

Whisper-Style Speech Encoders Align Semantically Beyond Phonetic Cues

Ryan Soh-Eun Shim^{1,2}, Domenico De Cristofaro³, Martin Chengzhi Hu⁴,
Alessandro Vietti³, Barbara Plank^{1,2}

¹LMU Munich, ²Munich Center for Machine Learning,

³Free University of Bozen-Bolzano, ⁴Technical University of Munich

Correspondence: s.shim@lmu.de

Abstract

Recent work suggests that Whisper-style speech encoders exhibit semantic alignment across languages, based on spoken translation retrieval using representational similarity. If so, this would enable Whisper-style encoders as a viable way to obtain acoustic representations with semantic knowledge, enabling downstream semantic tasks with speech input. However, prior work does not control for phonetic overlap between equivalent utterances, which may artificially support retrieval. We therefore conduct pronunciation-controlled experiments to test whether cross-lingual alignment arises from semantic rather than phonetic similarity, and further probe semantic information in the encoder by way of lexical semantic analogy tasks. Results show that spoken translation retrieval remains strongly above chance with substantially reduced phonetic cues in the final layers of Whisper’s encoder. Furthermore, for both word-level and utterance-level input, Whisper’s encoder representations exhibit non-trivial analogy performance, attaining as high as 25.98% accuracy in an analogy test set. Motivated by this layerwise semantic specialization, we test whether earlier encoder layers preserve representations better suited to ASR in unsupported low-resource languages. Our results confirm it to enable performance gains in automatic speech recognition on low-resource languages unseen during training.

1 Introduction

In speech, a growing body of work has shown speech foundation models to exhibit emergent multilingual capabilities (Peng et al., 2023; Yang et al., 2024), implying the existence of cross-lingual alignment in such models. Prior work probes such alignment through spoken translation retrieval (Abdullah et al., 2024; Ma et al., 2025), where Ma et al. (2025) find Whisper’s encoder to have an accuracy of up to 80% on the FLEURS dataset for language pairs such as English and French. If such retrieval

English:	You can use a boda-boda to get around Goma . The normal price is ~ 500 Congolese Francs for a short ride.
Italian:	È possibile usare un boda-boda per muoversi a Goma . Il prezzo normale è di circa 500 franchi congolesi per una corsa breve.
Chinese:	你可以乘坐 boda-boda 游览戈马(gema)。短途车程的正常价格是500 刚果法郎(gangguo falang)。

Table 1: Example parallel utterance from FLEURS. Highlighted items share pronunciation similarities that spoken translation retrieval could rely on instead of semantic cues. Colors indicate phonetically similar items.

is driven by semantic similarity, this would suggest that Whisper-style encoders provide a viable way to obtain acoustic representations with semantic knowledge, enabling downstream semantic tasks directly from speech.

However, a possible confounding factor in using speech retrieval as a proxy for measuring cross-lingual alignment is the existence of *pronunciation-level* cues such as cognates, loanwords, numbers, and proper nouns, which could serve as shortcuts even in the absence of a shared semantic space for retrieving semantically equivalent utterances in a different language (Table 1). As such, in this paper we ask: *to what degree does spoken translation retrieval actually rely on semantic features?* To answer this question, we explicitly control for such pronunciation-level cues by constructing a challenge set with substantially reduced pronunciation overlap between typologically distant languages, with the goal of understanding whether and to what extent such shortcuts impact cross-lingual alignment in speech. In addition, to observe whether such an alignment is likely to be induced by the speech translation objective or whether it can arise through the next-token prediction objective in ASR alone, we perform a controlled study on compa-

rable Whisper-style speech encoders trained with and without the speech translation objective. We find that the latter exhibits weak but non-trivial spoken translation retrieval capabilities, while the model trained with a speech translation objective shows substantially stronger retrieval in the final layers, suggesting speech translation to be a potential driver in cross-lingual alignment of such models.

We further ask whether the semantic structure observed at the utterance level is also reflected in word-level acoustic representations. To this end, we construct acoustic word embeddings by averaging Whisper encoder representations over word occurrences, and evaluate them on lexical semantic analogy tasks. We additionally compare these contextualized acoustic word embeddings against a single-word input control, finding non-trivial performance across all setups. Finally, leveraging our findings above, we perform early exiting experiments in the encoder layers to induce representations less tied to language-specific semantics, yielding performance gains for speech recognition on low-resource languages unseen during training. The contributions of our paper are as follows:

Cross-lingual speech retrieval is possible after controlling for pronunciation cues. We propose a challenge set where we aim to mitigate *pronunciation*-level cues, showing cross-lingual retrieval accuracy can remain strong particularly in the final layers, despite noticeable performance drops.¹ In addition, we show such retrieval capabilities in Whisper-style speech encoders to potentially be driven the existence of a speech translation objective, rather than arising from the next-token prediction objective in multilingual ASR alone.

Whisper encoder representations support acoustic word embeddings with lexical structure. To further confirm semantic structure in the Whisper encoder, we show that averaged encoder representations over word occurrences exhibit non-trivial analogy performance.

Early exiting the encoder generalizes better to related languages for ASR. Leveraging our insights, we early exit the ASR encoder to induce representations less tied to

language-specific semantics. Applying such an early exiting approach to ASR on seven low-resource languages, we observe consistent improvements in word and character error rates over using the final encoder layer on 5 out of 7 low-resource languages examined.

2 Related Work

In this section, we review prior work that analyze cross-lingual alignment in audio representations, and early exiting of neural models.

2.1 Cross-Lingual Alignment in Speech

Cross-lingual alignment is well established in multilingual text encoders such as mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019). For speech models, however, it remains unclear to what extent semantic alignment emerges, especially when training is dominated by multilingual ASR, where next-token prediction is tightly coupled to the input signal. Recent work on probing speech representations suggests that self-supervised models are predominantly phonetic (Choi et al., 2024), though such analyses often rely on short word-level inputs. In contrast, models trained with speech translation objectives show strong spoken translation retrieval performance (Abdullah et al., 2024; Ma et al., 2025), suggesting the emergence of a shared semantic subspace. Further, Ògúnremí et al. (2025) observe that Whisper-style encoders exhibit an English-centered interlingua, hypothesized to arise from the speech translation objective. However, retrieval-based evaluations may be inflated by pronunciation-level cues such as cognates and proper nouns, and it remains unclear whether semantic alignment is primarily driven by speech translation supervision or can emerge from multilingual ASR alone. We address both questions by (i) constructing a pronunciation-controlled challenge set across typologically distant languages, and (ii) comparing Whisper-style encoders trained with and without speech translation objectives.

2.2 Acoustic Word Embeddings

The idea of learning word embeddings from speech is initially explored in Chung and Glass (2017) and Chung and Glass (2018), where an RNN-based encoder-decoder is used to learn fixed-length vector representations of audio segments from a speech corpus, where such embeddings are argued to have access to alternative sources of information that do

¹<https://github.com/mainlp/pronunciation-challenge-set>

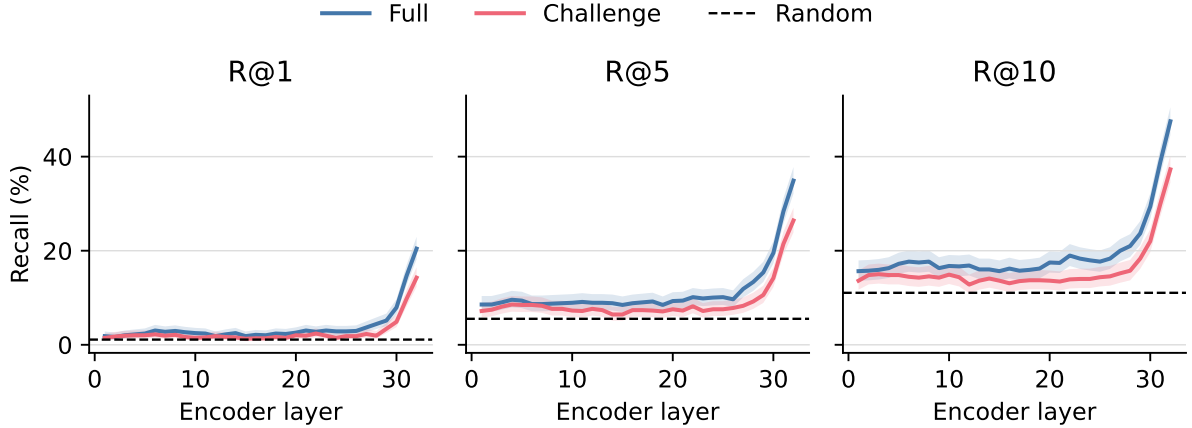


Figure 1: Spoken translation retrieval results in Whisper-large-v2. Plot shows Recall@1 ($R@1$), $R@5$, and $R@10$, micro-averaged across language pairs. Shaded regions indicate 95% Wilson confidence intervals. Dashed line is a random baseline. We observe that even after filtering out potential pronunciation shortcuts, semantic-based retrieval remains strongly above chance (dashed baseline) towards the later layers for all $R@K$ values.

not exist in text. Such fixed-length word representations are in practice challenging to construct, due to factors such as speaker variation, phonetic variation, requirement of word boundaries (Porjazovski et al., 2024). Our study extends upon this line of work by showing acoustic word embeddings can be extracted from Whisper’s encoder representations.

2.3 Early Exiting

Prior work has shown that representations at earlier layers of a neural network are often already adequate for making a correct prediction (Kaya et al., 2019), which allows for early exiting strategies that minimize compute (Schuster et al., 2022) and mitigate the influence of false demonstrations (Halawi et al., 2024). Recent work in mechanistic interpretability (Rai et al., 2025) extend this line of research by using early exiting as a way to understand model internals. For instance, the logit lens method (nostalgebraist, 2020) interprets the incremental layer updates of large language models by projecting intermediate hidden states into the vocabulary space. However, the work above has mainly been applied to decoder-based text models. (Langedijk et al., 2024) adapt the method to encoder-decoder architectures and also apply it to Whisper. However, their work aims to identify at what encoder layer stable predictions emerge, and focuses on the tasks Whisper was explicitly trained on. We extend this line of work and apply early exiting to strengthen model generalization, providing concrete applications as to how it can aid low-resource speech recognition.

3 Experimental Setup

In this section, we describe the methods we employ, along with our dataset construction procedure.

Langs	Original pool	Challenge	Sampled original
eng-zho	427	113	113
fra-zho	427	109	109
deu-zho	427	101	101
eng-jpn	427	80	80
fra-jpn	427	71	71
deu-jpn	427	69	69

Table 2: FLEURS dataset statistics. The sampled original condition is drawn without replacement from the original pool to match the challenge-set size within each language pair.

3.1 Model

In our experiments, we follow Ma et al. (2025) in examining Whisper (Radford et al., 2023). Whisper is an encoder-decoder Transformer model trained on the tasks of language identification, speech recognition, and $X \rightarrow$ English speech translation. It is trained on 680,000 hours of data, where 10% is for the task of $X \rightarrow$ English translation. Audio is first converted to a mel spectrogram, then passed through convolutional layers to extract features that are then passed to the Transformer’s blocks. In this paper, we use the whisper-large-v2 model for our main experiments, which has been shown to exhibit stronger cross-lingual alignment than later variants (Ma et al., 2025). In addition, to observe the effect of the speech translation objective on encoder representations, we employ Open Whisper-style Speech Models (OWSM) (Peng et al., 2024),

which provide a series of open-source, Whisper-style encoder-decoder models. OWSMs provide an ideal testing ground for this hypothesis due to the open source nature of its training data; their adherence to Whisper-style training specifications; and they provide models of various sizes and setups. OWSM includes two models of comparable size and architecture, OWSM v3.1 Small and Small Low-Restriction, of which the latter is not trained on speech translation. As such, we repeat our spoken translation retrieval experiments on the two models to gain insights on the impact of the speech translation objective.

3.2 Datasets

For our speech retrieval experiments, we follow Ma et al. (2025) in employing FLEURS (Conneau et al., 2023), a multilingual parallel speech dataset. They build their test set by combining the dev and test sets of English, French, German, Mandarin and Japanese, where only utterances common to all five languages are kept. This is followed by a deduplication step to remove utterances with the same content spoken by different speakers. We build our challenge set on top of this procedure.

We employ the spaCy (Honnibal et al., 2020) pipeline to filter out utterances containing proper nouns (such as *Goma* in Table 1). For pairs with Japanese, we further remove samples containing katakana script, which are largely reserved for loanwords in Japanese. We pair together typologically and phylogenetically distant languages to avoid the influence of cognates between related varieties, resulting in six language pairs (Table 2). Proficient in-house speakers of the five languages then manually inspect the resulting dataset to confirm their quality, which further removes utterances with similar sounding measurement units (e.g. GHz). As cross-lingual retrieval is direction-dependent, we perform our experiments on both src-trg and trg-src retrieval directions. To keep language pair proportions comparable, we sample the full set to contain the same language pair proportions as the challenge set.

To verify that our filtering procedure reduces pronunciation-level shortcuts, we measure local phonetic similarity directly from the audio. We first obtain IPA phone sequences using a state-of-the-art phone recognition model (Bharadwaj et al., 2026), and represent each phone using PanPhon articulatory features (Mortensen et al., 2016). For each utterance pair, we search for the most similar

Window	Original Δ	Challenge Δ	Filter Δ
4 phones	-0.35	0.00	+0.35
95% CI	[-0.47, -0.24]	[-0.11, 0.11]	[0.19, 0.51]
5 phones	-0.43	-0.04	+0.39
95% CI	[-0.57, -0.30]	[-0.17, 0.07]	[0.21, 0.57]
6 phones	-0.41	0.02	+0.44
95% CI	[-0.57, -0.27]	[-0.10, 0.14]	[0.25, 0.62]

Table 3: Local phonetic similarity before and after challenge-set filtering, measured using PanPhon features (values multiplied by 100). Original and Challenge report the difference (Δ) between translation pairs and length-matched random pairs; negative values indicate that translations are more phonetically similar than expected by chance. Filter Δ measures the change from Original to Challenge; positive values indicate that filtering reduces this excess phonetic similarity.

short pronunciation segment (4–6 phones) across the two languages. We then compare this similarity against randomly paired utterances from the same language pair. Because longer utterances have more opportunities for accidental matches, random pairs are matched approximately by utterance length. We repeat this randomization 1,000 times and compute 95% bootstrap confidence intervals. Finally, we apply the same procedure to size-matched samples from the original FLEURS data, allowing us to measure how much phonetic similarity is removed by our filtering procedure.

As shown in Table 3, original aligned pairs exhibit lower local distances than their length-controlled random baselines across all window sizes. Challenge-aligned and random pairs, in contrast, obtain nearly identical distances. Accordingly, the filtering effect is consistently positive, with confidence intervals excluding zero for 4-, 5-, and 6-phone windows. This indicates that the challenge set substantially reduces the short pronunciation matches expected from proper names, loanwords, and similar cues. Effects are strongest for Japanese-target pairs, while Chinese-target pair estimates are less conclusive due to phone recognition models not modeling tone, thereby losing information in our ngram estimates. These results establish a reduction in local pronunciation cues in our challenge set.

For our early exiting experiments, we use the FLEURS test sets of the same 7 low-resource languages examined in Ma et al. (2025), which include Cebuano (ceb), Irish Gaelic (ga), Javanese (jv), Asturian (ast), Kabuverdianu (kea), Sorani Kurdish (ckb), and Kyrgyz (ky).

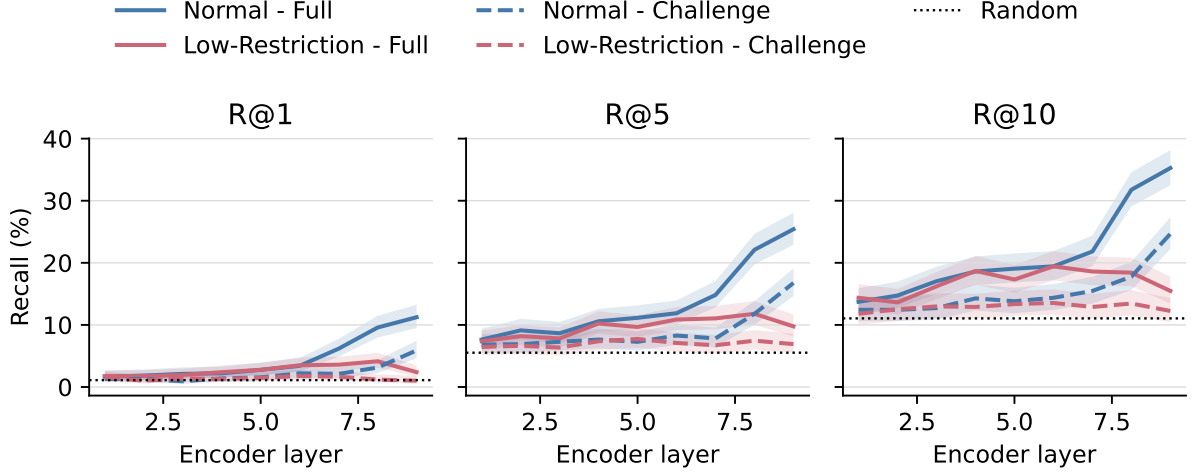


Figure 2: Spoken translation retrieval results in models with and without (LowRestriction) speech translation objective. Plots show Recall@1 ($R@1$), $R@5$, and $R@10$, micro-averaged across language pairs. Shaded regions indicate 95% Wilson confidence intervals. Dashed line is a random baseline. The model with an additional speech translation objective (Normal) shows stronger retrieval accuracy than the model without such an objective in the final layer, although the latter model (LowRestriction) also exhibits non-trivial retrieval accuracy at $R@10$.

4 Methodology

4.1 Cross-Lingual Speech Retrieval

To quantify cross-lingual alignment in the encoder of Whisper-style models, we follow prior work in employing translation retrieval as a proxy task. Ma et al. (2025) propose SeqSim to quantify the similarity between two sequences of audio embeddings $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ by measuring how well each frame in one sequence aligns with the most similar frame in the other, where the process is repeated in both directions. The benefit of this measure is that it is simple and allows frames in two audio embeddings to match one another regardless of their position, where mean pooling frame representations in contrast risks cramming too much information due to the large number of frames. Formally, SeqSim is defined as:

$$\begin{aligned} \text{Re}_{\text{seq}}(X, Y) &= \frac{1}{|X|} \sum_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} \mathbf{x}^\top \mathbf{y}, \\ \text{Pr}_{\text{seq}}(X, Y) &= \frac{1}{|Y|} \sum_{\mathbf{y} \in Y} \max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{y}, \\ \text{SeqSim}(X, Y) &= 2 \cdot \frac{\text{Pr}_{\text{seq}} \cdot \text{Re}_{\text{seq}}}{\text{Pr}_{\text{seq}} + \text{Re}_{\text{seq}}}. \end{aligned} \quad (1)$$

Ma et al. (2025) show SeqSim outperforms mean pooling and dynamic time warping for spoken translation retrieval. We thus leverage this metric in our work.

4.2 Word-Level Analogy Structure in Whisper Encoder Representations

To test whether Whisper encoder representations encode word-level relational structure, we construct acoustic word embeddings from LibriSpeech train-clean-100 (Panayotov et al., 2015) and evaluate them on the word analogy dataset employed in Mikolov et al. (2013), using 3CosMul retrieval (Levy and Goldberg, 2014). For the acoustic word embeddings, we follow prior work in leveraging Whisper’s cross-attention for forced alignment to obtain word-level timestamps (Yeh et al., 2025), which has been shown to perform roughly on par with forced alignment models, with the additional benefit of leveraging the model’s internal representations. As such, we first obtain word-level timestamps using Whisper’s teacher-forced decoder cross-attention, where for each utterance we run Whisper with the ground-truth transcription, aggregate cross-attention weights, and apply dynamic time warping to align text tokens to encoder frames. Given the resulting word boundaries, we pool encoder hidden states over the frames corresponding to each word span to obtain word representations. We use mean pooling over frames within each word boundary, and aggregate occurrences of the same word type over the whole dataset by L2-normalizing each occurrence vector and averaging across occurrences. Word types observed fewer than five times are discarded, yielding a vo-

Category	Acc.	Correct/Total	OOV
Overall	25.98	1243/4785	–
Semantic	44.10	142/322	–
Syntactic	24.67	1101/4463	–
<i>Semantic categories</i>			
capital-common	35.00	7/20	486
capital-world	16.67	4/24	4500
city-in-state	0.00	0/6	2461
currency	–	0/0	866
family	48.16	131/272	234
<i>Syntactic categories</i>			
adj.-to-adv.	14.77	96/650	342
opposite	24.44	22/90	722
comparative	20.22	188/930	402
superlative	17.92	43/240	882
present-participle	41.77	193/462	594
nationality-adj.	15.24	16/105	1494
past-tense	38.61	383/992	568
plural	13.67	111/812	520
plural-verbs	26.92	49/182	688

Table 4: R@5 word analogy accuracy (%) for Whisper-large-v2 encoder acoustic word embeddings using 3CosMul. The vocabulary contains 11,734 word types. Embeddings are L2-normalized. OOV counts indicate questions excluded because at least one word is absent from the vocabulary.

cabulary of 11,734 word types.

To confirm that the analogy structure is not solely an artifact of averaging over contextualized utterances, we additionally run an input-level control with the same Whisper-medium model. In the contextual condition, the encoder processes the full utterance, and word vectors are obtained by pooling frames of a given word from the contextualized utterance-level encoder states. In the input-level condition, we use the same word boundaries, but cut each word segment from the waveform and feed the isolated clip through the encoder on its own. Thus, the audio segment and timestamp source are held fixed, while surrounding-word context is removed from the encoder representation. Since this substantially increases the number of encoder forward passes, we restrict the control to Whisper-medium.

4.3 DecoderLens

In this work, we follow [nostalgebraist \(2020\)](#) in viewing the layers of a transformer as performing incremental updates to latent predictions of the next token. This assumption implies that the hidden states can be decoded to gain insights as to how the input is being processed in said layer. This method has gained widespread usage in text-based

decoder-only language models, where it is used to illustrate what the model captures in terms of vocabulary space at each internal layer ([Belrose et al., 2023](#); [Yom Din et al., 2024](#)). [Langedijk et al. \(2024\)](#) extend this method to encoder-decoder models, where the decoder is used as a probe over intermediate encoder representations. Concretely, let $\mathcal{M}$ be an encoder-decoder model with n encoder layers. At inference time, the model typically produces an output by passing the final encoder layer’s representation to the decoder (after a non-linear transformation such as layer normalization), i.e.,

$$\mathcal{M}(\mathbf{w}) = \text{Dec}(f(\text{Enc}(\mathbf{w})_n)), \quad (2)$$

where $\text{Enc}(\mathbf{w})_n$ denotes the hidden states at the final encoder layer and f captures the non-linear operation applied prior to decoding. DecoderLens modifies this procedure by replacing the final-layer encoder representation with that of an intermediate layer i . Formally, for encoder layer $i \in \{1, \dots, n\}$, DecoderLens computes:

$$\text{DecoderLens}(\mathbf{w}, i) = \text{Dec}(f(\text{Enc}(\mathbf{w})_i)). \quad (3)$$

Building on the findings that the final encoder layer’s representations are more semantically aligned, we early exit the encoder to induce representations less tied to language-specific semantics, which we hypothesize to generalize better to speech recognition on low-resource languages by weakening such target-language specialization. For our early exiting experiments on the low-resource languages examined in [Ma et al. \(2025\)](#), we select the best performing layer based on WER and CER on the dev set of FLEURS, and use that layer to perform inference on the FLEURS test sets of the respective language. As our goal is only to reduce the influence of language-specific semantic information, we perform this experiment on the 9 encoder layers of 24 to 32.

5 Results

In this section, we detail our results on spoken translation retrieval, acoustic word embedding extraction, and early exiting the encoder.

5.1 Cross-lingual speech retrieval remains possible after accounting for pronunciation cues.

First, we compare the results on the full data to our challenge set with the Whisper encoder. [Figure 1](#) shows our spoken translation retrieval results

Last: Ubelak lorok bubuk garing **itu** bebarengan **dengan** **banjur**. Kandai tangan **telah** sengerasik, remes **itu** **sudah** tebal.
Best: Uplik lorok bubuk garing **iku** bebarengan **lant** **banjur**. Kanti tangan **telas** yang resik, remes **iku** **dati** bal.
Gold: ublek loro bubuk garing **iku** bebarengan **lan** **banjur** kanthi tangan **teles** sing resik remes **iku** **dadi** bal

Table 5: Example prediction of early exiting Whisper-large-v2 on Javanese, a language not explicitly supported by the model. Colored words indicate comparability. In comparison against the reference, standard inference (layer 32) heavily uses Indonesian words, whereas early exiting (layer 29) results in transcriptions more phonetically faithful to Javanese, resulting also in lower error rates.

Category	Contextual	Single-word
Overall	17.95	9.78
Semantic	38.51	16.15
Syntactic	16.47	9.32
<i>Semantic categories</i>		
capital-common	15.00	15.00
capital-world	4.17	12.50
city-in-state	16.67	0.00
currency	—	—
family	43.75	16.91
<i>Syntactic categories</i>		
adj.-to-adv.	10.00	6.92
opposite	12.22	11.11
comparative	11.72	10.97
superlative	10.83	6.67
present-participle	28.79	19.91
nationality-adj.	11.43	5.71
past-tense	28.63	8.97
plural	9.11	4.19
plural-verbs	11.54	12.09

Table 6: R@5 analogy accuracy (%) for Whisper-medium acoustic word embeddings under contextual and single-word input conditions. Embeddings are L2-normalized.

in Recall@K across different language pairs using the Whisper-large-v2 encoder. For all R@K setups, performance is significantly above the random baseline (dashed black line in Figure 1) in the final layers. This indicates that although phonetic cues are indeed useful for retrieval, the models examined are able to rely more on semantic cues. We note however that as the metric employed does not take into account word order, the retrieval results primarily establish cross-lingual alignment on a lexical level; to what extent syntactic level information is encoded in the encoder remains undetermined, which we leave to future work.

5.2 Spoken translation retrieval in Whisper-style models is potentially driven by the speech translation objective.

Figure 2 shows our experiment in comparing OWSM Small and OWSM Small Low-restriction, where the latter differs from the former in lacking

Query	Nearest neighbors
<i>king</i>	<i>queen</i> (.986), <i>men</i> (.971), <i>man</i> (.969), <i>wind</i> (.969), <i>bill</i> (.969)
<i>man</i>	<i>men</i> (.994), <i>woman</i> (.989), <i>women</i> (.983), <i>mind</i> (.982), <i>girl</i> (.980)
<i>woman</i>	<i>women</i> (.993), <i>men</i> (.989), <i>man</i> (.989), <i>girl</i> (.981), <i>children</i> (.981)
<i>paris</i>	<i>france</i> (.976), <i>forest</i> (.973), <i>palace</i> (.973), <i>house</i> (.971), <i>darkness</i> (.970)
<i>france</i>	<i>forest</i> (.979), <i>earth</i> (.977), <i>house</i> (.977), <i>paris</i> (.976), <i>ours</i> (.976)
<i>good</i>	<i>great</i> (.980), <i>real</i> (.977), <i>fine</i> (.976), <i>little</i> (.974), <i>bad</i> (.974)
<i>bad</i>	<i>good</i> (.974), <i>sad</i> (.974), <i>high</i> (.972), <i>hot</i> (.972), <i>quiet</i> (.971)
<i>run</i>	<i>running</i> (.977), <i>gone</i> (.975), <i>ran</i> (.973), <i>come</i> (.971), <i>turn</i> (.971)

Table 7: Example nearest neighbors in the Whisper-large-v2 encoder acoustic word embedding space. Parentheses show cosine similarity.

the speech translation objective and speech translation data, while employing the same architecture, training setup, parameter count, and a subset of the ASR training data (Peng et al., 2024). We observe that on both the full and the challenge set, the model not trained on speech translation (OWSM Small Low-restriction) exhibits weaker retrieval capabilities in comparison with the model with such an objective (OWSM Small). For OWSM Small, the gap is particularly pronounced for the final layers, where we observe stronger performance under all retrieval settings. Although differences between the available checkpoints prevent a clean causal attribution, the results here are consistent with the speech translation objective contributing semantic cross-lingual alignment capabilities in speech models, which nevertheless can still arise but to a much lower degree when trained only on multilingual ASR.

5.3 Whisper-style speech encoders contain non-trivial lexical semantic knowledge.

As shown in Table 4, the Whisper-large-v2 encoder achieves a R@5 analogy accuracy of 25.98%.

Lang	CER (%)			WER (%)		
	Best	L32	Δ	Best	L32	Δ
ky	100.5 (24)	138.2	37.7	106.6 (24)	194.3	87.7
ga	79.2 (31)	85.5	6.3	101.3 (24)	121.8	20.5
jv	36.1 (29)	41.3	5.2	87.3 (29)	90.8	3.5
ceb	17.6 (29)	19.2	1.7	52.9 (29)	54.4	1.5
kea	34.4 (29)	35.1	0.8	91.6 (29)	92.6	1.0
ckb	57.2 (32)	57.2	0.0	102.6 (25)	122.4	19.8
ast	16.4 (32)	16.4	0.0	59.6 (29)	58.7	-0.9

Table 8: Early exiting results on FLEURS test set of low-resource languages. **Best** indicates performance and best layer selected (in parenthesis) on the dev set; **L32** is the final encoder layer performance. Δ shows performance improvement.

Performance is higher on semantic analogies than on grammatical/syntactic analogies, where we obtain 44.10% for semantic categories versus 24.67% for syntactic categories. The per-category results show particularly strong performance for family relations (48.16%), present participles (41.77%), and past tense (38.61%). Results for geography-related categories are less stable because only a small number of questions remain after excluding out-of-vocabulary items. These results suggest that Whisper encoder spaces preserve non-trivial lexical regularities, even though the model is trained on speech recognition and translation rather than explicit word-level text objectives. This further supports our overall hypothesis that Whisper-style speech encoders align more semantically. Nearest-neighbor inspection further supports this interpretation (Table 7). For example, the representation for *king* retrieves semantically related words such as *queen* and *man* among its nearest neighbors. As shown in Table 6, embeddings extracted from utterance-level input outperform embeddings obtained by presenting words in isolation: R@5 analogy accuracy decreases from 17.95% to 9.78%. The largest difference appears for semantic analogies, which decrease from 38.51% to 16.15%, while syntactic analogy accuracy decreases from 16.47% to 9.32%. This indicates that contextual utterance-level encoding contributes substantially to the relational structure observed in the acoustic word embedding space. Nevertheless, isolated-word representations retain non-trivial analogy performance.

5.4 Encoder representations in earlier layers generalize better to low-resource languages.

Finally, we evaluate early exiting, to measure to what degree our findings may impact automatic speech recognition quality by leveraging representations in pre-final layers. Table 8 shows our results in using encoder representations in earlier layers (given in parenthesis) to perform speech recognition on low-resource languages (compared to the last layer 32). We observe that interestingly 5 out of 7 languages examined show improved CER and WER using some earlier encoder layer’s representation, implying that earlier encoder layers may generalize better to low-resource languages not explicitly supported by Whisper. Table 5 provides an example of our early exited prediction on Javanese, where we observe the final layer to produce transcriptions with significantly more Indonesian words than the early exited transcriptions, which are more phonetically faithful to the Javanese gold transcriptions.

6 Conclusion

In this work, we revisit the question of cross-lingual alignment in Whisper-style speech foundation models. Through a series of controlled experiments, we demonstrate that Whisper-style speech encoders align semantically beyond phonetic clues, as tested through our challenge dataset, analogy tests and early exiting in decoding. We demonstrate that spoken translation retrieval is potentially driven by the additional speech translation pre-training objective, although it also arises weakly through speech recognition as an objective alone, which underscores the role of supervised multilingual multi-task signals in pre-training to aid shaping useful semantic representations. Complementing the cross-lingual retrieval results, we further find Whisper’s encoder to encode lexical semantic structure, obtaining non-trivial word analogy performance. Finally, by early exiting the encoder to induce representations less tied to language-specific semantics, we observe that using earlier encoder layer representations results in performance gains in speech recognition on low-resource languages, demonstrating the downstream applicability of our findings. As Whisper-style phone recognition models have been shown to perform weaker on sociolinguistic variation and unseen languages (Li et al., 2026), future work can look into whether our early-exiting findings extend

also to such phone recognition models.

Generative AI Use Disclosure

The authors acknowledge the usage of ChatGPT as an assistant tool in part of the source code’s development and in enhancing the coherence of parts of the manuscript.

Limitations

A limitation of our work is our focuses on Whisper and Open Whisper-style Speech Models, but it may be interesting compare whether our findings hold also for speech foundation models of different architectures, such as SeamlessM4T ([Communication et al., 2023](#)). Furthermore, due to the limited availability of models fully comparable in parameter count and training data that are trained with and without speech translation, our results for the effect of speech translation on cross-lingual alignment remains suggestive, instead of point to a causal effect. Future work should establish more comparable training conditions to determine if there is a causal relation between speech translation and cross-lingual alignment. In addition, although our retrieval results establish cross-lingual alignment on a lexical level, the metric we employ for retrieval does not establish utterance-level semantic alignment.

References

- Badr M. Abdullah, Mohammed Maqsood Shaik, and Dietrich Klakow. 2024. [Wave to interlingua: Analyzing representations of multilingual speech transformers for spoken language translation](#). In *Interspeech 2024*, pages 362–366.
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#). *Preprint*, arXiv:2303.08112.
- Shikhar Bharadwaj, Chin-Jou Li, Kwanghee Choi, Eunjung Yeo, William Chen, Shinji Watanabe, and David R. Mortensen. 2026. [An empirical recipe for universal phone recognition](#). *Preprint*, arXiv:2603.29042.
- Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. [Self-supervised speech representations are more phonetic than semantic](#). In *Interspeech 2024*, pages 4578–4582.
- Yu-An Chung and James Glass. 2017. Learning word embeddings from speech. *arXiv preprint arXiv:1711.01515*.
- Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. [Seamlessm4t: Massively multilingual & multimodal machine translation](#). *Preprint*, arXiv:2308.11596.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.
- 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 *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *The Thirteenth International Conference on Learning Representations*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2019. Shallow-deep networks: Understanding and mitigating network overthinking. In *International conference on machine learning*, pages 3301–3310. PMLR.
- Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. [Decoder-Lens: Layerwise interpretation of encoder-decoder transformers](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4764–4780, Mexico City, Mexico. Association for Computational Linguistics.
- Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of CoNLL*, pages 171–180.

- Chin-Jou Li, Calvin Chang, Shikhar Bharadwaj, Eunjung Yeo, Kwanghee Choi, Jian Zhu, David Mortensen, and Shinji Watanabe. 2026. [Powsm: A phonetic open whisper-style speech foundation model](#). *Preprint*, arXiv:2510.24992.
- Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2025. [Cross-lingual transfer learning for speech translation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 33–43, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. [Linguistic regularities in continuous space word representations](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. [Pan-Phon: A resource for mapping IPA segments to articulatory feature vectors](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.
- nostalgebraist. 2020. [Interpreting gpt: The logit lens](#). <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2025-05-19.
- Tolúlopé Ògúnremí, Christopher D Manning, Dan Jurafsky, and Karen Livescu. 2025. Transcribe, translate, or transliterate: An investigation of intermediate representations in spoken language models. *Proceedings of IEEE ASRU 2025*.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. [Librispeech: An asr corpus based on public domain audio books](#). In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- 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 *Interspeech 2023*, pages 396–400.
- Yifan Peng, Jinchuan Tian, William Chen, Sidhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. [Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer](#). *Preprint*, arXiv:2401.16658.
- Dejan Porjazovski, Tamás Grósz, and Mikko Kurimo. 2024. [From raw speech to fixed representations: A comprehensive evaluation of speech embedding techniques](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3546–3560.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. [A practical review of mechanistic interpretability for transformer-based language models](#). *Preprint*, arXiv:2407.02646.
- Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.
- Chih-Kai Yang, Kuan-Po Huang, Ke-Han Lu, Chun-Yi Kuan, Chi-Yuan Hsiao, and Hung-Yi Lee. 2024. [Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision](#). In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 540–544.
- Sung-Lin Yeh, Yen Meng, and Hao Tang. 2025. [Whisper has an internal word aligner](#). *Preprint*, arXiv:2509.09987.
- Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. [Jump to conclusions: Short-cutting transformers with linear transformations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9615–9625, Torino, Italia. ELRA and ICCL.