Supplementary Material of "OptiSub: Optimizing Video Subtitle Presentation for Varied Display and Font Sizes via Speech Pause-Driven Chunking"

A DETAILS OF WORD BOUNDARY DETECTION

We obtain the duration of each spoken word by performing an audio-text alignment process with Massive Multilingual Speech (MMS) [3] model and Dynamic Time Warping (DTW) [2]. Specifically, we used the MMS-1B-all checkpoint for MMS model. The process begins with segmenting the audio based on the start and end timings provided in the subtitle file. To account for potential inaccuracies (that might be caused by human or the time alignment from auto-transcription systems) in the start and end timings, we add temporal padding (1 second) before and after each segment. Because the MMS model supports multiple languages, selecting the target language (by exchanging the language adapter layer weights) in the model enables applications to different languages. From the segmented audio, logits are extracted using the MMS model's Connectionist Temporal Classification (CTC) head. These logits are then converted into log probabilities using a softmax function. Next, using the text corpus from the subtitle file, we obtain tokens using the MMS model's text tokenizer. To align the audio and text tokens, we perform DTW by constructing a trellis matrix that computes alignment costs between the time frames and tokens, followed by backtracking to find the optimal alignment path. On top of this timing information of characters, we find where each word starts and ends in seconds.

B DETAILS OF SYNTAX REFLECTION ANALYSIS

In Section 5.2 of the main paper, we evaluated syntax reflection by analyzing the hierarchy steps in the syntax tree [1] for *chunk-to-chunk* and *word-to-word* word pairs. A hierarchy step refers to the number of steps required to traverse the syntax tree to find a common ancestor between two neighboring words. For example, consider a subtitle case represented as `{[A, B], [C, D, E, F], [G, H, I]}`, where each letter denotes a word, and square brackets indicate chunks. In this case, there are three chunks. Chunk-to-chunk refers to the pairs of words at the boundaries of adjacent chunks, such as `(B, C)` and `(F, G)`. Word-to-word refers to pairs of adjacent words within the same chunk, such as `(A, B)`, `(C, D)`, `(D, E)`, `(E, F)`, `(G, H)`, and `(H, I)`. For each pair, we calculated the hierarchy step in the syntax tree and averaged these values.

REFERENCES

[1] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (01 2014). https://doi.org/10.3115/v1/P14-5010

[2] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[3] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling Speech Technology to 1,000+ Languages. arXiv:2305.13516 [cs.CL] https://arxiv.org/abs/2305.13516