Voice-Activated Teaching Assistant in Persian and Hindi

Bahareh Arghavani Nobar, Devnath Reddy Motati*

January 2024

Proposal Study Planner - Collaboration

1 Project overview

This project proposes the development of a Voice-Activated Teaching Assistant system capable of understanding and responding to user queries in both Persian and Hindi about the Natural Language Processing course. The system will leverage cutting-edge advancements in text-to-speech (TTS), automatic speech recognition (ASR), and large language models (LLMs) to provide a seamless and natural user experience for speakers of these languages.

1.1 Project Goals and Objectives

1. **Multilingual Teaching Support:** Enable students to interact with the voice assistant in Persian and Hindi, offering a seamless and interactive learning experience for speakers of these languages.
2. **Academic Inclusivity:** Bridge the language gap in NLP education, ensuring that students proficient in Persian and Hindi can access educational resources effectively.
3. **Seamless Integration of Technologies:** Integrate XTTS, Whisper ASR, and the fine-tuned Llama 2 model seamlessly to create a cohesive and efficient voice-activated teaching assistant system.
4. **Advance multilingual NLP research:** Contribute to the development of robust TTS, ASR, and LLM models for languages like Persian and Hindi.

*Advisor: Dr Vahid Behzadan

1.2 Motivation

This project is driven by the dual objectives of democratizing NLP education for Persian and Hindi speakers and harnessing cutting-edge technology to enhance the learning experience. The Voice-Activated Teaching Assistant integrates XTTS Text-To-Speech(TTS), Whisper Large Speech Model (ASR), and Llama 2, offering a sophisticated educational tool.

XTTS facilitates the creation of highly realistic and personalized speech outputs, Whisper ASR ensures accurate transcription of diverse accents, and Llama 2 enhances contextual understanding, providing precise explanations. This amalgamation of technologies not only ensures a natural and immersive learning experience but also supports personalized learning, accommodating individual preferences.

Beyond individual academic growth, the project's technological capabilities contribute to the broader mission of digital education. By making NLP education accessible to diverse linguistic backgrounds, we aim to inspire a passion for learning and technology. This initiative extends its social impact by empowering entire communities with knowledge relevant to the digital era. In essence, the Voice-Activated Teaching Assistant, with its advanced technologies, fosters a more equitable, accessible, and engaging educational experience for learners in Persian and Hindi-speaking communities.

2 Technical Approach

Our system seamlessly integrates three core components:

1. **XTTS Text-to-Speech (TTS):**

   XTTS stands for eXtended Text-to-Speech Synthesis. It's a powerful text-to-speech model developed by Coqui AI with some impressive features.

   Voice Cloning: XTTS can clone a voice with just a short 6-second audio sample. This allows you to generate speech that sounds exactly like someone else, in different languages.

   Multi-lingual: It supports 17 languages, including English, Hindi, Spanish, French, Chinese, and more.

   Emotion and Style Transfer: You can control the emotion and style of the generated speech, making it sound happy, sad, angry, or even sarcastic.

   Open-Source: The core model is open-source and available for anyone to use and experiment with.

   High Quality: Despite being open-source, XTTS generates high-quality, natural-sounding speech that meets production standards.

2. **Whisper Large Speech Model (ASR):** For the ASR component, we will integrate Whisper Speech Large, a state-of-the-art speech recognition

system. This system is known for its robustness and accuracy in transcribing speech, making it ideal for understanding diverse accents and dialects in Persian and Hindi.

3. **Llama-2 based Large Language Model (LLM):** In this project, the focus is on fine-tuning the Llama 2 model specifically for Persian and Hindi languages. This process involves adapting the model, originally trained on diverse languages and datasets, to better understand and generate text in Persian and Hindi. Fine-tuning is a crucial step in machine learning that tailors a pre-trained model to specific linguistic characteristics, idiomatic expressions, and cultural contexts of Persian and Hindi. This adaptation aims to significantly improve the performance of the voice assistant in terms of understanding user queries, generating relevant responses, and ensuring natural interaction in these languages. The success of this fine-tuning process will enhance the model's effectiveness in handling the unique nuances and complexities of Persian and Hindi, thus offering a more inclusive and accessible technology for these language communities.

3 Benefits and Impact:

- **Bridge the digital divide for Persian and Hindi speakers:** Overcome language barriers and provide equal access to technology for underserved communities.
- **Enhanced Linguistic Proficiency:** Students experience more accurate and culturally tailored responses, fostering a deeper comprehension of NLP concepts in their native languages.
- **Empower individuals with a user-friendly interface:** Create a seamless and intuitive experience, enabling individuals to confidently interact with technology in their native language.
- **Inclusivity and Accessibility:** Learners in Persian and Hindi-speaking communities gain equal access to advanced educational technology, promoting inclusivity and educational equity.
- **Advance NLP research for Hindi and Persian:** Contribute to the development of robust TTS, ASR, and LLM models for Hindi and Persian, benefiting millions of speakers worldwide.

4 Evaluation and Continuous Improvement

4.1 Performance Metrics

The performance of our voice assistant will be rigorously evaluated across key areas:

- **Automatic Speech Recognition (ASR) accuracy:**
  - Word error rate (WER) and character error rate (CER) on unseen speech data from diverse accents and speaking styles.
  - Tracking performance across different confidence levels to improve user awareness of potential errors.
- **Text-to-Speech (TTS) naturalness:**
  - Mean opinion score (MOS) and perceptual evaluation of speech quality (PESQ) to measure naturalness and intelligibility.
  - Subjective user feedback to refine pronunciation and intonation.
- **Large Language Model (LLM) response relevance:**
  - BLEU and ROUGE score comparisons with reference texts.
  - Task completion success rate and user satisfaction with response content.
- **User satisfaction:**
  - Regular surveys and A/B testing to gather feedback on usability, intuitiveness, and feature performance.

4.2 Continuous Improvement Cycle

This data will inform a continuous improvement cycle, where identified issues will be addressed through:

1. Model fine-tuning
2. Data augmentation
3. Interface adjustments
4. Dialogue management optimization

Through this iterative process, we aim to continually enhance the performance and user experience of our multilingual voice assistant.

5 Deliverables

5.1 Development of Voice-Activated Teaching Assistant

1. Fine-tuning and implementing Whisper Speech Large for ASR in Persian and Hindi.
2. Customizing and deploying XTTS for TTS in Persian and Hindi.

5.2 Integration of Advanced LLM

1. Adapting and fine-tuning Llama 2 for natural language processing tasks in Persian and Hindi, including dialogue generation, question answering, and intent classification.

5.3 Performance Evaluation

1. Comprehensive testing of ASR and TTS components for accuracy, naturalness, and user experience.
2. Evaluation of Llama 2's efficiency in understanding and responding accurately in NLP content Persian and Hindi.

5.4 User Experience, Feedback Analysis, and Continuous Improvement

1. Gathering and analyzing user feedback to measure satisfaction and identify areas for improvement.
2. Regular updates and improvements to the system based on feedback, technological advancements, and research findings.

5.5 Documentation and Reporting

1. Detailed documentation of the methodologies, processes, and outcomes.
2. Regular reporting on progress, challenges, and milestones achieved.

5.6 Dissemination of Research

1. Sharing findings and advancements through publications in peer-reviewed journals, presentations at conferences and workshops, and collaborations with academic institutions and technology companies.

These deliverables aim to ensure the development of a robust, efficient, and user-friendly multilingual voice assistant system, contributing significantly to digital inclusion and advancing NLP research in Persian and Hindi languages.

6 Resources

1. **Llama 2: Open Foundation and Fine-Tuned Chat Models.** `https://arxiv.org/abs/2307.09288`
2. **Robust Speech Recognition via Large-Scale Weak Supervision.** `https://ai.googleblog.com/2022/11/whisper-large-scale-weakly-supervised.html`

3. **Neural Codec Language Models are Zero-Shot Text-to-Speech Synthesizers.** https://arxiv.org/abs/2301.09754