results on the 4 validation sets released for the campaign in Table 4. For the sake of a fair comparison, we report the SubER with and without casing and punctuation, as Bahar et al. (2023) report the latter.

Even in this testing condition, our systems consistently outperform the others, surpassing even the top-performing cascade model on both language pairs. The superiority of our solution is particularly pronounced in en-de, where it achieves the lowest SubER in all conditions by a large margin (-3.4 SubER on average over the best cascade and -4.2 over the direct). On en-es, our solution and the cascade are, instead, close, with our models emerging on TED and ITV, whereas the cascade models are better on on EPTV and PELOTON. On average, however, our solution results slightly superior in this language pair with a 0.3 SubER reduction. All in all, our experiments evidence that our proposed

[9]Note that, as we did not test such production systems, they may have been improved at the time of writing this paper.

solution can effectively close the gap between cascade and direct subtitling systems for the first time.

## 6 Manual Evaluation

To corroborate the findings from our automatic evaluation, we conducted the first-ever manual evaluation of timestamp quality in subtitling. This evaluation was specifically conducted on the two language pairs addressed in our high-resource experiments (en-de, en-es).

To focus only on timestamp quality, we compared different timestamp estimation techniques on the same automatic translations. Specifically, we selected the outputs of our high-resource systems with joint CTC rescoring and confronted the baseline LEV method and our proposed SBAAM.

For each language pair, the manual evaluation set – on which the two systems were run – is composed of subsets of the EC Short Clips and MuST-Cinema test sets, for a total of 22 videos corresponding to approximately 1 hour of audio (see Appendix C.1 for details on the selected videos).

The evaluation was carried out by two annotators per language pair, who are proficient in English (C1) as well as native speakers or very proficient (C2) in the target language (German or Spanish).[10] We instructed the annotators with ad-hoc guidelines (described in Appendix C.2) to adjust the start, end, or both timestamps when the generated subtitles are not synchronized with the speech. To collect this information, we used ELAN (Wittenburg et al., 2006), an audio/video annotation tool commonly employed in the literature (Sloetjes and Wittenburg, 2008). To cope with the inherent assessors' subjectivity (i.e. being more aggressive/tolerant in deciding whether a timing is acceptable or has to be edited) the outputs of the two timestamp estimation methods were randomly assigned to the two annotators, ensuring that each one worked on all 22 audios and annotated half of the outputs from both methods. To prove the quality of the manual evaluation and better understand the difficulty of the task, for each language pair, 25% of the test set was double-annotated. We calculated Cohen's Kappa (Cohen, 1960) to measure the inter-annotator agreement on whether a start/end timestamp has to be edited or not. The resulting values – 0.65 for en-de and 0.61 for en-es – indicate a "substantial" agreement (Landis and Koch, 1977), confirming both the feasibility of this new evaluation task and the reliability of its outcomes.

[10]The annotators were paid 22€/h gross, in accordance with the average salary of data annotators (https://www.talent.com/salary?job=data+annotator).

| **en-de** | | | | |
|---|---|---|---|---|
| *model* | *time shift (ms)* | *% edited ts* | | |
| | | start | end | avg |
| LEV | 544 ± 718 | 38.47 | 40.74 | 39.61 |
| SBAAM | **321 ± 453** | **18.43** | **18.81** | **18.62** |
| **en-es** | | | | |
| *model* | *time shift (ms)* | *% edited ts* | | |
| | | start | end | avg |
| LEV | 520 ± 626 | 38.01 | 39.45 | 38.73 |
| SBAAM | **347 ± 570** | **11.17** | **12.89** | **12.03** |

Table 5: Results of the manual evaluation in terms of time shift in milliseconds (mean and standard variation), and percentage of the number of edited timestamps (divided in start and end timestamp, and their average).

Table 5 shows the results of the manual analysis in terms of start/end time shifts (in milliseconds) and percentage of modified timestamps. The superiority of SBAAM over LEV is evident across all metrics, domains, and language pairs. The percentage of modified timestamps is less than half in en-de and 3 times lower in en-es. Given that correcting automatic timestamp errors is a major concern for professionals in their subtitling experience with AI-based systems (Koponen et al., 2020a), the reduced error rates of SBAAM are likely to significantly alleviate post-editing effort. Furthermore, SBAAM exhibits a notably lower average time shift (33-40%), indicating not only less frequent but also less severe errors. The lower average time shift is complemented by a significantly lower standard deviation, which further evidences that SBAAM errors are less likely to be large.

Lastly, in Appendix C.3 we present the results of this manual analysis after excluding edits under 120 ms. This evaluation caters to user experience, as users typically perceive audio-visual stimuli under 120 ms as instantaneous (Efron, 1970), while the annotators – aided by ELAN that shows the text and waveform – made numerous fine-grained timestamp edits (even <20 ms). This evaluation accentuates the differences between the two methods and shows that SBAAM requires editing for only 12% of timestamps in en-de and 9% in en-es.

We can conclude that our manual analysis not only confirms the results obtained through the automatic SubSONAR metric, but also further highlights the efficacy of our proposed timestamp estimation method and its substantial superiority over previous solutions.

## 7 Conclusions

In recent years, research on automatic subtitling has shifted toward direct models that do not rely on the transcription of the input audio. The potential advantages of a transcription-free approach have motivated the use of direct models for the translation and segmentation steps, leaving unexplored the direct generation of timestamps on the target text. In this paper, we filled this gap by introducing the first direct system that does not require the generation of transcripts/captions in any phase of the process, including subtitle timestamp estimation. With the introduction of a new metric, SubSONAR, dedicated to evaluating timestamp quality, and experiments on different domains and 7 language pairs, we demonstrated the effectiveness of our solution, which was further validated by a dedicated manual evaluation. Lastly, we showed that our model closed the gap with the cascade paradigm, approaching and even outperforming the best cascade architectures from the last IWSLT campaign, thereby setting a new state of the art for most of the publicly available benchmarks.

## Limitations

While our direct AS model achieves significant advancements, there are still some limitations or problems that have not been addressed in this work and should be the focus of further research on the topic.

First of all, although our models are easily applicable to unwritten languages (the only required change is either to use speech units as targets for the CTC compression module or to remove the module itself), our experiments have not included unwritten languages due to the lack of available benchmarks. Moreover, for the same reason, we have not experimented with source languages different from English. Despite this, we do not foresee any specific problem that could arise in this condition, except for a slight drop in translation quality if the CTC compression and its auxiliary loss are removed (Zhang et al., 2022a).

Another limitation of this work regards the model compliance with constraints posed on CPL and CPS. While our model is trained on subtitles that mostly adhere to spatio-temporal requirements, no specific strategies are adopted for their actual fulfillment. Modifying the model or the training strategy to consistently meet these constraints, especially in complex audiovisual content, should be the topic of future works.

Regarding SubSONAR, its language coverage is currently limited to the languages supported by the SONAR speech encoder.

Lastly, due to the large computational costs required and for the sake of a fair comparison with previous works, we did not experiment with using our proposed method on top of large foundational ST models such as SeamlessM4T (Seamless Communication et al., 2023) and Whisper (Radford et al., 2023), although the latter can not be tested in our settings as it does not support translating from English into other languages. To adopt these models for the subtitling task, line and block boundary tokens should be added to their vocabulary and these models should be then fine-tuned on subtitling data (i.e. translations with line and block boundaries). No other modification is needed as SBAAM requires only looking at the cross-attention. Optionally, a CTC on Target module has to be trained on top of their encoder for the joint CTC rescoring. This line of research represents a natural next step of this work.

## Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The work is co-funded by the European Union under the project *AI4Culture: An AI platform for the cultural heritage data space* (Action number 101100683). We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. We also thank Giorgia Pucker, Lea Glaubig, and two anonymous annotators for the manual evaluation.

## References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega,

Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online).

Belen Alastruey, Aleix Sant, Gerard I. Gállego, David Dale, and Marta R. Costa-jussà. 2024. SpeechAlign: a Framework for Speech Translation Alignment Evaluation.

Aitor Alvarez, Carlos-D. Martnez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. *Speech Commun.*, 88(C):83–95.

AppTek. 2022. SubER - Subtitle Edit Rate. `https://github.com/apptek/SubER`.

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.

Parnia Bahar, Patrick Wilken, Javier Iranzo-Sánchez, Mattia Di Gangi, Evgeny Matusov, and Zoltán Tüske. 2023. Speech translation with style: AppTek's submissions to the IWSLT subtitling and formality tracks in 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 251–260, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Łukasz Bogucki. 2004. The constraint of relevance in subtitling. *The Journal of Specialised Translation*.

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.

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.

Xianzhao Chen, Yist Y. Lin, Kang Wang, Yi He, and Zejun Ma. 2023. Improving Frame-level Classifier for Word Timings with Non-peaky CTC in End-to-End Automatic Speech Recognition. In *Proc. INTERSPEECH 2023*, pages 2908–2912.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online.

Colin Cherry, Naveen Arivazhagan, Dirk Padfield, and Maxim Krikun. 2021. Subtitle Translation as Markup Translation. In *Proc. Interspeech 2021*, pages 2237–2241.

Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-yi Lee. 2021. Investigating the Reordering Capability in CTC-based Non-Autoregressive End-to-End Speech Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1068–1077, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-Many Multilingual End-to-End Speech Translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

Robert Efron. 1970. The minimum duration of a perception. *Neuropsychologia*, 8(1):57–63.

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.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. 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.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China.

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, page 369–376, New York, NY, USA.

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 *Proc. Interspeech 2020*, pages 5036–5040.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pages 198–208, Cham.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

T. Huang, G. Yang, and G. Tang. 1979. A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):13–18.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual End-to-End Speech Translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577.

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.

B.-H. Juang. 1984. On the hidden markov model and dynamic time warping for speech recognition — a unified view. *AT&T Bell Laboratories Technical Journal*, 63(7):1213–1243.

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.

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

Bilal Khalaf. 2016. An introduction to subtitling: Challenges and strategies. *International Journal of Comparative Literature and Translation Studies*, 3.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020a. MT for subtitling: Investigating professional translators' user experience and feedback. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 79–92, Virtual. Association for Machine Translation in the Americas.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020b. 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.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia.

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.

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.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. 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.

Antoni Oliver Gonzalez. 2006. Automatic multilingual subtitling in the eTitle project. In *Proceedings of Translating and the Computer 28*, London, UK. Aslib.

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.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct Speech Translation for Automatic Subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, and Matteo Negri. 2023b. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Sara Papi, Matteo Negri, and Marco Turchi. 2023c. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada.

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 *Proc. Interspeech 2019*, pages 2613–2617.

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.

David Ponce, Thierry Etchegoyhen, and Victor Ruiz. 2023. Unsupervised subtitle segmentation with masked language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 771–781, Toronto, Canada.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding Non-Autoregressive Neural Machine Translation Decoding with Reordering Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13727–13735.

Tara N. Sainath, Ruoming Pang, David Rybach, Basi García, and Trevor Strohman. 2020. Emitting Word Timings with End-to-End Models. In *Proc. Interspeech 2020*, pages 3615–3619.

Haşim Sak, Andrew Senior, Kanishka Rao, Ozan İrsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk. 2015. Learning acoustic frame labeling for speech recognition with recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4280–4284.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isochrony-Aware Neural Machine Translation for Automatic Dubbing. In *Proc. Interspeech 2022*, pages 1776–1780.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Jonathan H Venezia, Steven M Thurman, William Matchin, Sahara E George, and Gregory Hickok. 2016. Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, perception & psychophysics*.

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.

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.

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.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

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.

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).

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On decoder-only architecture for speech-to-text and large language model integration.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021a. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. 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 2619–2630, Online. Association for Computational Linguistics.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent Advances in Direct Speech-to-text Translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6796–6804. ijcai.org.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021b. How Does Distilled Data Complexity Impact the Quality and Confidence of Non-Autoregressive Machine Translation? In *Findings of*

*the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, Online. Association for Computational Linguistics.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. CTC alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia.

Zhang Yuxin and Yoshikazu Miyanaga. 2011. An improved dynamic time warping algorithm employing nonlinear median filtering. In *2011 11th International Symposium on Communications & Information Technologies (ISCIT)*, pages 439–442.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022a. Revisiting End-to-End Speech-to-Text Translation From Scratch. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26193–26205.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1663–1676, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

# A Experimental Settings

## A.1 Data

For the multilingual model, we leverage all the 7 language pairs of MuST-Cinema v1.1 (Karakanta et al., 2020b), namely: English to German, Spanish, French, Italian, Dutch, Portuguese, and Romanian. These texts already contain line and block segmentation, i.e., <eol> and <eob> tags are present in both transcripts and translations.

For the bilingual models, we use the same datasets of (Papi et al., 2023b), encompassing most of the training data admitted by the IWSLT 2023 Evaluation Campaign on automatic subtitling. We collect all the available ST corpora, namely MuST-Cinema, EuroParl-ST (Iranzo-Sánchez et al., 2020), and CoVoST v2 (Wang et al., 2020b). Also, we leverage most of the available ASR datasets (CommonVoice (Ardila et al., 2020), LibriSpeech (Panayotov et al., 2015), TEDLIUM v3 (Hernandez et al., 2018), and VoxPopuli (Wang et al., 2021)), by automatically translating the transcripts into the two target languages (German and Spanish) using the NeMo MT models.[11]

<eol> and <eob> tags are added to both transcripts and translations of all datasets, except for MuST-Cinema that already include them, using the multimodal segmenter by Papi et al. (2022).

## A.2 Model and Training

Our systems are implemented with fairseq-ST (Wang et al., 2020a) using the default settings unless specified otherwise. The input comprises 80 Mel-filterbank audio features extracted every 10 milliseconds, employing a sample window of 25. The input features are then preprocessed with two 1D convolutional layers with stride 2, reducing input length by a factor of 4.

The model architecture follows an encoder-decoder design, consisting of a Conformer encoder (Gulati et al., 2020) and a Transformer decoder (Vaswani et al., 2017), with a total number of 133M for both multilingual and bilingual models. The vocabularies are based on unigram SentencePiece (Kudo, 2018), with size 8,000 for the English source and 16,000 for the target (either German, Spanish, or multilingual). Specific hyperparameters are presented in Table 6.

[11] Publicly available at: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/machine_translation/machine_translation.html

| **Encoder** | |
|---|---|
| Layer type | Conformer |
| Total number of layers | 12 |
| $N_a$ layers | 8 |
| $N_s$ layers | 4 |
| Embedding dimension | 512 |
| FFN dimension | 2,048 |
| Convolutional Module kernel size (point- and depthwise) | 31 |
| **Decoder** | |
| Layer type | Transformer |
| $N_d$ layers | 6 |
| Embedding dimension | 512 |
| FFN dimension | 2,048 |

Table 6: Hyperparameters for the proposed model.

| | |
|---|---|
| Optimizer | AdamW |
| Optimizer Momentum | $\beta_1, \beta_2 = 0.9, 0.98$ |
| Source CTC weight ($\lambda_1$) | 1.0 |
| Target CTC weight ($\lambda_2$) | 2.0 |
| CE weight ($\lambda_3$) | 5.0 |
| CE label smoothing | 0.1 |
| Learning Rate scheduler | Noam |
| Learning Rate | 2e-3 |
| Warmup steps | 25,000 |
| Weight Decay | 0.001 |
| Dropout | 0.1 |
| Clip Normalization | 10.0 |
| Training steps | 200,000 |
| Maximum tokens | 40,000 |
| Update frequency | 2 |

Table 7: Model detailed training settings. The total batch size (in number of tokens) is obtained by multiplying maximum tokens for update frequency.

Label-smoothed CE is computed after the decoder, target CTC loss is computed on the output of the semantic encoder (full encoder), and the source CTC loss is computed on the output of the acoustic encoder. The values of the weights for all three losses ($\lambda_1, \lambda_2, \lambda_3$) are set to (1.0, 2.0, 5.0) according to (Yan et al., 2023). Utterance-level Cepstral Mean and Variance Normalization (CMVN) and SpecAugment (Park et al., 2019) are applied during training and segments longer than 30 seconds are filtered out (fairseq-ST default) to avoid excessive VRAM requirements. All model checkpoints are obtained by averaging the last 7 checkpoints obtained from the training. All trainings are executed on 4 NVIDIA Ampere GPU A100 (64GB VRAM).

For the multilingual case, we train the model in one step on all the MuST-Cinema languages by pre-pending the language tag to the subtitle texts, as already explained in §2. For the bilingual case, we first train the model on the ST and the machine-translated ASR datasets without <eol> and <eob> tags, and then we continue the training starting from the averaged checkpoint by including <eol>

| decoding | de | es | fr | it | nl | pt | ro | AVG |
|---|---|---|---|---|---|---|---|---|
| standard | 89.6 | **94.7** | **91.7** | 88.7 | 84.1 | 89.1 | 92.0 | 90.0 |
| joint CTC | **90.1** | 94.6 | 91.0 | **89.3** | **85.1** | **89.4** | **93.7** | **90.5** |

Table 8: Results of the CPL conformity (↑) in percentage (%) with and without joint CTC decoding for all the 7 languages of MuST-Cinema.

| decoding | MSTCIN | | ECSC | | EPI | | AVG |
|---|---|---|---|---|---|---|---|
| | de | es | de | es | de | es | |
| standard | 88.2 | **95.3** | **85.0** | 90.9 | **82.3** | **90.8** | **88.8** |
| joint CTC | **88.3** | 94.6 | 84.0 | **91.1** | 82.2 | 89.7 | 88.3 |

Table 9: CPL conformity results (↑) in percentage (%) of our high-resource models with and without joint CTC generation on MuST-Cinema (MSTCIN), EC Short Clips (ECSC) and EP Interviews (EPI) en-de and en-es.

and <eob> tags in the texts. For both cases (and steps, in the case of bilingual models), we use the training settings provided in Table 7.

For inference, we set the beam size to 5 and, according to (Yan et al., 2023), the joint CTC decoding weight $\alpha$ to 0.2. For the SBAAM timestamp estimation method, we extract the cross-attention from the 4$^{th}$ layer and average the scores across the attention heads, following (Papi et al., 2023c).

## B CPL and CPS Conformity

Table 8 and 9 report the CPL conformity of, respectively, our multilingual and bilingual models with and without the joint CTC decoding strategy. As we can see, the results are very close and no clear and coherent trend across the different languages and test sets emerges. We can conclude that the decoding strategy does not significantly impact CPL conformity.

Switching to analyze the CPS conformity, Table 10 and 11 show the related percentages for the multilingual and bilingual models. In this case, we do not only study the decoding strategy but also the timing estimation method used, as CPS is influenced both by the generated subtitles and the assigned timestamps. While also in this case the difference between the decoding strategies is limited, on average the joint CTC rescoring leads to a lower conformity in both cases. Looking at the differences between the timing estimation methods (Table 10), we notice that SubCTC is by far the worst method (with a ∼7% degradation). ATTN DTW and SBAAM, instead, lead to subtitles with higher conformity than the LEV baseline, although SBAAM is consistently slightly superior in all languages. The benefits of SBAAM over LEV

| model | de | es | fr | it | nl | pt | ro | AVG |
|---|---|---|---|---|---|---|---|---|
| LEV | 74.1 | 77.4 | 68.7 | 76.9 | 79.2 | 79.8 | **84.1** | 77.2 |
| - joint CTC | 73.6 | 75.8 | 67.5 | 77.8 | 77.5 | 79.8 | 83.9 | 76.6 |
| SubCTC | 68.3 | 70.6 | 63.7 | 69.6 | 72.9 | 72.8 | 77.1 | 70.7 |
| ATTN DTW | 75.6 | 79.3 | 71.3 | 78.0 | 80.5 | 81.4 | 83.8 | 78.6 |
| SBAAM | **75.7** | **79.7** | **72.5** | **78.5** | **81.7** | **82.1** | 84.0 | **79.2** |

Table 10: Results of the timestamp estimation methods in terms of CPS conformity (↑) in percentage (%) for all the 7 languages of MuST-Cinema.

| align. | joint | MSTCIN | | ECSC | | EPI | | AVG |
|---|---|---|---|---|---|---|---|---|
| | | de | es | de | es | de | es | |
| LEV | ✗ | **78.8** | **81.7** | 82.3 | 83.7 | 74.1 | 77.4 | **79.7** |
| | ✓ | 77.8 | 76.0 | **82.8** | 85.2 | **74.6** | 76.6 | 78.8 |
| SBAAM | ✓ | 76.7 | 77.5 | 82.0 | **87.4** | 73.4 | **78.7** | 79.3 |

Table 11: CPS conformity results (↑) in percentage (%) of our high-resource models with and without joint CTC generation (joint) and with the LEV or SBAAM timestamp alignment methods (align.) on MuST-Cinema (MSTCIN), EC Short Clips (ECSC) and EP Interviews (EPI) en-de and en-es.

are confirmed also by the bilingual systems, where SBAAM has a 0.5% higher compliance on average, although the gains are not coherent across the two language pairs.

We can conclude that we can reject the hypothesis about the potential benefits of the joint CTC rescoring in producing outputs of the correct length. The better timings assigned by the timestamp estimation methods, instead, provide little benefits in terms of CPS conformity.

## C Human Evaluation

### C.1 Audio Selection

Table 12 lists the audios included in the manual evaluation carried out by the annotators.

### C.2 Guidelines for Subtitle Timestamps Evaluation

The following italic text has been given to the annotators as guidelines for their work.

*You are asked to check if the source language speech and its corresponding translated subtitles are synchronized, that is if the subtitle remains on screen for the right amount of time with respect to the corresponding speech.*

*If the timing is wrong, move the timestamp of each subtitle such that the timing of the subtitle matches the timing of the audio. The timestamp can be adjusted at the beginning by changing the start timestamp, at the end by changing the end timestamp, or both. Change the timestamp only*

| Test Set | Audio Name | Duration |
|---|---|---|
| ECSC | I118982 | 0:02:38 |
| | I200637 | 0:01:30 |
| | I203453 | 0:01:38 |
| | I205895 | 0:01:31 |
| | I207338 | 0:01:29 |
| | I207340 | 0:01:35 |
| | I207805 | 0:01:52 |
| | I212173* | 0:03:01 |
| | I207806 | 0:02:00 |
| | I207807 | 0:01:54 |
| | I207810 | 0:01:59 |
| | I207813 | 0:02:21 |
| | I207814 | 0:01:58 |
| | I209528 | 0:02:08 |
| | I211037 | 0:01:58 |
| | I212173* | 0:03:01 |
| MSTCIN | 13518 | 0:05:23 |
| | 20008 | 0:05:03 |
| | 28521* | 0:04:41 |
| | 22979 | 0:05:02 |
| | 23129 | 0:05:35 |
| | 28521* | 0:04:41 |
| | | 1:02:58 |

Table 12: Audio selection for the manual analysis from EC Short Clips (ECSC) and MuST-Cinema (MSTCIN). * Repeated two times for computing the inter-annotator agreement for the two methods.

*when the content is not aligned or only partially aligned. Please notice that the content of the subtitles, even if not correct, must not be changed. Only modifications to the timestamps are allowed.*

*To summarize, please follow these guidelines:*

- *Adjust the timestamp such that the subtitle is synchronized with the corresponding audio: for this purpose, you can adjust the timestamp either at the beginning, at the end, or both;*
- *Make only necessary changes to align the partially or completely misaligned subtitle with the audio;*
- *Make only necessary changes to align the partially or completely misaligned subtitle with the audio;*
- *Do not change the subtitle content.*

*You will be given 22 videos to annotate and the work consists of around 6 hours of annotation. Please annotate the audio following the given order (from 1 to 22). To carry out your evaluation work at best, you should work on around 5 minutes of audio without interruptions. Then, you should take a break of around 5 minutes. We suggest to take a break after file numbers: 03, 06, 08, 11, 14, 16, 17, 18, 19, 20, 21, 22.*

| **en-de** | | | | |
|---|---|---|---|---|
| *model* | *time shift* | *% edited ts* | | |
| | | start | end | avg |
| LEV | $606 \pm 743$ | 33.18 | 36.58 | 34.88 |
| SBAAM | $\mathbf{422 \pm 503}$ | **12.95** | **13.52** | **13.24** |
| **en-es** | | | | |
| *model* | *time shift* | *% edited ts* | | |
| | | start | end | avg |
| LEV | $629 \pm 1015$ | 28.56 | 33.43 | 31.00 |
| SBAAM | $\mathbf{460 \pm 632}$ | **8.31** | **9.65** | **8.98** |

Table 13: Results of the manual evaluation in terms of time shift in milliseconds (mean and standard variation), and percentage of the number of modified timestamp (divided by start and end timestamp, and their average) with filtered shifts under 120 milliseconds.

*It can happen that you have to annotate an audio that you have already annotated, please annotate it as if it were the first time.*

In addition, specific guidelines on the usage of the annotation tool, ELAN, were provided (Figure 2 shows the interface). An initial training day, conducted with all the annotators and consisting of 3 hours of work, was held to explain the guidelines to the annotators, annotate a pilot audio, and discuss the results and common questions. This, with the 6 hours of annotation, resulted in 9 hours of work.

### C.3 Results with perception filtering

As works in the field of human perception (Efron, 1970; Venezia et al., 2016) have demonstrated that humans perceive acoustic and visual stimuli with duration inferior to 120ms as instantaneous, we re-compute the statistics of the manual analysis by filtering out time shifts of less than 120 ms. With this change, the resulting inter-annotator agreement increases to a Cohen's Kappa of 0.71 for en-de, and 0.68 for en-es, which is again "substatial" in both languages but higher than the one obtained in §6.

The results of the two timestamp estimation methods (LEV and SBAAM) with the filtered shifts are shown in Table 13. First of all, we notice that the time shift mean of SBAAM is $\frac{2}{3}$ of LEV, with important differences also in terms of the standard deviation, resulting in a reduction of 240ms for en-de and 383ms for en-es. Therefore, with filtered shifts, the gap between the two methods is more exacerbated than that shown in Table 5, although with similar conclusions. Regarding the percentage of edited timestamps, we observe a relevant decrease in the number of edits for both methods, with SBAAM achieving less than 9% of edited timestamps in en-es. Also in this case, the superior-

ity of the SBAAM method over LEV is widened, as the number of edits for LEV is slightly less than 3× those for SBAAM in en-de, and nearly 3.5× in en-es.

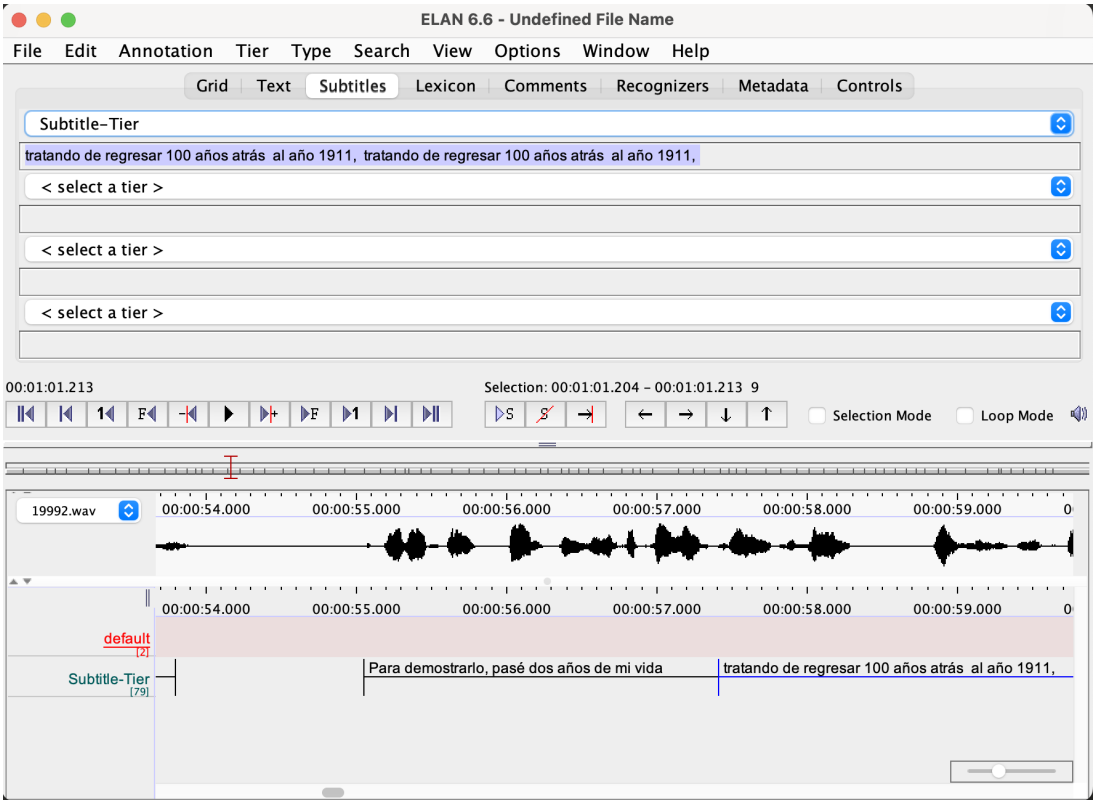


Figure 2: ELAN interface.