Multimodal Language Learning (MLL)

by
Azfar Sulaiman

Envisioning the Future of Computing Prize
Social and Ethical Responsibilities of Computing
Massachusetts Institute of Technology

Multimodal Language Learning (MLL)

Summary

The document elaborates on the Multimodal Language Learning (MLL) tool, focusing on its comprehensive approach to language learning through diverse communication modes, including text, graphics, voice, and video. It aims to provide culturally relevant learning experiences tailored to international students and non-native speakers. The MLL's design incorporates a varied media corpus and utilizes Retrieval Augment Generation for contextually accurate responses. The tool's effectiveness is showcased in scenarios demonstrating improved language skills, cultural understanding, and professional adaptability. Challenges such as potential oversimplification, dependency on technology, and data privacy concerns are prevalent though can be minimal. The paper suggests MLL as an enhancement to traditional language learning methods and underscores its potential in making language education more accessible and equitable.

Disclaimer: GPT-4 was only used to verify French and Japanese translations into English. The ideas, research, examples, and content printed here is my own, unless referenced explicitly.

Word Count: 3,000

Introduction

Gaining proficiency in a new language, particularly for adult learners, presents a formidable challenge. Yet, each year hundreds of thousands of students pursuing education in different countries have to navigate this complex landscape, compounded immensely if they are non-native speakers. Even after grueling applications process, many foreign students are asked to adapt to their new ecosystems, usually without adequate. Many students rely on either sporadic and expensive tutoring sessions or digital apps focused on habit formation[1] than on practical conversational styles. Furthermore, the adjustment is intensified by cultural shock, as local norms and values can significantly differ from those in their home countries. Drawing from my experiences as an international student and interactions with fellow non-native speakers, I have observed firsthand the impact of these challenges on natural social integration. People have different conversation styles, accents, expressions, and communicate using cultural and regional expressions that places additional strain on foreign students.

In this paper, I propose and evaluate multimodal language learning (MLL) tool that aims to help people learn foreign languages through contextual settings. The need to incorporate many forms of communication to improve the learning experience is emphasized by several theoretical frameworks that are the foundation of multimodal approaches in language instruction[2]. While this tool hasn't been developed yet, I outline possibilities in developing this technology, how it supports user's social and linguistic development and how it could prove to be net-positive benefits.

What is Multimodal Language Learning?

MLL is a personalized language reference mobile app that uses multimodal AI generation (such as computer vision, text-to-speech, image-to-text etc.) mechanisms to provide integrated, guided, and contextualized language skills development, geared towards international college

[1] https://blog.duolingo.com/how-duolingo-streak-builds-habit/
[2] Anis & Khan. "Integrating Multimodal Approaches in English Language Teaching for Inclusive Education: A Pedagogical Exploration" 2023. https://philarchive.org/archive/ANIIMA

students and non-native speakers. The goal is to supplement individual's learning capabilities through nuanced, culturally relevant, and multi-modal aspects to ensure users are more informed and appreciative of various norms. Leading academics have shown that meaningful connections among students are facilitated by collaborative activities that use a variety of communication channels, which improved language learning outcomes and created a supportive learning community[3]

For a particular language, the app iteratively learns the person's skill level to design journeys, visualizes scenarios for the user and provides region specific recommendations on user interactions such as common expressions, accents, and cultural values.

Architecture and Data

MLL would utilizes a sophisticated computational layer and advanced data training. It processes various media to communicate effectively with users, adapting its output to the most relevant mode. Multimodality will represent a key milestone in the coming years with reasoning and reliability being key differentiators[4] and MLL aims to advance that vision. This ambition necessitates the integration of advanced models like GPT-4, which contain trillions of parameters. While this is a computationally expensive pursuit, research is ongoing on open-source models nearing GPT–4 levels[5]. Moreover, MLL's architecture would be designed to learn not only from user interactions but also to gain deeper insights into language nuances.
MLL would also use the individuals 's geographical location via phone GPS to provide region-specific information.

For the training and fine-tuning of the generative AI models, the goal is to be contextually relevant and updated data as possible. MLL would require a wide publicly available corpus of

[3] Wang, S. H., & Lin, S. H. (2016). Collaborative multimodal learning in an online language course: A case study of an adult English as a second language course. Computers & Education, 92-93, 130-143
[4] Sam Altman, Episode 6: Sam Altman, 2024 https://www.youtube.com/watch?v=PkXELH6Y2lM
[5] Investor, The Pareto. "Mistral CEO Confirms 'Leak' of New Open-Source AI Model Nearing GPT-4 Performance." Medium (blog), February 2, 2024. https://medium.com/@pareto_investor/mistral-ceo-confirms-leak-of-new-open-source-ai-model-nearing-gpt-4-performance-82653b2b36e2

media such as literature, folklore, music, art, and video data, geographically labelled to provide region-specific context. This ensures diversity and depth of information available for the models. To manage high computational costs and keep MLL current, academic research has explored Retrieval Augment Generation (RAG) as a promising technique. Retrieval-augmented generation (RAG) improves the quality of responses by grounding AI models on external knowledge sources to update its internal information repository[6] and enhance trustworthiness. Advanced techniques such as text clustering and hierarchical tree generation, allow models to answer questions effectively and efficiently at different levels[7].

MLL System in Practice

I present three settings that illustrate the importance of MLL in supporting social and linguistic endeavors.

A. Refine linguistic skills through real-time feedback

Consider, for instance, a foreign exchange student Maria, grappling with the intricacies of French subjunctive mood—a notoriously difficult aspect for non-native speakers – while at a dinner party in Paris. Her regular apps lack the capability to provide guided, specific feedback on her difficulties with the subjunctive mood, particularly in identifying the scenarios where she most frequently makes mistakes. MLL would help her identify this specific area through its analysis of Maria's previous interactions and exercises. Additionally, MLL would develop a profile of her strengths and weaknesses as embeddings to then guide her learning and skill development in a more customized manner.

Maria would activate the MLL system to provide her real-time feedback. MLL would have the option to listen in through a speech-to-text modality and can check either only for errors or follow entire conversation. Here is how MLL might help:

---

[6] IBM Research Blog. "What is retrieval-augmented generation?" (2021, February 9). https://research.ibm.com/blog/retrieval-augmented-generation-RAG

[7] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval (arXiv:2401.18059). arXiv. http://arxiv.org/abs/2401.18059

- **Host**: *Maria, pensez-vous qu'il soit important de parler plusieurs langues?* (Maria, do you think it's important to speak multiple languages?)
- **Maria**: *Oui, je pense que c'est important.* (Yes, I think it is important.)
- **MLL:** Try using the subjunctive mood after *'pensez-vous que'*. Your sentence would be: '*Oui, je pense qu'il soit important.*' This is because in French, expressing an opinion, requires the subjunctive.

This scenario exemplifies the tool's ability to provide a rich, immersive learning experience. Maria is not just memorizing rules; she is actively using and understanding the subjunctive in a culturally relevant context. MLL's feedback is immediate and specific, allowing Maria to learn from her mistakes in real-time and gradually build confidence using the right linguistic expression. Maria could quickly learn to speak French fluently, allowing her to make native French friends easily.

B. Enhance cultural curiosity via visual depiction

MLL could augment Maria's learning by visualizing social interaction scenarios of social through video generation. It can offer a virtual avatar feature that represents her and allows her to observe and analyze her own expressions, actions, and interactions in a third-person view. For instance, Maria is planning to visit a bistro but is unsure of waiter interactions and, tipping practices. While she wants to appear respectful, as a graduate student, she is wary of budgetary constraints.

- **Maria:** I need to understand dining and tipping etiquettes at a traditional Parisian restaurant [and selects Image Gen feature]
- **MLL:** The tool provides a realistic simulation of the dining experience.
    i. Meal Process: MLL can offer features such as subtitles and pronunciation tips to assist her conversation. MLL also educates Maria on cultural nuances specific to dining in France, such as the norm of leisurely enjoying meals and the tendency for waiters to offer more space to diners compared to practices in other countries.

ii. Tipping Process: When the meal is over, the simulation demonstrates how the waiter presents the bill. MLL includes an overlay explaining that in France, service charges are usually included in the prices (*service compris*), but a small additional tip is appreciated for excellent service. MLL Cultural Insight: The tool notes that saying "*merci*" or "*c'est pour vous*" (this is for you) when leaving a tip adds an appreciative touch to the gesture.

Through contextualized simulations, MLL aims to reduce Maria's anxiety and potential awkwardness as she can practice her French and gain valuable dining etiquette insights. Maria can watch the simulation at her own pace, rewinding and pausing and inquiring at moments that need further clarity. Additionally, MLL provides valuable insights such as commonality and acceptance of credit card payments in French restaurants, common ingredients in French cuisine, which are beneficial for those with allergies or dietary preferences and typical dress codes for various types of dining establishments.

C. Professional skills development

MLL can assist Maria in preparing for a presentation in a university setting. She can record herself delivering a business presentation titled "Innovations in Sustainable Energy." Maria's unsure about her delivery's effectiveness and audience expectations. MLL processes the video and provides structured feedback:

1. Formality:

- **Maria** (starting her presentation casually): *Salut, je vais parler sur l'énergie durable...*" (Hi, I'm going to talk about sustainable energy...)
- **MLL** (shares a more formal greeting to align with French business etiquette): *Bonsoir, je suis ici pour vous présenter sur les innovations dans l'énergie durable...* (Good evening, I am here to present innovations in sustainable energy...)

2. Tonal Emphasis:

- **Maria** consistently maintains the same tone and pace, making her speech monotonous
- **MLL:** Maybe if you could vary your tone and slowing down for emphasis, while introducing new concepts. Here's an example with pause

- **MLL:** *L'énergie solaire,* pause *une forme d'énergie révolutionnaire...* (Solar energy, *pause* a revolutionary form of energy...)

3. Non-Verbal Communication:

- **Maria** uses hand gestures freely, which may be perceived as overly casual in a French business setting and avoids eye contact with the camera, looking down at her notes.
- **MLL**: Try minimizing hand gestures and maintain a formal posture to align with typical French business presentation styles. Also, eye contact is important in engaging with your audience and reading from a script might counter-productive.

4. Audience Considerations:

- **MLL:** The audience may provide feedback that is curt and direct, different to what you may be used to so this will be important to consider and understand you have done a good job!

MLL could provide Maria with specific and tailored advice in refining her presentation abilities, aligned with the standard expectations of a French professional audience. MLL's scope extends beyond linguistic proficiency, encompassing both verbal and non-verbal communication cues. Consequently, this ensures that her presentation is not only comprehensible but also elicits the desired reception, a testament to the multifaceted capabilities of MLL.

Potential Concerns

Understandably, challenges persist in devising an alignment process that is equitable, devoid of bias, transparent, and encompasses effective accountability mechanisms [8] and MLL is also not immune to potential hallucinatory outputs, algorithmic biases, and unethical AI utilization. Recent scholarly research underscores the efficacy of such a holistic and conscientious approach in mitigating the risks associated with model tuning and progression[9]. While each of

[8] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155

[9] X. Ferrer, T. v. Nuenen, J. M. Such, M. Coté and N. Criado, "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," in *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 72-80, June 2021, doi: 10.1109/MTS.2021.3056293

these concerns is worthy of an exhaustive analysis, my analysis will explore specific instances salient and unique to MLL technology.

A. Oversimplification and Cultural Stereotyping

A critical risk of MLL is the potential for oversimplifying complex cultural nuances and inadvertently promoting stereotypes. The diverse tapestry of human traditions and behaviors is challenging to encapsulate fully in a single AI-driven tool. Current ML approaches tend to suffer from misalignment between the objectives of resulting systems and human values[10] and MLL might favor easily digestible content over rich exploration of a particular culture. his risk is particularly poignant in contexts where native speakers or members of the represented culture might find the depictions lacking in depth or accuracy. For example, for Chinoy, an international student in Japan preparing for a business meeting, MLL provides him with generic guidelines on Japanese business etiquette. However, it may overlook the importance of after-meeting socializing, which is often a crucial aspect of business relationships in Japan. This oversight could lead to misunderstandings or perceived disrespect, running counter to MLL's objective of fostering cultural understanding and competence.

B. Over-reliance on Technology

While MLL enhances performance with human-like feedback, over-reliance on it could lead epistemic complacency[11]. For example, if Maria relies solely on MLL for her dining experiences, she might become overly dependent on the tool and less inclined to engage directly with locals or navigate the situation without technological assistance. Maria, being a non-native speaker, lacks a critical understanding of French and would come to believe MLL's output as ground truth. This could hinder robustness of her critical thinking ability to learn from peers. MLL may be useful, but it wouldn't replace the nuances and effectiveness of human-led teaching and

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems 2017.
[11] Thwaites & Walkins. "The Future of Fact-Checking". 2023. https://computing.mit.edu/wp-content/uploads/2023/06/The-Future-of-Fact-Checking.pdf

cultural immersion. Despite its multi-generational capabilities, MLL cannot replace human personalization and dynamics of natural social interactions.

C. Data Privacy and Copyright Attribution Concerns

There are ongoing concerns and lawsuits over data privacy and misattribution in training generative AI. MLL development would underscore the importance of tackling these responsibly. For Maria, since she uploads videos of herself for analysis by the model, there are risks of data privacy breaches. MLL would have to define clear and foolproof protocols around managing user private information and consent of processing data. Intentions of using user data to refine model training should be spelled out clearly and revenue streams via profile specific ads should be avoided. If such sensitive information is not properly secured, it could lead to unauthorized access and misuse of personal data. Given that MLL requires users to upload videos for feedback, Maria's personal data, including her image and voice, are at risk of being misused or breached. If the tool's data security measures are not robust, there's a potential for sensitive information to be accessed by unauthorized parties.

Additionally, since MLL is trained on widely available data, there may be significant concerns of attributing data. If these materials are sourced from copyrighted works without proper attribution or permission, there's a risk of copyright infringement. For example, if the tool uses a clip from a French movie to demonstrate a conversation style without securing the rights to that footage, it could lead to legal issues for the developers. There are complexities to content even if the engineering team consciously want to address it. Differentiating between general knowledge, which may not require citations, and specialized knowledge, which should ideally be attributed, is a nuanced task[12]. However, recent efforts of providing digital watermarks seem promising i.e. enabling an LLM to generate synthetic texts with embedded watermarks that contain information about their source(s) to address generation of a synthetic text by an LLM

[12] Li, Dongfang, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. "A Survey of Large Language Models Attribution." arXiv, December 14, 2023. https://doi.org/10.48550/arXiv.2311.03731

(source attribution) and (b) verify whether the text data from a data provider has been used to train an LLM (data provenance) [13].

MLL Prevails as Net Socially Positive

While these concerns are valid and substantiate further exploration, my own experiences and empirical research underscore the need and net positive social contributions of such a technology. First, I advocate that MLL would enhance empathy in society and deepen bonds. For example, MLL could provide suggestion in my textual correspondences with peers, especially around controversial topics to handle them with respect and openness. In Chinoy's example of learning about Japanese culture, MLL could simulate a traditional Japanese dining setting, including etiquettes on how to use chopsticks, the custom of saying "*itadakimasu*" before eating. These simulations would help students like Chinoy, and I enhance our cultural sensitivity. This fosters empathy and bridges communication gaps, as understanding the cultural context behind words can be as crucial as the language itself. Peers would appreciate such efforts. MLL can be a platform for sharing and preserving cultural heritage, offering insights that users may employ on their travels into lesser-known languages and traditions, thus contributing to cultural preservation and appreciation.

Second, MLL will not supplant replace traditional language learning mechanisms. Rather, I posit that MLL would supplement such methods. MLL's design philosophy is to be adaptable to a spectrum of learning preferences. For visual learners, MLL provides rich visual contexts; for auditory learners, it emphasizes listening comprehension and spoken interactions. To illustrate, as a kinesthetic learner, I foresee MLL's potential in integrating interactive modules that encourage physical engagement. For example, a virtual module where learners prepare a traditional Mexican dish, such as enchiladas, thereby learning the relevant vocabulary in an interactive manner. In this scenario, my digital avatar could provide feedback on my physical movements during the cooking process. Akin to GPT-3.5's effect democratizing student

[13] Anon. "WASA: WAtermark-Based Source Attribution for Large Language Model-Generated Data," October 13, 2023. https://openreview.net/forum?id=FDfq0RRkuz

learning[14], MLL can offer a viable option to individuals who may not have access to traditional language learning resources or the opportunity to travel It democratizes language learning, making quality education more accessible to a wider audience, regardless of geographical or socioeconomic barriers.

Conclusion

This essay presents and explores Multimodal Language Learning (MLL), a cutting-edge educational technology, that transcends and augments traditional methods to create bespoke, immersive scenarios, targeting specific linguistic challenges of individual learners. Specifically, building on recent technological and academic advancements, I demonstrate that this system is possible and promises socially beneficial prospects. While the issues characteristic of generative AI such as hallucinations and algorithmic bias are ever-present and this system can create additional concerns around oversimplification, overreliance, data privacy and copyright, I believe that the necessity of aiding cultural integrations, communications and professional development far exceed such hindrances. Provided it is developed, deployed, and used responsibly, MLL represents an exciting and essential opportunity in advancing more curiosity and empathy around cross-cultural experiences.

[14] Mollick. "How Does AI Impact Education?" Nov 2023. https://knowledge.wharton.upenn.edu/article/how-does-ai-impact-education/