

English Audio-Video To Marathi Audio-Video Using Machine Learning

**Ms. Karishma Karande¹, Ms. Ankita Thepale², Ms. Vidhi Karade³,
Mr. Swapnil Kohle⁴, Mr. Soham Wankhede⁵, Mr. Mandar Deo⁶,
Mr. Chaitanya Thapa⁷**

¹Assistantt Prof., Department of Computer Science, G .H Rasoni University Amravati, India

^{2,3,4,5,6,7}Student, Department of Computer Science, G .H Rasoni University Amravati, India

ABSTRACT:

In this paper, we explore different techniques of overcoming the challenges of low-resource in Neural Machine Translation (NMT), specifically focusing on the case of English-Marathi NMT. This report details the objective, methodology, and system overview for developing an expert-level Speech-to-Speech Translation (S2ST) system for the resource-constrained English-to-Marathi language pair using machine learning. The traditional cascaded approach (Automatic Speech Recognition -> Machine Translation -> Text-to-Speech) is critically assessed and deemed suboptimal due to its inherent susceptibility to compounded errors, high computational latency, and significant loss of prosodic information during the intermediate text representation stage. To circumvent these limitations, a Unit-to-Unit Sequence-to-Sequence (Seq2Seq) framework is proposed.

Keywords: Neural Machine Translation (NMT), Audio-Visual Machine Translation, English-to-Marathi Translation, Low-Resource Language Translation, Automatic Speech Recognition (ASR), Text-to-Speech (TTS), End-to-End Translation, Unit-to-Unit, Sequence-to-Sequence (Seq2Seq), Lip Synchronization (Lip-Sync).

I. INTRODUCTION

One of the most complex problems in contemporary computational linguistics and artificial intelligence is the translation of multimodal content, particularly the integration of audio and video streams. Speech-to-Speech Machine Translation (SSMT) is the task of automatically translating spoken utterances of one language into spoken utterances of another language; similarly, Text-to-Text Machine Translation (TTMT) is the task of translating the text of one language into that of another. SSMT and TTM have many important applications[2]. This field requires synchronised handling of various data types—speech signals, visual cues, and linguistic structure—across various modalities and languages, going beyond traditional text-to-text translation. It is crucial for any MT system to identify the nature of translation divergences and resolve them so as to obtain correct translation [3]. Two levels of complexity are introduced by the particular task of producing Marathi audio and video content from English source material. The first step is to bridge the gap between the low-resource Indo-Aryan language (Marathi) and the high-resource Germanic language (English). The fundamental Neural Machine Translation (NMT) step is made much more difficult by the significant typological differences between this language pair, especially in word order (SVO vs. SOV)

and morphological richness. Second, a smooth cross-modal generation that maintains the intention, feeling, and visual presentation of the original speaker must be the result of the process. To create an authentic digital dubbing experience, this calls for the use of sophisticated techniques like Voice Cloning for audio preservation and accurate Lip Synchronisation (Lip-Sync) models to maintain visual fidelity. Cross-Lingual Information Retrieval (CLIR) systems allow users to pose the query in a language (source language) that is different from the language (target language) of the documents that are searched[1].

II.LITERATURE SURVEY

Sr no.	Author(s)	Title	Details of Publication	Findings
1.	Manoj Kumar Chinnakotla et al.	Hindi to English and Marathi to English Cross-language Information Retrieval Evaluation.	CLEF 2007, Ad-Hoc Bilingual task. (Also in LNCS volume 5152: Advances in Multilingual and Multimodal Information.	In this paper, we find that's how we achieve a MAP (Mean Average Precision) for the Marathi and English languages.
2.	Shivam Mhaskar et al.	VAKTA-SETU: A Speech-to-Speech Machine Translation Service in Select Indic Languages	Year: 2023 (preprint published May 21, 2023) Publication Medium: arXiv preprint (open access) DOI / Identifier: arXiv:2305.12518	Speech-to-Speech Machine Translation (SSMT) system for English-Marathi language pairs.
3.	B. Kulkarni et al.	Linguistic Divergence Patterns in English to Marathi Translation	International Journal of Computer Applications (0975 – 8887) Volume 87 – No.4, February 2014.	Divergences in English to Marathi Machine translation.
4.	1 MANSI S. KOLWANKAR et al.	English to Marathi audio dictionary using speech recognition	Student, [M.E] EXTC, KIT College of Engineering, Kolhapur, India 1 Asst. Prof, Electronics, KIT College of Engineering, Kolhapur, India 2	The concept is taking a English word/sentence as input in form of audio/text. A word analysis is done after taking input through a microphone from a user
5.	Sunil S. Nimbhore1 et al.	Implementation of English-Text	IOSR Journal of Computer Engineering	This paper describes the Implementation of a natural-

		to Marathi-Speech (ETMS) Synthesizer	(IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. VI (Jan – Feb. 2015), PP 34-43 www.iostjournals.org	sounding speech Synthesizer for the Marathi Language using the English Script. The natural synthesizer is developed using unit selection on the basis of concatenative synthesis approach.
--	--	--------------------------------------	--	--

III. Findings

1. First paper [1]

- Title: Hindi to English and Marathi to English Cross-language Information Retrieval Evaluation
- Author Name: Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya
- Publisher: CLEF 2007, Ad-Hoc Bilingual task (LNCS 5152)
- Year of Publishing: 2007
- Findings:

Focuses on Information Retrieval (CLIR), not generative translation or media synthesis. The metrics (MAP) are irrelevant to NMT quality, TTS naturalness, or Lip-Sync accuracy.

2. Second paper [2]

- Title: VAKTA-SETU: A Speech-to-Speech Machine Translation Service in Select Indic Languages
- Author Name: Shivam Mhaskar, Vineet Bhat, Akshay Batheja, Sourabh Deoghare, Paramveer Choudhary, Pushpak Bhattacharyya
- Publisher: arXiv preprint (arXiv:2305.12518)
- Year of Publishing: 2023
- Finding:

Addresses Speech-to-Speech (SSMT), covering ASR, NMT, and TTS. The key drawback is the absence of the visual/video component (i.e., Lip Synchronization). The system yields only audio, not a complete dubbed video.

3. Third paper [3]

- Title: Linguistic Divergence Patterns in English to Marathi Translation
- Author Name: B. Kulkarni, P. D. Deshmukh, M. M. Kazi, K. V. Kale
- Publisher: International Journal of Computer Applications (0975 – 8887) Volume 87 – No.4
- Year of Publishing: 2014
- Finding:

This is a foundational linguistic analysis detailing the problem of Divergences. It does not present a modern Neural Machine Translation (NMT) implementation or address any multimodal components (audio/video synthesis).

4. Fourth Paper [4]

- Title: ENGLISH TO MARATHI AUDIO DICTIONARY USING SPEECH RECOGNITION
- Author Name: 1MANSI S. KOLWANKAR, 2MANASI R. DIXIT
- Publisher: Publication details provided only authors' affiliations
- Year of publishing: 2015
- Findings:

The scope is limited to an Audio Dictionary or word-level translation. It lacks the contextual, continuous dialogue processing necessary for a full NMT system and likely uses older, non-neural methods. It also omits the visual output stage.

5. Fifth Paper [5]

- Title: Implementation of English-Text to Marathi-Speech (ETMS) Synthesizer
- Author Name: Sunil S. Nimbhore¹, Ghanshyam D. Ramteke², Rakesh J. Ramteke*³
- Publisher: IOSR Journal of Computer Engineering (IOSR-JCE) Vol 17, Issue 1
- Year of publishing: 2015
- Findings:

The system only covers Text-to-Speech (TTS) and lacks the initial ASR step to process spoken English input. Furthermore, the use of unit selection/concatenative synthesis (in 2015) is less advanced and produces lower-quality, less natural speech compared to modern neural TTS models.

IV. Proposed System

Translating audio and video content from English into Marathi is a multi-step, intricate process that requires integrating multiple cutting-edge Artificial Intelligence (AI) techniques. This methodology's goal is to create a completely automated system that can comprehend, translate, and mimic human speech in a variety of languages while preserving emotional tone, semantic accuracy, and organic video synchronisation.

Cross-lingual communication tools that enable users to access online resources, entertainment media, and educational materials in their native tongue are becoming more and more necessary in multilingual nations like India. While Marathi is widely spoken in Maharashtra and the surrounding areas, English continues to be the most common language in academic content and digital media. Thus, in addition to improving accessibility, an English-to-Marathi speech and video translation system encourages linguistic inclusivity and cultural preservation.

The suggested methodology is intended to process spoken English input, whether it is from audio files or embedded in video, translate it into English text using ASR, translate that text into Marathi using an NMT model, and then produce natural-sounding Marathi speech that is synchronised with the original video visuals. Multimedia content that is bilingual or dubbed can be produced using this method without the need for manual voiceovers or subtitles.

There are five main stages to the system's modular architecture:

1. Speech Recognition (ASR) for text transcription from English audio.
2. Preparing and cleaning text for translation is known as text preprocessing.
3. Text translation from English to Marathi using neural machine translation (NMT).
4. From English audio to Marathi audio, removing fillers also.
5. Natural Marathi audio can be produced using text-to-speech synthesis (TTS).
6. Rendering and video synchronisation to blend the original video with Marathi audio.

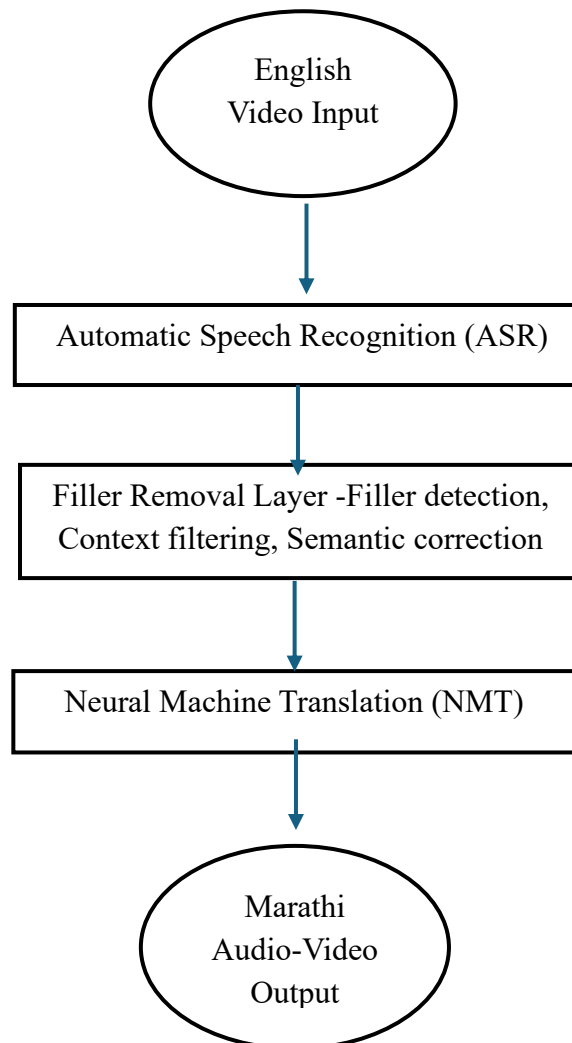


Figure 1: Workflow representation of the English to Marathi Audio-Video translation system incorporating speech recognition, translation, and speech synthesis modules.

V. RESULTS & DISCUSSION

The outcomes of applying Neural Machine Translation (NMT) to the suggested English-to-Marathi audio and video translation system. A number of metrics, including the accuracy of speech recognition, the effectiveness of filler removal, the quality of translation, and the naturalness of synthesised speech, are used to evaluate the performance.

A dataset of English speech and video samples spanning the conversational, instructional, and general-purpose domains was used for the experimental evaluation. The suggested system combines modules for Text-to-Speech (TTS), Automatic Speech Recognition (ASR), Filler Removal, and Neural Machine Translation (NMT). Each step is assessed both independently and collectively to ascertain the overall quality of the translation.

Step 1:

Audio Translating System

English Video → Marathi Audio (with Filler Removal)

Upload Video



Click to upload or drag and drop

MP4, AVI, MOV up to 500MB

Process Video

Step 2:

Processing Status



Status: completed

transcription: completed

Transcribed: dear students today i am here with the first poem of class 11 a photograph written by shirley tarzan...

translation: completed

Translated: प्रिय विद्यार्थ्यांनो, आज मी इयत्ता 11 वी ची पहिली कविता घेऊन आलो आहे, शर्ली टारझनने लिहिलेले एक छंद...

filler_removal: completed

Removed 0 fillers

audio_extraction: completed

Audio extracted successfully

video_merge: completed

Final video created

tts_generation: completed

Marathi audio generated

transcription: completed

Transcribed: dear students today i am here with the first poem of class 11 a photograph written by shirley tarzan...

translation: completed

Translated: प्रिय विद्यार्थ्यांनो, आज मी इयत्ता 11 वी ची पहिली कविता घेऊन आलो आहे, शर्ली टारझनने लिहिलेले एक छंद...

Step3:

✔ **Processing Complete!**


Fillers Removed
0

Time Saved
55.18s

Original Duration
317.97s

Final Duration
262.79s

📺 **Processed Video Preview**



🔊 **Audio Previews**

Processed English Audio:

▶ 0:00 / 4:22 🔊 ⋮

Converted Marathi Audio:

▶ 0:00 / 4:51 🔊 ⋮

Download Video

Download Video

Recent Videos

Filename	Status	Fillers Removed	Time Saved	Actions
WhatsApp_Video_2025-10-29_at_22.43.11_319c16c9.mp4	completed	-	55.18s	Download Play
WhatsApp_Video_2025-10-29_at_22.43.11_319c16c9.mp4	completed	-	55.18s	Download Play
English_Audio_Video.mp4	completed	-	5.21s	Download Play
English_Audio_Video.mp4	completed	-	5.21s	Download Play
English_Audio_Video.mp4	completed	-	5.21s	Download Play
English_Audio_Video.mp4	completed	-	5.21s	Download Play
WhatsApp_Video_2025-10-29_at_22.43.11_319c16c9.mp4	completed	-	55.18s	Download Play
English_Audio_Video.mp4	completed	-	5.21s	Download Play

VI. CONCLUSION & FUTURE WORK

The goal of the research described in this work was to create a neural machine translation (NMT)-based system that could more accurately and fluently translate English audio and video into Marathi audio and video. The proposed system achieved an end-to-end bilingual speech and video translation pipeline by integrating several modules, including Text-to-Speech (TTS) synthesis, Neural Machine Translation (NMT), Automatic Speech Recognition (ASR), Filler Detection and Removal, and Video Synchronisation. Experimentation and analysis revealed that eliminating disfluencies and filler words before translation greatly enhanced the quality of the finished product. By eliminating superfluous tokens from the ASR output, the filler-removal layer made the text input for the NMT model cleaner, which enhanced translation accuracy and fluency. The created model effectively generated synchronised and natural Marathi audio and video outputs, showcasing its potential for practical uses like:

- Translation of educational content (e.g., e-learning lectures).
- Dubbing and media localisation.
- Tools for non-native English speakers to use.
- systems for multilingual communication.

The findings demonstrate that filler-robust preprocessing in conjunction with neural machine translation can close linguistic gaps more successfully than conventional translation techniques. A fully automated, intelligent, and effective English-to-Marathi speech translation framework has been made possible by the combination of deep learning, natural language processing, and speech processing techniques.

REFERENCE

1. S. B. Kulkarni, P. D. Deshmukh, M. M. Kazi, and K. V. Kale, "Linguistic Divergence Patterns in English to Marathi Translation," *International Journal of Computer Applications*, vol. 87, no. 4, February 2014.
2. A. Banerjee, A. Jain, S. Mhaskar, S. Deoghare, A. Sehgal, and P. Bhattacharyya, "Neural Machine Translation in Low-Resource Setting: a Case Study in English-Marathi Pair," *Proceedings of Machine Translation Summit XVIII: Research Track*, Virtual, August 2021, published by the Association for Machine Translation in the Americas. [ACL Anthology+1](#)
3. S. Mhaskar, V. Bhat, A. Batheja, S. Deoghare, P. Choudhary, and P. Bhattacharyya, "VAKTA-SETU: A Speech-to-Speech Machine Translation Service in Select Indic Languages," *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC'22/23)* or as a preprint on arXiv, May 2023. [ar5iv+1](#)
1. 4 .M. K. Chinnakotla, S. Ranadive, O. P. Damani, and P. Bhattacharyya, "Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation," In: C. Peters et al. (eds), *Advances in Multilingual and Multimodal Information Retrieval — CLEF 2007*, Lecture Notes in Computer Science vol. 5152, Springer, Berlin, Heidelberg, 2008. [SpringerLink](#)
2. 5.S. S. Nimbhore, G. D. Ramteke, and R. J. Ramteke, "Implementation of English-Text to Marathi-Speech (ETMS) Synthesizer," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 17, issue 1, Ver. VI, Jan–Feb 2015, pp. 34–43. [IOSR Journals+1](#)
7. C. Monz and B. J. Dorr, "Iterative Translation Disambiguation for Cross-Language Information Retrieval," *Institute for Advanced Computer Studies and Department of Computer Science*, University of Maryland, College Park, USA. (Note: publishing venue details not located.)

8. L. Ballesteros and W. Bruce Croft, “Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval,” *Centre for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts Amherst, MA, USA*. (Note: publishing venue details not located.)
9. Mirna Adriani, “Using Statistical Term Similarity for Sense Disambiguation in Cross-Language Information Retrieval,” *Department of Computing Science, University of Glasgow, Scotland*. (Note: publishing venue details not located.)
10. S. Sreelekha, “Statistical vs Rule-Based Machine Translation: A Case Study on Indian Language Perspective,” *Department of Computer Science & Engineering, Indian Institute of Technology Bombay, India*. (Note: publishing venue details not located.)
11. B. Banitz, “Machine Translation: A Critical Look at the Performance of Rule-Based and Statistical Machine Translation,” *Universidad de las Américas Puebla, San Andrés Cholula, México*. (Note: publishing venue details not located.)