Enhancing Alphanumeric Sequence Recognition In Speech-To-Text Using Multimodal Large Language Models: A Proposal

Leon Behrndt

Reviewer: Prof. Dr. Ulf Leser
Supervisor: Stefan Ostwald

Abstract

The accurate recognition of alphanumeric sequences, such as customer IDs or reference numbers, from low-quality phone audio is a significant challenge for traditional Speech-to-Text (STT) systems. This limitation frequently hinders the automation capabilities of AI agents in call center environments, particularly during user authentication. This thesis proposes to investigate the hypothesis that Multimodal Large Language Models (MLLMs) can achieve a higher accuracy in this task compared to traditional STT systems by leveraging contextual information about the expected sequence format. We will conduct a comparative analysis of state-of-the-art MLLMs against traditional STT baselines across English, German, and French. The primary metric for success will be analysis of latency, supplemented by the Named Entity Recogniton (NER) and cost trade-offs. This research aims to provide a robust evaluation of MLLMs for a critical real-world use case and quantify the performance gains achievable through context-awareness. This thesis will explore a novel approach to address this challenge by leveraging the advanced capabilities of Multimodal Large Language Models.

Motivation

In the domain of call center automation, AI agents are increasingly tasked with complex, multi-turn conversations. Current architectures typically employ a three-step pipeline: 1) a Speech-to-Text (STT) service transcribes audio, 2) a Large Language Model (LLM) processes the text for understanding and decision-making, and 3) a Text-to-Speech (TTS) service generates the response. Each step in this cascade adds latency, which is detrimental to the user experience in real-time conversations (Arora et al. 2022; Wang et al. 2024). A key motivation for this research is to investigate the potential of collapsing the first two steps into a single, efficient process using a Multimodal LLM. (Cui et al. 2025; Peng et al. 2025)

A critical step in many of these interactions is the authentication of the user, which almost invariably requires the user to state an alphanumeric identifier like a customer ID, case number, or booking reference. The audio quality in telephony is often poor due to factors like network compression, background noise, and diverse speaker characteristics.

Traditional STT systems, while proficient at transcribing general conversation, struggle significantly with the nuances of human-stated alphanumeric sequences. (de Zuazo et al. 2025) The challenge extends beyond simple recognition and into interpreting complex conversational behaviors. Common failure modes include:

- Homophone Errors: Misinterpreting a letter as a similar-sounding word, especially at the start of an utterance (e.g., transcribing the letter 'C' as the word 'see').
- Complex User Utterances: Users rarely state identifiers in a clean, isolated manner. They often use phonetic clarifications ("N as in North Pole"), add filler words ("...and the last number is 4"), or perform self-corrections ("my number is 100... oh, wait, sorry, it's 99").

To overcome these limitations, recent research has emphasized the advantages of multimodal integration. Surveys have shown that Multimodal Large Language Models (MLLMs) can leverage both audio and visual inputs to enhance transcription accuracy in complex settings such as scientific lecture videos, achieving substantial gains over unimodal baselines (Wang et al. 2024). Moreover, the latency and error propagation issues inherent in cascade architectures—particularly for knowledge-rich, real-world speech tasks—underscore the need for more compact and integrated approaches (Peng et al. 2025).

Research Question And Hypothesis

Research Question

Can a single–stage Multimodal Large Language Model (MLLM) replace the conventional two-stage STT $\rightarrow$ LLM pipeline for extracting alphanumeric identifiers from noisy telephony speech contribute to its performance?

Hypotheses

Hypothesis H1 tests the primary merging goal, while H2 evaluates schema conditioning. Table 1 summarizes the two core hypotheses and how they will be empirically evaluated.

State of the Art

Classical STT Architecture

Traditional STT systems typically consist of an acoustic model, a pronunciation model, and a language model. The language model's role is to ensure the transcribed output is linguistically probable. While some systems allow

ID	Focus	Precise Statement	Evaluation
H1 – Merging	System–level comparison	*An end-to-end MLLM operating directly on raw audio will achieve equal or higher extraction accuracy than the baseline STT → LLM pipeline while eliminating one inference step.*	Commpares baseline accuracy and latency with MLLM approach on matched test sets across all difficulty levels.
H2 – Schema Benefit	Structured prompt	*Providing the MLLM with an explicit regular-expression schema of the target identifier increases extraction accuracy compared with the same MLLM prompted only with a generic instruction.*	Measure improvement in both exact-match rate and NER Success Rate with vs. without schema-based context.

Table 1: Hypotheses derived from the research question.

for context adaptation through mechanisms like "decoding" or "boosting" (e.g., increasing the probability of specific phrases), they often lack the flexibility to handle complex structural patterns. The prompting capabilities of models like OpenAI's Whisper represent a step towards better context utilization, but they are still fundamentally specialized Automatic Speech Recognition (ASR) models. (Cui et al. 2025; de Zuazo et al. 2025; Peng et al. 2025)

Multimodal Large Language Models

MLLMs like GPT-4o represent a paradigm shift. By integrating different data modalities, including audio, into a single unified model, they can process sound directly in the context of a rich textual prompt. Audio is typically processed through a specialized tokenizer that converts the waveform into a sequence of embeddings that the model can understand (Peng et al. 2025; Cui et al. 2025). This architecture allows the model to perform tasks like transcription and NER simultaneously, theoretically enabling it to use the contextual information in the prompt more effectively to guide the audio recognition process (Jannet et al. 2015).

Metrics: WER vs. NER Success Rate

The standard metric for STT performance is Word Error Rate (WER), which measures errors at the word level (Peng et al. 2025; Cui et al. 2025). However, for use cases like authentication, the overall transcription's accuracy is less important than the accuracy of a single, critical entity. This proposal therefore focuses on NER Success Rate, defined as the system's ability to correctly identify and extract the specific alphanumeric sequence (the entity and its value) from the audio (Raghuvanshi et al. 2019; Cui et al. 2025).

Methodology

Data Generation

Sources and responsibility. Data will be prepared by Parloa and the author. We combine

1. *real "call-center" snippets*
   - already available in English, German and French, but with different patterns (not only alphanumerical sequences) and
2. a *synthetic corpus*. These resources are currently available:
   - In creation: alphanumeric sequence patterns with Azure TTS (without prompts) and
   - a repository that supports generation with prompting via GPT4o mini (OpenAI).

Difficulty levels and patterns. Three alphanumeric patterns provide rising complexity:

1. *Numeric* (e.g. "855432") – baseline difficulty.
2. *Simple alphanumeric* ("CFR789") – medium.
3. *Conversational mixed* – hard: embedded IDs with clarifications, fillers, self-corrections ("my code is 100 . . . oh sorry it's 99").

Speaker variety. Synthetic clips will be used with randomised styles (age, gender, accent, speaking rate). At least ten distinct TTS voices per language will be used to emulate realistic variability.

Degradation pipeline. All audio (real and synthetic) is passed through SoX to apply the G. 711 narrow-band codec and injected background noise at 6–12 dB SNR to mimic telephony conditions.

Model Selection

Baseline Pipeline

1. **STT stage.** The study will benchmark *the three best-performing commercial STT APIs* (to be finalized after internal consultation). All candidates will be invoked in their *general-purpose* mode, with optional phrase biasing enabled only when identical functionality is available between providers.
2. **Text-LLM stage.** GPT-4o is retained as the normalisation/extraction component; the RegExp schema is supplied in the system prompt.

Status. At the time of writing only Azure STT is fully connected in the repository. However,, the connection of these services via endpoints is considered a low-effort task.

Experimental MLLMs At least three state-of-the-art models that accept raw audio and a text prompt will be evaluated.

Experimental Setup & Evaluation

Context injection. For every call the regular-expression schema is concatenated to the system prompt:

```
    <SYS> Extract the customer ID
matching {REGEX}. Return only the
canonical form.
```

For the baseline pipeline the schema is passed only to the text-LLM; for MLLMs it accompanies the audio in one request.

Metrics

- **Primary**: *end-to-end latency*, defined as wall-clock time from API request to receipt of the *final token* for the turn. (Time-to-first-token is no longer reported; streaming is disabled for all runs.)
- Monetary cost is logged as before (API invoices or GPU-hour estimates).
- **Secondary**: exact-match NER success rate after canonical normalisation (uppercase, no spaces). Entity-level correctness of the extracted sequence (i.e., whether each expected token — letter or digit — is correctly recognized and ordered).

Analysis. Results will be reported per language and per difficulty level.

Discussion

A key part of the thesis will be a discussion on the real-world trade-offs between the different approaches. This analysis will go beyond pure accuracy and create a multi-dimensional comparison, plotting the latency against NER Success Rate and cost for each model. This will provide valuable insights for businesses looking to implement these technologies, as the optimal choice may involve a trade-off between performance and operational constraints.

Outlook & Future Research

This study is conducted in a relatively isolated experimental setup. Future research could investigate how the performance of the MLLM is affected when the audio recognition prompt is embedded within a much larger and more complex prompt for a full-fledged conversational agent. It is conceivable that prompt "crowding" could degrade the performance of this specific task, which would be a critical factor for real-world deployment.

Furthermore, if this study confirms the viability of the MLLM-based approach, it would provide a strong justification for re-architecting production AI agents. Future work would involve integrating this end-to-end model into a live system to quantify the real-world latency savings gained by moving from a three-step (STT $\rightarrow$ LLM $\rightarrow$ TTS) to a two-step (MLLM $\rightarrow$ TTS) pipeline.

References

Arora, S.; Dalmia, S.; Chang, X.; Yan, B.; Black, A.; and Watanabe, S. 2022. Two-Pass Low Latency End-to-End Spoken Language Understanding. arXiv:2207.06670.

Cui, W.; Yu, D.; Jiao, X.; Meng, Z.; Zhang, G.; Wang, Q.; Guo, Y.; and King, I. 2025. Recent Advances in Speech Language Models: A Survey. arXiv:2410.03751.

de Zuazo, X.; Navas, E.; Saratxaga, I.; and Rioja, I. H. 2025. Whisper-LM: Improving ASR Models with Language Models for Low-Resource Languages. arXiv:2503.23542.

Jannet, M.; Galibert, O.; Adda-Decker, M.; and Rosset, S. 2015. How to evaluate ASR output for named entity recognition? .

Peng, J.; Wang, Y.; Li, B.; Guo, Y.; Wang, H.; Fang, Y.; Xi, Y.; Li, H.; Li, X.; Zhang, K.; Wang, S.; and Yu, K. 2025. A Survey on Speech Large Language Models for Understanding. arXiv:2410.18908.

Raghuvanshi, A.; Ramakrishnan, V.; Embar, V.; Carroll, L.; and Raghunathan, K. 2019. Entity resolution for noisy ASR transcripts. In Padó, S.; and Huang, R., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 61–66. Hong Kong, China: Association for Computational Linguistics.

Wang, M.; Wang, Y.; Vu, T.-T.; Shareghi, E.; and Haffari, G. 2024. Exploring the Potential of Multimodal LLM with Knowledge-Intensive Multimodal ASR. arXiv:2406.10880.