VoxKrikri: Unifying Speech and Language through Continuous Fusion

Dimitrios Damianos[1], *Leon Voukoutis*[1], *Georgios Paraskevopoulos*[1], *Vassilis Katsouros*[1]

[1]Institute for Speech and Language Processing, Athena Research Center, Greece

{d.damianos, leon.voukoutis, vsk, g.paraskevopoulos}@athenarc.gr

Abstract

We present a multimodal fusion framework that bridges pre-trained decoder-based large language models (LLM) and acoustic encoder-decoder architectures such as Whisper, with the aim of building speech-enabled LLMs. Instead of directly using audio embeddings, we explore an intermediate audio-conditioned text space as a more effective mechanism for alignment. Our method operates fully in continuous text representation spaces, fusing Whisper's hidden decoder states with those of an LLM through cross-modal attention, and supports both offline and streaming modes. We introduce *VoxKrikri*, the first Greek speech LLM, and show through analysis that our approach effectively aligns representations across modalities. These results highlight continuous space fusion as a promising path for multilingual and low-resource speech LLMs, while achieving state-of-the-art results for Automatic Speech Recognition in Greek, providing an average $\sim 20\%$ relative improvement across benchmarks.

Index Terms: Speech LLMs, modality fusion, continuous latent space, causal masking, ASR

1. Introduction

Large language models (LLMs) have achieved remarkable success in natural language processing, inspiring the development of multimodal LLMs that combine pre-trained language models with modality-specific encoders such as those for audio and images [1, 2, 3, 4, 5]. This design departs from traditional end-to-end architectures [6, 7] by leveraging the strengths of large pre-trained models.

Recent work in the speech domain can be broadly divided into approaches that use continuous speech representations and those that rely on discrete features. Continuous representations offer an intuitive and simple way to integrate speech signals into large language models (LLM). For example, in SLAM-ASR [8], continuous features are extracted from a speech decoder, downsampled, and then projected into the LLM's embedding space. Similar strategies are employed by the Qwen-Audio series [9, 10] and SALMONN [11], where embeddings from the Whisper encoder [12] and BEATs [13] are combined via a Q-Former [14] and mapped into the LLM's embedding space. Whispering-Llama [5] takes a slightly different approach, inserting adapters and cross-attention layers into LLaMA [15] to fuse Whisper encoder embeddings with latent language representations. In contrast, discrete features do not require additional adapter modules for modality alignment and can be directly integrated into the LLM vocabulary. For instance, both SpeechGPT [16] and AudioPaLM [17] apply clustering techniques to discretize speech embeddings, while Kimi-Audio [4] uses a vector quantization layer to convert continuous representations into discrete token sequences.

Recent studies suggest that LLMs naturally operate in a continuous latent space. COCOMIX [18] extracts continuous semantic concepts using a pre-trained Sparse Autoencoder (SAE) [19] and identifies the most influential ones via attribution scores. Similarly, COCONUT [20] feeds the final continuous hidden LLM state directly as the input embedding for the next token, leveraging the full embedding space and enhancing reasoning capabilities. Additionally, [21] shows that LLMs capture time- and space-continuous patterns and learn language in a continuous manner, in contrast with how humans understand language.

otivated by the potential of these continuous latent spaces, we investigate whether an intermediate audio-conditioned text space can facilitate a more seamless multimodal fusion than raw audio embeddings. We propose a novel framework that aligns the continuous representation spaces of both Whisper and the LLM. To our knowledge, this is the first study to leverage the Whisper decoder's hidden states as a semantic bridge for Greek speech processing. Our primary contributions are summarized as follows:

- **Continuous Cross-Modal Adaptation:** We introduce a framework that fuses Whisper's decoder hidden states with a target LLM layer via a flexible cross-attention mechanism, supporting both offline and low-latency streaming fusion.
- **VoxKrikri:** We develop the first Greek-centric Speech LLM, achieving new state-of-the-art results in Greek Automatic Speech Recognition (ASR) across multiple competitive benchmarks.
- **Latent Alignment Analysis:** We conduct an in-depth analysis of the latent feature space using rCCA and SAE-inspired techniques, proving that our framework successfully aligns the continuous representations of the two models.

VoxKrikri and its associated weights will be made available under the Llama 3.1 Community License Agreement upon acceptance.

2. Proposed Method

We introduce a multimodal fusion approach designed to integrate high-level audio features with textual representations by leveraging the continuous hidden state spaces of an audio-conditioned decoder. Unlike traditional discrete bottleneck approaches or simple projection layers, our framework utilizes the rich semantic information present in pre-trained speech models to guide the language generation process. An overview of the architectural pipeline and the interaction between the modality-specific components is illustrated in Fig. 1.

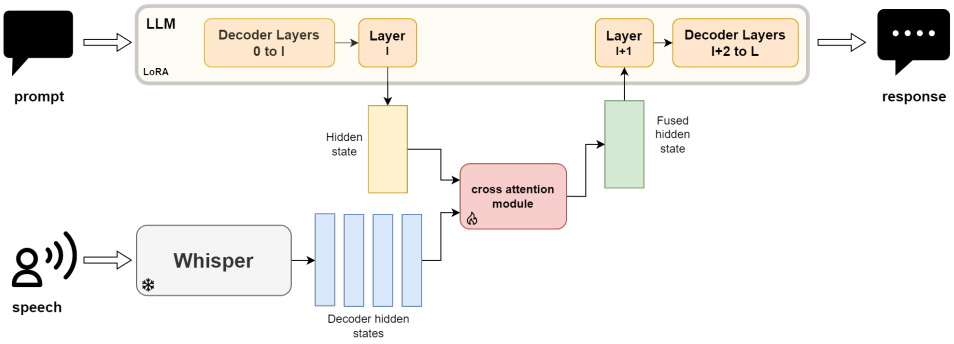


Figure 1: *VoxKrikri: Whisper's decoder hidden states are fused with LLM's hidden states at a designated injection layer I, using a cross-modality attention module. The attention module implements both full-sequence and causal fusion between the features, using a soft, proportional alignment method.*

2.1. Cross-modal Fusion Mechanism

The core of our method lies in the bridge between a pre-trained Whisper decoder and a Large Language Model (LLM). We hypothesize that the final hidden layer of Whisper's decoder contains a condensed, audio-conditioned textual representation, which we denote as A_s. Rather than treating these as simple tokens or discrete embeddings, we treat them as a continuous sequence of acoustic-semantic guides that provide context to the language model.

These representations are integrated into the LLM through a cross-attention module, which is inserted at a specific decoder layer I within the LLM's transformer stack. At this junction, the text-based hidden states Y_t serve as queries that attend to keys and values derived from the audio-conditioned states A_s. The depth of this insertion point, I, is a critical hyperparameter; by varying it between early layers (closer to the input embeddings), intermediate layers, or late layers (closer to the language head), we can empirically observe how fusion interacts with different levels of the LLM's internal hierarchy.

Furthermore, this modular design distinguishes our work from existing projection-based methods [8, 9, 11]. While those methods typically prepend audio embeddings to the LLM input—forcing the model to process the entire audio prefix before generating text—our cross-attention approach decouples the modality streams. This decoupling is what allows our framework to transition seamlessly between offline processing, using full-sequence fusion, and real-time streaming applications using causal fusion.

2.2. Full-sequence Fusion

In scenarios where latency is not a primary constraint, we employ full-sequence fusion. In this configuration, the LLM has global access to the entire audio sequence $A_{1:S}$ for every text token $Y_{1:T}$ being generated. This is mathematically formulated through the conditional probability:

$$p(Y_{1:T}|A_{1:S}) = \prod_{t=1}^{T} p(Y_t|Y_{<t}, A_{1:S}) \tag{1}$$

In this setting, the cross-attention mechanism acts as a global lookup table. For any given token t, the query Y_t can attend to any acoustic feature in the sequence, regardless of its temporal position, which is suitable for offline tasks where the full audio is available at inference time, such as audio question answering or speech summarization.

2.3. Causal Fusion

For real-time tasks such as streaming ASR or live captioning, temporal ordering must be preserved: each text token Y_t may only attend to audio frames observed up to its decoding step:

$$p(Y_t \mid Y_{<t}, A_{\leq s_t}), \tag{2}$$

where s_t denotes the last audio frame aligned with token t. This design ensures that predictions are based only on past and current context, maintaining temporal consistency and the autoregressive nature of both Whisper and the LLM. During training and inference, a causal attention mask enforces this restriction.

2.4. Causal Alignment Masking Strategy

To implement the causal constraint effectively without requiring explicit alignment , we introduce a soft, proportional alignment strategy. We assume a linear progression of information, where each text token $t \in \{0, \ldots, T-1\}$ corresponds to a proportional segment of the audio sequence. The alignment index s_t is computed as:

$$s_t = \left\lfloor \frac{S}{T} \cdot t \right\rfloor, \tag{3}$$

where T represents the total number of text tokens and S represents the total number of audio frames. This mapping enforces monotonicity: as the text generation progresses (increasing t), the window of accessible audio frames expands (increasing s_t), but never retreats. This uniform coverage allows the model to "pace" its attention across the audio signal, ensuring that information is consumed at a rate consistent with the generation of text.

To enforce this during the attention computation, we define a causal attention mask $M \in \mathbb{R}^{T \times S}$. The mask values are assigned as follows:

$$M_{t,s} = \begin{cases} 0, & s \leq s_t \\ -\infty, & s > s_t \end{cases}, \quad 0 \leq t \leq T-1, \quad 0 \leq s \leq S-1. \tag{4}$$

This mask is applied directly to the attention logits. Finally, the integration of the modalities into the LLM's hidden state h_t is performed via the modified cross-attention equation:

$$h_t = h_t + \text{softmax}\Big(\frac{(Y_tQ)(A_sK)^T}{\sqrt{d}} + M_{t,s}\Big)(A_sV), \tag{5}$$

where Q, K, and V are the learned projection matrices for queries, keys, and values, respectively. By making these parameters fully trainable, the model learns to selectively extract relevant acoustic features A_s that complement the linguistic context

Dataset	VoxKrikri-1		VoxKrikri-9		VoxKrikri-15		VoxKrikri-21		VoxKrikri-30		Whisper-large-v3
	causal	full	causal	full	causal	full	causal	full	causal	full	
GPC	14.88	14.81	13.06	14.26	12.65	12.75	12.27	**12.08**	12.64	12.6	17.27
Fleurs	10.85	10.5	11.13	10.21	10.2	10.3	9.52	**9.33**	9.48	10.2	11.02
CV	13.8	13.6	12.42	12.2	12.3	**11.97**	12.01	12.04	12.43	12.1	13.98
LG	10.55	10.41	10.03	9.89	9.78	9.54	9.47	**9.4**	9.66	9.52	10.86
HParl	14.7	15.01	13.9	13.8	13.55	13.32	**12.9**	13.01	13.41	13.3	16.99

Table 1: *ASR results on benchmark test sets (WER) using both causal and full-sequence fusion*

Y_t already present in the LLM. This formulation is flexible: setting $s_t = S-1$ for all t recovers the full-sequence fusion mode, while Equation (3) maintains the causal streaming mode.

3. Experimental Setup

3.1. Speech Model

We adopt Whisper-large-v3 as the acoustic backbone. This encoder–decoder model achieves state-of-the-art results on Greek ASR benchmarks and serves as a robust foundation. In our pipeline, Whisper acts as a frozen feature extractor, to preserves its highly generalized acoustic-linguistic representations. The extracted sequence of continuous representations already possess a degree of cross-modal alignment, serving as an ideal foundation for our fusion module.

3.2. LLM and Parameter-Efficient Adaptation

The language modeling component is built upon Llama-KriKri-8B [22], a Greek-adapted version of the Llama 3.1-8B architecture. which substantially outperforms the base Llama 3.1-8B on Greek benchmarks.[1] We apply LoRA [23] adapters for parameter-efficient adaptation during training.

Injection layer plays a central role in our framework, as different layers capture different levels of abstraction—ranging from low-level syntax in early layers to high-level semantics in the later stages. To investigate this, we experiment with five distinct injection layers out of the 32 decoder layers of the model: the 1st, 9th, 15th, 21st, and 30th layers, covering initial, middle-tier, and late processing. Each experimental configuration is referred to by its specific model identifier:

- **VoxKrikri-1**: Early-stage fusion
- **VoxKrikri-9 & VoxKrikri-15**: Intermediate-stage fusion.
- **VoxKrikri-21**: Late-intermediate fusion.
- **VoxKrikri-30**: Late-stage fusion

Dataset	#Hours	#Samples
GPC-2400	2447	430K
GPC-50	800	488K
Fleurs	13	3,2K
CV	12	10,8K
LG	72	23,5K
HParl	120	76K
Total	∼3300	∼1M

Table 2: *Training data analysis*

3.3. Datasets

For training, we employed the following datasets:

Logotypografia (LG) [24] is one of the earliest Greek corpora, comprising approximately 72 hours of speech for training and 9 hours for testing.

Common Voice (CV) [25] is a multilingual, crowd-sourced dataset developed by Mozilla. In our experiments, we use version 9.0 of the Greek subset, which contains 12 hours of training data, and 2 hours of test data.

HParl (HP) [26] consists of parliamentary recordings from the Hellenic Parliament, which consists of 120 hours of training and 11 hours of testing data.

Fleurs [27] is a multilingual speech corpus. The Greek portion comprises 13 hours of speech for training and 2 hours for testing.

Greek Podcast Corpus (GPC) [28] is a weakly supervised dataset of Greek podcasts, covering 16 diverse categories (e.g., *True Crime*, *Comedy*). We use the GPC-50 subset for training, containing 50 hours per category (800 hours total), and a test set of 1 hour per category (16 hours total). In addition, we collected 2,447 hours of transcribed Greek podcast audio (annotated as *GPC-2400*), totaling 434,530 samples. Table 3 details its category distribution, while Table 2 overviews the full training data.

Category	#Hours	#Samples
Arts	361	62K
Business	192	33K
Comedy	263	44K
Education	417	70K
Fiction	290	49K
Health	312	52K
History	170	48K
Kids	144	28K
Leisure	130	20K
Music	168	28K
Total	∼2400	∼430K

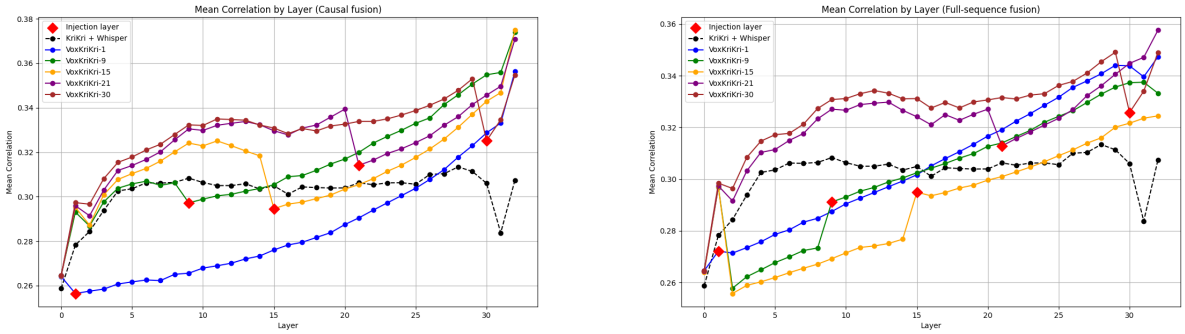
Table 3: *GPC-2400 categories breakdown*

3.4. Experimental & evaluation setting

For fine-tuning with the LoRA adapter, we set the hyperparameters to $r = 8$, $\alpha = 16$, and $dropout = 0.1$. This configuration resulted in approximately 744 million trainable parameters out of a total of 9.8 billion (around 7.53% of the full model). All experiments were conducted on NVIDIA A100 GPUs with 64GB memory, provided by the LEONARDO supercomputer [29].

Regarding our evaluation setting, we follow the established benchmarks from [11], [8], and [17], where the Automatic Speech Recognition (ASR) capabilities of the multimodal frameworks are assessed in direct comparison with dedicated speech-to-text models. This allows us to determine if the additional linguistic power of the LLM backbone provides a tangible advantage over traditional architectures. Specifically, we evaluate the *VoxKrikri-1, 9, 15, 21, 30* models using the Word Error Rate (WER) metric. We benchmark our results against the state-of-the-art Whisper-large-v3 baseline across five test sets.

[1] https://huggingface.co/collections/ilsp/ilsp-greek-evaluation-suite-6827304d5bf8b70d0346b02c



(a) (b)

Figure 2: *Mean CCA correlation per LLM layer between the hidden states and the Whisper decoder features for the base models and VoxKrikri-1,9,15,21,30. (a) Using causal fusion, (b) Using full-sequence fusion. Overall, intermediate-to-late fusion increases cross-modal correlation.*

4. Results

The results of evaluation on the five test sets can be seen at Table 1. Both full-sequence and causal fusion strategies are considered. Across all benchmarks, every *VoxKrikri* variant consistently outperforms Whisper, with improvements ranging from 0.6 to over 5 WER points depending on the dataset. Full-sequence fusion generally yields slightly lower WER than causal fusion, as seen for example on GPC-test (12.08 vs. 12.27) and Fleurs-test (9.33 vs. 9.52). Moreover, intermediate-to-late fusion layers (15, 21, and 30) tend to outperform early fusion (1, 9), with *VoxKrikri-21* achieving the best overall results on three datasets (GPC, Fleurs, and LG) and *VoxKrikri-15* excelling on CV-test. Notably, *VoxKrikri-21 (full)* reduces WER by over 5 points on GPC-test compared to Whisper (12.08 vs. 17.27).

These results highlight the robustness of our continuous latent fusion approach and demonstrate that carefully chosen fusion depths enable substantial performance gains. The observed synergy between Whisper's acoustic features and the LLM's linguistic priors suggests that mid-layer integration effectively balances perceptual input with high-level semantic reasoning.

5. Cross-modal Alignment

To study cross-modal alignment, we measure the representational similarity between the LLM's hidden states and Whisper's decoder features using rCCA [30]. We analyze 1,000 samples from the GPC test set, subsampling 20,000 features with 1,000 components and regularization $\lambda = 10^{-4}$ to handle dimensionality mismatch. The results, illustrated in Fig. 2, show that intermediate-to-late fusion consistently strengthens correlation, confirming superior representational alignment in deeper layers.

We observe that early-to-intermediate layers are highly sensitive to sequence length, showing sharp correlation increases when granted access to the full audio context. Conversely, late-fusion layers remain stable under both causal and full-sequence regimes, suggesting they are more resilient to temporal constraints once semantic integration is established. Ultimately, the peak alignment observed in these deeper layers directly corroborates the WER improvements, proving that stronger feature congruence drives higher transcription accuracy.

6. Conclusion & Future Work

In this work, we proposed a novel framework for narrowing the modality gap between pre-trained language and acoustic models by operating directly in their textual, continuous embedding spaces. Rather than relying on raw audio embeddings, our approach leverages an intermediate audio-conditioned text space, providing a linguistically grounded representation for more effective multimodal fusion. Building on this framework, we introduced *VoxKrikri*, the first Greek Speech-LLM, which achieves state-of-the-art performance on ASR tasks. Our analysis further showed that the proposed methodology effectively improves the alignment between speech and language representations. We believe this framework offers a strong foundation for future research on multimodal alignment in continuous latent spaces. Moreover, our introduction of causal fusion via causal cross-modal masking paves the way for streaming and real-time applications of SpeechLLMs.

Looking ahead, we plan to extend our work in several directions. First, we aim to incorporate general-purpose audio capabilities, for instance by integrating a BEATs encoder, and to evaluate our framework on more challenging tasks such as speech and audio question answering, or real-time translation. In addition, we are interested in exploring multi-layer injection strategies and experimenting with combinations of early and late fusion.

7. References

[1] B. et al., "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[2] L. et al., "Improved baselines with visual instruction tuning," 2023.

[3] W. et al., "Cogvlm: visual expert for pretrained language models," in *NIPS 2024*, 2024, pp. 121 475–121 499.

[4] D. et al., "Kimi-audio technical report," *arXiv e-prints*, pp. arXiv–2504, 2025.

[5] R. et al., "Whispering llama: A cross-modal generative error correction framework for speech recognition," in *EMNLP 2023*. ACL, 2023, pp. 10 007–10 016.

[6] ——, "Learning transferable visual models from natural language supervision," in *ICML*. PmLR, 2021, pp. 8748–8763.

[7] E. et al., "Clap learning audio concepts from natural language supervision," in *ICASSP*, 2023.

[8] M. et al., "An embarrassingly simple approach for llm with strong asr capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[9] C. et al., "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[10] ——, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[11] T. et al., "Salmonn: Towards generic hearing abilities for large language models," *CoRR*, 2023.

[12] R. et al., "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28 492–28 518.

[13] C. et al., "Beats: Audio pre-training with acoustic tokenizers," in *ICML*. PMLR, 2023, pp. 5178–5193.

[14] L. et al., "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*. PMLR, 2023, pp. 19 730–19 742.

[15] T. et al., "Llama: Open and efficient foundation language models."

[16] Z. et al., "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[17] R. et al., "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2021.

[18] T. et al., "Llm pretraining with continuous concepts," *arXiv preprint arXiv:2502.08524*, 2025.

[19] S. et al., "A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models," *arXiv preprint arXiv:2503.05613*, 2025.

[20] H. et al., "Training large language models to reason in a continuous latent space," *arXiv preprint arXiv:2412.06769*, 2024.

[21] M. et al., "Language models are implicitly continuous," in *The Thirteenth International Conference on Learning Representations*.

[22] R. et al., "Krikri: Advancing open large language models for greek," *arXiv preprint arXiv:2505.13772*, 2025.

[23] H. et al., "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[24] D. et al., "Large vocabulary continuous speech recognition in greek: corpus and an automatic dictation system." in *Interspeech*, 2003, pp. 1565–1568.

[25] A. et al., "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[26] P. et al., "Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for modern greek," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 286–299, 2023.

[27] C. et al., "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE SLT*. IEEE, 2023, pp. 798–805.

[28] P. et al., "The greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data," in *Proc. Interspeech 2024*, 2024, pp. 3969–3973.

[29] T. et al., "Leonardo: A pan-european pre-exascale supercomputer for hpc and ai applications," *JLSRF*, vol. 9, no. 1, 2024.

[30] ——, "Canonical correlation analysis in high dimensions with structured regularization," *Statistical modelling*, vol. 23, no. 3, pp. 203–227, 2023.