The FBK Sentence-Aware Subtitling System at the IWSLT 2026 Subtitling Track

Mauro Cettolo, Roldano Cattoni, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler, Trento, Italy
{cettolo,cattoni,negri,bentivo}@fbk.eu

Abstract

This paper describes the FBK submissions to the Subtitling track of the 2026 IWSLT Evaluation Campaign. The task requires automatically subtitling English audio-visual content across three domains (ITV entertainment series, Asharq-Bloomberg news programs, and YouTube recordings from the YODAS dataset), into up to four target languages per domain, chosen from a pool of five (Arabic, Chinese, German, Japanese, and Spanish). All submitted systems are based on an ASR-MT cascade framework built exclusively from freely available open-source components usable without restrictions, including for commercial purposes. Our primary system implements a two-stage pipeline: the first stage produces time-aligned subtitles via voice activity detection, automatic transcription, and subtitle-level translation, while the second refinement stage re-processes the audio at a longer context level, combining long-form transcription with sentence-level translation, and re-aligning the resulting output to the original subtitle timing. This design preserves synchronization constraints while leveraging broader context to improve both transcription and translation quality. We also submitted two contrastive systems: one corresponding to the first-stage baseline pipeline, and another sharing the same baseline architecture but using alternative components.

1 Introduction

The paper describes the FBK submissions to the Subtitling track of the 2026 IWSLT Evaluation Campaign, which asked participants to automatically subtitle three kinds of audio-visual documents, where the spoken language is always English: ITV entertainment series, to be subtitled in Chinese, German, Japanese, and Spanish (Europe); news programs from the Asharq-Bloomberg platform, to be subtitled in Arabic, Chinese, German, and Japanese; and audio recordings from the YODAS YouTube dataset, to be subtitled in Chinese, German, and Japanese.

We submitted runs from three subtitling systems based on an ASR-MT cascade framework, all built using freely available open-source components that can be used without restrictions, including for commercial purposes. The contrastive-1 system corresponds to a baseline pipeline in which audio is segmented via SpeechBrain voice activity detection, transcribed with a Whisper-based ASR model producing time-aligned subtitles, and translated at the *subtitle* level using a multilingual MT model (MADLAD-400 10B).

The primary system extends this approach with a second refinement stage aimed at improving textual quality. In this stage, adjacent audio segments are aggregated into longer units, transcribed with a more accurate ASR model for long-form speech (VOXTRAL), segmented into *sentence* level units, and translated accordingly. The resulting translations are then re-segmented to match the original subtitle timing and inserted into the same SRT template generated in the first stage. This design allows for preserving subtitle synchronization while leveraging sentence-level context to improve both transcription and translation quality.

In addition, we submitted a second contrastive system that follows the same baseline architecture but differs in the models employed, using SHAS for audio segmentation, Faster Whisper for ASR, and a smaller MADLAD-400 (3B) model for MT.

Results on development and evaluation sets demonstrate the effectiveness of the proposed two-stage framework, demonstrating that revisiting segmentation at the sentence level is a key factor for improving our baseline subtitling performance.

2 Two-Stage Subtitling Framework

In this section, we present the overall architecture of our subtitling system. First, we briefly describe

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 68–77

the models we used in developing the subtitle system (Section 2.1). We then illustrate the system architecture (Section 2.2), detailing how these models are employed across the different processing steps (Sections 2.3 and 2.4). The description ends with a list of some additional modules that automatically edit the subtitles before the final release (Section 2.5).

2.1 Models

SHAS (Tsiamas et al., 2022) is an audio segmentation method that leverages pre-trained classification models to perform segmentation. In general, it cannot be assumed that each audio segment corresponds to a single subtitle; therefore, an additional segmentation step is required to split each segment into subtitle units. SHAS code and models are released under the MIT License.

SpeechBrain VAD (SB VAD) is a ready-to-use model for Voice Activity Detection (VAD) released inside SpeechBrain, an open-source[1] and all-in-one conversational AI toolkit (Ravanelli et al., 2021). The pre-trained VAD model can process both short and long speech recordings and outputs segments where speech activity is detected. As with SHAS, audio segments generated by SB VAD typically need to be further segmented to match individual subtitles.

Whisper (Radford et al., 2023) (WH) is a family of open-source models[2] for speech-related tasks trained on 680,000 hours of labeled audio data collected from the web. Built upon a Transformer-based encoder-decoder architecture, WH is capable of performing ASR (supporting nearly 100 languages), ST (from those languages into English), language identification and timestamp estimation. This latter functionality enables WH to produce time-aligned transcripts, including output in SRT format. We employed the 1.55B parameters `large-v3 multilingual` model[3] to perform the speech transcription.

Faster Whisper [4] (FW) is a reimplementation of WH using CTranslate2, which is a fast inference engine for Transformer models. This implementation is faster and less memory demanding than WH for an equivalent accuracy. It is licensed under the MIT License. The models, being just a converted version of the original ones, inherit their license (Apache 2.0).

VOXTRAL (Liu et al., 2025) (VX) is an open-source (Apache 2.0) automatic speech recognition model belonging to the Mistral AI family. Its performance is comparable to WH large models on VAD-segmented audio, while it achieves better results on longer audio segments than those typically produced by VAD. However, unlike WH, it cannot generate time-aligned transcripts in SRT format. As a result, it cannot directly replace WH in our subtitling pipeline. In Section 2.2, we describe how VX, in particular the 3B parameters `Voxtral-Mini-3B-2507` model,[5] is integrated into our architecture.

MADLAD (Kudugunta et al., 2023) is a family of large-scale, pre-trained Transformer models for MT, licensed under the Apache License 2.0. These models are designed to support translation across over 400 languages, many of which are low-resource and underrepresented in existing MT systems. The training corpus for MADLAD models comprises approximately 18 billion sentence pairs, mined and filtered from a wide variety of multilingual web and public sources to ensure broad linguistic and domain diversity. In our experiments, we used `MADLAD-400-10B-MT`,[6] a 10.7B-parameters MT model capable of performing bidirectional translation between all supported languages.

2.2 Architecture

Figure 1 illustrates the architecture of our subtitling system. It consists of two main stages. Stage 1 follows a conventional ASR–MT pipeline; the subtitles it generates serve as the baseline and constitute our `contrastive-1` submission. Stage 2 revisits segmentation to better match sentence-level translation units; its subtitles constitute our `primary` submission. The two stages are described in the following sections.

2.3 Stage 1: Baseline Subtitling

The first stage, shown on the left side of Figure 1 and highlighted in blue, closely resembles the cascade we used in the 2024 edition of the shared task (Gaido et al., 2024; Ahmad et al., 2024), with the main differences lying in the adopted models.

[1] SpeechBrain is licensed under the Apache License 2.0.

[2] Whisper is licensed under the MIT License while the models are under the Apache License 2.0.

[3] https://github.com/openai/whisper

[4] https://github.com/SYSTRAN/faster-whisper

[5] https://huggingface.co/mistralai/Voxtral-Mini-3B-2507

[6] https://huggingface.co/google/madlad400-10b-mt

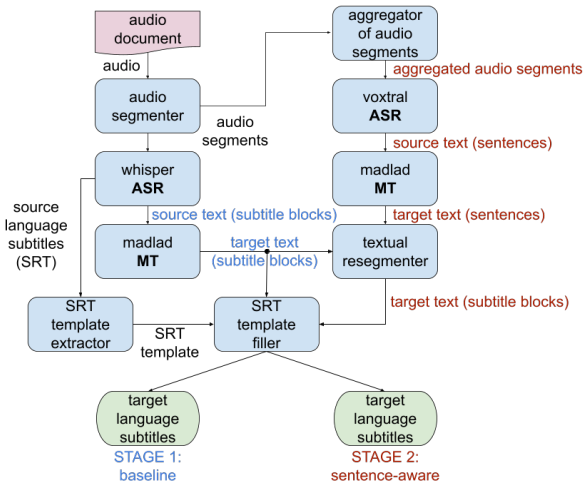

Figure 1: Two-stage subtitling architecture.

The audio extracted from the input audiovisual content is first segmented using SB VAD. Each segment is then transcribed with WH, enabling SRT output generation and thus producing a further segmentation into subtitle units. The text of each subtitle is subsequently translated using MADLAD and used to replace the original transcript in the SRT file.

The SRT of the translated subtitles is then released as the output of the first stage, after applying minor post-processing steps described in Section 2.5.

2.4 Stage 2: Sentence-aware Refinement

The second stage represents the main novelty with respect to our previous participation in the subtitling task. The key insight underlying its design is that subtitle-level segments are often suboptimal for both ASR and MT, while longer, sentence-like units lead to better transcription and translation quality. It is shown on the right side of Figure 1 and highlighted in red.

We improve over the baseline by aggregating adjacent audio segments generated by SB VAD, according to heuristics that consider both segments' time proximity and segments' duration. Thresholds for each heuristic were tuned on the ITV dev26 set. For proximity, we experimented with gaps of 1, 2, 5, 10, and 20 seconds between consecutive segments, selecting 10 seconds as the optimal value. For maximum duration, we started from the 30-minute upper limit supported by VOXTRAL for transcription and progressively reduced it until out-of-memory errors were fully eliminated across all development sets, arriving at a final threshold of 600 seconds.

The resulting source text is then segmented into sentences in correspondence of strong punctuation marks (.!?:). These sentences are translated using MADLAD, whose task is facilitated compared to the baseline setting, as it is trained to translate well-formed sentences rather than the partial fragments typically found in subtitles.

The sentence-level translations are then re-segmented at the word level to match the baseline subtitle segmentation using `mweralign` (Post and Hoang, 2025), a Python reimplementation of the method proposed by Matusov et al. (2005), originally distributed as an executable binary without source code.

Finally, the re-aligned translation is inserted into the same SRT template generated during Stage 1, preserving the original subtitle timing.

2.5 Post-processing of Subtitles

The SRT files generated by both stages are post-processed before release to improve subtitle compliance in terms of reading speed, maximum line length, and maximum number of lines per subtitle.

Reading speed. If a subtitle exceeds the predefined reading speed threshold (4 cps for Japanese, 9 cps for Chinese, and 21 cps for other languages), its end timestamp is extended to the minimum value that satisfies the constraint, if possible; otherwise, it is set to coincide with the start time of the following subtitle.

Line length and number of lines. If the text of a subtitle exceeds the maximum allowed length (13/16/42 characters for Japanese, Chinese, and other languages, respectively), it is split into lines of compliant and balanced length, prioritizing strong punctuation (.!?;) and then weak punctuation (,:-). If more than two lines would be required, the subtitle is further split into multiple subtitles. In this case, the end time of each subtitle coincides with the start time of the next one, and intermediate timestamps are computed based on an average character duration:

$$\texttt{char\_dur} = \frac{\texttt{duration}_{\texttt{original subtitle}}}{\texttt{total\_chars}_{\texttt{original subtitle}}}$$

which is then used to assign durations to the newly created subtitles, proportional to the number of characters present in each of them.

Hallucination filtering. A known failure mode of both ASR and MT models is the generation of nonsensical repetitions, a form of hallucination (Ji

et al., 2023) in which the model loops over the same output fragment rather than producing meaningful content. To address this, we apply a post-processing script that removes consecutive repetitions of n-grams from the generated text. Specifically, the script scans the output for sequences in which the same n-gram (for n ranging from 4 down to 1) appears at least four times consecutively, retaining only a single occurrence. The detection is applied iteratively from the largest n to the smallest, so that longer repeated patterns are resolved before shorter ones.

3 Results

In this section, we present and discuss our experimental results. In Section 3.1, we report the results obtained on the development sets, which guided our choice of models for the final system. Computational costs of each processing step are quantified and discussed in Section 3.2. In Section 3.3, we describe the submitted runs and present a selection of the most relevant official results provided by the task organizers.

3.1 System Development

Model selection mainly concerns the early stages of the subtitling pipeline, namely audio segmentation and speech transcription; it is discussed in the next section. As for machine translation, we rely on the `MADLAD-400-10B-MT` model (Section 2.1), which has previously demonstrated strong performance while being freely usable; in Section 3.1.2, we compare different setups to determine which one allows us to use the MT model most effectively. The validation of the overall two-stage subtitling architecture, as well as the effectiveness of the selected models, is addressed in Section 3.1.3.

3.1.1 Audio segmentation and ASR

Regarding audio segmentation, the system we used for the 2024 edition of the task (Gaido et al., 2024) relied on SHAS. Given the widespread use of SB VAD and its known robustness in noisy conditions, we compared SHAS and SB VAD on the ITV dev2026 set.

The first two rows of Table 1 report the main statistics of the segmentation step performed by the two segmenters. SB VAD identifies about 20% more speech content than SHAS, while the average segment length remains comparable (approximately 8.6 seconds).

Audio Segments by	duration	#segs
SHAS	3445.2	401
SB VAD	4132.4	478
aggregation of SB VAD segs	5403.1	104
total	8090.1	3

Table 1: Total duration (in seconds) of audio segments of the ITV dev2026 set generated by SHAS, SB VAD and after the aggregation of SB VAD segments. The last row provides the total gross duration of the 3 audio docs of the set.

	Audio Segments by		
	SHAS	SB VAD	aggregation of SB VAD segs
FW	31.45	19.72	21.51
WH	29.79	18.48	19.60
VX	31.90	19.04	16.86

Table 2: %WER scores of transcriptions of the ITV dev2026 set generated by FW, WH, and VX ASR models starting from audio pre-segmented by three different segmenters.

The two columns SHAS and SB VAD of Table 2 report the performance of the three ASR systems considered in this study (FW, WH, and VX) when transcribing speech segments generated by the two segmenters. The %WER scores obtained on SB VAD segmentation are significantly better than those obtained on SHAS segmentation, mainly due to a lower number of deletions and a higher number of correctly recognized words. For example, the two %WER values for WH are obtained from these counts:

$$\mathrm{WER}_{\mathrm{WH}}^{\mathrm{SHAS}} = \frac{\mathtt{S}+\mathtt{D}+\mathtt{I}}{\mathtt{S}+\mathtt{D}+\mathtt{C}} = \frac{725+2192+127}{725+2192+7302} = 0.2979$$

$$\mathrm{WER}_{\mathrm{WH}}^{\mathrm{SB\ VAD}} = \frac{\mathtt{S}+\mathtt{D}+\mathtt{I}}{\mathtt{S}+\mathtt{D}+\mathtt{C}} = \frac{801+868+219}{801+868+8550} = 0.1848$$

These differences are clearly due to the erroneous removal of speech portions by SHAS.

The aggregation step described in Section 2.4 leads to a substantial reduction in the number of segments, an increase in their duration, and a larger portion of audio being transcribed (Table 1). As shown by the %WER scores in Table 2, this aggregation does not benefit FW or WH, and may even degrade their performance. In contrast, VX significantly benefits from longer segments, achieving the best transcription accuracy (%WER = 16.86) among those observed in this study. This behavior is consistent with VX's architecture: unlike WH-based models, VX features a 32k token context win-

dow capable of processing up to 30 minutes of audio, which allows it to exploit the broader acoustic and linguistic context made available by segment aggregation. Conversely, this same characteristic explains VX's comparatively weaker performance on the shorter segments produced by SHAS and SB VAD: without sufficient audio context to exploit, its large context window becomes irrelevant, and VX trails WH on both segmenters, even ranking last on SHAS-generated segments (%WER=31.90).

These experiments lead to the following conclusions: (i) SB VAD is preferable to SHAS for audio segmentation; (ii) WH is preferable to FW for baseline transcription and SRT generation; (iii) aggregating SB VAD segments and transcribing them with VX leads to improved transcription quality.

3.1.2 MT

As already mentioned, the translation model we selected is `MADLAD-400-10B-MT`, which provides strong performance and is freely usable. In addition to the 10B model, the family also includes smaller variants, notably 3B and 7B. For all of them, generation can be performed using different decoding strategies: Greedy Decoding, which ensures determinism (i.e., identical outputs for the same input); Multinomial Sampling, which instead introduces randomness; and Beam Search, which allows enabling a `length_penalty` (`lp`) option to exponentially penalize or encourage longer sequences, provided that the `beam_size` is greater than 1.

Table 3 reports subtitle evaluation scores, computed with all official metrics used by the shared task organizers, obtained with different `MADLAD-400-MT` configurations. All translations are generated starting from the same automatic transcript of the ITV dev2026 set, namely the one produced by WH on SB VAD segmentation, which achieves a %WER of 18.48 (Table 2). Table 3 rows are grouped into blocks of five; each block reports results for a given language pair and MT model size, varying the decoding strategy. The results for the intermediate model 7B have been omitted, as they provide limited additional information, given that they fall between those of the other two models, which are already quite close to each other. We also exclude the stochastic decoding strategy, as it is not suitable for a subtitling scenario.

The first two blocks compare the performance of the 3B and 10B models on the same language pair (en-de). As expected, the larger model consistently outperforms the smaller one, although the margin is relatively small across all decoding strategies. Regarding decoding methods, greedy decoding is almost always inferior to beam search in all blocks. Decreasing the value of `lp`, which favors shorter sequences, leads to improvements in SubER and CPS, initially substantial, then progressively smaller, but does not necessarily improve translation quality, which tends instead to peak around `lp` = 0.

Based on these results, we eventually selected the 10B model with beam search decoding and `lp` = 0, slightly prioritizing translation quality over subtitle compliance.

3.1.3 Subtitling

In this section, we evaluate the impact of the proposed two-stage architecture on subtitling quality. Table 4 reports automatic evaluation results on the Asharq-Bloomberg, ITV, and YODAS dev2026 sets.

The subtitling systems we evaluate are:

`Baseline Subtitling`: It features the Stage 1 of our two-stage framework, described in Section 2.3; audio segmentation is performed by SB VAD, ASR by WH, MT by MADLAD-400-10B-MT with 4-beams decoding and `lp`=0.0. It represents our **contrastive-1** submission;

`Sentence-aware Subtitling`: It consists of the full two-stage framework, described in Section 2.2; starting from the Stage 1 output, ASR is performed by VX, MT again by MADLAD-400-10B-MT with 4-beams decoding and `lp`=0.0. It represents our **primary** submission;

`Light Subtitling`: the architecture is that of Stage 1, but segmentation is carried out by SHAS, ASR by FW, MT by MADLAD-400-3B-MT with Greedy decoding. It is used for our **contrastive-2** submission.

All in all, **the primary system consistently improves over the contrastive systems across most language pairs and datasets**. These improvements are particularly evident in translation quality metrics (BLEU, ChrF, and BLEURT), confirming the effectiveness of the sentence-aware refinement stage. The gains can be attributed to two main factors. First, the use of longer, aggregated audio segments leads to more accurate transcriptions, as discussed in Section 3.1.1. Second, translating well-formed sentence units, rather than subtitle fragments, allows the MT model to better exploit

MT (MADLAD)			Sub. qual.	Translation quality			Sub. compl.
decoding	size	pair	SubER	BLEU	ChrF	BLEURT	CPS
greedy	3b		81.28	16.43	43.70	.4772	77.27
lp=1.0	3b		73.06	18.32	44.73	.4875	77.77
lp=0.0	3b	en-de	70.19	19.13	44.23	.4811	80.29
lp=-1.0	3b		68.71	19.64	44.20	.4819	81.93
lp=-2.0	3b		68.23	19.41	43.91	.4791	82.13
greedy	10b		78.44	16.93	43.31	.4775	77.60
lp=1.0	10b		72.08	18.76	44.73	.4845	78.26
lp=0.0	10b	en-de	69.32	20.12	44.97	.4875	80.43
lp=-1.0	10b		68.03	20.35	44.57	.4835	82.40
lp=-2.0	10b		67.55	20.43	44.20	.4825	83.27
greedy	10b		64.19	20.88	43.40	.4985	84.03
lp=1.0	10b		64.40	20.80	43.55	.5021	83.04
lp=0.0	10b	en-es	62.67	21.02	43.51	.5014	85.35
lp=-1.0	10b		62.13	20.67	43.15	.4990	86.59
lp=-2.0	10b		61.56	20.87	43.11	.4975	86.96
greedy	10b		113.53	5.59	12.73	.2486	43.56
lp=1.0	10b		110.93	5.36	12.76	.2491	43.29
lp=0.0	10b	en-ja	99.29	6.11	12.23	.2476	52.50
lp=-1.0	10b		97.17	6.17	11.97	.2419	54.62
lp=-2.0	10b		96.77	5.94	11.83	.2367	55.65
greedy	10b		71.59	22.23*	19.35	.4499	84.70
lp=1.0	10b		66.61	23.18*	20.20	.4647	95.72
lp=0.0	10b	en-zh	64.53	22.77*	20.13	.4566	96.65
lp=-1.0	10b		63.69	21.94*	19.62	.4454	97.39
lp=-2.0	10b		63.76	21.59*	19.41	.4403	97.83

Table 3: Comparison of MT setups on the ITV dev2026 set. Decoding is either Greedy or Beam Search; for the latter, the number of beams is set to 4, while length_penalty (`lp`) varies in [-2.0, 1.0]. `CPL` and `LPB` are always 100.00, therefore are not shown. (*) The BLEU scores for Chinese are computed at character level.

contextual information, resulting in more fluent and adequate translations.

As expected, considering the results presented in Section 3.1.2, improvements in translation quality sometimes come at the cost of reduced subtitle compliance (e.g., CPS or CPL), due to the use of longer and more syntactically complete translations. However, the post-processing steps described in Section 2.5 help mitigate this effect, ensuring an overall good balance between quality and compliance.

Focusing on contrastive systems, **while contrastive-2 may outperform contrastive-1 (and even the primary run) in some cases** (e.g., Asharq-Bloomberg en-ja)**, it remains consistently inferior to contrastive-1 overall**, highlighting how model selection and system setup also impact performance, in addition to the proposed refinement strategy.

The relatively bad scores observed in general for en-ja require comment. The low translation quality scores may be due to the word-level segmentation of Japanese, which does not use whitespace to delimit word boundaries. Therefore, this segmentation is necessary both for subtitle alignment (via `mweralign`) and for calculating the scores themselves. Inappropriate segmentation could negatively impact both operations, and this may be the reason for the observed problem. Indeed, by bypassing word-level segmentation and evaluating hypothesis against reference each as a single long string at the character level, both BLEU and ChrF for Japanese become comparable to, or even slightly better than, those for Chinese (BLEU: 29.77 vs. 29.05; ChrF: 29.38 vs. 27.11, for Japanese and Chinese respectively, on our ITV dev2026 primary run), confirming that the underlying translation quality is similar between the two languages. The low CPS and CPL values, on the other hand, are a direct consequence of the extremely strict thresholds imposed for Japanese: for example, a modest relaxation of the reading speed threshold, from 4 to 6, increases the CPS of our ITV dev2026 primary run from 57.06 to 78.74; notably, the same relaxation raises the CPS of the reference subtitles from 44.89 to 95.74, confirming that these scores reflect the severity of the threshold

rather than subtitle non-compliance.

Anyway, **results on all domains and language pairs, en-ja included, confirm the effectiveness of the proposed two-stage framework, demonstrating that revisiting segmentation at the sentence level is a key factor for improving subtitling performance.**

Example. A representative example of the qualitative advantage of sentence-level translation over subtitle-level translation is illustrated in Table 5. The reference subtitles render the English sentence

We've got these expectations around Fed cuts very much being solidified.[7]

as a coherent two-block unit:

Die Erwartungen hinsichtlich Zinssenkungen der Fed / haben sich deutlich verfestigt.

The baseline system, however, does not have access to the full sentence at translation time. Instead, it translates two subtitle-level fragments independently: *"We've got these expectations around"* and *"Fed cuts very much being solidified."* The first fragment yields a semantically impoverished subtitle: *"Wir haben diese Erwartungen"* (lit. *"We have these expectations"*), which conveys no information about what those expectations concern. The second fragment, translated in isolation, produces *"Die Kürzungen der Fed verfestigen sich."*, a grammatically well-formed sentence but that loses the logical connection with the preceding subtitle and adopts a stylistically marked present tense (*verfestigen sich*) rather than the more appropriate present perfect found in the reference (*haben sich verfestigt*).

The sentence-aware system, by contrast, translates the full sentence as a single unit, correctly recovering both the subject (*"Die Erwartungen in Bezug auf die Kürzungen der Fed"*) and the predicate (*"haben sich sehr verfestigt"*). The resulting translation is then re-segmented and inserted into the SRT template of the baseline, yielding two subtitles that together form a coherent and complete sentence, closely mirroring the structure of the reference.

3.2 Computational Costs

The development experiments presented above were run on a single Quadro RTX 8000 46GB GPU, recording execution times and memory usage. Table 6 collects the execution time of core operations (excluding overheads such as model loading) and the recorded peaks of GPU memory usage for some of those experiments. The first two lines compare the computational costs of the two audio segmentation algorithms. SHAS pays the price, especially in terms of computational time, for using a pretrained model. However, the time to segment the entire ITV dev2026 set for both methods is at least an order of magnitude less than that of the other operations, as we will see shortly, and therefore does not represent a bottleneck in the overall process.

The three middle rows compare the ASR models used for transcribing audio segments generated by SHAS on the ITV dev2026 set. The time and memory footprint of FW, WH, and VX are increasing, but less than expected. In particular, FW is faster than WH by less than 12%, well below what was reported by the model developers (actually for models other than the large-v3 used here).[8] Even though its model has twice as many parameters as WH and FW's large-v3 model, the transcription process performed with VX is only slightly more demanding: about 25% slower and 12% larger than that of WH.

The last lines compare the costs of the various MADLAD setups when used to translate the ITV dev2026 transcripts obtained by WH on the audio segmentation performed by SHAS. The values are the averages measured for the translation of the entire set in all four target languages (de, es, ja, and zh). First of all, we note that, given the same model, greedy decoding is slightly faster than beam search, especially for the 10B model; on the contrary, the memory usage does not depend perceptibly on decoding type. Instead, looking at the model size, it impacts linearly on the memory consumption, given that the ratio between the number of parameters (about 3.5) is practically identical to that between the memory footprints (about 3.4); on the other hand, its influence on execution times is smaller but still high: using the 10B models leads to 20% longer execution times for greedy decoding, and 30% longer when performing beam search.

Ultimately, **the best performing transcription and translation models are also the most expensive ones, while for audio segmentation, the SB VAD option does not appear to have any drawbacks**, at least in the context of these experiments.

[7] In this case, the ASRs did not make any mistakes, therefore the automatic transcripts coincide with the reference one.

[8] https://github.com/SYSTRAN/faster-whisper

domain	run	pair	Sub. qual.	Translation quality			Subtitle compliance		
			SubER	BLEU	ChrF	BLEURT	CPS	CPL	LPB
Asharq-Bloomberg	primary		72.83	13.85	44.87	.5217	89.53	94.48	90.12
	contrastive-1	en-ar	74.78	12.89	43.04	.4918	89.32	100.00	100.00
	contrastive-2		74.54	12.05	42.20	.4861	91.53	100.00	99.97
	primary		58.77	30.98	57.32	.5972	70.07	94.06	82.65
	contrastive-1	en-de	64.16	26.75	54.32	.5618	68.30	100.00	100.00
	contrastive-2		63.85	26.02	53.55	.5455	64.59	100.00	100.00
	primary		94.61	10.33	15.38	.2411	19.57	75.14	100.00
	contrastive-1	en-ja	117.86	7.10	13.30	.1730	13.57	100.00	100.00
	contrastive-2		78.51	12.62	19.25	.3295	12.57	36.83	100.00
	primary		80.50	27.75*	23.82	.4763	95.14	89.01	100.00
	contrastive-1	en-zh	89.91	26.21*	22.56	.4389	95.43	100.00	100.00
	contrastive-2		71.49	24.48*	21.26	.4661	81.99	63.85	100.00
ITV	primary		72.55	20.09	46.68	.5154	77.81	95.46	98.67
	contrastive-1	en-de	69.32	20.12	44.97	.4875	80.43	100.00	100.00
	contrastive-2		76.71	16.69	37.81	.3894	78.46	100.00	100.00
	primary		64.39	22.94	45.29	.5369	83.21	96.89	99.09
	contrastive-1	en-es	62.67	21.02	43.51	.5014	85.35	100.00	100.00
	contrastive-2		69.18	17.69	37.92	.4052	82.57	100.00	100.00
	primary		92.33	7.14	14.01	.2865	57.06	70.05	100.00
	contrastive-1	en-ja	99.29	6.11	12.23	.2476	52.50	100.00	100.00
	contrastive-2		97.32	4.96	9.02	.1720	21.69	62.34	100.00
	primary		64.63	23.38*	20.35	.4770	96.59	92.96	100.00
	contrastive-1	en-zh	64.53	22.77*	20.13	.4566	96.65	100.00	100.00
	contrastive-2		74.07	18.38*	16.72	.3839	79.27	90.67	100.00
YODAS	primary		58.07	32.33	55.78	.6186	66.16	95.36	97.09
	contrastive-1	en-de	60.14	29.81	53.32	.5877	64.39	100.00	100.00
	contrastive-2		57.74	27.39	50.76	.5527	65.48	100.00	100.00
	primary		80.26	10.66	16.21	.3224	42.22	78.46	100.00
	contrastive-1	en-ja	85.27	9.03	13.77	.2740	32.99	100.00	100.00
	contrastive-2		75.84	10.45	16.06	.3293	23.52	65.84	100.00
	primary		72.79	25.05*	21.67	.5091	87.99	91.65	100.00
	contrastive-1	en-zh	76.70	23.94*	20.55	.4721	88.26	100.00	100.00
	contrastive-2		74.95	20.44*	17.68	.4379	74.82	88.97	100.00

Table 4: Automatic evaluation scores on the Asharq-Bloomberg, ITV, and YODAS dev2026 sets. (*) The BLEU scores for Chinese are computed at character level.

3.3 Official Submissions

For each domain and language pair included in the IWSLT 2026 Subtitling shared task, we submitted the subtitles automatically generated by the three systems evaluated on the dev2026 sets in Section 3.1.3. Table 7 reports a subset of the official results provided by the organizers, focusing on the ITV test2023 set. This dataset, originally introduced in the 2023 edition of the shared task, has been consistently used in subsequent editions, including this one, to measure progress over time. The remaining results obtained by our systems confirm the trends observed on the dev2026 sets and are therefore omitted for brevity; they are available in (Adelani et al., 2026). In addition to our results, the table includes the primary scores achieved this year by AppTek, as well as those obtained in previous editions by different participants. This comparison enables us to quantify overall progress in the subtitling task, track improvements of systems across years, and better contextualize the performance of our current models.

Focusing on our own systems, **the improvements enabled by the proposed two-stage framework are clearly evident when compared to our previous submissions.** In particular, FBK26 shows a substantial gain over FBK24 (single-stage cascade) and FBK23 (direct, trained under constrained conditions), especially in terms of translation quality (BLEU, ChrF, BLEURT) and overall subtitle quality (SubER).

When compared to the AppTek26 runs, our two-stage system proves competitive in both German and Spanish, although it exhibits noticeably lower subtitle compliance.

Compared to the HW-TSC24 systems, our results are consistently better, often by a large margin, with the only exception being SubER on en-es, where our score is slightly worse (67.28 vs. 66.78). **Finally, the commercial MATESUB23 system produced more compliant subtitles, particularly in terms of CPS, but with substantially lower translation quality** (our BLEURT is more than 10 points higher on German and over 8 points higher on Spanish).

Overall, these results trace a clear upward trajec-

	Block	Timing	Text
English	1452	0S:52,039 → 0S:53,519	We've got these expectations around
source	1453	0S:53,759 → 0S:55,640	Fed cuts very much being solidified.
reference	1765	0S:52,219 → 0S:54,474	Die Erwartungen hinsichtlich Zinssenkungen der Fed
SRT	1766	0S:54,507 → 0S:55,960	haben sich deutlich verfestigt.
contrastive-1	1917	0S:52,039 → 0S:53,519	Wir haben diese Erwartungen
	1918	0S:53,759 → 0S:56,010	Die Kürzungen der Fed verfestigen sich.
primary	1704	0S:52,039 → 0S:53,519	Die Erwartungen in Bezug auf
	1705	0S:53,759 → 0S:56,010	die Kürzungen der Fed haben sich sehr verfestigt.

Table 5: Example of subtitle-level vs. sentence-level translation, from the dev2026 Asharq-Bloomberg en-de set. The baseline system translates each subtitle fragment independently, producing a semantically incomplete first block and a second block disconnected from the preceding subtitle. The sentence-aware system translates the full sentence and re-segments the output, yielding a coherent two-block subtitle that closely matches the reference. 0S stands for a hh:mm offset of 02:25.

step	model		time	GPU memory
Audio Seg.	SHAS		48s	7.5GB
	SB VAD		34s	0.6GB
ASR	FW		8m00s	6.9GB
	WH		9m03s	9.6GB
	VX		12m07s	9.7GB
MT	MADLAD	3B greedy	13m12s	12.2GB
		3B beam	14m43s	12.2GB
		10B greedy	16m31s	41.8GB
		10B beam	21m06s	41.8GB

Table 6: Costs in terms of total execution time (excluding the model loading) and GPU memory usage of the various processing steps run for the ITV dev2026 set on a single GPU Quadro RTX 8000 46GB.

tory in FBK's subtitling performance across successive shared task editions, with the 2026 two-stage system closing much of the gap with respect to well-established commercial solutions.

4 Conclusions

We presented FBK's submissions to the automatic subtitling tasks of the IWSLT2026 evaluation campaign. We proposed a two-stage framework that, by reviewing transcription and translation at the *sentence* level, consistently improves, across all domains and language pairs proposed in this edition of the shared task, the subtitling performance of a baseline system operating at *subtitle* level, especially in terms of translation quality. Specifically, the official results on the oldest legacy test set (ITV test2023) show not only the impressive improvements of our two-stage system compared to our previous subtitling systems, but also its competitiveness versus commercial systems, thus demonstrating the feasibility of automatically generating good subtitles with models freely available.

Acknowledgments

The work presented in this paper has received funding from the DVPS project, under the European Union's Horizon Europe Framework, Grant Agreement No. 101213369.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech Translation and Metrics in 2026: Findings of the IWSLT Campaign. In *Proc. of IWSLT*, San Diego, US-CA.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John P. McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Marco Gaido, Sara Papi, Mauro Cettolo, Roldano Cattoni, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2024. Automatic subtitling and subtitle compression: FBK at the IWSLT 2024 subtitling track. In *Proc. of IWSLT*, pages 86–96, Bangkok, Thailand (in-person and online).

ITV team	pair	Sub. qual. SubER	Translation quality BLEU	ChrF	BLEURT	Subtitle compliance CPS	CPL	LPB
AppTek23		69.15	14.42	35.51	.4023	86.01	100.00	100.00
AppTek24		68.70	17.96	41.40	.4720	67.64	100.00	99.96
AppTek25		65.26	18.80	41.83	.5012	93.32	100.00	100.00
AppTek26		**67.91**	**20.09**	**46.26**	**.5520**	**81.72**	**100.00**	**93.67**
Fbk23	en-de	81.45	8.03	26.42	.2283	67.75	85.12	100.00
Fbk24		74.25	16.16	36.10	.3928	54.70	92.97	100.00
Fbk26		**69.50**	**20.25**	**44.81**	**.5492**	**67.84**	**95.65**	**95.92**
Hw-Tsc24		70.97	18.33	42.97	.5057	60.15	62.37	100.00
Matesub23		71.35	14.90	37.26	.4438	80.21	99.47	100.00
AppTek23		80.33	11.23	29.87	.2478	94.67	100.00	100.00
AppTek24		66.55	22.05	45.49	.4782	77.61	100.00	100.00
AppTek26		**64.67**	**23.65**	**49.70**	**.5514**	**88.12**	**100.00**	**97.05**
Fbk23	en-es	81.41	9.23	27.44	.2083	74.67	92.94	100.00
Fbk24		70.35	19.15	40.08	.3959	62.11	94.22	100.00
Fbk26		**67.28**	**23.16**	**47.69**	**.5356**	**73.84**	**96.22**	**97.20**
Hw-Tsc24		66.78	22.44	46.67	.5098	68.95	67.58	100.00
Matesub23		69.34	18.52	41.41	.4530	81.93	99.51	100.00

Table 7: Official automatic evaluation scores on the ITV test2023 set of AppTek and Fbk 2026 primary runs (in **bold**) and of primary runs submitted by participants to past editions of the shared task.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):1–38.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In *Proc. of NeurIPS*, New Orleans, US-LA.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proc. of IWSLT*, Pittsburgh, US-PA.

Matt Post and Hieu Hoang. 2025. Effects of Automatic Alignment on Speech Translation Metrics. In *Proc. of IWSLT*, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proc. of ICML*, volume 202, Honolulu, US-HI.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

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. In *Proc. Interspeech 2022*, pages 106–110.