TRAVID: An End-to-End Video Translation Framework

Prottay Kumar Adhikary
NIT Silchar, Assam, India
prottay71@gmail.com

Sugandhi Bandaru
NIT Silchar, Assam, India
bsugandhi99@gmail.com

Subhojit Ghimire
NIT Silchar, Assam, India
subhojitg@live.com

Santanu Pal
Wipro AI Research (Lab45), London, UK
santanu.pal.ju@gmail.com

Partha Pakray
NIT Silchar, Assam, India
parthapakray@gmail.com

Abstract

In today's globalized world, effective communication with people from diverse linguistic backgrounds has become increasingly crucial. While traditional methods of language translation, such as written text or voice-only translations, can accomplish the task, they often fail to capture the complete context and nuanced information conveyed through nonverbal cues like facial expressions and lip movements. In this paper, we present an end-to-end video translation system that not only translates spoken language but also synchronizes the translated speech with the lip movements of the speaker. Our system focuses on translating educational lectures in various Indian languages, and it is designed to be effective even in low-resource system settings. By incorporating lip movements that align with the target language and matching them with the speaker's voice using voice cloning techniques, our application offers an enhanced experience for students and users. This additional feature creates a more immersive and realistic learning environment, ultimately making the learning process more effective and engaging.

1 Introduction

Face-to-Face (F2F) translation is a sub-field within the research domain of Machine Translation (MT). MT refers to the process of utilizing machines to translate text or speech from one language to another (Somers, 1992). F2F translation specifically focuses on translating spoken language in real-time during face-to-face conversations or interactions. The objective is to bridge language barriers and facilitate seamless communication between individuals who speak different languages.

F2F translation is also a part of the broader field of multi-modal machine translation, which integrates videos or visual information along with translation. This approach aims to enhance engagement among native language speakers during sessions.

Visual cues, such as lip synchronization according to the native languages, contribute to a more realistic and immersive translated lecture session. These visual elements provide valuable context information that aids in the translation process. Compared to image-guided multi-modal machine translation, videos provide visual and acoustic modalities with rich embedded information, such as actions, objects, and temporal transitions. From the past few years, image-based multi-modal models (Chen et al., 2022b) only had marginal performance gains compared to their text-only counterparts, although very few of them are F2F translation (Chen et al., 2022a).

F2F translation goes beyond traditional text-to-text or speech-to-speech translation methods. In a simple cascade-based F2F translation approach, several steps are involved: (i) **Capturing original speech:** the source video of a person delivering a speech is recorded or obtained, (ii) **Translating the captured speech:** the captured spoken language in the source video is translated to the desired language using machine translation techniques, (iii) **Generating an output video:** Based on the translated text, an output video is generated where the same person appears to be speaking in the translated language, and (iv) **Maintaining lip synchronization:** during the generation of the output video, efforts are made to ensure that the lip movements of the person in the video match the target language, providing lip synchronization as per the translated language. By following these steps, cascade-based F2F translation aims to deliver translated videos with synchronized lip movements, enhancing the authenticity and naturalness of the translated speech (K R et al., 2019). The intermediate steps i.e., **Translating the captured speech** can be modelled either direct (Etchegoyhen et al., 2022) or cascade-based approach (Bahar et al., 2020). The cascade-based approach first performs a speech-to-text through an automatic speech

1

recognizer (ASR) then the transcribed source text to desired target text using a text-to-text machine translation system and finally a text-to-speech system transform the translated text to speech in the desired language.

In addition to managing the individual components of our cascade-based system, we face significant challenges with F2F translation, particularly in the areas of lip synchronization and voice or tone alignment. The process involves recording a speech, converting it to text using speech-to-text technologies, translating this text from the original to the target language, and then converting the translated text back to speech via text-to-speech systems. This process can be achieved using either a cascade or direct approach. A major challenge in this end-to-end F2F translation framework is ensuring that the lip movements sync with the translated speech track. This can be complex, as the duration of the translated speech may be longer or shorter than the original, depending on the distinct grammatical structures of the two languages. Additionally, the lips must move in a manner consistent with the frequency of the generated sound and must maintain the speaker's original voice or tones. Failing to do so can result in dubbing that appears off and unrealistic (Prajwal et al., 2020a).

F2F translation can have a huge impact on bridging the language gap in the educational sector. Numerous educational organisations create content to reach a global audience. Unfortunately, the lack of language intelligibility often prevents content consumers from fully utilizing the material at hand. While some videos provide manually executed dubs, these have their own set of challenges. It's true that manual translation tends to be more accurate than machine translation, but it also faces unavoidable limitations. These include cost, availability, efficiency, and most importantly, the quality of lip synchronization, which often falls short of the mark (Chung and Zisserman, 2017). Additionally, manual dubbing may be available in many but not all languages. The goal of the F2F translation system is to automate this dubbing process effectively and efficiently and make the online content available in whichever preferred language, thus overcoming the linguistic barrier between audio-visual content and the corresponding non-native consumer. This technology could also be used to assist language learning by giving students realistic and immersive opportunities to practise speaking and listening in a foreign language (Jha et al., 2019). Through this paper, we contribute to creating a more equitable and accessible education landscape that enables native individuals to learn and grow without any language barrier. Our main objective is to motivate every individual by providing a platform, through which one can grasp knowledge from videos in an unfamiliar language. To the best of our knowledge, our F2F translation framework is the first online end-to-end video translation system we bring up to the community.

2 Related Work

In this section, we present part of previous studies conducted in this field and summarise our learning and inspiration to better complement our research. Prajwal et al. (2020b) in their study explores the use of machine learning algorithms for lip-to-speech synthesis. The authors propose a new approach that takes into account individual speaking styles, resulting in increased accuracy. They use audio-visual data to train deep neural networks to capture unique lip movements and speaking styles, resulting in speech synthesis that is close to the original. The results show that their method outperforms current methods and produces speech that is similar to natural speech. K R et al. (2019) outlines a system for automatically translating speech between two people speaking different languages in real-time. The authors propose a multi-modal approach to translation that makes use of both audio and visual cues. This is accomplished by incorporating a novel visual module, LipGAN, for generating realistic talking faces in real-time from the translated audio. Their approach outperforms existing methods, demonstrating the potential for real-time F2F translation in practical applications. Ritter et al. (1999) in their research examines the development of a translation agent capable of performing real-time F2F speech translation. The authors present a multi-modal approach to translation that combines audio and visual information. They use machine learning algorithms to analyse each speaker's lip movements, speech, and facial expressions to produce a real-time audio-visual output with the speaker's face and synchronised lip movement. The results show that their method produces accurate translations and has the potential for practical applications in real-world scenarios. For translation, Chitralekha[1] is a valuable tool because it efficiently

---

[1] https://github.com/AI4Bharat/Chitralekha

creates multi-lingual subtitles and voice-overs for informative videos. However, it may not be a efficient for longer-length videos. Lastly, Huang et al. (2017) presented a novel problem of unpaired face translation between static photos and dynamic videos, which could be used to predict and improve video faces. To accomplish this task, the authors propose using a CycleGAN model with an identity-aware constraint. The model is trained on a large face dataset and tested on a variety of face images and videos. The results show that the proposed method can effectively translate faces between images and videos while preserving the individual's identity, outperforming existing methods.

3 The TRAVID Framework

Our framework 'TRAVID' is capable of generating translated videos from English to four Indian languages: Bengali, Hindi, Nepali, and Telugu. Flask[2] has been used as the foundation of our application, providing various built-in functionalities for building a Python-based web application. For the server-side and database, we utilize Python 3.9. In terms of audio and video processing, we primarily rely on the libraries Librosa[3] and ffmpeg[4]. These libraries provide extensive capabilities for audio and video processing, manipulation, and rendering. The primary objective of this work is to effectively and efficiently translate spoken language from an input video. Additionally, we aim to generate audio that resembles the speaker's voice and synchronize the translated speech with the speaker's lip movements. The entire process begins by obtaining the source video, target language, and speaker's gender (for voice model selection) as input from the user through our web interface. Behind the scenes, the task is divided into three sub-tasks: (1) Audio-to-Text Processing, (2) Text-to-Audio Processing, and (3) Video Processing. The steps involved in this process are depicted in Figure 1.

3.1 Audio to Text Processing

The input video, in our case an MPEG-4 (.mp4) file, is initially converted to a Waveform Audio (.wav) file using FFmpeg. This conversion enables us to perform text detection from the audio rather than the video file. Subsequently, we employ Librosa to identify non-mute sections within the 'start' and
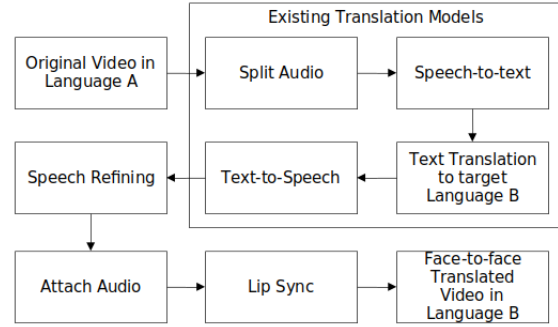

Figure 1: Steps involved in Video Translation System

'end' frame indexes, which are stored in a silence array. Each element of the silence array represents a small audio chunk, aiding in reducing system load and enhancing the overall efficiency of the framework during audio processing. Next, we convert each audio chunk from the silence array into an individual text chunk using Speech Recognition[5]. This library utilizes Google's Cloud speech API[6] to covert text from speech. Finally, Deep-translator[7] is employed to translate the generated text into the target language. Deep Translator utilizes the state-of-the-art Google Translate Ajax API[8] to generate the desired target language translation. The translated texts are stored and subsequently passed to the audio speech engine for further processing.

3.2 Text to Audio Processing

The translated text in the target language is inserted to the gTTS[9] library, which converts the text into speech and saves it as an audio file. This marks the completion of the speech generation process and initiates the speech refinement process. In order to match the audio target length with the source audio length, adjustments are made. The length of the translated speech may differ from that of the original speech. To address this, the speech speed is modified to align with the original audio file. The "Fixed Pitch-Shifting" technique is employed to ensure that the generated speech closely resembles the voice of the original speaker. Librosa provides the capability to detect the frequency of the audio and shift the pitch of the audio time series from one musical note to another (Rosenzweig et al., 2021). In the context of voice cloning, the mean frequency of the audio is determined, with the lower note con-

---

sidered as F2 (87.31 Hz) and the higher note as G6 (98.00 Hz). This frequency range represents the average range of human speech. The calculation of the steps required for shifting (n_steps) is performed using Equation 1.

$$n\_steps = \log_2(\frac{f_{src}}{f_{tgt}})^2 \qquad (1)$$

The variable f_{src} refers to the frequency of the source audio and f_{tgt} refers to the corresponding target audio. With this, the Text to Audio Processing engine gives the desired audio to the video-processing engine.

3.3 Video Processing for Lip Synchronization

We have utilized a lip-synchronization network called Wav2Lip (Prajwal et al., 2020a) for the purpose of lip-syncing and generating talking face videos. This model has been trained on the LRS2 training set and demonstrates an approximate accuracy of 91% on the LRS2 test set. The video sub-network of the model examines each frame of the source video and identifies faces, with a particular focus on the lip region. The relevant audio segment is then fed into the speech sub-network component of Wav2Lip, which modifies the input face crop to emphasize the lips area and produces the final video output. Throughout this process, the lip portion of the source video is replaced by concatenating the current face crop with the lower half of the detected face. By leveraging the translated speech and the source video, Wav2Lip generates lip-synced translated videos. The resulting translated video is subsequently presented on our front-end for display.

4 Demo Scenarios

Our framework TRAVID has a visually appealing landing page[10], which has an overview of the framework (cf. Figure 2). A demonstration video of our system is available on YouTube[11]. The demo User Interface (UI) is designed a landing page, have been carefully crafted to provide a seamless and intuitive navigation experience. The landing page is effectively communicates its purpose and functionality without the need for extensive instructions or guidance. Users can easily understand what the page offers and how to navigate it intuitively. The page is organized into distinct sections that make it

---

easy for users to locate and access the information they are looking for. This organization achieved through the use of clear headings, visually distinct sections, or a logical flow of content. The top menu bar is mentioned as a key element of the landing page, providing menu options that direct users to different feature pages. This menu bar remains accessible and visible to users across different sections of the landing page, allowing them to navigate to specific areas of interest easily. Upon signing in or signing up as a new user, the statement states that the user will be directed to the core section of the demo. This core section is the central part of the landing page, where users can access the main functionality and key features of TRAVID.
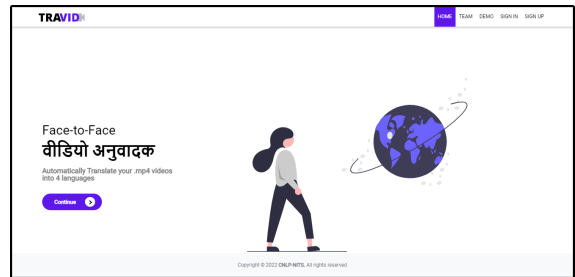

Figure 2: Homepage of TRAVID

The upload page shown in Figure 3 includes two drop-down menus: one for selecting the desired language for translation and another for choosing the output voice model (speaker). There are two options available for video input on the upload page: live recording, which allows users to capture real-time audio-visual input using their device's camera and microphone, or accepting pre-saved audio-visual content from the system. After receiving the input, the back-end framework, discussed in Section 3, initiates the translation process. Once the text, audio, and video processing are complete, the output page displays the translated video alongside the source video. Users have the option to download the source video and translated text. Additionally, they can provide reviews based on the output they received, which can help us improve and enhance the user-friendliness of our system.

The output page, depicted in Figure 4, provides a clear presentation of the original input video and the generated output video side-by-side. It offers convenient options to play and review both videos simultaneously. Additionally, users can save and download both the translated video and a translated text document. Furthermore, users can explore the demo section, which displays test case videos, to
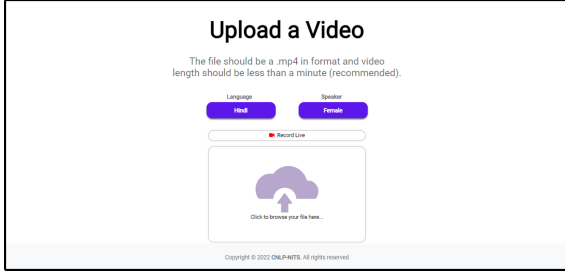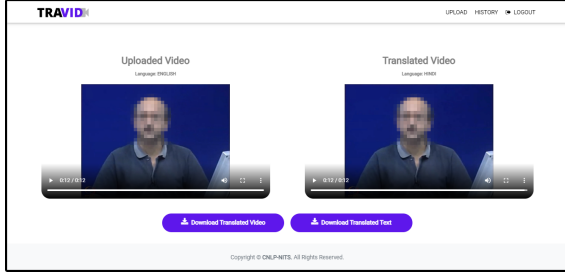
Figure 3: Upload page of TRAVID

Figure 4: Output page of TRAVID

gain an understanding of the translation quality that TRAVID's translation model can produce. In addition to the demo section, TRAVID includes a feedback page where users can rate videos alongside their translated output according to specific criteria and provide feedback to enhance our framework (refer to Figure 5).
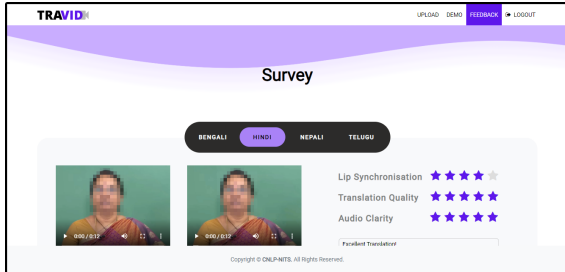

Figure 5: Survey page of TRAVID

5 Evaluation

To gauge the effectiveness of our method, We conducted a user study to assess the quality of our lip-synced translations, with participants asked to rate the translation quality, lip synchronization, and audio clarity. Evaluators compared the target video with the source language video clip and provided rankings for the quality of the output video on a scale of 1 to 5. The collected ratings were used to calculate inter-annotator agreement using Cohen's κ (Cohen, 1960), Fleiss' κ (Fleiss, 1971), and Pearson's r (Pearson, 1895) scores. Inter-annotator

agreements were computed for all four languages: Bengali, Hindi, Nepali, and Telugu. Table 2 displays the agreement scores for each language based on Lip Synchronization (Lip Sync), Translation Quality (TQ), and Audio Quality (AQ). The ratings were collected by comparing the translated videos to source videos from 5 indigenous users for each of the selected languages. Moreover, a manual examination was conducted by professional evaluators, and the results are presented in Table 1. Further details regarding inter-rater agreements can be found in Appendix A, B, C.

The core component of TRAVID is based on the CNLP_NITS system, which emerged as a top performer in the Lip-Sync 2021 Challenge shared task[12]. The objective of this challenge was to convert English input videos into Hindi or Tamil output videos while ensuring lip synchronization. The quality of the Hindi Task-1 was assessed using various evaluation metrics such as Lip-Sync Quality (LSQ), Fluency Consistency (FC), Semantic Consistency (SC), and Overall User Experience (UX). Evaluators rated the quality of the output videos on a scale of 1 to 5, with higher scores indicating better quality when compared to the source language video clip. Our system, CNLP-NITS (NIT Silchar), achieved the top position with a final score of 3.84, surpassing the Baseline system (IIT Madras) with a score of 3.68 and TeamCSRL (CS RESEARCH LABS) with a score of 3.46. The comparison of the three evaluation matrices revealed a high degree of similarity. The results indicate that the translations were perceived as reasonable and easy to understand by the majority of participants, leading to fair to moderate agreement and a positive correlation among their assessments. The overall scores of the Lip-Sync Challenge 2021 are presented in Table 1.

Hindi Task 1			Hindi Task 2
CNLP_NITS	Baseline	TeamCSRL	CNLP-NITS
3.86	3.49	3.08	3.37
3.63	3.52	3.32	3.29
3.94	3.87	3.92	3.40
3.94	3.83	3.51	3.38
3.84	3.68	3.46	3.36

Table 1: Leadership Positions Based on NLP Challenges

6 Limitation

There is a constraint when uploading huge videos; the system may require a lot of computational re-

---

[12]Leaderboard of Lip-Sync Challenge 2021

Language	Cohen's κ			Fleiss' κ			Pearson's r		
	Lip Sync	TQ	AQ	Lip Sync	TQ	AQ	Lip Sync	TQ	AQ
Bengali	0.600	0.276	0.296	0.698	0.379	0.215	0.586	0.295	0.258
Hindi	0.510	0.390	0.391	0.595	0.291	0.345	0.400	0.292	0.318
Nepali	0.170	0.576	0.328	0.112	0.541	0.218	0.171	0.501	0.256
Telugu	0.287	0.330	0.388	0.214	0.291	0.376	0.212	0.271	0.331

Table 2: Average agreement Scores for evaluation of TRAVID generated Videos

sources and data to process and render the translated video. Also, so far we have trained our models only on a single speaker, so videos with multiple speakers may yield poor results. The quality of speech recognition and translation may vary depending on factors such as noise, accent, dialect, etc. The generated faces may not look natural or convincing enough for some applications or scenarios such as low lighting, moving background, etc. Considering the state-of-the-art ASR system in use, the ASR results were already deemed satisfactory, thus not necessitating the utilization of lip sync from the video as an additional multimodal input for accuracy enhancement. Still, the system may be unable to handle linguistic challenges such as idioms, metaphors, slang, etc. The method may not be able to capture cultural nuances and context that affect the meaning and tone of speech, as the syntheses is machine generated. The biggest bottleneck in our current system which uses cascade approach is time complexity, due to the need for extensive computation and audiovisual processing.

7 Conclusion

In this paper, we presented an end-to-end video translation system that effectively translates the speaker's native language into the local language of the audience while synchronizing the translated speech with the speaker's lip movements. Our proposed system demonstrates the potential of lip-synced Face-to-Face video translation in enhancing communication between individuals from diverse linguistic backgrounds.

Moreover, our video translation system represents a significant advancement in overcoming the limitations of traditional language translation methods. By incorporating lip synchronization and matching translated speech with the speaker's lip movements, we created an immersive and realistic experience for users. This additional feature, along with the ability to capture nonverbal cues, adds depth and context to the translated content, making it more effective and engaging, especially in educational settings.

Through our system's participation and success in the Lip-Sync 2021 Challenge, we have demonstrated its capability and superiority in achieving accurate lip synchronization and high-quality translations. The evaluations and ratings obtained from both users and professional evaluators validate the effectiveness of our approach, further emphasizing its potential for real-world applications. The positive feedback received through human assessments, as discussed in the evaluation section above, validates the effectiveness of our system. However, further research is necessary to enhance the quality of lip-syncing and explore the system's applicability in different languages and more naturalistic settings. With the ongoing advancements in technology and the increasing demand for multilingual communication, our system has the potential to revolutionize the way language translation is approached. Its adaptability to low-resource system settings makes it accessible and valuable in diverse environments.

Moving forward, we envision further enhancements and refinements to our video translation system, leveraging the advancements in natural language processing, computer vision, and machine learning. To boost video translation efficiency, the videos can be broken into smaller segments, leverage GPUs for parallel processing, batch translate frames, subsample for reduced load, and implement caching for reused translations. By continuously improving the accuracy, fluency, and naturalness of translated content, we aim to provide an unparalleled experience for users, fostering effective cross-cultural communication and knowledge sharing.

In summary, our video translation system stands as a promising solution to the challenges of multilingual communication, offering a comprehensive and immersive experience that unlocks new possibilities for global connectivity and understanding.

Ethics Statement

We honour the Code of Ethics set by IJCNLP-AACL in our paper and abide by them. We have

used open-source materials in our development to produce new, better and useful resources, which will be made open-source for the keen mind to feed upon and make more improvements in the future. We have not written or in any way propagated false knowledge, hateful speech and anything controversial that may give rise to conflict. We intend good for the brighter future of mankind. We have not stolen anybody's work, and have properly cited and credited where credit is due. Our website does not feature any harmful content or advertisement. Our website is solely educational and is to be used for educational purposes only. Even though we reserve the right to use our paper and production in any way we see fit, we promise to extend them ethically and in an innovative manner.

References

Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. 2020. Tight integrated end-to-end training for cascaded speech translation. *CoRR*, abs/2011.12167.

Shiyu Chen, Yawen Zeng, Da Cao, and Shaofei Lu. 2022a. Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems*, 245:108598.

Shiyu Chen, Yawen Zeng, Da Cao, and Shaofei Lu. 2022b. Vision talks: Visual relationship-enhanced transformer for video-guided machine translation. *Expert Systems with Applications*, 209:118264.

Joon Son Chung and Andrew Zisserman. 2017. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pages 251–263, Cham. Springer International Publishing.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G. Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. 2022. Cascade or direct speech translation? a case study. *Applied Sciences*, 12(3).

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhiwu Huang, Bernhard Kratzwald, Danda Pani Paudel, Jiqing Wu, and Luc Van Gool. 2017. Face translation between images and videos using identity-aware cyclegan. *arXiv preprint arXiv:1712.00971*.

Abhishek Jha, Vikram Voleti, Vinay Namboodiri, and C. V. Jawahar. 2019. Cross-language speech dependent lip-synchronization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7140–7144.

Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1428–1436, New York, NY, USA. Association for Computing Machinery.

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020a. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, New York, NY, USA. Association for Computing Machinery.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020b. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805.

Max Ritter, Uwe Meier, Jie Yang, and Alex Waibel. 1999. Face translation: A multimodal translation agent. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*.

Sebastian Rosenzweig, Simon Schwär, Jonathan Driedger, and Meinard Müller. 2021. Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings. In *2021 24th International Conference on Digital Audio Effects (DAFx)*, pages 121–128. IEEE.

Harold Somers. 1992. An introduction to machine translation.

A Cohen's Kappa

The definition of κ is

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o denotes the relative observed agreement among raters and p_e denotes the hypothetical probability of chance agreement, calculated by using observed data to compute the odds of each observer randomly seeing each group.

B Fleiss' Kappa

The Fleiss' kappa κ can be defined as,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

The factors $1 - \bar{P}_e$ and $\bar{P} - \bar{P}_e$ indicate the level of agreement that can be reached above chance and the level of agreement that has actually been attained.

C Pearson Correlation Coefficient Analysis

Given a pair of random variables (X, Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where cov is the covariance, σ_X is the standard deviation of X and σ_Y is the standard deviation of Y.

Table [3, 4, 5] displays the calculated Pearson Correlation Coefficient (PCC) values for Bengali, Table [6, 7, 8] displays the calculated PCC values for Hindi, Table [9, 10, 11] displays the calculated PCC values for Nepali, Table [12, 13, 14] displays the calculated PCC values for Telugu with respect to all three assessment areas, namely Lip Synchronization, Translation Quality, and Audio Quality. Where R1-R5 (Rater 01-05) denote five raters who assessed our system in all four languages.

	R1	R2	R3	R4	R5
R1	1	-0.2	0.81	0.72	0.46
R2	-0.2	1	0.75	0.81	0.85
R3	0.81	0.75	1	0.84	-0.02
R4	0.72	0.55	0.84	1	0.84
R5	0.46	0.85	-0.02	0.84	1

Table 3: PCC for Bengali Lip Synchronization

	R1	R2	R3	R4	R5
R1	1	0.32	0.86	0.21	0.87
R2	0.32	1	0.68	0.12	0.38
R3	0.86	0.68	1	-0.64	0.57
R4	0.21	0.12	-0.64	1	-0.42
R5	0.87	0.38	0.57	-0.42	1

Table 4: PCC for Bengali Translation Quality

	R1	R2	R3	R4	R5
R1	1	0.79	-0.18	0.25	0.33
R2	0.79	1	0.42	0.39	0.12
R3	-0.18	0.42	1	-0.19	0.39
R4	0.25	0.39	-0.19	1	0.26
R5	0.33	0.12	0.39	0.26	1

Table 5: PCC for Bengali Audio Quality

	R1	R2	R3	R4	R5
R1	1	-0.12	0.26	-0.06	0.21
R2	-0.12	1	0.55	0.39	0.65
R3	0.26	0.55	1	0.66	0.72
R4	-0.06	0.39	0.66	1	0.74
R5	0.21	0.65	0.72	0.74	1

Table 6: PCC for Hindi Lip Synchronization

	R1	R2	R3	R4	R5
R1	1	-0.42	-0.42	0.51	0.66
R2	-0.42	1	0.88	0.56	-0.12
R3	0.36	0.88	1	0.76	0.62
R4	0.51	0.56	0.76	1	-0.11
R5	0.66	-0.12	0.62	-0.11	1

Table 7: PCC for Hindi Translation Quality

	R1	R2	R3	R4	R5
R1	1	0.22	0.68	-0.01	0.72
R2	0.22	1	-0.19	0.3	0.59
R3	0.68	-0.19	1	0.12	0.26
R4	-0.01	0.3	0.12	1	0.49
R5	0.72	0.59	0.26	0.49	1

Table 8: PCC for Hindi Audio Quality

	R1	R2	R3	R4	R5
R1	1	-0.2	-0.01	0.52	0.11
R2	-0.2	1	0.16	0.29	0.21
R3	-0.01	0.16	1	-0.22	0.34
R4	0.52	0.29	-0.22	1	0.51
R5	0.11	0.21	0.34	0.51	1

Table 9: PCC for Nepali Lip Synchronization

	R1	R2	R3	R4	R5
R1	1	0.69	0.58	-0.09	0.66
R2	0.69	1	0.49	0.63	-0.21
R3	0.58	0.49	1	0.66	0.79
R4	-0.09	0.63	0.66	1	0.81
R5	0.66	-0.21	0.79	0.81	1

Table 10: PCC for Nepali Translation Quality

	R1	R2	R3	R4	R5
R1	1	0.11	-0.02	-0.15	0.29
R2	0.11	1	0.59	0.63	0.82
R3	-0.02	0.59	1	0.26	-0.09
R4	-0.15	0.63	0.26	1	0.12
R5	0.29	0.82	-0.09	0.12	1

Table 11: PCC for Nepali Audio Quality

	R1	R2	R3	R4	R5
R1	1	0.19	0.26	-0.02	-0.16
R2	0.19	1	-0.05	0.19	0.22
R3	0.26	-0.05	1	0.39	0.51
R4	-0.02	0.19	0.39	1	0.59
R5	-0.16	0.22	0.51	0.59	1

Table 12: PCC for Telugu Lip Synchronization

	R1	R2	R3	R4	R5
R1	1	-0.18	-0.18	0.44	-0.13
R2	-0.18	1	0.63	0.31	0.58
R3	0.39	0.63	1	0.58	0.67
R4	0.44	0.31	0.58	1	-0.01
R5	-0.13	0.58	0.67	-0.01	1

Table 13: PCC for Telugu Translation Quality

	R1	R2	R3	R4	R5
R1	1	0.83	0.11	0.32	0.19
R2	0.83	1	0.59	0.3	0.63
R3	0.11	0.59	1	-0.19	0.61
R4	0.32	0.41	0.3	1	-0.08
R5	0.19	0.63	0.61	-0.08	1

Table 14: PCC for Telugu Audio Quality