Do Factual Recall Mechanisms Carry over from Text to Speech in Multimodal Language Models?

Luca Modica[1,3,4*] **Filip Landin**[2,3,4*] **Mehrdad Farahani**[3,4*] **Livia Qian**[5]
Gabriel Skantze[5] **Richard Johansson**[3,4]

[1]Zenseact [2]Unbox AI [3]Chalmers University of Technology
[4]University of Gothenburg [5]KTH Royal Institute of Technology

mehrdad.farahani@chalmers.se

Abstract

In recent years, several Speech Language Models (SLMs) that represent speech and written text jointly have been presented. The question then emerges about how model-internal mechanisms are *similar* and *different* when operating in the two modalities. We focus on how these systems encode, store, and retrieve factual knowledge, which has previously been investigated for text-only models. To investigate mechanisms behind the storage and recall of factual association in SLMs, we leverage Causal Mediation Analysis, a technique previously applied to text-based models.

Initial results using `SpiritLM`, a multimodal model integrating discrete speech tokens reveal discrepancies between text-to-text and speech-to-text results, suggesting that the emergent mechanisms for factual recall are only partially carried over from the text to the speech modality. These results advance our understanding of how internal mechanisms encode factual associations in SLMs while contributing insights for improving speech-enabled AI systems.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in various NLP tasks, including answering factual questions such as *"the capital of Italy is"* by relying on information stored in their parameters (Petroni et al., 2019). However, these systems still suffer from hallucination and are prone to committing factual errors, which limits their trustworthiness and usability (Kandpal et al., 2023): this motivates further investigation into the mechanisms behind knowledge recall and factual memory. Research using intervention-based methods reveals that factual knowledge can be *localized* within text-based LLMs, particularly in mid-layer feed-forward networks (MLP) (Geva et al., 2021; Meng et al., 2022; Geva et al., 2023). These findings are being used to develop methods that edit model parameters, allowing for precise intervention on factual associations, representing a step forward in more accurate and steerable models (Meng et al., 2022, 2023).

Speech-language models trained directly on audio without text supervision – such as those using GSLM-style training (Lakhotia et al., 2021) – have shown promise in speech understanding tasks (Lin et al., 2025; Basu et al., 2024; Peng et al., 2026; Hassid et al., 2023; Zhang et al., 2023). Since they do not leverage text-based knowledge, their factual understanding is more limited. On the other hand, speech models built on top of LLMs, like `SpiritLM` (Nguyen et al., 2025), might retain or develop a deeper understanding of knowledge encoded in the text-based model. What is less understood is whether this behavior, if it exists, originates from the separate training on speech data or whether it comes from mechanisms learned from text. This opens up interesting research questions:

- Are the mechanisms behind factual recall modality-independent?
- Does factual localization in speech-based inputs emerge independently without reliance on the backbone architecture?

In this paper, we investigate where and how factual associations are stored and recalled in `SpiritLM` by using Causal Tracing (CT), one of the intervention-based techniques used to study the causal effect of components within a neural network (Meng et al., 2022). We focus on two specific input-wise settings:

1. T→T (text-to-text): where the model receives and produces text.
2. S→T (speech-to-text): where the model takes audio as input but still generates a text output.

[*]Equal contribution.

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 401–409

By extending CT to analyze factual recall in a multimodal setting, we show that speech input leads to weaker but detectable traces of factual localization.

2 Methodology

This section first describes the causal mediation analysis (CMA) framework and the mathematical underpinnings of causal tracing, followed by the `SpiritLM` model, dataset preparation, and the experiment design.

2.1 Preliminaries: Causal Mediation Analysis

CMA is a framework for investigating questions about the relative contributions to an overall effect of individual components in a complex system (Pearl, 2001). Following Vig et al. (2020), it has emerged as part of the standard toolbox for the analysis of LLMs; in mechanistic interpretability, it is also known as *activation patching* (Heimersheim and Nanda, 2024). Meng et al. (2022) applied CMA to investigate factual recall in LMs. Their approach consists of three steps:

Clean run. The LM is provided a clean prompt $X = x$, producing a probability $\mathbb{P}_x[o]$, where o denotes the expected decoded token. The corresponding hidden states from this inference are cached.

Corrupted run. The model receives a corrupted input prompt $X = x^*$, resulting in a new predicted output probability $\mathbb{P}_{x^*}[o]$. Meng et al. (2022) carried out the corruption intervention by obfuscating the subject tokens with noise proportional to the standard deviation over all input embeddings.

Corrupted-with-restoration run. The same corrupted prompt $X = x^*$ is passed to the model, but with the activation value of selected component C_i restored (patched) from the *clean run*. The result is denoted $\mathbb{P}_{x^*,\text{clean } C_i}[o]$, where "clean C_i" refers to the value of the component C_i from the clean inference.

The results of the three runs allow us to quantify the mediated effects of interventions. The relative contribution of a hidden-state mediator is measured by the *Indirect Effect* (IE), defined as the difference between the corrupted-with-restoration run and the corrupted run:

$$\text{IE} = \mathbb{P}_{x^*,\text{clean } C_i}[o] - \mathbb{P}_{x^*}[o].$$

By averaging over multiple prompts, we obtain Average Indirect Effect (AIE) at different levels of the model components and then visualize the contribution results.

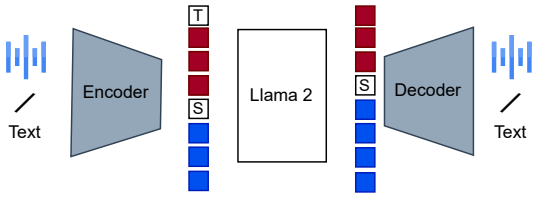

Figure 1: The SpiritLM architecture.

2.2 The multimodal large language model under study: The `SpiritLM` model

Our work examines `SpiritLM` (Nguyen et al., 2025) as a case of a multimodal (speech) language model that can generate text and audio language content. Furthermore, `SpiritLM` uses discrete speech tokens and is trained on interleaved speech and text token sequences for better generalization and alignment across modalities, making it well-suited for the proposed study.

We illustrate the high-level architecture of `SpiritLM` in Figure 1. The model handles mixed speech and text inputs using special *modality declaration tokens* ("T" for text, "S" for speech). Audio is discretized into tokens with HuBERT (Hsu et al., 2021) and text with the Llama2 tokenizer. The interleaved sequence, with each portion prefixed by its respective modality tokens, is input to Llama2. The model predicts the next tokens based on the most recent modality token: a "T" token prompts text generation, while "S" prompts discrete speech tokens. At inference, speech tokens are decoded with HiFi-GAN (Kong et al., 2020).

The employed speech representation allows targeting and analyzing specific speech tokens in a Causal Mediation Analysis experiment in both a uni-modal and cross-modal context.

2.3 Dataset and data preparation

For our study, we use the *Known* dataset (Meng et al., 2022): It contains almost 1000 factual prompts that the GPT2-XL model knows,[1] with the annotated subject and object (expected correct answer).

Starting from the available text data, we introduce the speech modality counterpart for each information in Known (prompt, subject, and object). The utterances are generated using the TTS model

[1]Link to the collection of factual prompts: `https://rome.baulab.info/data/dsets/known_1000.json`

MeloTTS (Zhao et al., 2023), which is based on architectures that leverage adversarial learning to improve expressive power and high-quality speech synthesis (Kim et al., 2021; Kong et al., 2023). We assess the reliability of the curated speech modality through two complementary approaches: manual inspection of challenging samples, particularly prompts, and automatic transcription of the generated audio using *Whisper-small*,[2] a lightweight ASR model. The prompt transcription results yield a Word Error Rate of 19%: this demonstrates good performance despite the inherent difficulty of transcribing proper nouns, and the reliability of the TTS model.

To further ensure the dataset quality for the subsequent experiment, we filter the original collection factual statements based on the model performance in the 2 different modalities as input, resulting in 2 datasets: *Known-t2t* and *Known-s2t*. *Known-t2t* includes datapoints where the model readily generates either an exact correct answer or a close variant in a Text → Text scenario. For instance, "Rome" is correct for the prompt "The capital of Italy is ___", while answers like "Rome, Italy" or "the city of Rome" are considered partially correct. *Known-s2t* follows the same selection criteria, but in a Speech → Text setting.

2.4 Experiment design

Factual associations in `SpiritLM` are investigated through two CMA experiments, in text and speech domain, in order to determine causal effects of network components: single transformer layers, MLP, and attention sub-layers.

Experiments are conducted on prompts from the datasets introduced in 2.3. Similarly to Meng et al. (2022), the corrupted run is done by adding noise to the representation of the subject tokens.

Experiment 1: Within-modality factual recall (Text → Text) In the first experiment, a text prompt is fed into the model and the log probability of predicting the corresponding attribute is computed for each of the three CMA iterations – clean, corrupted, and corrupted-with-restoration runs – described in Section 2.1. The IE is aggregated by the position of the token in the sentence: *first subject token, middle subject tokens, last subject token, first subsequent token, further tokens, and last token*, averaged over all prompts (AIE), and presented as log AIE for readability and comparison.

Experiment 2: Cross-modality factual recall (Speech → Text) The second experiment is similar, but uses the synthesized version of the dataset, where prompts are converted to audio. Each utterance is encoded and discretized by `HuBERT`), and the resulting tokens are fed into the language model, where the CMA pipeline is applied as in the previous experiment. An additional challenge here is that, in the corrupted run, it is no longer obvious how to localize the subject token(s) in the input prompt. Connectionist Temporal Classification (CTC)-based forced alignment (Kürzinger et al., 2020) is therefore used to find the target time range of the subject in the utterance, and thus the related range of the speech tokens (see Appendix A for more details). The same technique admits a mapping between speech tokens and the corresponding text ones, which is used to post-process the CMA results. The causal traces of the speech (HuBERT) token are aggregated, as for text, by the corresponding text tokens, which facilitates direct comparison and interpretation of causal influence across modalities. The quality of the forced alignment is validated by manually inspecting the speech segments corresponding to text tokens, ensuring that token boundaries are properly aligned without overlaps or significant gaps throughout the prompt utterance.

3 Results and Discussion

We begin our experiments at two levels to examine whether factual associations – previously shown to localize around subject tokens in text-only models – can also be recalled and expressed in other modalities. In our case, we focus on `SpiritLM` and its speech modality. To measure this, we use causal mediation analysis to compute the Average Indirect Effect (AIE) for each layer and token across all filtered query prompts. The AIE captures the marginal contribution of an internal component to the final factual prediction under intervention. Higher AIE values indicate which layers and positions influence factual recall more.

As a baseline, we perform CT on the backbone model used in `SpiritLM`, using text-to-text prompts. As expected from prior work by Meng et al. (2022), we observe strong causal signals (AIE) centered on subject tokens at early layers, especially in mid-layer MLPs (see Figure 2). We also detect notable effects at the final token po-

[2]Link to the model checkpoint used: `https://huggingface.co/openai/whisper-small`

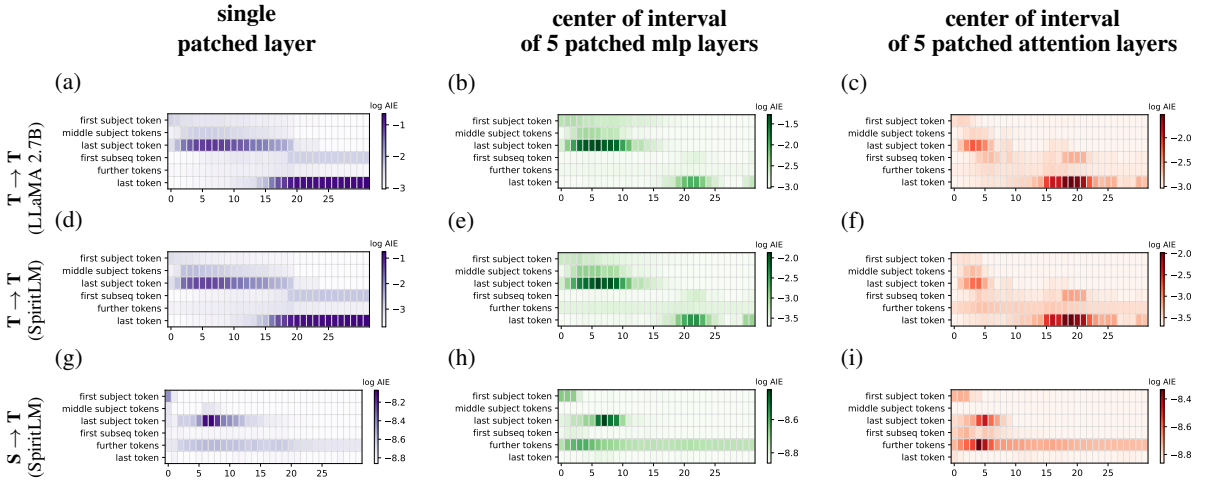

Figure 2: Log-scaled AIE across different modules and modalities over 754 prompts. In each subfigure, the x-axis represents the layers and the y-axis shows the tokens of interest.

sition in upper layers, where strong causality is typically observed. Extending the same CT analysis to SpiritLM in the T→T setting reveals nearly identical behavior: consistent causal signals around subject tokens across hidden States, MLP, and attention layers (Figure 2). This confirms that text-processing pathways in SpiritLM preserve their original capabilities after fine-tuning on speech.

On the other hand, we observe different results when the input is speech (S→T). The AIE drops drastically, showing a much more diffuse and lower-magnitude signal; however, we can still observe an effect around the subject tokens in the MLP and attention layers (see Figure 2).

Our experiments suggest that factual associations in spoken language models like SpiritLM are not strictly modality-dependent. Although the model is capable of retrieving knowledge from both modalities – including, to some extent, speech – the factual recall mechanisms are much more readily activated when the model is provided with text input than with speech. For SpiritLM, text serves as a more structured and reliable trigger for recalling facts, suggesting that the speech-based fine-tuning in this model does not fully utilize the fact-recalling mechanisms learned by the text-based backbone model. However, we do not have sufficient evidence to conclude whether the partial transfer of factual capabilities from text to speech in SpiritLM arises from noise introduced by controlled conditions or limitations of the mapping between speech and text tokens. Based on recent studies (Xiang et al., 2025; Cuervo et al., 2026), we also hypothesize that a likely cause is the semantic gap between the two modalities, possibly arising from the post-training speech adaptation of SpiritLM's text-only backbone: while the two modalities may become increasingly aligned in direction (cosine similarity) across deeper layers, a divergence in magnitude (Euclidean distance) can still persist, ultimately compromising the transfer of factual knowledge to the speech modality.

4 Related Work

Interpretability for speech LLMs remains under-explored compared to their text counterparts: in particular, investigating to what extent multimodal text/speech systems share underlying mechanisms remains unexplored. Recent works (Pasad et al., 2024, 2021; Shen et al., 2024) have explored speech model interpretability, considering speech features at different granularity levels (e.g., word boundaries, pronunciation), finding, for example, that frame-level representations within each word segment are not all equally informative. These studies lead to open questions more related to how sentence-level properties (e.g., subject) are encoded, a gap our work seeks to fill. More recently, Glazer et al. (2026) explored mechanistic interpretability for ASR systems, applying logit lens and activation patching to reveal internal model dynamics responsible for repetition hallucinations and semantic biases within acoustic representations: these findings suggest promising directions for similar investigations in speech LLMs and beyond the ASR setting.

5 Conclusion

The study investigates the mechanisms of factual recall in Speech LLMs, focusing on whether this process in the speech modality operates independently or relies on the text modality and the capabilities of the original text model. By using the CMA framework with `SpiritLM`, we show that the model preserves the same text-based computation pathways as its corresponding text-only counterpart, while the speech modality leads to a considerably weaker causal effect at the level of the MLP and attention layers. Although the latter does not conclusively prove if knowledge from text is transferred to the speech modality, these preliminary insights hint that speech LLMs are not strictly modality-dependent when recalling facts.

Limitations and Future Directions

Dataset selectivity. We conducted our experiment on a single synthesized speech dataset – *Known* (Meng et al., 2022), which might not capture all the nuances of how modality interactions affect factual memory tracing. Using other datasets, from *Spoken SQuAD* (Li et al., 2018) to a synthesized version of *PopQA* (Mallen et al., 2023), can provide a valuable contribution in this field.

Model generalization. Although the insights obtained using `SpiritLM`, their compatibility with other Speech LLMs that employ discrete speech tokens needs to be explored (Zhang et al., 2023; Rubenstein et al., 2023). Furthermore, replicating our experiments within `SpiritLM` through different text-only backbones (Yang et al., 2025; Jiang et al., 2023) or a joint speech-text training strategy would be crucial to assess the generalizability of our findings.

Discrete speech tokens limitations. Using discrete tokens represents an interesting strategy to integrate the modality with text-based tokens seamlessly; however, recent studies have shown performance limitations on semantic understanding tasks, which might affect results of factual recall studies of speech Large Language Models (Wang et al., 2025). Considering speech LLMs that employ different strategies to convey a richer speech representation, such as training on continuous speech representation (Peng et al., 2026; Tang et al., 2024), can represent a promising future direction of this investigation.

Ethical Considerations

This study focuses on the interpretability of speech-language models. As part of our research, we do not release any new models or datasets; therefore, we do not implicate any potential risks or concerns related to the misuse of our results.

Acknowledgments

This research was funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partly funded by the Swedish Research Council through grant agreement no. 2022-06725. We also acknowledge the Computer Science and Engineering department at Chalmers and the University of Gothenburg, which funds part of the conference costs through the *Lars Pareto travel grant*.

References

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. Understanding information storage and transfer in multi-modal large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. 2026. Closing the gap between text and speech understanding in LLMs. In *Proceedings of the The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya,

Joseph Keshet, and Aviv Navon. 2026. Beyond transcription: Mechanistic interpretability in asr. In *The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)*, Singapore.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Honolulu, Hawaii, USA. JMLR.org.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. In *Proceedings of INTERSPEECH*, Dublin, Ireland.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. *CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition*, page 267–278. Springer International Publishing.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Proceedings of INTERSPEECH*, Hyderabad, India.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, and 2 others. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *Preprint*, arXiv:2502.17516.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass editing memory in a transformer. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, Seattle, Washington. Morgan Kaufmann Publishers Inc.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2026. A survey on speech large language models. *IEEE Journal of Selected Topics in Signal Processing*, 20(1).

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. AudioPaLM: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2024. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4250–4261, Mexico City, Mexico. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *Preprint*, arXiv:2004.12265.

Dingdong Wang, Junan Li, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen M. Meng. 2025. Speech discrete tokens or continuous features? A comparative analysis for spoken language understanding in SpeechLLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24913–24924, Suzhou, China. Association for Computational Linguistics.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. MeloTTS: High-quality multi-lingual multi-accent text-to-speech.

A Forced Alignment for Cross-modal Token Mapping: Implementation Details

Text preprocessing for CTC. For the transcription to be compatible with the forced alignment, a text preprocessing is necessary to ensure all characters are included in the CTC model vocabulary. For example, digits and special characters such as "%" are converted to their written format (e.g, "0" becomes "zero", or "%" becomes "percent"). On the text token-level, preprocessing can lead to a longer or shorter sequence of tokens, compared to the original text being tokenized. We later refer to the preprocessed text tokens as *spoken text tokens*. The preprocessing step concludes by joining the spoken text tokens into a single string, using the word boundary character defined by the CTC model as a separator.

Frame-wise label probability estimation from audio waveform. We generate emission probabilities per audio frame, using the pre-trained `HuBERT-LARGE`[3] model as a speech tokenizer. This model is fine-tuned for automatic speech recognition (ASR) with CTC loss, representing a suitable candidate for this use case (Hsu et al., 2021).

Trellis matrix generation with log-probability of label alignments at each time step. Given an audio input sequence $\mathbf{X} = (x_1, \ldots, x_T)$ and transcript labels $(c_1, \ldots, c_N)$ at the character level, we compute through dynamic programming and map all possible joint probabilities in the trellis diagram matrix $K \in \mathbb{R}^{T \times N}$; $K_{(t,j)}$ represents the maximum log-probability of aligning the first labels j up to time t. To compute the probability at time step $t+1$ for label c_{j+1}, we consider two possible transitions: either we stayed on the same label c_{j+1} or transitioned from c_j to c_{j+1}. Based on these criteria, the trellis is updated as follows:

$$K_{(t+1,j+1)} = \max \begin{cases} K_{(t,j)}\, p(t+1, c_{j+1}) \\ K_{(t,j+1)}\, p(t+1, \text{repeat}) \end{cases}$$

where $p(t+1, c_{j+1})$ is the probability of emitting label c_{j+1} at time $t+1$, and $p(t+1, \text{repeat})$ is the probability of emitting no label change.

Find the most likely path from the trellis matrix. Once the trellis is generated, we will traverse it following the elements with the highest probability. Starting from the last label index belonging to the last time step, we progress in the matrix backwards, choosing to keep the current label or move to the previous label based on the highest probability for each time step. The process ends when we reach the beginning of the sequence, obtaining the most likely path that aligns text and audio.

Merge repetitions and segments into words (spoken text tokens). The final step involves postprocessing the output from the optimal path. Because the path may contain consecutive repetitions of the same label, we merge path points corresponding to repeated characters into a single segment to make it close to the original transcript.[4] Similarly, we group segments that correspond to the same spoken text token, using the word boundary character as a guide. The result is a sequence of segments, each representing a spoken text token from the transcript and annotated with the corresponding range of audio frames and average emission probability.

From this segmentation, we can derive the time range of each text token, which can also be used to obtain the speech token range, using the token rate of the speech tokenizer. The last result allows us to directly map text tokens and related speech tokens. Figure 3 illustrates an example of the final output from CTC-based forced alignment, demonstrating this alignment process. Beginning with the preprocessed transcript "THE|CAPITAL|OF|ROMAN|REPUBLIC|IS," we align each text token with its corresponding segment in the audio. For each text token, we label the aligned speech segment with the average probability over the merged segment, clearly indicating its position within the utterance as a highlighted segment on the spectrogram, with boundaries marking its start and end. This segmentation process allows us to determine the precise time range for each text token.

- Given the frame range (f_{start}, f_{end}) of a spoken text token segment, we can first compute its time

[3]Link to the model checkpoint used: `https://huggingface.co/facebook/hubert-large-ls960-ft`

[4]When merging path points into a single segment, we use the average probability of all frames in that segment.

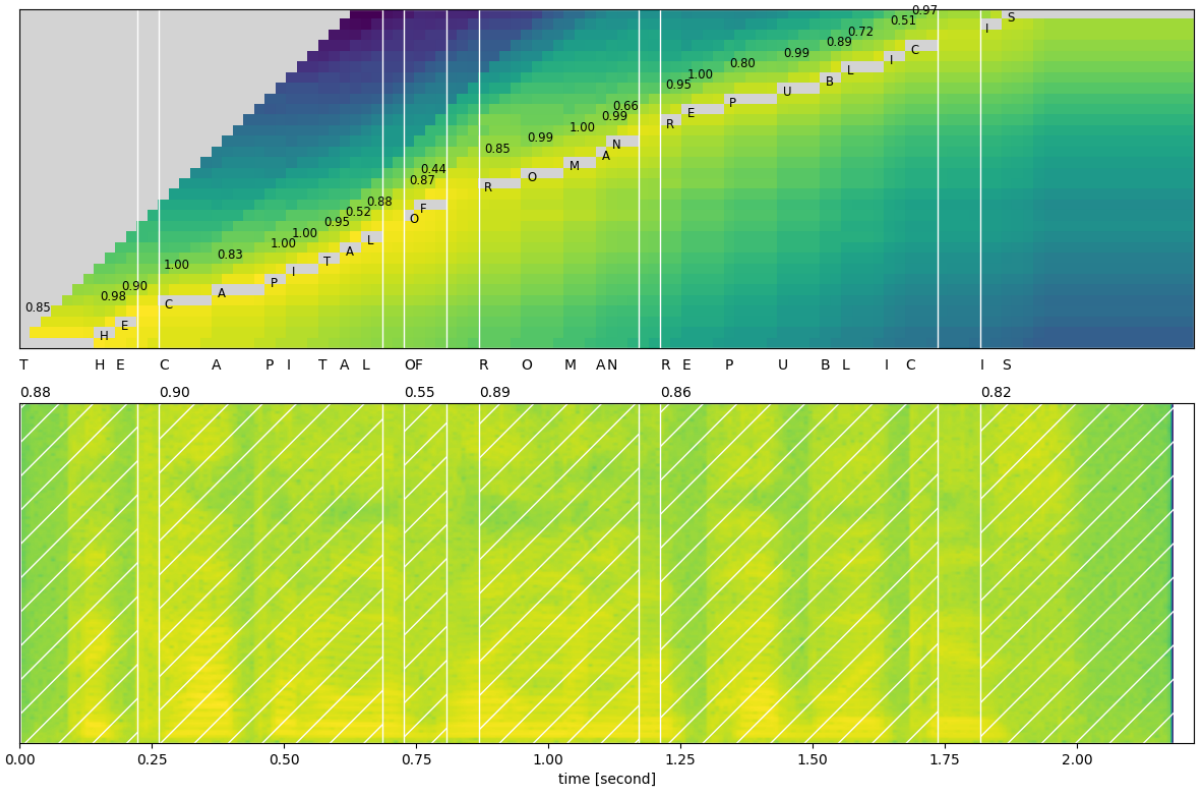

Figure 3: Results of the forced alignment for a speech utterance (transcript: "The capital of Roman Republic is"). The plot on top shows the trellis matrix, with the highlighted optimal path and score for each labeled letter; on the bottom, instead, we show the mel spectrogram of the spoken utterance, with the corresponding boundaries between each (spoken) text token.

range in seconds (s_{start}, s_{end}) in the utterance with the following formula:

$$s_{start} = \frac{\lfloor ratio \cdot f_{start} \rfloor}{sr}, s_{end} = \frac{\lfloor ratio \cdot f_{end} \rfloor}{sr}, \quad (1)$$

Where sr represents the sample rate of the original sampled waveform $Z = (z_1, \ldots, z_M)$, while $ratio = \frac{M}{T}$ represents the number of samples contained in a frame.

- Then, considering the token rate of the speech tokenizer tr and the previously computed time range (s_{start}, s_{end}), the corresponding speech token range (stk_{start}, stk_{end}) is given by:

$$stk_{start} = \lfloor s_{start} \cdot tr \rfloor, stk_{end} = \lceil s_{end} \cdot tr \rceil. \quad (2)$$

This direct mapping provides a clear correspondence between each text token and its associated speech tokens, linking elements of the transcript to their acoustic realizations in the audio.