

Leveraging Synthetic Audio Data for End-to-End Low-Resource Speech Translation

Yasmin Moslem

Bering Lab

Abstract

This paper describes our system submission to the International Conference on Spoken Language Translation (IWSLT 2024) for Irish-to-English speech translation. We built end-to-end systems based on Whisper, and employed a number of data augmentation techniques, such as speech back-translation and noise augmentation. We investigate the effect of using synthetic audio data and discuss several methods for enriching signal diversity.

1 Introduction

Resource scarcity and the scattered nature of the data are crucial challenges for low-resource languages (Lankford et al., 2021; Haddow et al., 2022; Lovenia et al., 2024). In this sense, Irish is considered a low-resource language and significantly lacking in speech and language tools and resources (Barry et al., 2022; Lynn, 2022). Researchers have been employing various data augmentation techniques to improve the quality of low-resource textual machine translation (MT) systems. Among these techniques is using synthetic data generated by back-translation (Sennrich et al., 2016; Edunov et al., 2018; Dowling et al., 2019; Poncelas et al., 2019; Haque et al., 2020), or large language models (Moslem et al., 2022). Similarly, in the area of speech, Lee et al. (2023) showed that models trained solely on synthetic audio datasets can generalize their performance to human voice data. Nevertheless, Guo et al. (2023) revealed a consistent decrease in the diversity of the outputs of language models trained on synthetic textual data. We observe that leveraging synthetic audio data generated by text-to-speech (TTS) models can be beneficial for training speech translation models, especially for low-resource languages. However, it can lack the diversity found in authentic audio signals in terms of pitch, speed, and background noise.

Speech translation systems can be cascaded systems or end-to-end systems (Agarwal et al., 2023).

Cascaded systems use two models, one for automatic speech recognition (ASR) and one for textual machine translation (MT). End-to-end speech translation systems use one model for the whole process; hence, it is more challenging. In this work, we present end-to-end speech translation models.

In addition to describing our system submitted to IWSLT 2024, this work presents the following contributions:

- Showcasing “speech back-translation” as an effective data augmentation technique for speech translation. In other words, just as back-translation can improve the output quality of text-to-text MT, generating source-side synthetic audio data can considerably enhance the performance of speech translation systems, especially for low-resource languages.
- Introducing a collection of datasets for Irish-to-English speech translation, three of which comprise 196 hours of synthetic audio.
- Exploring diverse training settings and data processing techniques such as noise augmentation and voice audio detection (VAD).
- Releasing versions of Whisper models, specifically fine-tuned for Irish-to-English speech translation.

2 Authentic Data

The organizers of the IWSLT shared task, provided the IWSLT-2023 dataset, which consists of training, dev, and test portions. We used both the training and dev portions for training, and the test portion for evaluation. We also used the Irish portion of the FLEURS datasets. Moreover, we employed the bilingual audio-text data available at the Bitesize website for teaching Irish.¹

¹<https://huggingface.co/datasets/yoslem/BitesizeIrish-GA-EN>

Dataset	Audio	Translation	Train Hours (H:M)	Train Segments	Test Segments
👤 IWSLT-2023	Authentic	Authentic	8:25	8,598	347
👤 FLEURS	Authentic	Authentic	16:45	3,991	0
👤 Bitesize	Authentic	Authentic	5:15	6,149	0
👤 SpokenWords	Authentic	MTed	3:02	10,925	0
🌐 EUbookshop	Synthetic	Authentic	159:45	67,268	0
🗣️ Tatoeba	Synthetic	Authentic	2:39	3,966	0
W Wikimedia	Synthetic	Authentic	34:23	15,090	0
Authentic (👤)			33:27	29,663	347
Synthetic (🌐 🗣️ W)			196:47	86,324	0
Authentic (👤) + Synthetic (🗣️ W)			70:29	48,719	347
Authentic (👤) + Synthetic (🌐 🗣️ W)			229:14	115,987	347

Table 1: Data Statistics: “Audio” and “Translation” columns refer to whether the data is human-generated or machine-generated. “Train Hours” and “Train Segments” refer to the size of the training data in terms of duration and number of utterances, respectively. Finally, “Test Segments” refer to the number of utterances in the test dataset.

3 Synthetic Data

This section explains diverse approaches for creating synthetic data for speech translation. We describe each approach, as well as its advantages and disadvantages.

3.1 Machine Translation

When both audio and transcription are available, but there is no translation, forward MT can be useful as a data augmentation technique. However, there is the risk of feeding incorrect target translations into the training process. Forward MT is more sensitive to the quality of the system used to produce the synthetic data. Compared to back-translation, biases and errors in synthetic data are intuitively more problematic in forward-translation, since they directly affect the gold labels (Bogoychev and Sennrich, 2019). Hence, the used MT system must be of high quality.

We automatically translated the Irish portion of the Spoken Words dataset into English using the Google Translation API. For quality considerations, we decided to use this dataset for training only, but not for evaluation. The dataset consists of 10,925 utterances. Some words are spoken by multiple narrators.²

3.2 Synthetic Audio Data

OPUS (Tiedemann, 2012) hosts several bilingual textual datasets. We extracted portions of the

²<https://huggingface.co/datasets/yomoslem/SpokenWords-GA-EN-MTed>

Tatoeba, Wikimedia, and EUbookshop datasets, comprising 1,983, 7,545 and 33,634 segments, respectively. We extensively filter the datasets based on the following criteria: removing duplicates, removing segments longer than 30 words,³ language detection with fastText (Joulin et al., 2017) (both sides), and Seamless toxicity filtering (Barrault et al., 2023). Finally, we used Azure Speech service to generate two sets of audio data, one with a female voice (OrlaNeural) and the other with a male voice (ColmNeural). As an outcome of this process, we introduce three new datasets, Tatoeba-Speech-Irish,⁴ Wikimedia-Speech-Irish,⁵ and EUbookshop-Speech-Irish,⁶ which together comprise 196 hours of synthetic audio. Table 1 illustrates the statistics of our datasets.

3.3 Audio Signal Processing Augmentation

Synthetic audio data generated by TTS models can have different characteristics than authentic audio. In addition to quality considerations, we observe that among the features that distinguish data generated by TTS systems from authentic data are: 1) lack of noise, and 2) silence differences.

Lack of noise: TTS systems try to mimic studio settings, and produce very clean audio signals.

³<https://github.com/yomoslem/MT-Preparation>

⁴<https://huggingface.co/datasets/yomoslem/Tatoeba-Speech-Irish>

⁵<https://huggingface.co/datasets/yomoslem/Wikimedia-Speech-Irish>

⁶<https://huggingface.co/datasets/yomoslem/EUbookshop-Speech-Irish>

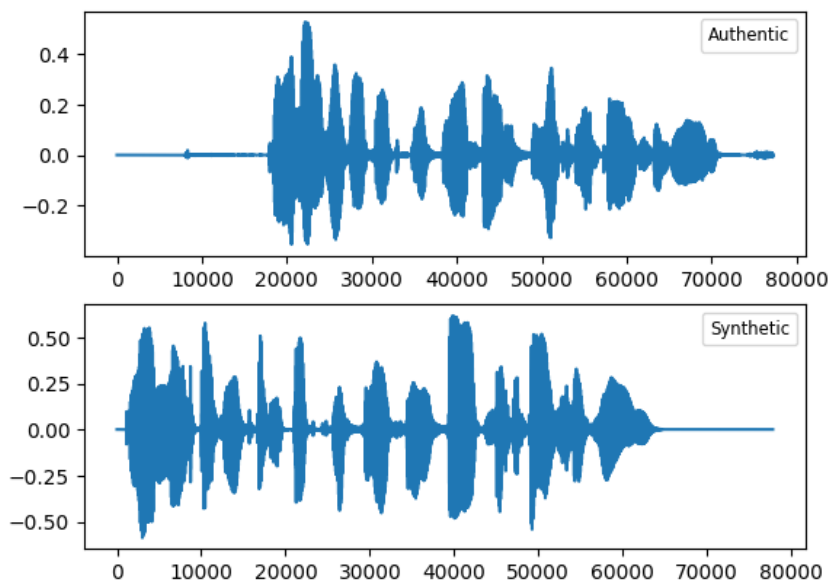


Figure 1: Comparing authentic (top) and synthetic (bottom) audio signals

However, authentic audio signals can include all sorts of environmental noise, ranging from white noise to background voices of people and cars. Even in studio settings, some breath signals can occur unless the audio is extensively edited.

Silence variances: All the synthetic audio signals we generated start at a similar point, with almost no silence at the beginning of the audio (probably to facilitate mixing tracks). However, authentic audio signals can start at any point depending on the recording and processing settings, or whether a signal is truncated from a longer one.

Figure 1 illustrates an example sentence from the Common Voice dataset uttered by a human female (non-studio settings) and its synthetic equivalent generated by Azure TTS system.⁷ The Irish sentence represented here is “*Go raibh maith agaibh as ucht na fíorchaoín fáilte a d’fhear sibh romham.*” It can be translated into English as “*Thank you all for that very generous welcome.*” The authentic signal has more noise (both white background noise and sounds of starting/stopping the recording software), while the synthetic signal does not show any noise occurrence. Moreover, unlike the authentic signal, the synthetic data starts almost immediately. Another observation is that this specific authentic signal has a lower volume than synthetic signals.

⁷Voice name: “Microsoft Server Speech Text to Speech Voice (ga-IE, OrlaNeural)”

3.3.1 Voice Activity Detection

One of the most common audio preprocessing techniques is Voice Activity Detection (VAD). The main idea of VAD is to remove low-amplitude samples from an audio signal. Low-amplitude samples might represent science or noise samples of audio signals, which usually occur at the beginning and end of an audio signal, but can also happen in the midst of longer audio signals. In its basic form, this can be achieved by removing any sample below an absolute value of a threshold (e.g. ± 0.001). However, advanced models like *Silero VAD*⁸ can be used as part of the *torchaudio* framework, and include more sophisticated options (e.g. minimum silence duration) to avoid removing important low-amplitude samples like breath and natural silent durations.

During training, data processed with VAD can either substitute the original data or augment it, i.e. both processed and unprocessed data can be used during training. In one of our experiments (cf. Section 4), we used basic VAD with a threshold of ± 0.001 as a data augmentation technique. When basic VAD is used (i.e. without taking a minimum silence duration into account), this can also speed up the audio signal; in other words, the utterance is spoken faster. At inference time of all the models, we used Silero VAD within Faster-Whisper based on CTranslate2 (Klein et al., 2020).

⁸<https://github.com/snakers4/silero-vad>

3.3.2 Noise Augmentation

Mimicking the effect of white noise can take diverse forms, ranging from using real noise to generating random arrays. To simulate light white noise, we generated a random array with a distribution scale 0.002 and added it to all the audio signals in the dataset.

4 Experiments

Our experiments fine-tune Whisper (Radford et al., 2022) for the task of Irish-to-English speech translation. We experiment with a number of data augmentation techniques, such as speech back-translation (source-side synthetic audio data generation), and audio data augmentation with noise and VAD.

4.1 Speech Back-Translation

By the term “speech back-translation”, we refer to generating source-side synthetic audio for data augmentation of speech translation systems, in the same manner that back-translation is employed in text-to-text MT systems. Section 3.2 explains how we created these synthetic audio datasets. In this set of experiments, we built 3 systems by fine-tuning Whisper Medium. We use different types of datasets as outlined by Table 1.

- **Model A:** It uses the authentic data only, namely IWSLT-2023 dataset, FLEURS, Bite-size, and SpokenWords.
- **Model B:** It uses the same authentic data used in Model A as well as two synthetic audio datasets, namely Tatoeba-Speech-Irish, and Wikimedia-Speech-Irish.
- **Model B++:** In addition to the authentic and two synthetic datasets used in the aforementioned models, Model B++ uses a third synthetic dataset, namely EUbookshop-Speech-Irish.

4.2 Noise and VAD Augmentation

- **Model C:** It uses the same data as Model B, as well as two versions of the data augmented with basic VAD, and white noise. In other words, we fine-tuned Whisper-Medium on all the authentic data and two synthetic data as well as two augmented datasets, one with low-amplitude sample removal, and one with noise augmentation, as described in Section 3.

4.3 Training Arguments

We tried different learning rates and warm-up values. Specifically, we experimented with warm-up ratios 0%, 1%, and 3% out of 3000 steps, which corresponds to 0, 30, 90 warm-up steps, respectively. As Table 5 and Table 4 demonstrate, when fine-tuning Whisper Small, changing the warm-up ratio does not seem to lead to a consistent improvement for the first two sizes of data used in Model A and Model B. However, increasing the warm-up ratio to 3% when the size of data is larger as in Model C, seems to slightly improve the performance. For the learning rate, we used 1e-4 across all the experiments for the sake of consistency. The batch size was decided based on the compute capacity of one A100-SXM4-80GB GPU. Hence, we used a batch size of 64 examples when fine-tuning Whisper Small and a batch size of 16 examples when fine-tuning Whisper Medium. The max length of generation was set to 225. As this is an Irish-to-English translation task, both the tokenizer language and model generation language were set to English. We train the main models with Whisper Medium for at least two epochs, and save the best performing checkpoint based on the chrF++ score on the validation dataset. Section 5 elaborates on the results of these experiments.

4.4 Training Epochs

As we reported in the previous section, we used 3000 steps for all the experiments with Whisper Small, as further training did not seem to improve the output quality when more than one epoch of data is already reached. However, Whisper Medium was trained with a smaller batch size due to computing constrains. We wanted to see the effect of training for at least two epochs. Hence, we report different step milestones in Table 6. In deep learning training in general, it is a common practice to use early stopping. However, for low-resource languages, a smaller value for early stopping can result in the model not seeing the whole data, which can affect the robustness of the model. This is especially true if we are not sure if the validation dataset is well-representative of the task that the model will be actually required to tackle in the real world. While there is no one rule that applies to all cases, we recommend taking this point into consideration when training generic models for low-resource languages.

Whisper	Model	Datasets	Data Size	BLEU \uparrow	chrF++ \uparrow	WER \downarrow	Semantic 1 \uparrow	Semantic 2 \uparrow
Medium	A	authentic	29,663	32.38	48.95	58.85	62.09	63.28
	B	A + synthetic (2d)	48,719	<u>36.34</u>	<u>54.08</u>	<u>53.35</u>	<u>68.31</u>	<u>69.93</u>
	B++	A + synthetic (3d)	115,987	38.41	57.18	51.10	69.72	71.13
	C	B + augmented	146,157	34.09	51.40	55.83	64.26	65.56

Table 2: Evaluation Results: Model B++ that uses both authentic data and 3 synthetic audio datasets achieved the best results across all the systems. The results show that augmenting the training data with synthetic audio (i.e. Model B and Model B++) outperforms using authentic data only (Model A), while further signal processing augmentation with white noise and VAD (Model C) did not help. Moreover, increasing the amount of high-quality synthetic audio data in Model B++ resulted in better quality than Model B that uses a less amount of synthetic data.

5 Evaluation and Results

To evaluate our systems, we calculated BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and TER (Snover et al., 2006), as implemented in the sacreBLEU library⁹ (Post, 2018). For semantic evaluation, we used an embedding-based approach, calculating and comparing cosine similarity between the vector embeddings of each reference and the equivalent translation generated by the model. We report the average of semantic similarity across utterances. We used two models with Sentence-Transformers (Reimers and Gurevych, 2019), “*all-mpnet-base-v2*” (Semantic 1) and “*all-MiniLM-L12-v2*” (Wang et al., 2020) (Semantic 2). As we fine-tuned all the models for approximately two epochs, we report the evaluation of the best performing checkpoint.

For inference, we used Faster-Whisper¹⁰ with the default VAD arguments. We also compared the results without VAD, and found that applying VAD at inference time is better for all the models (cf. Appendix A). We used 5 for “beam size” and 2 for “no repeat ngram size”.

As Table 2 shows, after fine-tuning Whisper Medium on both the authentic and synthetic audio data (Model B), there are consistent improvements across all metrics compared to when we fine-tuned it on the authentic audio data only (Model A). Moreover, Model B++ that uses three synthetic datasets outperforms Model B that uses only two synthetic datasets. This demonstrates that augmented authentic audio data with high-quality synthetic audio data can enhance end-to-end speech translation systems, especially for low-resource languages like Irish.

Model C uses the same training data as Model B

as well as two augmented versions, one version that applies basic VAD, removing low-amplitude samples (cf. Section 3.3.1) and another version that injects white background noise into the data (cf. Section 3.3.2). Although Model C that uses noise and VAD augmented data still outperforms Model A that uses authentic training data only, both Model B and B++ that combines authentic data with synthetic data outperform Model C.

While the choice of augmentation techniques were based on manual observation of the characteristics of the authentic data and the synthetic data, the achieved improvements encourage further investigation. In the future, we would like to conduct more experiments that employ other data augmentation techniques. Moreover, we would like to measure the effect of adding synthetic audio data compared to augmenting the authentic data only. Finally, as the main purpose of this research is to understand the best practices of using synthetic audio data (i.e. data generated by TTS models) to improve speech translation quality, we will conduct further study on mimicking authentic data characteristics to enhance the effect of data augmentation with synthetic audio data.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastopoulos, 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

⁹<https://github.com/mjpost/sacrebleu>

¹⁰<https://github.com/SYSTRAN/faster-whisper>

- 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). Association for Computational Linguistics.
- Loic Barrault, Andy Chung, David Dale, Ning Dong (ai), Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng-Jen Chen, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Abinash Ramakrishnan, Alexandre Mourachko, Amanda Kallet, Ann Lee, Anna Sun, Bapi Akula, Benjamin Peloquin, Bernie Huang, Bokai Yu, Brian Ellis, Can Balioglu, Carleigh Wood, Changhan Wang, Christophe Ropers, Cynthia Gao, Daniel Li (fair), Elahe Kalbassi, Ethan Ye, Gabriel Mejia Gonzalez, Hirofumi Inaguma, Holger Schwenk, Igor Tufanov, Ilia Kulikov, Janice Lam, Jeff Wang (pm Ai), Juan Pino, Justin Haaheim, Justine Kao, Prangthip Hasanti, Kevin Tran, Maha Elbayad, Marta R Costa-jussa, Mohamed Ramadan, Naji El Hachem, Onur Çelebi, Paco Guzmán, Paden Tomasello, Pengwei Li, Pierre Andrews, Ruslan Mavlyutov, Russ Howes, Safiyah Saleem, Skyler Wang, Somya Jain, Sravya Popuri, Tuan Tran, Vish Vogeti, Xutai Ma, and Yilin Yang. 2023. **SeamlessM4T—Massively Multilingual & Multimodal Machine Translation**. *Meta AI*.
- James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J Ó Meachair, and Jennifer Foster. 2022. **gaBERT — an Irish Language Model**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4774–4788, Marseille, France. European Language Resources Association.
- Nikolay Bogoychev and Rico Sennrich. 2019. **Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation**.
- Meghan Dowling, Teresa Lynn, and Andy Way. 2019. **Leveraging backtranslation to improve machine translation for Gaelic languages**. In *Proceedings of the Celtic Language Technology Workshop*, pages 58–62, Dublin, Ireland. European Association for Machine Translation.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. **Understanding Back-Translation at Scale**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. **The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text**. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. **Survey of Low-Resource Machine Translation**. *Computational Linguistics*, 06:1–67.
- Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. **Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT’s Submission to the AdapMT 2020 English-to-Hindi AI Translation Shared Task**. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAD).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. **Bag of Tricks for Efficient Text Classification**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. **The OpenNMT Neural Machine Translation Toolkit: 2020 Edition**. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.
- Seamus Lankford, Haithem Alfi, and Andy Way. 2021. **Transformers for Low-Resource Languages: Is Féidir Linn!** In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.
- Jihyun Lee, Yejin Jeon, Wonjun Lee, Yunsu Kim, and Gary Geunbae Lee. 2023. **Exploring the Viability of Synthetic Audio Data for Audio-Based Dialogue State Tracking**. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Taipei, Taiwan.
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhillah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johannes Lee, R Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi

- Leong, Quyet V Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. [SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages](#).
- Teresa Lynn. 2022. [Report on the Irish Language](#). Technical report, European Language Equality.
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. [Domain-Specific Text Generation for Machine Translation](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.
- 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.
- Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. [Adaptation of Machine Translation Models with Back-Translated Data Using Transductive Data Selection Methods](#). In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing CILing 2019: Computational Linguistics and Intelligent Text Processing*, pages 567–579, La Rochelle, France. Springer Nature Switzerland.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving Neural Machine Translation Models with Monolingual Data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A Study of Translation Edit Rate with Targeted Human Annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Jörg Tiedemann. 2012. [Parallel Data, Tools and Interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 485 in NIPS'20, pages 5776–5788, Red Hook, NY, USA. Curran Associates Inc.

A Appendix: Arguments

A.1 Inference VAD

Argument	Type	Value	Argument	Type	Value
threshold	float	0.5	min_silence_duration_ms	int	2000
min_speech_duration_ms	int	250	window_size_samples	int	1024
max_speech_duration_s	float	float("inf")	speech_pad_ms	int	400

Table 3: Default VAD values of *Faster-Whisper*.

A.2 Training Warm-up Ratio

Whisper	Model	Datasets	Data Size	Warm-up	BLEU	chrF++	WER	Semantic 1	Semantic 2
Small	A	authentic	29,663	0.00	31.49	45.59	59.66	58.23	60.35
				0.01	30.97	46.19	59.57	59.69	61.09
				0.03	31.43	46.71	61.14	60.48	61.59
	B	A + synthetic	48,719	0.00	34.09	50.79	55.47	65.64	66.66
				0.01	31.92	47.32	58.31	62.56	63.57
				0.03	34.15	49.81	56.87	65.09	66.43
	C	B + augmented	146,157	0.00	30.75	45.87	61.37	60.51	61.98
				0.01	32.82	48.31	57.95	63.26	64.72
				0.03	35.07	50.23	56.73	63.33	64.80

Table 4: **Comparing diverse values of warm-up ratio at training time.** Ratios are out of 3000 steps. Hence, 0.01 and 0.03 correspond to 30 steps and 90 steps, respectively. The results here are **with VAD at inference time**, using the default VAD arguments of *Faster-Whisper*. The highest score in each group is displayed in a bold font.

Whisper	Model	Datasets	Data Size	Warm-up	BLEU \uparrow	chrF++ \uparrow	WER \downarrow	Semantic 1 \uparrow	Semantic 2 \uparrow
Small	A	authentic	29,663	0.00	29.14	43.34	60.51	56.96	58.14
				0.01	30.66	45.41	62.09	58.69	59.79
				0.03	30.68	45.36	62.09	57.82	59.29
	B	A + synthetic	48,719	0.00	32.05	48.32	58.44	62.51	63.72
				0.01	31.94	46.81	59.93	61.57	62.36
				0.03	31.61	47.74	59.16	62.49	64.09
	C	B + augmented	146,157	0.00	30.51	44.52	63.48	59.6	60.71
				0.01	32.58	47.65	59.39	62.86	63.72
				0.03	31.89	48.83	59.84	62.32	63.17

Table 5: **Comparing diverse values of warm-up ratio at training time.** Ratios are out of 3000 steps. Hence, 0.01 and 0.03 correspond to 30 steps and 90 steps, respectively. The results here are **without VAD at inference time**. The highest score in each group is displayed in a bold font.

A.3 Training Epochs

Whisper	Model	Datasets	Data Size	Warm-up	Steps	Epoch	Best Epoch	BLEU \uparrow	chrF++ \uparrow	WER \downarrow	Semantic 1 \uparrow	Semantic 2 \uparrow
	A	authentic	29,663	0.03	2,000	1.08	1.02	29.14	47.03	63.17	60.78	62.11
				cont.	4,000	2.16	1.83	<u>32.38</u>	<u>48.95</u>	<u>58.85</u>	<u>62.09</u>	<u>63.28</u>
	B	A + synthetic (2d)	48,719	0.03	4,000	1.31	1.22	36.02	53.73	<u>53.26</u>	66.86	68.16
				cont.	7,000	2.30	2.27	<u>36.34</u>	<u>54.08</u>	53.35	<u>68.31</u>	<u>69.93</u>
Medium				0.03	4,000	0.55	0.55	38.41	57.18	51.10	69.72	71.13
	B++	A + synthetic (3d)	115,987	cont.	8,000	1.10	0.55	~	~	~	~	~
				cont.	15,000	2.07	0.55	~	~	~	~	~
	C	B + augmented	146,157	0.03	4,000	0.44	0.38	33.46	50.72	57.59	63.01	64.56
				cont.	10,000	1.09	1.05	<u>34.09</u>	<u>51.4</u>	<u>55.83</u>	<u>64.26</u>	<u>65.56</u>
				cont.	19,000	2.08	1.05	~	~	~	~	~

Table 6: Investigating the effect of training for 1-2 epoch(s). It seems that smaller amounts of training data can benefit from training for 2+ while larger amounts of data can benefit from training for only 1 epoch or less. The first row of each section starts the training with warm-up ratio 0.03, then the next 1 or 2 row(s) continues training for more steps without changing any training arguments. The reported scores are for the best step, based on training validation with 100-step intervals. That is why some rows are marked with the “~” sign, as the best step was still the same as the one reported in the previous row.