Cross-Lingual Transfer Learning for Speech Translation

Anonymous ACL submission

Abstract

There has been increasing interest in building multilingual foundation models for NLP and speech research. This paper examines how to expand the speech translation capability of these models with restricted data. Whisper, a speech foundation model with strong performance on speech recognition and English translation, is used as the example model. Using speech-to-speech retrieval to analyse the audio representations generated by the encoder, we show that utterances from different languages are mapped to a shared semantic space. This shared embedding space can then be leveraged for zero-shot cross-lingual transfer in speech translation. By fine-tuning the Whisper decoder with only English-to-Chinese speech translation data, improved performance for translation to Chinese can be obtained for multiple languages, in addition to English. Furthermore, for languages related to those seen in training it is possible to perform speech translation, despite the model never seeing the language in training, or being able to perform transcription.

1 Introduction

Speech translation (ST) systems directly generate transcriptions in the target language from spoken utterances in a different language and have various applications (Inaguma et al., 2019; Nakamura, 2009). With the growing demand for multilingual models, it is crucial to develop translation systems that support multiple languages, both as source and target. However, data collection for training ST systems is more challenging than for Neural Machine Translation (NMT) and Automatic Speech Recognition (ASR) tasks. Unlike NMT, where the same text corpus can be used for both translation directions (Artetxe and Schwenk, 2019), ST systems face challenges due to their asymmetric input-output nature. For instance, data for translating audio in language X into text in English ($X \rightarrow en$) would be easier to collect than $en \rightarrow X$ data, largely due to the higher global demand for English translations. Moreover, high-resource language pairs have more available data than low-resource pairs.

Given the high cost of collecting diverse data pairs for ST systems, understanding what is required to build a multilingual ST model and expand its capability to more languages is essential. In this work, we use OpenAI's Whisper (Radford et al., 2023) as a case study to explore the behavior of multilingual speech foundation models. Whisper is pre-trained to support speech recognition in 100 languages and translation from 99 languages into English ($X \rightarrow en$). The encoder can extract semantic information from the acoustic features. We hypothesise that the features in different languages are aligned within a shared semantic space, and this alignment could enable the model to support translation from multiple source languages, a key feature for expanding multilingual ST capabilities. Whisper's decoder acts as a language model that generates tokens conditioned on the encoder outputs. By supporting multiple languages at the token level, the decoder facilitates translation into various target languages. This flexibility allows us to test and expand its ST capabilities to new target languages, which we verify through zero-shot and fine-tuning experiments.

In this work, we explore how to extend Whisper's capability in speech translation, expanding its supported translation language pairs. First, we evaluate the level of language invariance in the embeddings produced by the Whisper encoder using a speech-to-speech retrieval task (Lee et al., 2015). Second, we expand the translation to a new target language by fine-tuning Whisper, the results show a level of cross-lingual transferability among the source languages. Third, we show that Whisper can translate spoken utterances from previously unseen languages into English texts, indicating its ability to map unseen languages into a shared speech embedding space.

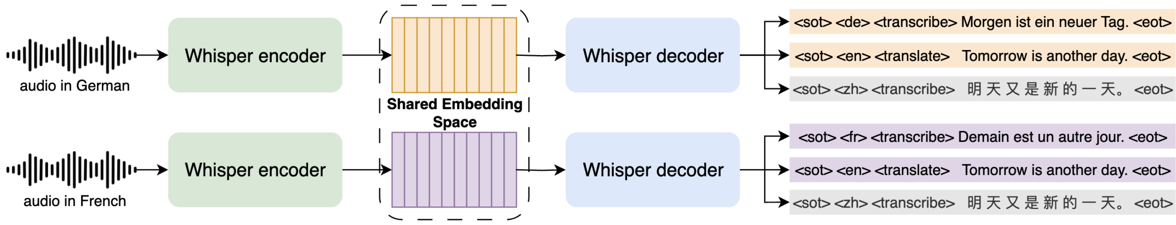

Figure 1: Illustration of the decoding process of Whisper for ASR and speech translation tasks.

2 Related Works

Prior work has shown that multilingual text models, such as M-BERT (Pires et al., 2019), produce language-invariant embeddings, mapping the same semantic information from different languages to a similar embedding space. This language invariance enables cross-lingual text retrieval (Pires et al., 2019; Wu and Dredze, 2019; Cao et al., 2019) and boosts the model performance in other languages, when fine-tuned only on English corpus (Pires et al., 2019). This transfer learning capability is particularly beneficial in low-resource settings. (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019) have shown that using machine translation as the training objective can effectively generate language-invariant embeddings. In this work, we examine whether it applies to speech-based models with speech translation as the training objective. Unlike text models, Whisper's pre-training for speech translation only uses English as the target language, and its utterance embeddings are not explicitly aggregated. Additionally, speech representations are much longer than text tokens. These differences add to the difficulty of the auto-alignment in the speech encoder space.

3 Speech Translation

Whisper Model

The Whisper models are trained in a weakly supervised way and come in various sizes, from the tiny model with 39M parameters to the large model with 1550M parameters (Radford et al., 2023). During pre-training, the model learns in a multi-task fashion on automatic speech recognition, speech translation, voice activity detection, and language identification. In decoding, it generates different outputs based on the "context" tokens given to the decoder. For ASR, Whisper converts an utterance in language L into its corresponding transcription, $\text{Utt}_{\mathsf{L}} \rightarrow \text{Text}_{\mathsf{L}}$. For speech translation, it supports translation from any supported language to English, represented as $\text{Utt}_{\mathsf{L}} \rightarrow \text{Text}_{\mathsf{EN}}$. Figure 1 illustrates the example decoding process and the associated context tokens in orange and purple text blocks.

Audio Embeddings

Given that multilingual text models like M-BERT generate language-invariant embeddings, it's reasonable to investigate whether Whisper, a multilingual speech model, exhibits similar properties. If Whisper's encoder produces language-invariant speech embeddings, it would be a significant advantage for handling multiple source languages in speech translation. This cross-lingual capability enables Whisper to effectively translate between various language pairs by aligning speech representations across different source languages.

To assess the cross-lingual alignment of Whisper, we use zero-shot speech-to-speech retrieval tasks (Boito et al., 2020; Duquenne et al., 2023) as an evaluation method. In this task, given a query audio q, the goal is to retrieve an utterance $\hat{r}_q$ in the target language that conveys the same meaning as q from a set of R candidates. We measure the performance of the speech retrieval task using the recall rate, $\text{R@1} = \frac{1}{|Q|}\sum_{q \in Q} \mathbb{I}(r_q, \hat{r}_q)$ where r_q is the retrieved result and $\hat{r}_q$ is the reference. For each query q and candidate audio r, we extract the encoder output sequences from Whisper, denoted as E_q and E_r. The retrieved utterance r_q is then determined as the one with the highest similarity score, $r_q = \arg\max_{r \in R} \text{Sim}(E_q, E_r)$.

We propose **SeqSim**, a metric inspired by BERTScore (Zhang et al., 2019), to compute similarity between two speech embedding sequences:

$$\text{Re}_{\text{seq}} = \frac{1}{|X|}\sum_{\boldsymbol{x} \in X} \max_{\boldsymbol{y} \in Y} \boldsymbol{x}^\top \boldsymbol{y};\ \text{Pr}_{\text{seq}} = \frac{1}{|Y|}\sum_{\boldsymbol{y} \in Y} \max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top \boldsymbol{y}$$
$$\text{SeqSim} = 2 \cdot \frac{\text{Pr}_{\text{seq}} \cdot \text{Re}_{\text{seq}}}{\text{Pr}_{\text{seq}} + \text{Re}_{\text{seq}}} \quad (1)$$

While BERTScore evaluates text generation tasks by comparing embeddings of individual tokens, SeqSim adapts this concept for audio frames. It computes the cosine similarity between embeddings of audio frames from one speech utterance $\boldsymbol{X}$ and those from another speech utterance $\boldsymbol{Y}$. Specifically, SeqSim measures how well each audio frame in $\boldsymbol{X}$ matches with the most similar frame in $\boldsymbol{Y}$.

New Target Languages

Although Whisper was trained to translate speech into English, its decoder has been exposed to a diverse range of languages and their corresponding tokens throughout its training. This extensive multilingual exposure suggests that the model might also be capable of translating into other languages. To investigate this potential, we evaluate Whisper's baseline translation performance for languages beyond English. Following (Peng et al., 2023), which demonstrated that the `<transcribe>` task token can outperform `<translate>` in the translation task, we compare these tokens in the zero-shot experiments to test translation into new target languages. Fine-tuning the model for a new target language is also compared. Figure 1 shows the decoding process with an added target language: Chinese, zh.

We discuss above the potential of Whisper's encoder generating embeddings within a shared semantic space, facilitating cross-lingual transferability. This feature allows Whisper to handle multiple source languages in speech translation. When fine-tuning Whisper for a specific language pair to expand the speech translation to a new target language (e.g. $en \rightarrow zh$), we expect improved performance for other source languages translating into the same target language ($X \rightarrow zh$). This aspect will be examined in Section 4.3.

New Source Language

Low-resource languages not seen during Whisper's training have different lexical representations compared to the languages the model was trained on. However, they may share similar acoustic features. It remains to be seen whether speech embeddings for these low-resource languages also fall within the model's shared semantic space. If so, this alignment could enable Whisper to effectively expand its speech translation capabilities to include these new source languages. Section 4.4 will explore this possibility through experiments.

4 Experimental Results

4.1 Setup

The Whisper large-v2 model is selected for the multilingual speech translation experiments, which shows superior performance compared to other model sizes (Radford et al., 2023). We evaluate speech translation on the FLEURS dataset (Conneau et al., 2023), which provides n-way parallel speech data. For the main experiments, we selected 5 languages: English (en), French (fr), German (de), Chinese (zh), and Japanese (ja), chosen for their wide usage and representation of different language families. To extend Whisper's ability to translate into a new target language, we use the en-to-zh subset from the CoVoST dataset (Wang et al., 2021), totalling 428 hours, in supervised training. For experiments in Section 4.4 evaluating new source languages, we choose 6 languages unsupported by Whisper: Kabuverdianu (kea), Asturian (ast), Cebuano (ceb), Kyrgyz (ky), Sorani Kurdish (ckb), and Irish (ga). Detailed descriptions of the datasets and the experimental setup are provided in Appendix A.1 and A.2.

Query	R@1 [%]				
	en	fr	de	zh	ja
en	-	80.0	80.0	46.2	45.5
fr	73.2	-	64.8	42.0	48.1
de	70.4	62.2	-	42.7	48.1
zh	26.5	25.4	19.0	-	43.2
ja	18.1	22.3	16.4	35.2	-

Table 1: Zero-shot speech-to-speech retrieval results measured with SeqSim on FLEURS.

4.2 Results on Speech-to-Speech Retrieval

In preliminary experiments, we compared various similarity measures on three language pairs from FLEURS. SeqSim consistently outperformed other measures in capturing speech embedding similarity. Consequently, SeqSim is adopted for the retrieval experiments presented in this paper. Detailed comparison and results are discussed in Appendix B.2.

Using SeqSim, we conduct experiments on 20 language pairs from the FLEURS dataset, with results detailed in Table 1. On all 20 language pairs, SeqSim consistently achieved remarkably higher recall rates compared to a random baseline of 0.2%. This suggests that these languages share a common embedding space, where semantically similar speech utterances are mapped to close regions. Notably, retrieval performance is better when both the query and the candidate utterances belong to the same language family. For instance, retrieval between English (en), French (fr), and German (de) – all Indo-European languages – show higher performance. This is likely due to greater overlap in phoneme representations among these languages, which facilitates the model's ability to align and match audio frames effectively.

BLEU / COMET		Zero-shot		Fine-tune
Dataset	src	Translate	Transcribe	en-to-zh
FLEURS	en	1.0 / 58.8	10.3 / 66.3	**29.1 / 78.4**
	fr	0.9 / 56.2	15.7 / 66.7	**23.0 / 74.1**
	de	1.0 / 57.2	16.8 / 67.1	**24.0 / 74.7**
	ja	1.0 / 59.3	15.9 / 70.7	**19.2 / 74.7**
CoVoST	en	1.8 / 59.0	3.8 / 61.2	**31.9 / 76.3**

Table 2: Zero-shot and fine-tuning results (BLEU/COMET) for Whisper speech translation into Chinese.

src	code	WER	R@1	ST (en)
kea	pt	89.5	85.4	32.6
ast	es	47.8	72.8	27.9
ceb	en	98.1	37.9	10.0
ky	ru	103.2	21.0	4.2
ckb	fa	107.1	19.1	1.9
ga	en	105.9	11.0	2.6

Table 3: ASR, retrieval (R@1), and ST (BLEU score) into English for 6 unsupported languages on FLEURS data, with Whisper decoding language code specified.

4.3 New Target Language

Whisper is originally designed for speech translation into English. This section explores methods to extend its capabilities to translate into other target languages, using Chinese as an example.

Zero-shot: Following research by (Peng et al., 2023), we tested two sets of context tokens in the zero-shot experiments: `<sot><zh><translate>` and `<sot><zh><transcribe>`. The first set follows Whisper's default decoding process. Since Whisper was initially trained to produce English translations, it outputs English words even when a target language code zh is used. In contrast, using the `transcribe` token gave a large performance improvement, as the results show in Table 2. This suggests that Whisper has learned to handle tokens of multiple languages through its multilingual speech recognition training, suggesting its potential for translating into languages beyond English.

Fine-tune: We fine-tune Whisper on English-to-Chinese speech translation data from CoVoST, freezing the encoder to preserve the audio embedding space and updating only decoder parameters with the context tokens `<sot><zh><transcribe>`. This improved English-to-Chinese translation on the FLEURS and CoVoST datasets, as shown in Table 2. Testing French, German, and Japanese utterances from FLEURS revealed that fine-tuning also improved BLEU and COMET scores for these languages. Although these source languages were not included in fine-tuning, the improvements in English translation capabilities benefited them due to the cross-lingual alignment feature of Whisper.

4.4 New Source Languages

We have shown that Whisper features a shared semantic embedding space across languages. This section explores whether this cross-lingual transferability extends to low-resource languages that Whisper has not been directly trained on. To test this, we select 6 unsupported languages from the FLEURS dataset and used a language code from their most similar language, chosen based on vocabulary overlap, for decoding (Qian et al., 2024). While Whisper struggles with accurate ASR transcriptions for these low-resource languages, as shown by high WER in Table 3, some languages exhibit high recall (R@1) rates when retrieving English speech (such as kea and ast). This suggests that even though these languages were unseen during training, their audio embeddings are mapped to the shared semantic space. This effectiveness likely results from the audio similarities between these low-resource languages and those in Whisper's training data.

Utilising these speech embeddings, the Whisper decoder can translate these languages into English. The results in Table 3 reveal surprisingly good BLEU scores for languages like Kabuverdianu (kea) and Asturian (ast) (only BLEU scores are given as some languages are not supported by COMET). This suggests that Whisper's cross-lingual alignment enhances performance in both retrieval and translation tasks for languages not explicitly included in its training.

5 Conclusions

This work demonstrates how to extend speech translation capabilities in Whisper. Whisper's decoder, supporting diverse language tokens, allows for effective expansion to new target languages. Our experiments reveal high recall rates in speech-to-speech retrieval, indicating that Whisper's encoder captures language-invariant features across languages. Fine-tuning Whisper on en-to-zh data improved BLEU scores by 5.9 for three other source languages. Furthermore, Whisper can successfully translate speech from some previously unseen languages into English, despite high WERs. These results confirm that Whisper maps utterances into a shared embedding space, enabling effective cross-lingual transfer for speech translation.

6 Limitations

Despite promising results, this work has several limitations. First, fine-tuning Whisper on $en \to zh$ translation data led to performance degradation on $X \to en$ translations, highlighting a common issue of catastrophic forgetting. Additionally, our experiments focused on one new target language. While we believe the findings are applicable to other target languages, evaluating the model across a broader range of target languages would provide a more comprehensive assessment of its capabilities. Lastly, although Whisper shows potential for unseen languages, there is room for improvement in handling low-resource languages more effectively, such as Irish (ga). Future work will explore these aspects.

7 Risks and Ethics

There are no known ethical concerns or risks associated with the findings of this work.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proc. of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 4218–4222.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In *Proc. of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 6486–6493.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *Proc. International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proc. 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*, pages arXiv–2308.

David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 649–665.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations (ICLR)*.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Viacheslav Klimkov, Srikanth Ronanki, Jonas Rohnke, and Thomas Drugman. 2019. Fine-Grained Robust Prosody Transfer for Single-Speaker Neural Text-To-Speech. In *Proc. INTERSPEECH 2019*, pages 4440–4444.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: CTC meets optimal transport. In *Proc. of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lin-Shan Lee, James Glass, Hung-Yi Lee, and Chun-An Chan. 2015. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal. 2019. Towards Achieving Robust Universal Neural Vocoding. In *Proc. INTERSPEECH 2019*, pages 181–185.

Satoshi Nakamura. 2009. Overcoming the language barrier with speech translation technology. *Science & Technology Trends-Quarterly Review*, 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization. In *Proc. INTERSPEECH 2023*, pages 396–400.

Gabriel Peyré and Marco Cuturi. 2019. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Mengjie Qian, Siyuan Tang, Rao Ma, Katherine Knill, and Mark Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. In *to be published Proc. INTERSPEECH 2024*. Preprint arXiv:2407.06800.

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. International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Stan Salvador and Philip Chan. 2004. FastDTW: Toward accurate dynamic time warping in linear time and space. In *Proc. KDD Workshop on Mining Temporal and Sequential Data*, volume 6, pages 70–80. Seattle, Washington.

Holger Schwenk and Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proc. of the 2nd Workshop on Representation Learning for NLP*, pages 157–167.

Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 78–109, Virtual. Association for Machine Translation in the Americas.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. *Proc. INTERSPEECH 2021*, pages 2247–2251.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Experimental Setup

A.1 Data Details

Table 4 listed three public datasets we used in the experiments. For the FLEURS dataset (Conneau et al., 2023), we processed the data by retaining only the utterances that are available in all five selected languages. The original dev and test sets provided in the dataset are combined to create a bigger evaluation set. To increase the difficulty of the designed retrieval task, we randomly kept only one instance for utterances with the same transcription but recorded by different speakers. For the supervised experiments, we fine-tune the Whisper model on the CoVoST dataset (Wang et al., 2021), which is part of the Common Voice project (Ardila et al., 2020). In the speech retrieval experiments to demonstrate the alignment of the encoder outputs, an additional dataset MaSS (Boito et al., 2020) is used. The MaSS dataset contains parallel speech data extracted from verses in 8 languages: English (en), Spanish (es), Russian (ru), Romanian (ro), French (fr), Finnish (fi), Hungarian (hu), and Basque (eu). As the released Hungarian data is incomplete we discarded it in the experiments.

Dataset	Split	Langs	Utts	Hours	Words
FLEURS	test	5	426	1.1	9K
CoVoST	train	2	288,204	428	2.8M
	dev	2	1,000	1.6	9K
	test	2	1,000	1.6	9K
MaSS	test	7	814	8.3	18K

Table 4: Dataset description. The number of utterances, total duration of speech data, and word counts in the references are calculated based on the English data.

A.2 Training Details

In the training and evaluation of Whisper, the original audio is chunked or padded into segments with a length of 30 seconds. In our zero-shot speech-to-speech retrieval experiments, we only keep the embedding vectors that correspond to meaningful content in the original audio and remove the ones associated with the padded part. This practice proves to be effective in the retrieval experiments. To evaluate the model performance on ST, we use BLEU (Papineni et al., 2002) and COMET scores (Rei et al., 2020; Stewart et al., 2020; Rei et al., 2022) with the *Unbabel/wmt22-comet-da* model. In the supervised ST setting, the model parameters are updated on the training set of CoVoST for 220K steps with fine-tuning or LoRA tuning (Hu et al., 2022). The initial learning rate is $1e^{-5}$ for fine-tuning and $1e^{-3}$ for LoRA tuning and decays linearly. A batch size of 16 is used during training.

B Analysis of Audio Embeddings

B.1 Visualisation of Encoder Alignment

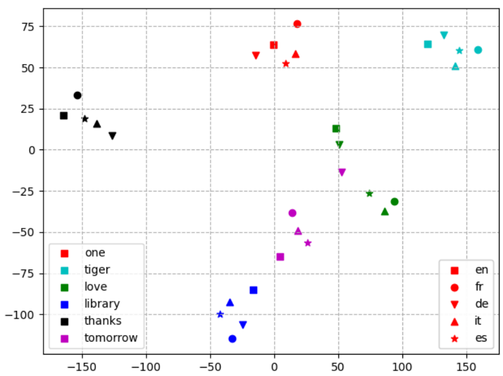

Figure 2: t-SNE visualization of contextual speech embeddings generated by Whisper large-v2 encoder for 6 word tuples across 5 languages.

To study the language-invariance of the Whisper encoder space, we use the Amazon text-to-speech service (Lorenzo-Trueba et al., 2019; Klimkov et al., 2019) to generate utterances for a set of words in different languages. From these utterances, the average speech embedding was computed using the Whisper large-v2 encoder. The resulting embeddings were reduced using t-SNE (Van der Maaten and Hinton, 2008) and plotted as shown in Figure 2. This initial analysis indicates that embeddings for words with the same meaning, such as *"thanks"* in different languages *(merci, danke, grazie, gracias)*, are closely aligned.

To further illustrate how languages share a common embedding space, we present an example of two parallel utterances from the FLEURS dataset, as shown in Figure 3. We computed average speech embedding vectors for each word based on word-level timestamp information. The figure reveals that words with similar meanings, even if they are in different languages and have different pronunciations, tend to be mapped to similar regions in the embedding space. For instance, *doorbell* (English) and *Türklingel* (German) show high cosine similarity scores despite their distinct pronunciations, indicating their embeddings are close due to their shared meaning. Additionally, the cosine similarity matrix also reflects word order changes. For exam-

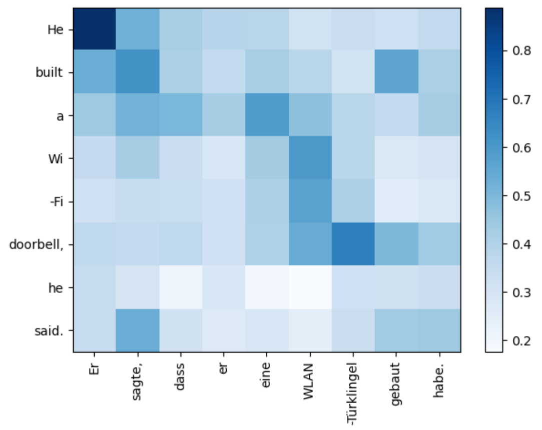

Figure 3: Cosine similarity matrix of utterance representations between an English sentence and its German counterpart selected from FLEURS test sets.

ple, *built* (English) and *gebaut* (German) have high cosine similarity because they convey the same concept, and *sagte* (German) aligns closely with *said* (English). This alignment in the embedding space supports the idea that semantically similar utterances across different languages are mapped to nearby regions in the embedding space, highlighting the shared nature of the embedding space.

B.2 Comparison of different similarity measures

To compute the similarity between two speech embedding sequences, we propose to use the AvgSim metric. The mean vector of embedding sequences X and Y are aggregated and then the cosine similarity between them is calculated to get an average similarity score. Compared to SeqSim, AvgSim captures the overall vector similarity rather than individual contextual speech embedding vectors.

$$\mathrm{AvgSim} = \mathrm{CosSim}\left(\frac{1}{|X|}\sum_{\boldsymbol{x}\in X}\boldsymbol{x}, \frac{1}{|Y|}\sum_{\boldsymbol{y}\in Y}\boldsymbol{y}\right) \tag{2}$$

In Table 5, different similarity measures are compared on three language pairs from the FLEURS data for the speech-to-speech retrieval task. Results from two additional metrics are listed here. In (Le et al., 2023), distance metrics based on Dynamic Time Warping (DTW) (Salvador and Chan, 2004) and Optimal Transport (OT) (Peyré and Cuturi, 2019) are used to measure the similarity, $\mathrm{Sim}(X, Y)$, between the contextual speech embeddings X and Y. Both metrics use cosine distance to derive an overall sequence similarity score.

While AvgSim is straightforward to compute, it overlooks the nuanced differences between the two sequences. DTWSim aligns the utterance representations in a monotonic fashion, which may not hold when the word order is different for the source and target sentence. To this end, we also use Optimal Transport (following (Le et al., 2023)) to compare individual embedding pairs. We do not add a cost associated with the embedding index to ensure OT can capture token re-orderings. As the results show, it outperforms the previous two methods. Across three retrieval settings, our proposed SeqSim better captures the speech embedding similarity and shows the best performance.

Method	R@1 [%]		
	en-fr	en-de	de-fr
Random	0.2	0.2	0.2
AvgSim	28.2	27.5	24.6
DTWSim	29.9	26.5	22.1
OTSim	72.3	66.7	55.2
SeqSim	**80.0**	**80.0**	**62.2**

Table 5: Comparison of different similarity measures for zero-shot speech-to-speech retrieval on FLEURS test sets.

B.3 Analysis of Speech-to-Speech Retrieval

In Figure 4 we alternate the speech embeddings using outputs from different encoder layers of Whisper. As shown, outputs from the last encoder layer consistently achieve the best retrieval performance. For bottom layers, the recall rate drops significantly. The results indicate that outputs from higher layers are better aligned and exhibit stronger cross-lingual characteristics.

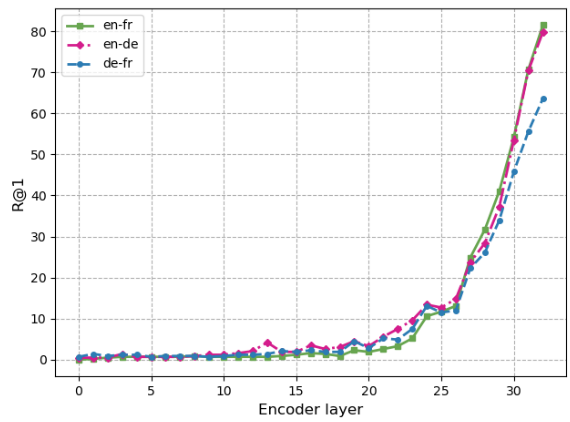

Figure 4: Speech-to-speech retrieval performance using outputs from different encoder layers of Whisper large-v2.

In Table 6 we show the retrieval performance using encoder outputs from different Whisper models on FLEURS test sets. Even for the tiny model with only 39M parameters, the recall rate is much better than the random baseline of 0.2%, suggesting that all models acquire the capability to do cross-lingual utterance alignment during pre-training. When the model size increases, the recall rate also improves. This implies that the retrieval performance will likely continue to improve if larger and more capable multilingual models are released in the future. For the Whisper large models (released at different times), the v2 model shows the best performance compared to the other two versions. Whisper large-v3 is trained on additional data (5M vs 680k hours) in the form of 320k hours of weakly and 4M pseudo-labeled training data. We believe the latter degrades performance here.

Model	Size	R@1 [%] en-fr	en-de	de-fr
tiny	39M	9.2	9.9	6.8
base	74M	16.7	16.0	11.0
small	244M	27.7	26.1	20.2
medium	769M	50.7	41.8	39.7
large-v1	1550M	59.9	51.6	48.8
large-v2	1550M	**80.0**	**80.0**	**62.2**
large-v3	1550M	59.9	50.5	47.2

Table 6: Ablation of R@1 against different model sizes.

In addition to FLEURS, we run speech-to-speech retrieval experiments on MaSS to validate the effectiveness of the aligned speech embedding space. Retrieval performance is presented in Table 7 across paired datasets in seven languages. The baseline for random selection is 0.1% in this setting. The supervised baseline is taken from (Boito et al., 2020) who built a system based on contrastive learning (Harwath et al., 2018). Excluding the low-resource language Basque (eu), the proposed zero-shot retrieval method outperforms the baseline and shows an average R@1 of 75.3%. Although Whisper is only trained using utterances in different languages translated to English, it demonstrates good retrieval performance between arbitrary language pairs, which can be seen as an emergent ability.

C Ablation of Speech Translation

Ablation results are shown in Table 8. For *FT (all)*, we fine-tune all the parameters of Whisper. For *LoRA (dec)*, trainable LoRA parameters with a rank of 8 are inserted in the decoder and updated on the training set. In both settings, performance in all languages improved compared to the zero-shot results

Query	R@1 [%] en	es	ru	ro	fr	fi	eu
en	-	79.5	66.8	71.7	86.6	64.1	7.6
es	71.9	-	62.7	83.4	87.5	62.9	13.4
ru	67.8	72.4	-	83.4	70.4	72.0	5.5
ro	65.5	84.8	79.1	-	85.1	69.0	9.7
fr	83.0	91.3	67.0	89.8	-	66.2	6.9
fi	70.1	74.2	77.4	81.6	71.7	-	11.2
eu	14.6	25.7	6.5	14.6	11.3	9.6	-

Table 7: Zero-shot speech-to-speech retrieval results on 42 language pairs measured with SeqSim on MaSS.

Dataset	src	BLEU / COMET FT (dec)	FT (all)	LoRA (dec)
FLEURS	en	29.1 / **78.4**	**29.3** / 77.8	23.3 / 73.1
	fr	**23.0 / 74.1**	21.5 / 72.3	19.5 / 69.3
	de	**24.0 / 74.7**	23.3 / 72.8	20.1 / 70.2
	ja	**19.2 / 74.7**	17.7 / 72.6	16.8 / 72.3
CoVoST	en	**31.9 / 76.3**	31.2 / 75.8	26.3 / 72.9

Table 8: Ablation of zero-shot cross-lingual transfer.

in Table 2, highlighting Whisper's effective cross-lingual transfer capability. LoRA shows worse performance compared to fine-tuning while being more parameter efficient. Moreover, compared to only fine-tuning the decoder part, fine-tuning all parameters shows similar performance on the English test set. Since the encoder parameters are changed in the adaptation, there is a shift in the speech embedding space, leading to a performance drop in languages not seen in the training. This suggests that only adapting the decoder parameters is a better strategy when extending Whisper's speech translation ability.

src	code	WER	ST (zh)
kea	pt	89.5	19.5
ast	es	47.8	18.7

Table 9: ASR and ST (BLEU score) into Chinese results on FLEURS data Kabuverdianu (kea) and Asturian (ast), with Whisper language code specified.

In Section 4.4, we showed that the audio embeddings for some previously unseen languages (e.g. kea and ast) align well in the shared semantic space, and these languages achieve good BLEU scores when translated into English using the baseline Whisper large-v2 model, as shown in Table 3. Table 9 demonstrates that these languages also achieve reasonable BLEU scores for Chinese translation with the fine-tuned model from Section 4.3 despite the high WERs.