Adapting Whisper for Japanese-to-Chinese Speech Translation on Lightweight Models

Jian Feng Wen Gao

June 2025

Abstract

This paper investigates the potential of **Whisper**[3]'s lightweight models for direct translation between non-English language pairs, with a focus on Japanese-to-Chinese speech translation. While Whisper supports multilingual transcription and English-targeted translation out of the box, it lacks support for other translation directions. We fine-tune the `base` and `tiny` variants of Whisper using a high-quality dataset of Japanese movies, TV shows, and anime aligned with simplified Chinese subtitles. Our results show that the base model achieves promising **BLEU**[4] scores on real-world data, while the tiny model struggles to produce semantically complete outputs. These findings suggest that Whisper's architecture can be adapted for non-English translation, though model size remains a critical factor. This study also highlights the potential for deploying compact ASR-translation models in GPU-limited or edge computing environments. These findings demonstrate Whisper's extensibility beyond its default English-centric design and offer insights into balancing accuracy and efficiency for multilingual translation in constrained settings.

1 Introduction

Speech-to-text translation has become a vital tool for cross-lingual communication, enabling applications in media localization, education, and accessibility. While models like Whisper, developed by OpenAI, have demonstrated strong multilingual transcription and English-targeted translation, their capabilities for direct translation between non-English languages remain underexplored.

This design constraint presents a gap for applications requiring translation between widely used language pairs such as Japanese and Chinese—particularly relevant in East Asian media and education. Moreover, many real-world use cases demand models that can run on CPU-only devices or edge environments where GPU resources are unavailable. This calls for compact and efficient architectures that still provide usable translation quality.

In this study, we examine whether Whisper's smaller variants—specifically the `base` and `tiny` models—can be adapted to support Japanese-to-Chinese

speech translation. We fine-tune both models using a curated dataset derived from Japanese movies, TV dramas, and anime, paired with professionally aligned simplified Chinese subtitles. This domain offers realistic, emotionally rich, and acoustically diverse content, which poses unique challenges for speech translation systems.

Our goal is twofold: to explore the feasibility of extending Whisper to support non-English translation directions, and to evaluate whether lightweight models can deliver acceptable performance in this setting. Through this, we aim to inform future development of deployable, low-resource speech translation solutions for underrepresented language pairs.

2 Related Work

Recent advancements in multilingual speech models have significantly expanded the scope of automatic speech recognition (ASR) and speech translation. Whisper, introduced by OpenAI, is a notable example, supporting both transcription and speech-to-text translation across multiple languages. However, its translation capability is limited to non-English speech into English text, with no support for direct translation between non-English language pairs.

Efforts to address multilingual translation at scale have focused on large models such as Meta's **NLLB**[8] and **SeamlessM4T**[7]. These models demonstrate strong cross-lingual performance but are resource-intensive and often impractical for deployment in CPU-only or edge environments. In contrast, lightweight or parameter-efficient models—via techniques such as quantization, distillation, and **LoRA**[9] have emerged as promising solutions for low-resource deployment.

Whisper's smaller variants (`tiny`, `base`) offer a potential balance between performance and efficiency, but few studies have examined their capacity to translate non-English when fine-tuned with aligned data. In particular, the Japanese-to-Chinese direction is underexplored despite its practical importance in media, education, and localization.

Additionally, domain adaptation plays a crucial role in translation quality. Previous research has shown that fine-tuning in-domain data, such as technical speech, legal transcripts, or movie dialogue, can substantially boost performance. Our work builds on this insight by focusing on Japanese audiovisual content, including films, TV series, and anime, which exhibit diverse linguistic styles and present a challenging but valuable benchmark for speech translation systems.

3 Dataset

To train and evaluate our Whisper models for Japanese-to-Chinese speech translation, we constructed a high-quality, domain-specific dataset named **ScreenTalk-JA2ZH** [1]. This dataset consists of Japanese audio paired with aligned Chinese

subtitles, focusing specifically on content from Japanese **films, TV dramas, and anime**. These domains are characterized by diverse speech styles, rich emotional expression, background music, and colloquial language, making them a valuable benchmark for real-world speech translation tasks.

Dataset Statistics

The dataset is split into the following subsets:

- **Training set:** 582 hours
- **Validation set:** 73 hours
- **Test set:** 73 hours

All audio was standardized to 16kHz mono WAV format. The content was segmented at the sentence level using timestamp metadata and manually aligned with simplified Chinese subtitles. Basic quality checks ensured that the segments featured clear speech and accurate alignment, making the dataset suitable for supervised training.

Example Alignment

To demonstrate the structure of the dataset, Figure 1 shows a sample mel spectrogram derived from a Japanese audio segment. Figure 2 shows the aligned Chinese subtitle.

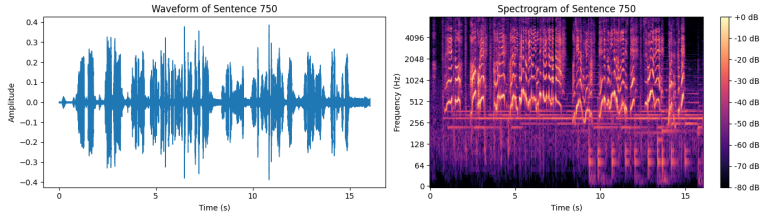

Figure 1: Mel spectrogram of a Japanese audio segment from the training set.

🎤 Sentence 750: 对了 差点就忘了 这副面具是我拿来准备送给你的 爷爷说这是一个传奇民族留下的 厉不厉害 你拿去用吧

Figure 2: Aligned Chinese subtitle corresponding to the audio segment.

This example illustrates the clean temporal and semantic alignment between speech and text—crucial for training effective supervised speech translation models.

Data Collection and Annotation

All source audio originates from professionally produced Japanese-language media, including films, TV dramas, and anime. The dialogue spans everyday conversations, dramatic monologues, emotional outbursts, and high-energy action sequences. This diversity in speech styles—combined with a variety of character voices and background sounds—offers rich acoustic variation, enhancing the model's exposure to real-world scenarios.

Subtitles were translated into simplified Chinese by fluent bilingual annotators and reviewed for consistency and accuracy. Audio segments were aligned at the sentence level based on timestamps, and noisy or unusable segments were manually filtered to ensure training quality.

Public Subset for Reproducibility

To support reproducibility and facilitate community research, we release a small-scale subset of the dataset named **ScreenTalk-JA2ZH-XS**[2], available on the Hugging Face. This subset includes a limited amount of manually aligned Japanese audio and simplified Chinese subtitles drawn from the same domains as the full corpus. It can be used for preliminary experimentation, benchmarking, or training lightweight prototypes.

4 Training Process

To explore Whisper's adaptability under constrained resources, we fine-tuned both the **Whisper tiny** and **Whisper base** models using the same dataset and training procedure.

Loss Curve Comparison

Both the `tiny` and `base` models exhibited a similar downward trend in training loss, suggesting comparable convergence behavior during the initial training stages.

The `tiny` model triggered early stopping sooner, resulting in slightly higher final loss. This is likely due to its lower capacity, which limits further learning progress under the same training setup. In contrast, the `base` model continued training for more steps and achieved a lower final loss.

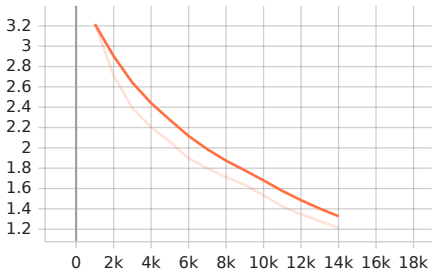

Figure 3: Whisper Tiny Training Loss

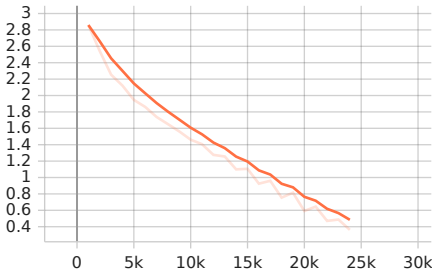

Figure 4: Whisper Base Training Loss

These results suggest that while both models learn effectively during early epochs, the `base` model benefits from greater capacity and a longer training schedule, allowing for more substantial loss reduction overall.

5 Validation Process

Model performance was monitored on the validation set using both BLEU score and evaluation loss. These metrics provide complementary perspectives on translation quality: BLEU reflects semantic adequacy, while loss captures token-level prediction consistency.

5.1 BLEU Score Evolution

Tiny model: The BLEU score for the tiny model initially climbed and reached a peak of approximately 0.72, demonstrating that even a lightweight model can partially learn the speech-to-translation mapping in early training. However, this gain was not sustained—the score declined rapidly after peaking, leading to an early stopping trigger. This sharp drop suggests that the model overfit quickly to patterns it could memorize but lacked the representational capacity to generalize more abstract or variable translation mappings. Given the limited number of parameters in the tiny model, it likely struggled to maintain stable performance in the presence of acoustic variability and linguistic complexity typical of Japanese audiovisual media. The model's small footprint makes it attractive for deployment, but this result highlights a trade-off between efficiency and robustness in cross-lingual translation.

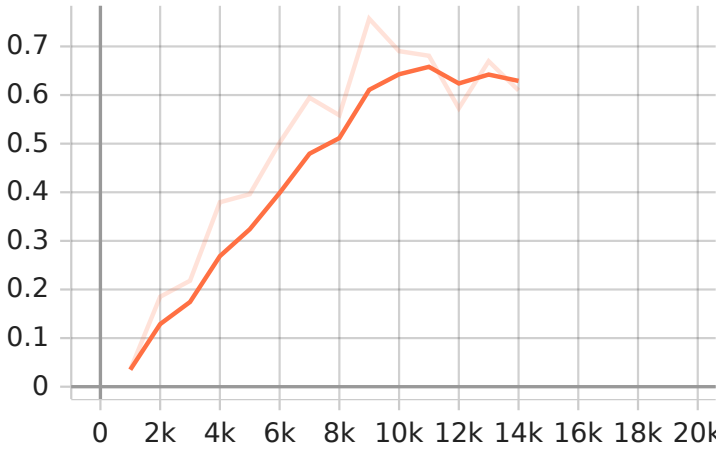

Figure 5: Validation BLEU Score — Whisper Tiny

Base model: The base model exhibited a smooth and progressive improvement in BLEU score, ultimately peaking around 0.96. In contrast to the tiny model, which experienced a sharp drop after its peak, the base model maintained high performance for a longer period. After reaching its maximum, the BLEU score oscillated within the 0.80–0.96 range over several evaluation intervals before triggering early stopping. This sustained performance suggests that the base model possessed sufficient capacity to generalize well across diverse and noisy inputs. The relatively stable BLEU behavior reflects the model's stronger ability to preserve semantic fidelity and resist overfitting, even in the context of expressive, domain-specific speech.

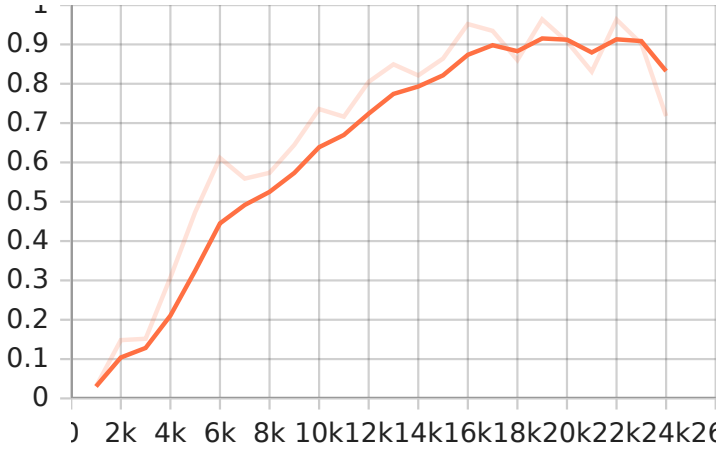

Figure 6: Validation BLEU Score — Whisper Base

Comparative Insights: The validation BLEU trajectories highlight dis-

tinct behavioral patterns between the tiny and base models. Although the tiny model achieved a peak BLEU of approximately 0.72, its performance was volatile—rising quickly but followed by a sharp drop that led to early stopping. This instability suggests difficulty in maintaining semantic consistency, likely due to limited model capacity.

In contrast, the base model reached a higher BLEU of around 0.96 and maintained stable performance across multiple evaluation steps before early stopping. Its smoother curve and extended high-score plateau indicate stronger generalization and better alignment with the validation set.

Overall, while both models showed learning ability, only the base model demonstrated resilience in preserving translation quality across training. The results underline the importance of model size when tackling complex, non-English translation tasks.

5.2 Evaluation Loss Trend

The evaluation loss curves of both models exhibit early signs of overfitting, but the patterns diverge due to their capacity and training durations.

The `tiny` model's loss decreased slightly in the early stages, then entered a phase of fluctuation without clear further gains. Since training was terminated early by the stopping criterion, the curve did not develop a full U-shape. This abrupt end suggests that while overfitting had likely begun, the model lacked sufficient capacity to sustain meaningful learning or to manifest a more stable pattern.

In contrast, the `base` model trained longer and demonstrated a more classic U-shaped validation loss curve. After a steady decline to a well-defined minimum, the loss began to rise gradually, signaling overfitting in the later stages—but only after several strong evaluation steps. The deeper minimum and longer stable period reflect better learning and generalization before degradation set in.

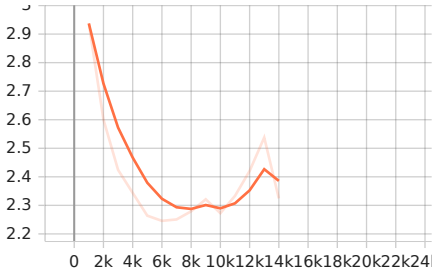

Figure 7: Whisper Tiny Validation Loss

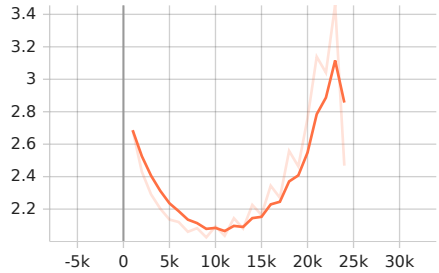

Figure 8: Whisper Base Validation Loss

Overall, while evaluation loss provides useful insights into training dynamics, it is not always a reliable indicator of translation quality. For speech translation

tasks, BLEU scores offer a more direct measure of semantic fidelity and are thus more informative for evaluating real-world performance.

6 Test Set Evaluation

After training, both models were evaluated on a held-out test set to assess their ability to generalize to unseen data. The `tiny` model achieved a BLEU score of approximately **0.60**, while the `base` model reached a higher score of **0.7179**.

Although the `tiny` model produced intelligible translations in many cases, its output was less consistent and prone to semantic drift, especially in longer or noisier audio segments. In contrast, the `base` model demonstrated stronger fluency and alignment with reference translations, particularly in maintaining sentence structure and preserving speaker intent.

It is worth noting that while BLEU provides a quantitative benchmark, it may not fully reflect qualitative aspects of translation, such as fluency or contextual appropriateness. In our experiments, the `base` model consistently outperformed the `tiny` model not only in BLEU score but also in subjective assessments of translation quality.

These results confirm that the `base` model offers a better balance between performance and model size, whereas the `tiny` variant, despite being more lightweight and suitable for deployment on resource-constrained devices, exhibits limitations in both stability and semantic accuracy.

7 Model Comparison

To explore the effect of model capacity on translation quality, we fine-tuned both the `Whisper tiny` and `base` models using the same dataset and training protocol. While the `tiny` model offers a significant advantage in terms of computational efficiency and deployment flexibility, its performance fell notably short of the `base` model across all evaluation metrics.

The `tiny` model peaked at a BLEU score of approximately 0.60 on the test set, compared to 0.7179 achieved by the `base` model. Qualitatively, the `tiny` model struggled with semantic coherence, often producing incomplete or imprecise translations—particularly in longer or context-rich segments.

We attribute this performance gap primarily to the limited parameter capacity of the `tiny` model. Its reduced representational power makes it less capable of modeling complex acoustic and linguistic patterns, which are especially prevalent in Japanese audiovisual content like films, TV dramas, and anime.

These findings underscore the importance of model expressiveness in non-English speech translation tasks. While smaller models may offer deployment advantages, achieving high-quality translation—especially in linguistically and acoustically diverse domains—may require larger architectures.

As part of future work, we plan to experiment with larger Whisper variants (e.g., `medium` and `large`) and explore efficient fine-tuning strategies such as

LoRA, with the goal of balancing performance and computational cost for real-world deployment.

8 Discussion

Our experiments provide insights into the feasibility and tradeoffs involved in adapting Whisper for non-English speech translation under resource constraints.

Evaluation Metrics and Semantic Alignment

A notable observation is the divergence between token-level evaluation loss and semantic-level translation quality. Toward the end of training for the `base` model, evaluation loss began to rise while BLEU scores continued to improve. This suggests that the model may have learned better semantic alignments even as it incurred greater token-level mismatches with the reference translations. Such behavior highlights a known limitation of using loss as a sole proxy for translation quality.

BLEU, while helpful, is itself a surface-level metric and may not fully capture nuanced meaning preservation. For future work, it would be valuable to incorporate additional evaluation signals such as **COMET**[5], **chrF**[6], or human assessments to provide a more comprehensive view of translation performance.

Balancing Performance and Deployability

The results also highlight the practical tradeoff between model size and deployment feasibility. Larger models typically offer higher accuracy, but often exceed the computational budgets of real-world applications. Interestingly, the Whisper `base` model strikes a promising balance—it can be deployed on modern CPUs while offering much stronger translation quality than the `tiny` variant. Although not as lightweight, its gains in robustness and stability may justify its use in edge environments where translation accuracy is a priority.

In summary, our findings point to the importance of aligning model capacity with deployment constraints, and of using task-appropriate metrics to evaluate performance. These considerations are essential for advancing Whisper's utility in low-resource, non-English translation tasks.

9 Future Work

Our study highlights several directions for future development.

First, we plan to scale up to larger Whisper variants such as `medium` and `large`. Given the limitations observed in the `tiny` model, we expect that increased model capacity will lead to more stable training dynamics and stronger translation quality. To offset the computational cost, we also intend to experiment with parameter-efficient fine-tuning strategies such as LoRA.

Second, we aim to enhance evaluation by adopting more comprehensive metrics. While BLEU provides a useful baseline, it does not fully capture semantic quality or fluency. Future work will incorporate metrics like COMET and chrF, as well as human assessments, to better understand model outputs in real-world applications.

Third, we plan to expand the dataset to include more diverse speech sources beyond Japanese films, TV dramas, and anime. This includes conversational speech, news broadcasts, and domain-specific content such as educational or technical dialogues. A broader dataset will allow the model to generalize better across speaking styles, acoustic environments, and linguistic registers.

Together, these efforts will further unlock the potential of Whisper-based translation models for non-English, low-resource settings.

10 Release and Availability

The fine-tuned Whisper models for Japanese-to-Chinese speech translation are available on the Hugging Face Hub:

- **Whisper Base**: `https://huggingface.co/Itbanque/whisper-ja2zh-base`
- **Whisper Tiny**: `https://huggingface.co/Itbanque/whisper-ja2zh-tiny`

We provide all relevant model files, training logs, and evaluation metrics to support transparency and reproducibility.

11 Limitations

Despite the promising results, this study has several limitations:

- **Domain specificity:** The training data is primarily drawn from Japanese films, television dramas, and anime. While rich in linguistic diversity, this focus may limit the model's generalizability to other domains such as academic lectures, news broadcasts, or spontaneous dialogue.
- **Language direction:** This work addresses only the Japanese-to-Chinese translation direction. The effectiveness of similar fine-tuning strategies on other non-English language pairs remains to be explored.
- **Computational cost:** Full fine-tuning, even on compact models like Whisper `base`, requires substantial GPU memory and training time, which may present a barrier to adoption in low-resource settings.
- **Lack of baseline:** The original Whisper model does not support Japanese-to-Chinese translation natively, making it difficult to compare against a standard zero-shot baseline for this task.

- **Evaluation Scope:** While BLEU provides a quantitative proxy for translation quality, it may not fully capture semantic fidelity or fluency, especially in colloquial or expressive speech common in media content. Other evaluation methods such as COMET, chrF, or human judgment were not used in this study, which may limit the completeness of the quality assessment.

Acknowledgements

We thank all contributors involved in the dataset preparation and annotation process. In particular, we acknowledge Wen Gao for curating and aligning the Japanese-Chinese corpus used in this study. We also appreciate the open-source community for developing the Whisper model and related tools.

A Appendix

A.1 Training Configuration Details

- Epochs: 20
- Learning rate: 3e-4
- Batch size: 96(tiny), 64(base)
- Gradient Accumulation Steps: 1
- Warm Up Steps: 1000
- Precision: fp16
- Early Stopping: 5
- Eval Strategy: Step-based

References

[1] Itbanque. *ScreenTalk-JA2ZH Dataset*. Hugging Face. Available at: `https://huggingface.co/datasets/Itbanque/ScreenTalk_JA2ZH`

[2] Itbanque. *ScreenTalk-JA2ZH-XS: A Sample Dataset for Whisper Fine-Tuning*. Hugging Face. Available at: `https://huggingface.co/datasets/Itbanque/ScreenTalk_JA2ZH-XS`

[3] OpenAI. *Whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. GitHub Repository: `https://github.com/openai/whisper`

[4] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. *BLEU: a method for automatic evaluation of machine translation*. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002.

[5] Rei, R., et al. *COMET: A Neural Framework for MT Evaluation*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). `https://github.com/Unbabel/COMET`

[6] Popović, Maja. *chrF: character n-gram F-score for automatic MT evaluation*. In Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT), pages 392–395, 2015. `https://aclanthology.org/W15-3049`

[7] Seamless Communication Team. *SeamlessM4T: Massively Multilingual and Multimodal Translation*. Meta AI, 2023. `https://github.com/facebookresearch/seamless_communication`

[8] Team, NLLB et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. Transactions of the Association for Computational Linguistics (TACL), 2022. `https://aclanthology.org/2022.tacl-1.114/`

[9] Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Lu, and Chen, Weizhu. *LoRA: Low-Rank Adaptation of Large Language Models*. In Proceedings of the 11th International Conference on Learning Representations (ICLR), 2022. `https://arxiv.org/abs/2106.09685`