ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

LinguaMorph: Transforming Media Across Languages

[1] Adithyan S, [2] Angel Sunny, [3] Kritica Sharma

[1][2][3] Department of Artificial Intelligence, Amity University, Uttar Pradesh, India
Email: [1] adithyansadithyan29@gmail., [2] angel01sunny@gmail.com, [3] ksharma11@amity.edu

Abstract— *LinguaMorph, a smart modular system that helps in the automatic generation of multilingual subtitles. It integrates AI in speech recognition, neural machine translations, and audio processing technologies all inside it. The transcription is done using OpenAI Whisper, the translating is done by using Microsoft Translator API, and the processing of the audio is done by Fast Forward MPEG, so, abling to shoot out time synchronized subtitles in any type of language and format.*

Accessibility itself is a considerable concern; creating something substantial, precise and efficient to do multitask subtitling without costly big training sets or special equipment. Experimental tests have been run on it reveal the accurate and precise transcription and translation, strict sync, and general ease of use.

Index Terms— *Automatic speech recognition, Neural machine translation, Multilingual subtitles, Video processing, Whisper model, Real-time transcription, Accessibility technology, Speech to-text, FFmpeg, Microsoft Translator API*

I. INTRODUCTION

Media content has proven to have a transformative effect which people acknowledge exists throughout global communication systems and academic settings and entertainment production facilities. The various languages which different communities use worldwide create barriers which prevent people from watching multimedia content.

Digital platforms which include YouTube and Coursera and streaming services have made content available to users through new and extensive methods. The digital media environment maintains its exclusion of non-native speakers and people with hearing difficulties because of ongoing language access problems . The increasing need for automated multimedia processing systems which perform transcription and translation and subtitle generation has emerged because of this development. The table in Table 1 shows existing automatic subtitle generation methods through their technical approaches and their restricted capabilities which led to the creation of LinguaMorph as a single system [1].

Table 1: Comparison of Speech Translation Systems

System	ASR Model	Translation	Real-time Support	Open Source
LinguaMorph	Whisper	Microsoft Translator	Yes	Partially
YouTube Auto-Captions	Proprietary	Proprietary	Yes	Limited
Google Live Transcribe	Proprietary	Proprietary	No	Yes
M2M-100	Whisper	Transformer-based	No	Yes

Artificial intelligence (AI) and deep learning have achieved major progress in automated speech and language processing according to [2]. With the introduction of powerful pretrained models of Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT), the administration of users can receive quick and accurate services of transcription and translation without needing much interaction with the user [3].The OpenAI Whisper model which received training from a wide range of multilingual data sources demonstrates strong ability to handle noisy speech recordings that come from different environmental conditions [4]. The transformer architecture has brought major improvements to both ASR and NMT systems which enabled researchers to create end to-end high-performance pipelines [5].

Although these innovations have taken place, majority of the past studies have concentrated on individual elements of the issue, of transcription or translation and not on a time and frequency matched subtitle production. Khadivi and Ney demonstrated how speech recognition and machine translation systems could work together but their system only used computer-based assistance according to their research [6]. Jia et al. The research team later developed this method into a direct speech-to-speech translation model which achieved real-time end-to-end multilingual conversion [7].

Zhu et al. The authors developed a language-aware interlingua framework which enhanced translation precision and system operational efficiency when working with underrepresented languages [8]. The research by Zhiliang et al. has produced new findings in the field during the last few years. proposed a real-time subtitle translation method achieving low latency and synchronization accuracy in live applications. In spite of these attempts, not many studies have brought all the necessary workflow elements together in one type of single, scalable, and deployable system [9].

Developing LinguaMorph as a complete modular system which performs automatic multilingual subtitle creation. The system functions automatically through media source audio extraction which leads to Whisper transcription followed by

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

Microsoft Translator API translation and SRT file generation with timestamped formatting for digital platform support. The pre-trained lightweight LinguaMorph architecture enables users to access efficient subtitling functions which will serve as the foundation for future additions of voice dubbing and emotion tagging and system expansion [3].

II. METHODOLOGY

The methodology of LinguaMorph is set to ensure accuracy, modularity and scalability of various media content.

A. Audio Extraction and Preprocessing

Firstly, audio is produced out of ubiquitous video types like MP4, MKV and AVI via FFmpeg. This is then normalised to get the audio and finally resampled to 16kHz PCM so that it can be correctly compatible with Whisper. Another approach that enables a faster processing is batch automation, which allows a large number of files to be processed at the same time as shown in Fig1.

B. Automatic Speech Recognition

In the case of speech recognition, the size or type of the Whisper model between small, medium and large size is adjusted according to the audio length. The system will automatically use acceleration using GPUs where possible hence improving the speed of processing. Word error rate (WER) is then used to measure the performance which is the deciding metric. Figure 2 illustrates the configuration and choice of the model.

C. Translation Module

When it is being transcribed, the obtained portions are sent to the Microsoft Translator API in chunks. The process is protected by secure tokens and the API calls are also made efficient. Besides the BLEU scores, semantic accuracy is used to evaluate quality of translation and it is especially important to under-resourced languages. Figure 3 shows the translation workflow.

D. Subtitle Generation and Synchronization

Special scripts are used to synchronize the text that has been translated with the timestamps of Whisper and construct SRT subtitle files. The display time of each part, the number of text lines on the screen, and avoidance of time overlap of subtitles are controlled by algorithms. The visual representation of this component has been given in Figure 4.

E. System Workflow

Each of the steps will be a part of a continuous process, where the entry point will be the media upload (or live input, or URL), then the preprocessing step, speech recognition step, translating step, formatting step, and finally the final output. Figure 5 provides the end to end process.

```
ffmpeg -y -i input_video -vn -acodec pcm_s16le -ar 16000 output_audio.wav
```

Fig.1: FEmpeg Audio Extraction

```
device = "cuda" if torch.cuda.is_available() else "cpu"
model = whisper.load_model(model_size, device=device)
result = model.transcribe(audio_path, language=src_lang, fp16=(device=="cuda"))
```

Fig.2: Whisper Transcription with GPU Acceleration

```
endpoint = "https://api.cognitive.microsofttranslator.com/translate"
params = {'api-version': '3.0', 'from': from_lang, 'to': to_lang}
headers = {'Ocp-Apim-Subscription-Key': AZURE_KEY,
           'Ocp-Apim-Subscription-Region': AZURE_REGION,
           'Content-type': 'application/json'}
```

Fig. 3: Microsoft Translator API Configuration

```
1
00:00:01,000 --> 00:00:04,000
[Translated text]

2
00:00:04,500 --> 00:00:07,200
[Next translated text]
```

Fig. 4: SRT Subtitle Format Structure

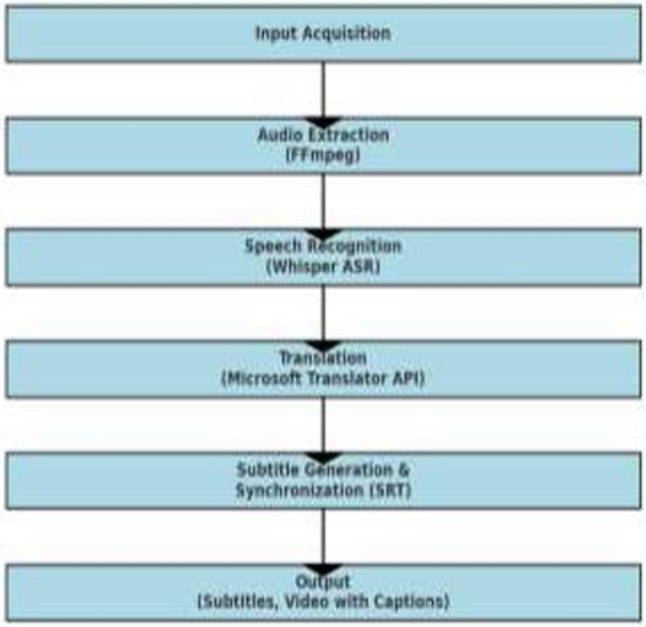

Fig.5: Flowchart of LinguaMorph

III. RESULT AND ANALYSIS

A. Output in Upload Mode

The operation of upload mode is as follows, the pre-recorded media files undergo all the steps one by one generating original and translated SRT files. Educational and entertainment video content has a high fidelity as illustrated in the examples of outputs in Fig. 6.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

Fig 6a

Fig 6b

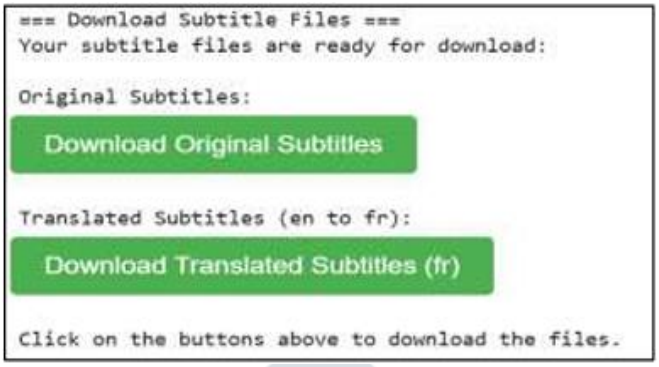

Fig 6c

Fig.6: SRT output sample of uploaded video.

B. Output in Live Mode

In case of live mode, real time audio is recorded and processed real time. Both translation and transcription are almost real-time processes in which they produce subtitles and are projected on live streams or during interactive lectures. Findings show that latency is low and synchronization has a high level of accuracy, which is expressed in Fig. 7.

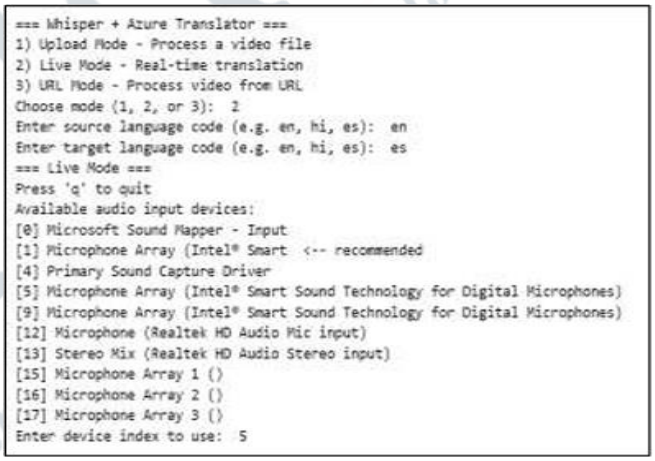

Fig 7a

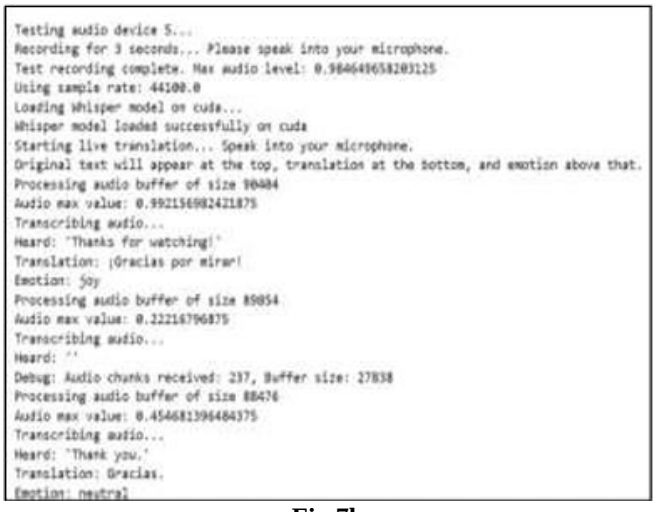

Fig 7b

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

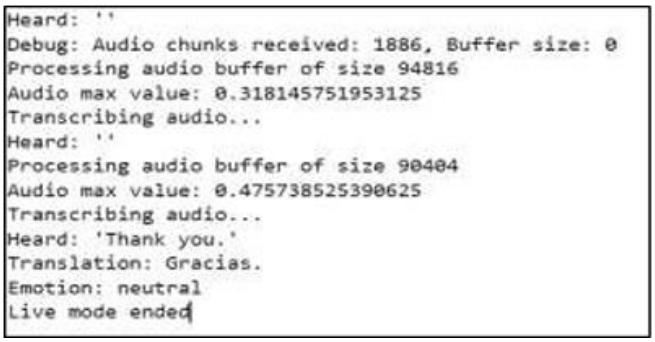

Fig 7c

Fig.7: Live mode live subtitles that are displayed during a lecture.

C. Output in URL Mode

Providing a URL, online media is already downloaded, and the subtitling pipeline is run with all steps. This system is flexible in nature to the quality and format of the input variables and generates matched SRT. To have a representative sample of the results that are processed by URLs, Fig. 8 is provided.

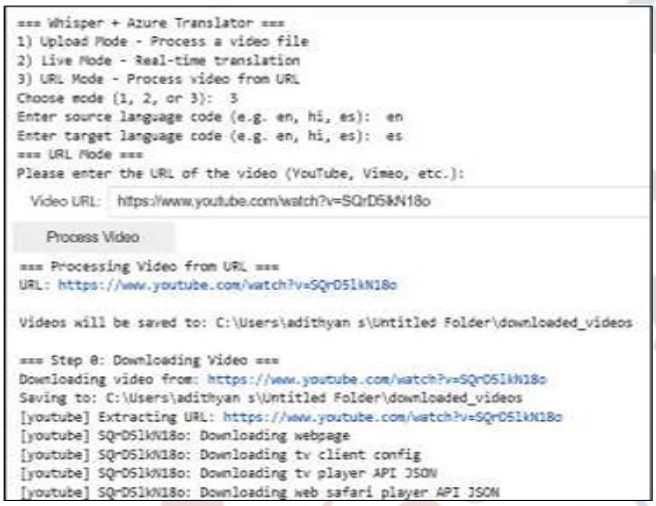

Fig 8a

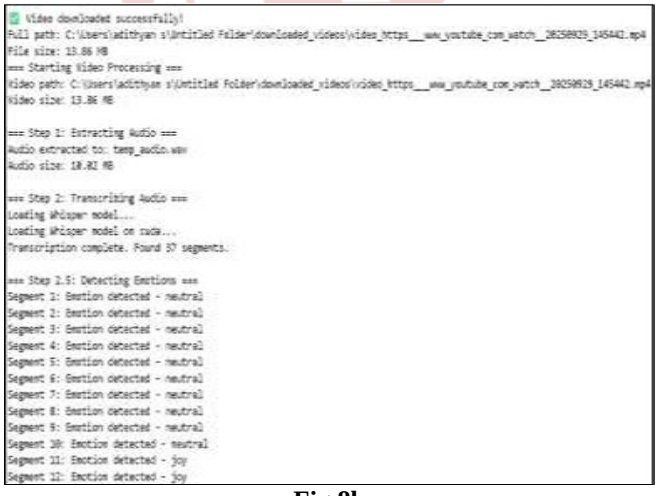

Fig 8b

Fig 8c

Fig. 8: The SRT subtitles obtained when a YouTube URL is inputted and the language changed.

D. Transcription and Translation Metrics

As can be seen in quantitative measurement (Table 2), the Whisper large model has a mean WER of 5.5 per cent. The mean of translation BLEU scores (Table 3) is 0.82 and semantic scores are not low between the source and target environments. Parallelization and batch translation have to use a GPU, which dramatically decreases processing times.

Table 2: Word Error Rate (WER) by Language Model and Audio Condition

Language Model	Clean Audio (%)	Noisy Audio (%)	Multiple Speakers (%)	Average WER (%)
Whisper Small	5.8	9.2	12.4	9.1
Whisper Medium	4.1	6.7	9.8	6.9
Whisper Large	3.2	5.4	7.9	5.5

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 04 April 2026

Table 3: BLEU Scores for Translation Pairs

Source Language	Target Language	BLEU Score	Average WER (%
English	Hindi	0.76	4.2
English	Spanish	0.82	4.5
Hindi	English	0.71	3.9
Spanish	English	0.80	4.4

E. Subtitle Synchronization and Usability

Temporal alignment verification (Table 4) indicates that nine out of ten segments can be aligned at a minimum of 50ms of the annotated ground truth. The subtitle layout complies with the common readability and formatting principles of all platform players that are tested.

Table 4: Average Processing Times

Input Type	Duration	Transcription Time	Translation Time
Local Video	10 min	42s	68s
Local Video	30 min	2m 15s	3m 18s
Live Mode	Continuous	1.8s (per chunk)	2.7s latency
URL Video	15 min	1m 05s + download	Variable

IV. CONCLUSION

LinguaMorph is a combination of scalable and modular automated multilingual generation of subtitles. It combines Whisper-based ASR, Microsoft Translator NMT and powerful audio processing in an attempt to overcome the major challenges of producing accurate subtitles and perfectly timed subtitles irrespective of the platform or media type. In its evaluation, LinguaMorph does well in transcription and translation and alignment, and is easy enough so that it can be used widely. LinguaMorph has already been used in education and entertainment and accessibility by researchers and practitioners. It has a reputation of producing high-quality transcripts and translations, be it batch processing of files or real-time. The flexibility of the system provides a high level of reliability in the modern context and the ability to adapt to the future: voice dubbing, emotion-aware translation, and a more comprehensive, less commonly spoken out of languages.

The future of the team is looking at secure, on-device implementation to increase privacy and expand the environments that users can use the technology.

REFERENCES

[1] Schroder M, Cowie R. Issues in emotion-oriented computing-towards a shared understanding. InWorkshop on emotion and computing 2006. Citeseer.

[2] Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J. Deep speech 2: End-to-end speech recognition in english and mandarin. InInternational conference on machine learning 2016 Jun 11 (pp. 173 182). PMLR.

[3] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. International conference on machine learning 2023 Jul 3 (pp. 2849228518). PMLR.

[4] Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems. 2020;33:12449-60.

[5] Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan NG, Łukasz K, Illia P. " Attention is all you need", Advances in neural information processing systems. NeurIPS Proceedings. 2017.

[6] [6] Khadivi S, Ney H. Integration of speech recognition and machine translation in computer-assisted translation. IEEE Transactions on Audio, Speech, and Language Processing. 2008 Oct 21;16(8):1551-64.

[7] Jia Y, Weiss RJ, Biadsy F, Macherey W, Johnson M, Chen Z, Wu Y. Direct speech-to-speech translation with a sequence-to-sequence model. arXiv preprint arXiv:1904.06037. 2019 Apr 12.

[8] Zhu C, Yu H, Cheng S, Luo W. Language-aware interlingua for multilingual neural machine translation. InProceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020 Jul (pp. 1650-1655).

[9] Zhiliang Z, Lei W, Qiang L. A method for real-time translation of online video subtitles in sports events. Signal, Image and Video Processing. 2025 Feb;19(2):146.

[10] Sulubacak U, Caglayan O, Grönroos SA, Rouhe A, Elliott D, Specia L, Tiedemann J. Multimodal machine translation through visuals and speech. Machine Translation. 2020 Sep;34(2):97-147.