BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

FIT-Voice: A RAG-based speech dialogue system for realtime academic information

Sathvik Udupa, Petr Schwarz

Abstract

Cascaded voice pipelines incur 1-2 second response latencies and hallucinate on domain-specific queries. We propose endpoint anticipation - speculative LLM generation triggered by learned turn-end forecasting -combined with context-aware RAG integration. Anticipation reduces latency by 33% (1.2s → 0.8s) and recovers RAG-augmented latency from 1.7s to 1.3s; context-aware chunking improves retrieval hit rate by up to 54 percentage points. These techniques are realised in FIT-Voice, a real-time academic voice assistant for FIT BUT, with a question answering benchmark.

*udupa@fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Advancements in spoken dialogue systems [1, 2] have driven widespread adoption of speech technologies across voice assistants [3], speech LLMs [4], customer support, and emergency services. A cost-effective, modular approach employs a cascaded pipeline: a streaming ASR continuously transcribes incoming speech, whose hypotheses are asynchronously prefilled into an LLM; upon end-of-turn detection [5], the LLM halts prefill and begins generation, streaming responses to a TTS system [6]. Despite heavy optimization, such pipelines incur response latencies of 1–2 seconds - far exceeding the 250 ms latency of natural human conversation [7]. Beyond latency, reliability is equally critical: relying solely on an LLM is suboptimal for domain-specific queries, where knowledge gaps can lead to hallucinations.

This work addresses both limitations with two concrete contributions. For latency, we propose the first anticipation-based speculative generation technique for cascaded voice pipelines: a model estimates end-of-turn probability within a short horizon and speculatively triggers LLM generation on the partial transcript, discarding incorrect anticipations silently and playing correct ones immediately. This requires no modification to the underlying ASR, LLM, or TTS components, making it broadly applicable to any cascaded pipeline. For accuracy, we integrate Retrieval-Augmented Generation (RAG) [8], grounding LLM responses in a curated domain-specific knowledge base with documents retrieved at runtime.

We implement both techniques within the Unmute [1] spoken dialogue system as modular APIs, and demonstrate FIT-Voice on real-time spoken question answering over FIT BUT academic content, with evaluation of response latency and a RAG-based text question answering (QA) benchmark.

2. Proposed methodology

This section details the endpoint anticipation module, RAG construction, and their integration into the voice system.

2.1 Endpoint Anticipation

Given a speaker turn, we train a model to forecast end-of-turn (EOT) within a defined horizon h (e.g., 960 ms). Incoming speech is processed by a streaming encoder - Mimi [9] at 12.5 Hz - and a frame-level binary classifier is trained on top: frames in $[0, \text{EOT} - h)$ are labeled 0 and frames in $[\text{EOT} - h, \text{EOT}]$ are labeled 1. At inference, a probability threshold triggers speculative generation.

2.2 RAG Data Curation

We identify FIT BUT webpages[2] across categories including personnel, courses, programmes, publica-

[1] https://github.com/kyutai-labs/unmute
[2] https://www.fit.vut.cz/

tions, and projects, and build a targeted web scraping pipeline handling their varied formats. Scraped content is structured into JSONs preserving hierarchical metadata (e.g., person →research group), enabling context-aware retrieval.

2.3 RAG System

We adopt a vector database [10] using pretrained English embeddings [11], with documents embedded in metadata-preserving chunks and hosted as a lightweight low-latency retrieval API. Retrieval optimisation is left to future work.

2.4 RAG benchmark

To evaluate retrieval quality, we construct a text question answering benchmark over the FIT BUT database. For structured data, questions are procedurally generated from JSONs with paraphrased variants (e.g., *Who teaches course X?*). For longform answers (e.g., *What are the learning objectives of course X?*), reference answers are summarized using Olmo 3.1 32B [12]. Evaluation is performed using an LLM-as-judge approach [13] with Qwen 2.5 4B [14], scoring responses on a 1–5 scale, where 5 denotes a correct and complete answer, and 1 denotes an incorrect or absent one.

2.5 System Integration

We build on the open-source Unmute spoken dialogue system, hosting endpoint anticipation and RAG as independent APIs integrated via WebSocket. Unmute provides a live voice-to-voice interface accessible over the browser. For reproducible evaluation, we additionally implement a file-based inference mode that processes audio from disk and saves generated responses, from which latency metrics are extracted.

3. Results and discussions

Endpoint anticipation outperforms the VAP baseline [15]: EPA-S achieves an MRA of 640 ms and HEA of 66.3% vs. VAP's 19.2% at the same horizon (Table 1 in poster). The large HEA gap reflects that VAP, being an unsupervised turn-taking module, does not learn precise anticipation timings - whereas EPA-S learns turn-final speech patterns, enabling reliable early triggering.

Table 2 shows system latency evaluated on Full Duplex Bench [16]. Anticipation reduces Unmute latency from 1.2s to 0.8s (33%), approaching the ∼250 ms of natural human conversation. Notably, RAG integration alone inflates latency to 1.7s due to retrieval overhead at the turn boundary; anticipation recovers this to 1.3s by overlapping retrieval with the remaining speech. This confirms that anticipation is practically beneficial for real-time RAG-augmented voice systems.

Table 3 presents RAG benchmark results comparing default vs. context-aware chunking. Context-aware chunking (preserves hierarchical metadata during document chunking) yields substantial hit rate gains across most categories. The improvement comes from richer chunk context, allowing the retriever to better discriminate relevant documents. Longform questions remain challenging regardless of chunking strategy, as answers span multiple documents and require synthesis rather than direct retrieval. LLM-as-judge scores improve from 1.92 to 2.65 (top-1) and 2.08 to 2.64 (top-5) overall, confirming that metadata-preserving chunking meaningfully improves end-to-end answer quality. Clear headroom remains through re-ranking and retrieval-aware generation. Note that the benchmark evaluates text-based retrieval; in practice, voice-based queries are additionally affected by ASR transcription errors, which represent a further source of degradation not captured here.

4. Conclusions

Spoken dialogue systems face two fundamental challenges: high response latency in cascaded pipelines, and limited reliability on domain-specific queries. This work presents targeted contributions to both. To address latency, we propose endpoint anticipation - a novel turn-end forecasting method that enables speculative LLM generation - reducing baseline response latency by 33% (1.2s → 0.8s) and keeping RAG-augmented systems within practical bounds (1.7s → 1.3s). To address reliability, we introduce context-aware RAG integration, improving retrieval hit rate by up to 54 percentage points in structured categories such as Courses and Projects, with overall LLM-as-judge scores improving from 1.92 to 2.65 (top-1). Together, these contributions are realised in FIT-Voice, a realtime voice question answering system over FIT BUT academic content, accompanied by a QA benchmark. Future work will focus on retrieval re-ranking and mesuring spoken QA correctness.

Acknowledgements

I thank Dr. Petr Schwarz for supervision and Dr. Shinji Watanabe for collaboration on endpoint anticipation research.

References

[1] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. WavChat: A Survey of Spoken Dialogue Models. *arXiv preprint arXiv:2411.13577*, 2024.

[2] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*, 2025.

[3] Lawal Ibrahim Dutsinma Faruk, Mohammad Dawood Babakerkhell, Pornchai Mongkolnam, Vithida Chongsuphajaisiddhi, Suree Funilkul, and Debajyoti Pal. A review of subjective scales measuring the user experience of voice assistants. *IEEE Access*, 2024.

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[5] Sathvik Udupa, Shinji Watanabe, Petr Schwarz, and Jan Cernocky. Streaming endpointer for spoken dialogue using neural audio codecs and label-delayed training. *IEEE ASRU*, 2025.

[6] Tatiana Likhomanenko, Luke Carlson, Richard He Bai, Zijin Gu, Han Tran, Zakaria Aldeneh, Yizhe Zhang, Ruixiang Zhang, Huangjie Zheng, and Navdeep Jaitly. Chipchat: Low-latency cascaded conversational agent in mlx. *IEEE ASRU*, 2025.

[7] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

[9] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. (arXiv:2410.00037), October 2024.

[10] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library.(2024). *arXiv preprint arXiv:2401.08281*, 2024.

[11] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. SIGIR '24, page 641–649, New York, NY, USA, 2024. Association for Computing Machinery.

[12] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

[14] A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint*, 2024.

[15] Erik Ekstedt and Gabriel Skantze. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, pages 5190–5194, 2022.

[16] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *IEEE ASRU*, 2025.