Contextual ASR with Retrieval Augmented Large Language Model

Cihan Xiao[†⋆], *Zejiang Hou*[‡], *Daniel Garcia-Romero*[‡], *Kyu J Han*[‡]

[†]Center for Language and Speech Processing, Johns Hopkins University, USA
[‡]AWS AI Labs, USA

***Abstract*—Automatic speech recognition (ASR) systems can benefit from incorporating contextual information to improve recognition accuracy, especially for uncommon words or phrases. Current approaches like custom vocabularies or prompting with previous transcript segments provide limited contextual control. Compared to existing context biasing methods, RAG promises more flexible and scalable contextual control by leveraging LLMs' broad knowledge. To this end, we propose leveraging large language models (LLMs) and retrieval-augmented generation (RAG) to enhance the contextual capabilities of ASR systems. Specifically, we propose systems based on text and audio LLMs to perform contextual error correction with context retrieved by querying a text-based retriever using the ASR module's first-pass ASR hypotheses and a frequency-based custom vocabulary (CV) list. Our experiments reveal that the fine-tuned system has effectively learned to extract the relevant context to perform error correction while maintaining robustness against noise.**

***Index Terms*—ASR, RAG, LLM**

I. Introduction

Automatic speech recognition (ASR) systems have made significant strides in recent years, thanks to advances in large neural models and the availability of large speech datasets. However, these systems still struggle with accurate recognition of uncommon words or phrases, especially when they lack sufficient context. Incorporating contextual information can enhance ASR accuracy, but existing approaches like custom vocabularies or providing previous transcript segments offer limited contextual control.

Recent developments in natural language processing, particularly the emergence of LLMs and retrieval-augmented generation (RAG), present an opportunity to address this limitation. LLMs can be leveraged to better understand, extract, and utilize textual context for post-processing and error correction of ASR output. Furthermore, RAG combines the strengths of LLMs and retrieval systems, allowing for more flexible and scalable retrieval of relevant context.

In this paper, we propose leveraging LLMs and RAG to enhance the contextual capabilities of ASR systems. Specifically, we developing systems based on LLMs to perform text-only and audio-grounded contextual error correction. The RAG context is retrieved by querying a text-based retriever using the ASR module's first-pass hypotheses. In addition, we incorporate custom vocabularies extracted via frequency-based methods into the intermediate representations for further enhancement. Our experiments, conducted on the Earnings21/22 and VoxPopuli datasets, demonstrate that the fine-tuned system has successfully learned to extract relevant context from the retrieved information and apply it to correct errors in the ASR hypothesis of a given speech utterance. Moreover, the proposed approach exhibits robustness against noisy context, a crucial factor in real-world scenarios.

⋆ Work done during internship at AWS AI Labs.

II. Related Work

Contextual ASR Previous research has made significant strides in improving contextual ASR through various methodologies. Studies such as [1], [2] have demonstrated notable advancements by leveraging contextual information to enhance ASR performance via different modeling or inference strategies. In addition, [3] introduced a novel approach that injects textual information to the audio encoder utilizing nonparallel text corpus and [4] proposed a method capable of handling a large-scale catalog with 20K custom entities. While these approaches have achieved commendable results, they inherently limit the versatility of the system due to their dependency on crafted custom vocabulary lists, thereby constraining their ability to process more complex and structured contexts.

Retrieval-Augmented Generation (RAG) The ability to dynamically incorporate external information makes RAG an ideal approach for contextual ASR, as it enables the system to handle structured contexts and integrate their rich semantic information. [5] has explored incorporating RAG into ASR systems without using LLMs, and [6] illustrates a cascaded method for constructing cross-modal audio-text RAG systems, showing that first-pass outputs from an ASR module can serve adequately as queries for retrieval.

Audio-LLM Recent progress in audio large language models (Audio-LLMs) has gained significant traction, driven by the remarkable success of text-based LLMs and the growing demand for cross-modal audio and language processing capabilities. [7] has attempted incorporating speech information using a T5-style model, whereas [8] utilizes cross-attention mechanisms to integrate audio signals. Building on top of [7], [9] combined RAG with audio LLMs with a focus on entity retrieval and prompting. Work such as [10] and [11] developed systems that allow interfacing audio signals and text information using LLMs. Additionally, [12] have explored retrieval augmented decoding, biasing the LLM with multi-head attention at different layers of the LLM's internal representations.

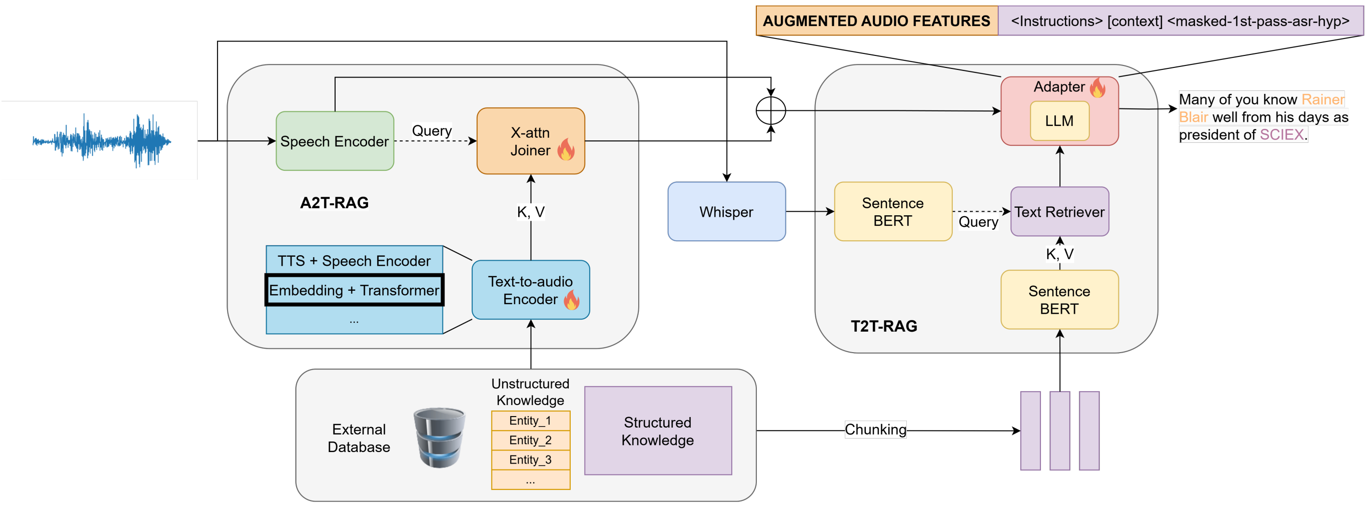

Fig. 1: Hybrid RAG-ASR System. The $\oplus$ symbol represents addition.

III. Methodology

A. Text-based System

Our approach enhances the contextual capabilities of ASR systems by leveraging text-based LLMs and RAG. We use the first-pass ASR hypothesis as a query to retrieve relevant textual context from a document pool, which is then combined with the ASR hypothesis in an instruction template for contextual error correction by an LLM.

We encode the ASR output with a text sentence embedding model and use FAISS to retrieve related context from a chunked text corpus encoded by the same model. The retrieved context is then concatenated with the ASR hypothesis and used to prompt the Mistral-7B-Instruct-v0.3 [13] model.

Our experiments involve two settings: (1) a zero-shot scenario using the pre-trained LLM without modifications, and (2) a fine-tuned approach where the LLM is adapted using LoRA (Low-Rank Adaptation) to reduce hallucination and improve task-specific understanding.

B. Audio-grounded System

In addition to the text-based approach, we propose an audio-grounded system that leverages the capabilities of audio LLMs to better integrate speech modality into the contextual error correction process. Our audio-grounded system utilizes the QWen-audio-chat [14] model, a cross-modal LLM that combines Whisper's [15] audio encoder with the QWen-7B language model as the decoder. The input sequence to the LLM decoder consists of a few key components: 1) the audio representations encoded by the ASR encoder, 2) the retrieved textual context and 3) the first-pass ASR hypothesis, along with the task instruction.

To enhance system robustness and generalization, we employ

- **Context dropout**, which randomly excludes the retrieved textual context from the prompt with a certain probability (default 0.2), forcing the model to rely more on the audio signal and the first-pass ASR hypothesis.
- **First-pass hypothesis masking**, which applies BERT-style masks to tokens in the first-pass ASR hypothesis, with a default probability of 0.4. By randomly masking input tokens, the model is encouraged to leverage the audio signal and retrieve context more effectively, preventing it from acting lazily and directly copy-pasting the first-pass hypothesis.

C. Hybrid System

The audio-grounded approach utilizes both the speech signal and retrieved textual context, but many ASR applications also require handling domain-specific terminologies that are challenging for semantic-based retrievers. To address this, we design a hybrid system that integrates unstructured knowledge sources directly into the audio representation before passing it to the LLM.

The hybrid system introduces a text representation encoder that transforms custom vocabularies into audio-like representations. This text encoder can be designed in various ways, such as using a cascaded TTS and ASR encoder. In our implementation, we use a design consisting of an embedding layer followed by 2 transformer encoder layers. A joiner block with 2 transformer decoder layers then takes in the audio encoder's outputs as queries to cross-attend to these encoded text representations, with a residual connection preserving the original audio information while integrating the additional text knowledge as illustrated in Figure 1. Similar to the audio-grounded system, we apply context dropout also to the CV context, replacing it with a special masking token randomly at training time, to handle situations where the CV context is unavailable.

This design allows the model to combine the speech signal with domain-specific text knowledge for more effective contextual error correction, while also being memory-efficient by avoiding the need to concatenate text representations with the audio sequence.

TABLE I: Experiment results. *ASR-HYP* refers to the input first-pass ASR hypothesis prepended to the decoder for inference. *Context* reflects the context type, e.g. retrieved by the RAG system or the frequency-based custom vocabulary (CV) list. The CV list is not available for the VoxPopuli corpus.

	System	Speech Encoder	Decoder	ASR-HYP	Context	Earnings21			VoxPopuli		
						WER↓	RWER↓	F1↑	WER↓	RWER↓	F1↑
1	Baselines	Whisper	Whisper	-	-	10.92	10.95	0.78	9.59	20.94	0.88
2				-	CV	28.97	27.17	0.64	-	-	-
3		QWen-audio-chat	QWen-7B	-	-	19.11	22.63	0.61	7.43	26.70	0.85
4			QWen-7B	-	RAG	42.48	28.73	0.73	9.28	22.51	0.88
5	Text-based	-	Mistral	Whisper	-	12.51	12.04	0.84	9.48	23.56	0.87
6			Mistral	Whisper	RAG	11.39	11.37	0.80	11.24	23.04	0.87
7			Mistral-FT	Whisper	-	11.36	12.14	0.74	8.68	21.47	0.88
8			Mistral-FT	Whisper	RAG	11.18	11.41	0.87	9.31	16.75	0.91
9	Audio-grounded	QWen-audio-chat	QWen-7B-FT	-	-	14.21	13.38	0.57	7.31	27.75	0.84
10			QWen-7B-FT	-	RAG	13.37	12.08	0.84	5.91	23.04	0.87
11			QWen-7B-FT	Whisper	-	8.99	9.46	0.80	5.65	17.80	0.90
12			QWen-7B-FT	Whisper	RAG	8.65	8.83	0.88	**5.17**	**12.57**	**0.93**
13	Hybrid	QWen-audio-chat	QWen-7B-FT	Whisper	CV	**8.13**	8.59	0.81	-	-	-
14			QWen-7B-FT	Whisper	RAG + CV	8.61	**8.37**	**0.91**	-	-	-

IV. Experiments

A. Experimental Setup

Datasets Our experiments are conducted on two contextual ASR datasets, including the Earnings21/22 [16], [17] and VoxPopuli [18] corpora. Authors of [19] augmented the Earnings21/22 dataset by incorporating real-world contexts. Following the settings in the paper, we use Earnings22 as the training data and Earnings21 as the test data. We concatenate utterances from the same recordings by their chronological order for efficient training.

Pre-processing For each utterance in the Earnings21/22 corpus, we use the context files associated with the corresponding recording as the RAG database for retrieval. In the case of VoxPopuli, we use a single corpus-level database, which consists of the English portion of the Europarl [20] corpus, a non-parallel corpus of parliamentary proceedings. Unless stated otherwise, we default to chunking the context file into segments of 320 characters with a 20-character overlap using LangChain. The chunked text is then embedded using the ALL-MPNET-BASE-V2 [21] model, retrieved based on the top-4 results according to FAISS's [22] cosine similarity metric.

Evaluation To compute the RWER and F1 scores, rare words are identified by their frequency count in the training set and their occurrence in the evaluation set, filtered by an occurrence threshold. The authors of [19] augmented the Earnings21 data by replacing the `inaudible` and `unk` tokens with transcribed words. As we do not have access to this data, we reproduce the Whisper baseline using the original dataset, applying word-level Whisper normalization without deletion to align word tags for compatibility, as done for all of our results. For the hybrid system, the training CV list is crafted based on each utterance's ground-truth transcript with dynamically sampled distractors, whereas the evaluation CV list is derived solely from the recording-level context document.

Training For the LLM fine-tuning experiments, we adopt the Low-Rank Adaptation (LoRA) [23] technique with rank $= 64$ and $\alpha = 128$ for our experiments, resulting in approximately 163 million trainable parameters out of 8.5 billion parameters for QWen-audio-chat LLM fine-tuning and 170 million out of 7.4 billion parameters for Mistral fine-tuning. For the hybrid system, we initialize the embedding layer of the text encoder with the weights from Whisper decoder's embedding layer and tokenize the input custom vocabularies using the Whisper tokenizer. The training process is conducted in two phases. In phase 1, we freeze the LLM and the audio encoder, training only the text encoder and the joiner. The combined text encoder and joiner have approximately 160 million parameters, comparable to the trainable parameters of the other systems. In phase 2, we apply LoRA fine-tuning on the augmented Earnings22 split detailed in Section IV-C.

B. Results

Our results, displayed in Table I, demonstrate the effectiveness of the proposed systems for contextual error correction in ASR systems. For the text-based system, direct inference using both the context and the first-pass hypothesis suffers from hallucinations. To address this, we employ a length-filtering mechanism that rejects the LLM hypothesis if its length deviates significantly from the first-pass ASR hypothesis. Using only the first-pass hypothesis, effectively allowing the LLM to perform standard error correction, shows some improvement (as seen in row 5). The post-fine-tuning improvements compared to the Whisper baseline on the VoxPopuli dataset suggest that text-only contextual error correction can be effective, as indicated by the performance gap between row 1 and row 8. However, these performance gains appear to be sensitive to the quality of the context, as seen in the degradation observed when the system is trained and evaluated on the Earnings21/22 corpus, whose context is significantly noisier and less structured. These findings suggest that the text-based system's performance is highly dependent on the quality and noisiness of the input context.

The audio-grounded system demonstrates greater robustness and superior performance compared to the text-based system. Despite the mismatch in training and inference setting, the system shows an improvement in the overall WER without any auxiliary information, as evidenced by the results in row 9 of the table. Additionally, the system maintains strong performance with only the retrieved context (row 10), showing improved results compared to the QWen baseline (row 3). The substantial F1 increase (0.27 for Earnings21 and 0.03 for VoxPopuli) showcases that the system is able to effectively perform one-pass contextual ASR as well. Row 11 assesses the model's error-correction capability without context, and a moderate performance gain is observed. Together, rows 11 and 10 provide insight into the contributions of the system's enhanced error correction and context extraction abilities, respectively. Row 12, with input aligning with the training setting, yields strong results across all metrics. In particular, for the VoxPopuli corpus, this system stands out with a 4.42 absolute reduction in WER, an 8.37 reduction in RWER, and a 0.05 increase in F1. As for the Earnings21 test set, the system achieves a further 0.01 F1 increase compared to the text-based system, while also out-performing the Whisper baseline with a 2.27 WER reduction and a 2.12 RWER reduction, suggesting the proposed system's strong overall performance and robustness to noise.

Row 13 evaluates the system when only the CV context is provided, whose biasing is driven solely by the unstructured entity list. This setup results in a 0.87 reduction in RWER and a 0.01 increase in F1 compared to the context-less error correction model in row 11. When the model is also provided with the RAG context during inference, it achieves the lowest RWER of 8.37 and the highest F1 score of 0.91. This suggests that the inclusion of the retrieved structured context leads to a further 0.22 reduction in RWER and a 0.03 increase in F1 score. However, this improvement comes with a 0.48 increase in general WER, likely caused by the additional noise introduced by the retrieved context.

TABLE II: Hit-rate analysis of Earnings21. By default, we set the retrieval top-k $= 4$.

Context	Chunk Size	Chunk Overlap	Query	Hit-rate↑
Oracle	-	-	-	69.02
RAG	50	0	Whisper-ASR	28.53
	320	20	Whisper-ASR	42.08
	500	20	QWen-Audio-Chat	42.98
	500	20	Whisper-ASR	44.22
	500	20	Oracle	45.50
	1000	20	QWen-Audio-Chat	47.71
	1000	20	Whisper-ASR	48.15
	1000	20	Oracle	48.44
CV + RAG	320	20	Whisper-ASR	**54.12**

C. Ablation Study

Hit-rate We compute the context hit-rate, defined as the ratio of the number of rare words that appear in the provided context to those present in the target utterance, under various settings for analysis, as shown in Table II. The first row of Table II represents the upper bound of the hit-rate for the Earnings21 dataset, where the numerator includes all rare words present in the full context document. This upper bound indicates that approximately 30% of the rare words are not covered by the context. As we increase the chunk size of the context document, there is a corresponding rise in the hit-rate. However, this improvement diminishes as the chunk size reaches a certain point, suggesting a saturation point beyond which larger chunks do not substantially enhance the contextual coverage. We also examine the impact of the retrieval query quality on the RAG system's context. Results show that queries with a higher WER cause a slight reduction in the hit-rate compared to using the oracle transcript as the query. However, this degradation is relatively minor. Finally, the last row of the table demonstrates that adding the CV list further boosts the hit-rate. This indicates that the proposed hybrid system is an effective strategy for incorporating contextual information.

TABLE III: Results of the context-augmented Earnings21/22 experiments.

Train Context	WER↓	RWER↓	P↑	R↑	F1↑
Original	**8.65**	**8.83**	0.96	0.82	0.88
Augmented	8.74	8.81	**0.97**	**0.86**	**0.91**

Context Augmentation Given the limited availability of context in the Earnings22 split (43 out of 125 recordings), we experiment with a context augmentation strategy that synthesizes additional context documents by prompting the Claude-3.5-Sonnet [24] LLM to generate background documents based on the ground-truth transcript. We then train an audio-grounded system with the augmented context documents, whose performance is presented in Table III. The augmented system shows improved performance in the contextual error correction task compared to its un-augmented counterpart in terms of the RWER and F1 metrics. However, this improvement comes with a minor growth in general WER by 0.09, likely due to the model's increased reliance on the context, including any noise that may be present.

V. Conclusions

Our work introduces a novel approach to enhancing ASR systems by integrating LLMs with RAG and a hybrid system capable of handling both structured and unstructured contexts. The proposed audio-grounded approach leverages the speech signal along with retrieved textual context, while the hybrid system integrates additional unstructured knowledge directly into the audio representation using a soft attention. Our key contribution lies in providing a scalable solution that is compatible with various types of context, both structured and unstructured, without relying on hand-crafted custom vocabulary lists, thus making it suitable for real-world applications. Experimental results on the Earnings21 and VoxPopuli datasets demonstrate that our method enables effective use of external text information, improving recognition accuracy for uncommon words and domain-specific terms.

REFERENCES

[1] Ruizhe Huang, Mahsa Yarmohammadi, Sanjeev Khudanpur, and Daniel Povey, "Improving neural biasing for contextual speech recognition by early context injection and text perturbation," 2024.

[2] Jiyang Tang, Kwangyoun Kim, Suwon Shon, Felix Wu, Prashant Sridhar, and Shinji Watanabe, "Improving asr contextual biasing with guided attention," 2024.

[3] Zhong Meng, Zelin Wu, Rohit Prabhavalkar, Cal Peyser, Weiran Wang, Nanxin Chen, Tara N. Sainath, and Bhuvana Ramabhadran, "Text injection for neural contextual biasing," 2024.

[4] Sai Muralidhar Jayanthi, Devang Kulshreshtha, Saket Dingliwal, Srikanth Ronanki, and Sravan Bodapati, "Retrieve and copy: Scaling asr personalization to large catalogs," in *EMNLP 2023*.

[5] Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Karen, Zeeshan Ahmed, Dinesh Manocha, and Xuedong Zhang, "DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, Eds., Torino, Italia, May 2024, pp. 5132–5139, ELRA and ICCL.

[6] Ryan O'Connor, "Retrieval augmented generation on audio data with langchain and chroma," https://www.assemblyai.com/blog/retrieval-augmented-generation-audio-langchain/, 2023.

[7] Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Yongqiang Wang, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul Rubenstein, Lukas Zilka, Dian Yu, Zhong Meng, Golan Pundak, Nikhil Siddhartha, Johan Schalkwyk, and Yonghui Wu, "Slm: Bridge the thin gap between speech and text foundation models," 2023.

[8] Zhehuai Chen, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Nithin Rao Koluguri, Piotr Żelasko, Jagadeesh Balam, and Boris Ginsburg, "Bestow: Efficient and streamable speech language model with the best of two worlds in gpt and t5," 2024.

[9] Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey, "Retrieval augmented end-to-end spoken dialog models," 2024.

[10] Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff, "Speechverse: A large-scale generalizable audio language model," 2024.

[11] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer, "Prompting large language models with speech recognition abilities," 2023.

[12] Jari Kolehmainen, Aditya Gourav, Prashanth Gurunath Shivakumar, Yi Gu, Ankur Gandhe, Ariya Rastrow, Grant Strimel, and Ivan Bulyko, "Multi-modal retrieval for large language model based speech recognition," in *ACL 2024*.

[13] 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, "Mistral 7b," 2023.

[14] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," 2023.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[16] Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté, "Earnings-21: A practical benchmark for asr in the wild," in *Interspeech 2021*.

[17] Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra, "Earnings-22: A practical benchmark for accents in the wild," 2022.

[18] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," 2021.

[19] Ruizhe Huang, Mahsa Yarmohammad, Jan Trmal, Jing Liu, Desh Raj, Leibny Paola Garcia, Alexei V. Ivanov, Patrick Ehlen, Mingzhi (Emily) Yu, Ariya Rastrow, Daniel Povey, and Sanjeev Khudanpur, "Conec: Earnings call dataset with real-world contexts for benchmarking contextual speech recognition," in *LREC-COLING 2024*.

[20] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X: Papers*, 2005, pp. 79–86.

[21] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu, "Mpnet: Masked and permuted pre-training for language understanding," 2020.

[22] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou, "The faiss library," 2024.

[23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," 2021.

[24] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," 2024.