VOICEBRIDGE: AI POWERED TULU TO ENGLISH AUDIO TRANSLATOR FOR VIDEOS

Project Reference No.: *49S_BE_4266*

College : Sahyadri College of Engineering and Management, Mangaluru
Branch : Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Guide : Dr. Gurusiddayya Hiremath
Student(s) : Ms. Meiha
Ms. Neha Anna Shaju
Ms. Neha Shetty

Keywords:
Automatic Speech Recognition, Neural Machine Translation, Text-to-Speech Synthesis, Low-Resource Language, Tulu, Video Dubbing, AI Pipeline

Introduction:

1. With the increased emergence of local video content, there is a growing need for language accessibility tools—specifically for languages that are often underrepresented in digital media such as Tulu. Despite being a culturally rich language spoken by over two million people in coastal Karnataka and parts of Kerala, Tulu has been largely excluded from the digital ecosystem due to the lack of linguistic resources and inadequate technological support.
2. VoiceBridge is an AI-powered system designed to address this accessibility gap by automatically transcribing spoken Tulu audio from videos, translating it into English, and generating natural-sounding, time-synchronized dubbed English audio. The system integrates three core AI components: Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) synthesis, forming an end-to-end pipeline capable of operating offline.
3. Tulu is identified by UNESCO as a vulnerable language, and unlike major Indian languages such as Hindi or Tamil, it lacks open-source corpora, standardized orthography, and pretrained NLP models. This project directly tackles these challenges by building a custom Tulu speech dataset, fine-tuning state-of-the-art multilingual models, and delivering a usable dubbing system that preserves emotional tone, timing accuracy, and contextual meaning.
4. The VoiceBridge system supports both real-time and batch processing, enabling applications in educational video translation, entertainment, public service announcements, and cultural documentation. Its modular architecture ensures that individual components can be independently upgraded as newer models become available, and the framework is designed to be scalable to other low-resource languages with minimal adjustments.

Objectives:

1. To develop an AI-powered end-to-end pipeline for automatic Tulu-to-English video dubbing, integrating ASR, NMT, and TTS modules.
2. To build a custom Tulu speech dataset from publicly available media, consisting of over 1,000 manually transcribed and bilingual-annotated audio segments.
3. To fine-tune a Whisper-based ASR model for Tulu speech recognition, achieving low Word Error Rate (WER) and Character Error Rate (CER) through cross-validation.
4. To evaluate and compare multilingual NMT models (MarianMT, mBART, NLLB) for Tulu-to-English translation quality using BLEU, METEOR, and TER metrics.
5. To synchronize synthesized English audio with the original Tulu video using FFmpeg and MoviePy, producing a coherent and time-aligned dubbed output.
6. To support the digital preservation of Tulu language and promote linguistic inclusion for low-resource languages through AI-based media localization.

Methodology:

The VoiceBridge system follows a modular, seven-stage pipeline as illustrated in Figure 1. Each stage operates independently and communicates through standard file formats (.wav, .txt, .mp4), enabling fault isolation, parallel development, and independent upgrades.

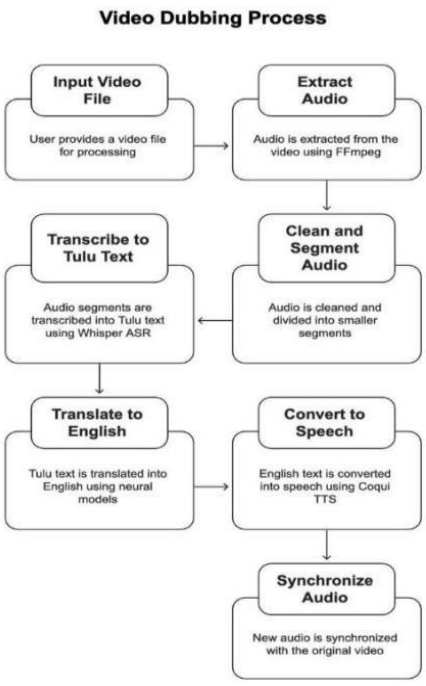

Figure 1: Workflow of the Tulu to English Video Translation System

1. **Video Input and Audio Extraction:**

   The user uploads an MP4 or compatible video file. FFmpeg extracts the audio stream and converts it to a normalized .wav format, also collecting metadata such as duration, frame rate, and bitrate for downstream synchronization.

2. **Audio Preprocessing:**

   Using PyDub and Librosa, the raw audio undergoes noise reduction (spectral gating), segmentation into 5–10 second clips, silence removal, and amplitude normalization. Each segment is stored with start/end timestamps for later alignment.

3. **Automatic Speech Recognition (ASR):**
   OpenAI's Whisper model is fine-tuned on a custom Tulu dataset of over 1,000 manually annotated audio-text pairs sourced from Tulu movies on YouTube. The dataset was created by segmenting videos, filtering non-speech content, and transcribing each segment in both Tulu (Latin script) and English. A 5-fold cross-validation training approach was employed, achieving an average WER of 4.056 and CER of 4.073.
4. **Neural Machine Translation (NMT):**
   Three multilingual transformer-based models were evaluated: MarianMT, mBART, and Meta's NLLB (No Language Left Behind). MarianMT demonstrated the strongest baseline performance for Tulu-to-English translation, leveraging multilingual transfer learning despite the absence of native Tulu training data. Translations are evaluated using BLEU, METEOR, and TER metrics, supplemented by human expert evaluation for fluency, adequacy, and cultural accuracy.
5. **Text-to-Speech Synthesis (TTS):**
   The Google Text-to-Speech API is used to synthesize natural English speech from translated text. Individual audio clips are generated per segment (aligned with ASR timestamps) to ensure precise synchronization. Parameters such as pitch, speaking rate, and voice selection are tunable to match tone and style.
6. **Audio-Video Synchronization:**
   FFmpeg and MoviePy are used to align synthesized English audio segments with their corresponding video timestamps. Crossfades and silence padding are applied for smooth transitions. The final output is a dubbed video where the new English audio replaces the original Tulu audio track while preserving all visual content.

Result and Conclusion:

The VoiceBridge system was evaluated across all three AI components—ASR, NMT, and TTS—yielding strong results for a low-resource language setting.

1. **ASR Performance:**
   The fine-tuned Whisper-based Tulu ASR model demonstrated reliable performance across five-fold cross-validation, achieving an average WER of 4.056 and CER of 4.073. On cleaned datasets, WER ranged between 1.75–2.20 and CER between 1.80–4.30. Cross-validation stability ($\sigma \approx 0.83$) confirms consistent performance. Remaining errors were primarily caused by repetition loops and script drift, consistent with known limitations of Whisper for lowresource, multilingual settings.
2. **Translation Quality:**
   Among the NMT models evaluated, MarianMT achieved the highest BLEU and METEOR scores for Tulu-to-English translation, benefiting from its multilingual transfer capability. mBART showed semantic inconsistencies due to limited Dravidian language exposure, and NLLB showed potential for improvement with further fine-tuning and data augmentation.
3. **Dubbed Video Output:**
   The system successfully produced English-dubbed videos from Tulu input, displayed sideby-side in the web interface (Figure 2). The user interface includes a landing page, video upload interface supporting MP4/AVI/MOV/MKV formats, and a download-enabled output page for the final dubbed video.

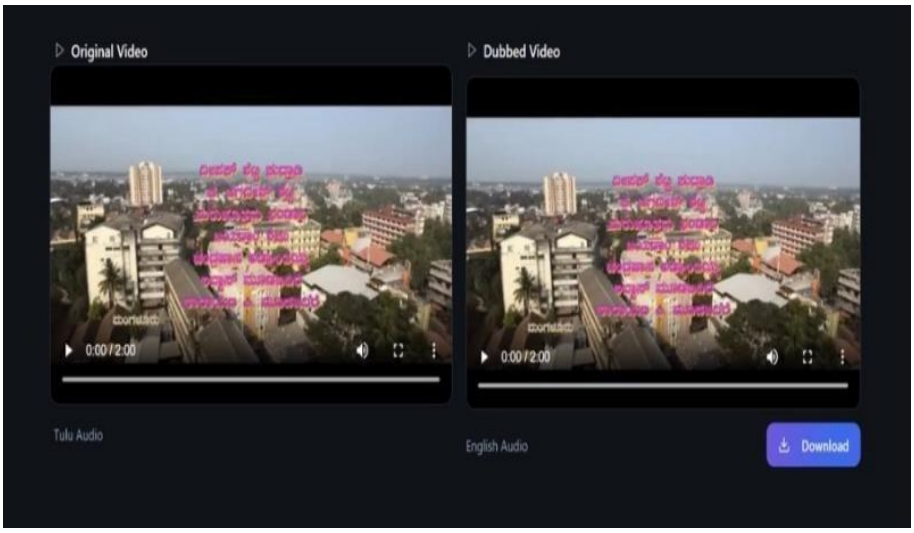

Figure 2: VoiceBridge Dubbed Video Output Interface

In conclusion, the VoiceBridge system demonstrates that AI-driven dubbing for low-resource languages is achievable with careful dataset construction, model fine-tuning, and modular pipeline design. The project advances digital inclusion by making Tulu video content accessible to non-Tulu-speaking audiences, and serves as a replicable framework for other endangered or low-resource languages.

References:

1. [1] Y. Wu et al., "VideoDubber: Machine Translation with Speech-Aware Length Control for Video Dubbing," in Proc. AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 13772–13779, 2023.
2. [2] P. Suresh et al., "Multilingual Video Content Transformation: Automated Subtitle Translation and Voice Integration," unpublished.
3. [3] E. Artioli et al., "Generative AI for Realistic Voice Dubbing Across Languages," in Proc. 4th MileHigh Video Conference, pp. 75–76, 2025.
4. [4] C. Le et al., "TransVIP: Speech-to-Speech Translation System with Voice and Isochrony Preservation," 2024.
5. [5] A. Dsouza et al., "SynthPipe: AI-Based Human-in-the-Loop Video Dubbing Pipeline," in Proc. ICAECT, pp. 1–5, 2022.
6. [6] C. Zhang et al., "UWSpeech: Speech-to-Speech Translation for Unwritten Languages," in Proc. AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14319–14327, 2021.
7. [7] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. ICML, pp. 28492–28518, 2023.

Future Scope:

The future scope of this project includes:

1. Integration of Wav2Lip for fine-grained lip synchronization, modifying mouth movements in video frames to precisely match dubbed English phonemes for a visually authentic experience.
2. Speaker diarization and multi-speaker adaptation, enabling the system to identify and assign distinct TTS voices to different speakers within a single video.
3. Extension of the framework to support other low-resource and endangered Indian languages such as Kodava, Konkani, and Santali, leveraging the same modular pipeline with language-specific parameters.

4. Expansion of the Tulu dataset through community-driven crowdsourcing, covering wider dialectal variation and speaking styles for more robust ASR and NMT performance.
5. Real-time translation capability optimized for on-device processing, enabling deployment in live classrooms, public communication systems, and mobile applications.
6. Bidirectional translation support (English-to-Tulu dubbing) to further promote intercultural communication and Tulu language revitalization.
7. Integration of ethical safeguards including culturally sensitive translation, speaker consent management, and privacy-centered data handling for responsible AI deployment.