Length-Aware NMT and Adaptive Duration for Automatic Dubbing

Zhiqiang Rao, Hengchao Shang, Jinlong Yang, Daimeng Wei, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhengzhe Yu, Zhanglin Wu, Yuhao Xie, Bin Wei, Jiawei Zheng, Lizhi Lei and Hao Yang
Huawei Translation Service Center, Beijing, China
{raozhiqiang,shanghengchao,yangjinlong7,weidaimeng,lizongyao,
guojiaxin1,lishaojun18,yuzhengzhe,wuzhanglin2,xieyuhao2,weibin29,
zhengjiawei15,leilizhi,yanghao30}@huawei.com

Abstract

This paper presents the submission of Huawei Translation Services Center for the IWSLT 2023 dubbing task in the unconstrained setting. The proposed solution consists of a Transformer-based machine translation model and a phoneme duration predictor. The Transformer is deep and multiple target-to-source length-ratio class labels are used to control target lengths. The variation predictor in FastSpeech2 is utilized to predict phoneme durations. To optimize the isochrony in dubbing, re-ranking and scaling are performed. The source audio duration is used as a reference to re-rank the translations of different length-ratio labels, and the one with minimum time deviation is preferred. Additionally, the phoneme duration outputs are scaled within a defined threshold to narrow the duration gap with the source audio.

1 Introduction

Automatic dubbing (AD) (Federico et al., 2020; Brannon et al., 2022; Chronopoulou et al., 2023) technology uses artificial intelligence (AI) to automatically generate dubbed audio for video content. Dubbing is the process of replacing the audio with a translation of the original audio in a different language. AI dubbing technology automates this process by using machine learning algorithms to translate the original audio and synthesize a new voice that sounds natural and resembles a human voice. The synthesized voice is then synchronized with the lip movements of the characters in the video to produce dubbed audio. This technology has the potential to significantly reduce the time and cost of creating dubbed audio and make it easier to reach a global audience by translating video content into multiple languages.

Recent advances in the field of automatic dubbing have contributed to the development of more efficient and cost-effective methods for producing localized content. Researchers have utilized various techniques and technologies, including machine translation (MT) (Lopez, 2008; Vaswani et al., 2017), speech synthesis (Wang et al., 2017b; Ren et al., 2022), and speech recognition (Gulati et al., 2020; Schneider et al., 2019), to improve the accuracy and quality of automatic dubbing systems.

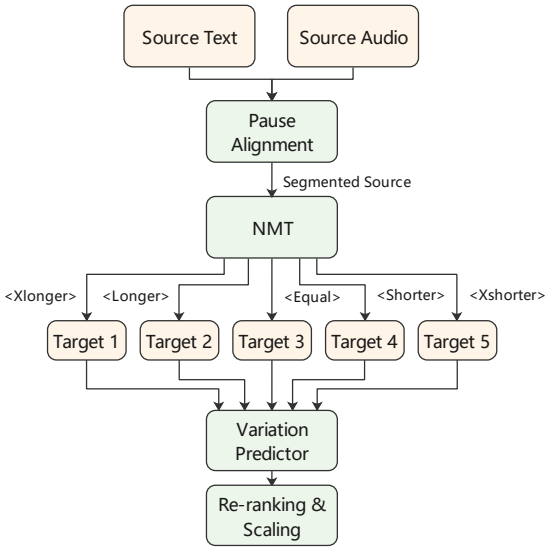

Figure 1: System pipeline.

Isometric machine translation (Lakew et al., 2022; Li et al., 2022) is a technique used in automatic dubbing where translations should match a given length to allow for synchronicity between source and target speech. For neural MT, generating translations of length close to the source length, while preserving quality is a challenging task. Controlling MT output length comes at a cost to translation quality, which is usually mitigated with a two-step approach of generating N-best hypotheses and then re-ranking based on length and quality.

Another area of research focuses on the synchronization of the dubbed audio with the original source audio. This is essential for ensuring that the dubbed audio matches the timing and intonation of the original speech. Researchers have developed various methods for achieving accurate synchro-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 138–143

nization, including the use of phoneme duration predictors and machine learning algorithms to detect and align speech segments (Virkar et al., 2021; Effendi et al., 2022; Virkar et al., 2022).

One of the latest developments in automatic dubbing research is the use of deep neural networks for speech synthesis (Chronopoulou et al., 2023; Ren et al., 2022). These networks enable the creation of more naturalistic and expressive speech, improving the overall quality of the dubbed audio. In conclusion, recent research in automatic dubbing has shown significant progress and promise for the future of localized content production. By combining advanced machine learning techniques with speech synthesis, speech recognition, and sentiment analysis, researchers are developing more accurate, efficient, and cost-effective automatic dubbing systems.

The IWSLT 2023 (Agarwal et al., 2023) dubbing task focuses on isochrony in dubbing, which refers to the property that the speech translation is time aligned with the original speaker's video. The task assumes that the front Automatic Speech Recognition (ASR) output text and subsequent Text-to-Speech (TTS) models already exist, and the goal is to predict the phonemes and their durations. Our proposed solution involves using a Transformer-based (Vaswani et al., 2017) machine translation model and a phoneme duration predictor. A Deep Transformer (Wang et al., 2017a, 2019) model is utilized to handle multiple target-to-source length-ratio class labels, which are used to control target lengths. The phoneme duration predictor is based on the variation predictor used in FastSpeech2 (Ren et al., 2022). To optimize isochrony in dubbing, the solution utilizes re-ranking and scaling techniques. The translations generated by different length-ratio labels are re-ranked based on their time deviation from the source audio duration, with the minimum deviation one preferred. The phoneme duration outputs are also scaled within a predefined threshold to narrow the duration gap with the source audio. These techniques help to ensure that the translated speech is synchronized with the original speaker's video.

2 Data

The data provided in the constrained setting is derived from CoVoST2 (Wang et al., 2020) De-En data, consisting of German source text, English target text, speech durations, and English phonemes and durations (Brannon et al., 2022). We additionally apply WMT2014 De-En data for training the MT model. The amount of data for both sets is shown in Table 1.

Data	Size
CoVoST2	0.289M
WMT2014	4.5M

Table 1: The bilingual data sizes.

To achieve better training results of the MT model, we used some data pre-processing methods to clean the bilingual data, including removing duplicate sentences, using Moses (Koehn et al., 2007) to normalize punctuation, filtering out overly long sentences, using langid (Lui and Baldwin, 2011, 2012) to filter out sentences that do not match the desired language, and using fast-align (Dyer et al., 2013) to filter out unaligned sentence pairs.

3 System

The system consists of four parts: Pause Alignment, Machine Translation, Phoneme Duration Variation Predictor, and Re-ranking and Scaling. Figure 1 shows the system pipeline. The following describes the four parts in detail.

3.1 Pause Alignment

During inference, we use a Voice Activity Detector (VAD) (Team, 2021) to obtain speech segments and their durations from the source audio. The test data for the task already provides text segments separated by pauses. However, we found that the number of speech segments obtained by VAD sometimes does not match the number of text segments provided, resulting in incorrect matching of pause counts. This can cause significant discrepancy between the synthetic dubbing and the lip movements of the character in the video when the pause duration is long.

To address this issue, we first perform pause alignment between the source text and the source audio. We use the proportion of tokens in each text segment to the total number of tokens, and the proportion of duration of each speech segment to the total duration, to find the best alignment between the text and speech segments. When the number of text segments is less than the number of speech segments, we merge the audio segments to reduce the number of speech segments. The final

speech segments that need to be retained are split at the following points:

$$i^{'} = \arg\min_{j} \left| \frac{|s_{1..j}|}{S} - \frac{|t_{1..i}|}{T} \right|; j \geq i$$

Where $|t_{1..i}|$ means total number of tokens from the first to the i-th text segment. $|s_{1..j}|$ means total duration from the first to the j-th speech segment. T and S represent the total number of tokens in the text and the total duration of the speech, respectively. $i^{'}$ is the i-th speech segmentation point after merging, corresponding to the i-th text segment.

Conversely, when the number of speech segments is less than the number of text segments, we merge the text segments. The final retained text segmentation points are:

$$j^{'} = \arg\min_{i} \left| \frac{|t_{1..i}|}{T} - \frac{|s_{1..j}|}{S} \right|; i \geq j$$

3.2 Machine Translation

We trained a Neural Machine Translation (NMT) model using Deep Transformer, which features pre-layer normalization, 25 encoder layers, and 6 decoder layers. Other structural parameters are consistent with the Transformer-Base model.

Following existing length control methods, we divided the bilingual data into 5 categories based on the target-to-source character length ratio (LR) for each sample (Lakew et al., 2022; Li et al., 2022). The labels were defined based on LR thresholds: $Xshorter < 0.8 < Shorter < 0.9 < Equal < 1.1 < Longer < 1.2 < Xlonger$. During training, we added a length tag <Xshorter/Shorter/Equal/Longer/Xlonger> at the beginning of each source sentence. In the inference process, text segments are sent to the translation model separately and the required tag is prepended at the beginning of each input segment.

3.3 Phoneme Duration Variation Predictor

As with FastSpeech2 (Ren et al., 2022), after using an open-source grapheme-to-phoneme tool (Park, 2019) to convert the NMT output translation sequence into a phoneme sequence, the pre-trained variation predictor module in FastSpeech2 was used to generate initial phoneme durations. The variation predictor takes the hidden sequence as input and predicts the variance of the mean squared error (MSE) loss for each phoneme's duration. It consists of a 2-layer 1D-convolutional network with ReLU activation, followed by layer-normalization and dropout layers, and an additional linear layer to project the hidden state into the output sequence. The final output is the length of each phoneme.

3.4 Re-ranking and Scaling

To select the best isochrony dubbing, we used source texts with 5 different tags prepended as inputs for the NMT model. After converting the output translations into phoneme durations using the phoneme duration variation predictor, we re-ranked them based on the source audio duration as reference, and selected the output with the least duration deviation.

Additionally, we used the ratio of the source audio duration to the total predicted phoneme duration as a reference, and scaled the predicted phoneme duration within a certain threshold to further optimize the synchronization between the synthesized dubbing and the source video.

$$s_j^{'} = \arg\min_{s_{jk}^{'}} (\left| s_{jk}^{'} \right| - |s_j|); k \in [1, 5]$$

$$s_j^{'} = s_j^{'} \cdot Scale(\frac{|s_j|}{\left| s_j^{'} \right|})$$

$$Scale(r) = \begin{cases} 1.1, & r > 1.1 \\ r, & 0.9 < r < 1.1 \\ 0.9, & r < 0.9 \end{cases}$$

Where $|s_j|$ is the total duration of source speech segment s_j, $\left| s_j^{'} \right|$ is the total duration of generated dubbing segment $s_j^{'}$. And $Scale()$ is a scaling function.

4 Experiments

We used SentencePiece (Kudo and Richardson, 2018) to process NMT bilingual text and obtain subword vocabularies, resulting in a German vocabulary of 29k and an English vocabulary of 25k. We trained a Transformer NMT model using fairseq (Ott et al., 2019), with an encoder of 25 layers, a decoder of 6 layers, 8 attention heads, embeddings of 512, and FFN embeddings of 2048. The model was optimized using Adam (Kingma and Ba, 2017) with an initial learning rate of 5e-4, and warmup steps of 4000. Dropout was set to 0.1. The model

was trained on 8 GPUs, with a batch size of 2048 tokens and an update frequency of 4.

During the inference phase, an open-source VAD tool was used to process the source speech and obtain speech segments and durations for subsequent selection of NMT translated text lengths and adjusting the duration of synthetic dubbings. The NMT translated text was then converted to phoneme sequences using an open-source grapheme-to-phoneme tool, and the initial phoneme durations were predicted using a pre-trained variation predictor module in FastSpeech2.

As the main evaluation method for this task is manual evaluation, and our method allows for adjustment of phoneme duration prediction, We mainly experiment and compare BLEU (Papineni et al., 2002) under different strategies of machine translation. To measure the synchronicity between source and dubbed speech, we use speech overlap (SO) (Chronopoulou et al., 2023) metric. It should be noted that the metrics presented don't take into account speech naturalness, which is extremely important to people viewing dubs. (Brannon et al., 2022) showed that human dubbers produces natural speech even at the cost of isochrony. The experimental results on the two test sets of the task are shown in Table 2.

Strategy	subset1		subset2	
	BLEU	SO	BLEU	SO
Xlonger	24.8	0.71	22.0	0.49
Longer	28.0	0.82	26.1	0.70
Equal	37.4	0.83	32.4	0.83
Shorter	42.7	0.79	37.4	0.85
Xshorter	45.7	0.73	43.3	0.83
Re-ranking	31.2	0.92	33.8	0.93
Scaling	31.2	0.97	33.8	0.98
- w/o PA	31.6	0.89	34.7	0.87

Table 2: Experimental results of NMT.

We present the BLEU and SO results using five different LR tags, re-ranking and scaling strategies. The results of the two test sets have the same trend in BLEU, that is, the shorter the generated translation, the higher the BLEU value. Since subset2 has pause punctuation, it is more difficult to translate, so under the same LR tag at all levels, the BLEU value of subset2 will be lower than that of subset1. In terms of SO, both too long or too short translations will cause SO to decrease. The results of medium LR settings can achieve the highest SO value.

Too long translations will result in lower quality of machine translation, while short translations will result in insufficient duration for generating dubbing. After re-ranking, the translations can achieve more moderate results in translation quality and duration. Moreover, by setting appropriate scaling thresholds, scaling operation can further improve the isochrony without affecting BLEU.

We also compared the results without pause alignment, as shown in the last row of Table 2. The SO of both test sets decreased significantly, but the BLEU increased slightly. After analysis, the MT translation is more likely to mismatch with the shorter segment duration, so the shorter translation is selected during re-ranking. While our results show that the shorter the translation, the higher the BLEU.

5 Conclusion

This paper describes the submission of Huawei Translation Services Center for the IWSLT 2023 dubbing task under the unconstrained setting. Our solution consists of four parts: pause alignment, machine translation, phoneme duration variation predictor, re-ranking and scaling. Pause alignment is used to align source audio and source text to improve synchronization between synthetic dubbing and source video. The machine translation model is trained using the Deep Transformer structure. To control the output translation length, multiple target-to-source length-ratio tags are used to adjust the length. Pre-trained variation predictor in FastSpeech2 is used to predict phoneme durations. In order to optimize the isochrony in dubbing, the results of different lengths of the machine translation output are re-ranked and scaled. Using the source audio duration as a reference, the translations with different length ratios are re-ranked, and the output with the smallest time deviation is preferred. In addition, the phoneme duration output is scaled within a defined threshold, further narrowing the duration gap from the source audio. We compare the experimental results of different length-ratio strategies, and our method can achieve a balanced result in BLEU and speech overlap.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda

Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, John Javorský, Dávid and Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail 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 Ojha Kr., 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)*. Association for Computational Linguistics.

William Brannon, Yogesh Virkar, and Brian Thompson. 2022. Dubbing in practice: A large scale study of human localization with insights for automatic dubbing.

Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel M. Lakew, and Marcello Federico. 2023. Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico. 2022. Duration modeling of neural tts for automatic dubbing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8037–8041.

Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020. From speech-to-speech translation to automatic dubbing. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 257–264, Online. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

Zongyao Li, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Minghan Wang, Ting Zhu, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC's participation in the IWSLT 2022 isometric spoken language translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 361–368, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3).

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jongseok Park, Kyubyong Kim. 2019. g2pe. https://github.com/Kyubyong/g2p.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

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. Curran Associates, Inc.

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

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. 2022. Prosodic alignment for off-screen automatic dubbing.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. 2017a. Deep neural machine translation with linear associative unit. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Vancouver, Canada. Association for Computational Linguistics.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017b. Tacotron: Towards end-to-end speech synthesis.