Towards Speech to Speech Machine Translation focusing on Indian Languages

Vandan Mujadia[1], **S. Umesh**[2], **Hema A. Murthy**[2], **Rajeev Sangal**[1], **Dipti Misra Sharma**[1]
[1]IIIT Hyderabad, India; [2]IIT Madras, India
vandan.mu@research.iiit.ac.in,umeshs@ee.iitm.ac.in
hema@cse.iitm.ac.in, sangal@iiit.ac.in, dipti@iiit.ac.in

Abstract

We introduce an SSMT (Speech to Speech Machine Translation, aka Speech to Speech Video Translation) Pipeline[1], as a web application for translating videos from one language to another by cascading multiple language modules. Our speech translation system combines highly accurate speech to text (ASR) for Indian English, pre-possessing modules to bridge ASR-MT gaps such as spoken disfluency and punctuation, robust machine translation (MT) systems for multiple language pairs, SRT module for translated text, text to speech (TTS) module and a module to render translated synthesized audio on the original video. It is user-friendly, flexible, and easily accessible system. We aim to provide a complete configurable speech translation experience to users and researchers with this system. It also supports human intervention where users can edit outputs of different modules and the edited output can then be used for subsequent processing to improve overall output quality. By adopting a human-in-the-loop approach, the aim is to configure technology in such a way where it can assist humans and help to reduce the involved human efforts in speech translation involving English and Indian languages. As per our understanding, this is the first fully integrated system for English to Indian languages (Hindi, Telugu, Gujarati, Marathi, and Punjabi) video translation. Our evaluation shows that one can get 3.5+ MOS score using the developed pipeline with human intervention for English to Hindi. A short video demonstrating our system is available at https://youtu.be/MVftzoeRg48.

1 Introduction

India writes in many languages and speaks in many more tongues [2]. It is a geographically vast multilingual society with 22 recognized languages. The languages constitute 1.17+ billion speakers across 28 states and 7 union territories. According to the 2011 Census, while 129 million (10.6%) Indians speak English, only 259,678 (0.02%) Indians speak it as their first language. And only 8% Indians read newspapers in English while others prefer news in their local languages. As stated in a report from karnataka (gfgc.kar.nic.in, 2014), about 40% of all enrolled students from non-metropolitan regions fail to achieve their educational goals because they are unable to cope with English and very few study materials are available in native Indian languages [3]. Same is true for medical and health awareness related content. There is a huge void of content in Indian languages that necessitates urgent action. One solution for this problem is to use translation. Such translations can help different language speakers to seamlessly communicate with each other. With this work, we aim to ease this language barrier through speech to speech machine translation (SSMT) by providing a baseline system for video translation.

Generally, there are two ways to implement speech-to-speech translation (SSMT): the first approach is to cascade systems of Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS); second, direct end to end speech translation. Cascaded SSMT systems have been successfully demonstrated for English and European languages, but one finds minimal work done for Indian languages. Recent work in direct end to end speech translation (Translatotron) (Jia et al., 2019, 2022) attempts to directly translate speech from one language into speech in another language with the source speaker's voice in the translated speech. It achieves high translation quality on two Spanish to English datasets, although the reported performance is poorer than a baseline cascade of speech translation and TTS models (Jia et al., 2019, 2022). We are not aware of any cascaded or direct speech-to-speech translation work involving Indian languages.

[1] https://ssmt.iiit.ac.in/ssmtiiith
[2] shorturl.at/dnSV8
[3] shorturl.at/crCJ7

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics
System Demonstrations, pages 161–168
May 2-4, 2023 ©2023 Association for Computational Linguistics

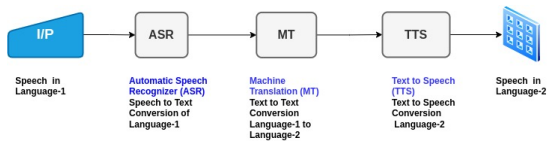

Figure 1: Speech to Speech Machine Translation: Cascading Approach

For the purpose of our work, we decided to implement a cascaded SSMT system. We also analyse gaps between the automatic modules and address them with pre-processing and post-processing tools. These gaps are :speech disfluencies, domain processing, and target language subtitling. Our system takes an English video as an input and outputs the same video in the chosen Indian language. Our proposed video translation pipeline is user-friendly, flexible, and easily accessible with following key modules:

1. highly accurate speech to text (ASR) for Indian English, 2. pre-possessing modules to bridge ASR-MT gaps such as spoken disfluency and punctuation, 3. robust machine translation (MT) systems for multiple language pairs, 4. SRT module for translated text, 5. text to speech (TTS) module, and 6. a module to render translated synthesized audio on the original video.

2 Approaches

The canonical approach as shown in Figure-1 includes automatic speech recognition (ASR) to transcribe source language speech to text and then, machine translation (MT) to translate transcribed text into target language, and at the end, text to speech synthesis (TTS) to generate speech in target language from the translated text (Sperber and Paulik, 2020; Wahlster, 2000; Lavie et al., 1997).

In this work, we aim to develop a system which can translate speech or video in English to selected Indian languages. While following a cascaded approach, one can not directly chain modules such as ASR, MT, and TTS as it is a well known fact that spoken language has various idiosyncrasies. These include lack of well-formed sentences and disfluencies (Rao et al., 2007). Traditional machine translation systems are trained on well formed, written, and grammatical pairs of sentences. Therefore, it is crucial to address these aspects before directly translating transcribed text using machine translation. Similarly, to sync original video with the translated text, time-stamping translated text is an essential step before text to speech synthesis. Also, video content syncing (speaker lip, video content) with the generated speech is another important factor in making system complete.

In recent times, technological advancements have enabled ASR, MT, and TTS to make quantum leaps. Today, computers are capable of doing these with greater accuracy and efficiency than ever before. But they cannot be expected to be 100% accurate. Human effort is still required to correct or edit these outputs. Therefore, in our pipeline, we also include steps where a human can intervene after each automatic module which eventually reduces the overall human effort for the task. For the translation of technical lectures, domain processing is one of the important steps before translating transcribed text into the target language. Domain processing is included as a pre-processing step, where identification of domain and domain terms are carried out before machine translation.

In this pipeline, we also aim to develop an interface for state-of-the-art video to video machine translation between English to 5 Indian languages (Hindi, Telugu, Gujarati, Marathi, and Punjabi) along with pre-processing of ASR output to make it translatable as well as post-processing of machine translation output to make it suitable for dubbing and video syncing. The subsequent section explains the SSMT pipeline, the interface, and the process in it. In section 4, we discuss the pipeline performance and conclude in section 5.

3 Process

Here, our task is to combine technologies to make video to video translation possible for English to Indian languages as shown in Figure-2. As discussed earlier, there are gaps between the components that need special processing. In order to fill these gaps, we link the major components with pre/post processing support tools with the provision of human interventions. These are before and after each major language component. Therefore, as visible in different colors in Figure-2, we categorize the overall process into 4 major parts: **Input/Output**, **Core technology**, **Pre/Post Processing Support Tools**, and **Pre/Post Editing** as Human Intervention.

3.1 Input/Output

In this, we deal with the input and output process, tools, and the user interface of the pipeline. This includes uploading a video, processing the video,

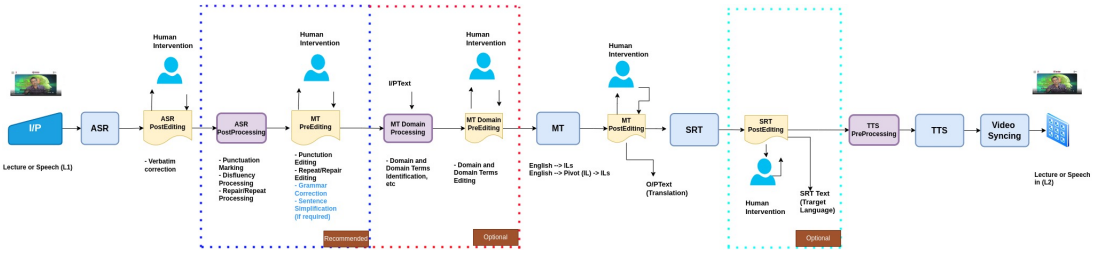

Figure 2: Worked out Speech to Speech Video Translation Process

displaying the translated video and subtitling in the target language. These are shown in dark blue color in Figure-2 at start and end. Figure-3 shows the application page where users can upload video, select one of the target languages, gender for target video voice, and start the process by clicking *START* button. Figure-4 shows the screenshot of the interface after it completes the entire process (visible as Speech to Text, English to Indian Language MT, and Text to Speech). Users can play the source language video and choose the subtitled language either as English or the opted one. Users can play the same video in the opted language by clicking the language button as shown in blue color Figure-4.

3.2 Core Technology

This category includes core components such as ASR, MT, SRT, TTS, and Video Syncing. They are in sky blue color boxes in Figure-2. We describe each of these core components in detail and also point out where human intervention is required.

3.2.1 ASR for Transcription

Transcription is the process of translating an audio (of a video lecture) into text. This is usually carried out using automatic speech recognition (ASR) technology, human transcriptionists, or a combination of the two. ASR refers to the technologies developed to process human speech automatically and convert it into text (Juang and Rabiner, 2005). For this work, we have integrated the ASR developed by IIT-Madras (Shetty et al., 2020; Arunkumar and Umesh, 2022). Along with the verbatim output, ASR is also tuned to provide timestamp for each token which is directly used to create subtitle for the video/audio. The ASR system achieves 6 and 13 WER (Favre et al., 2013)[4] on the general and technical domain, respectively. The different categories of errors in the automatically transcribed

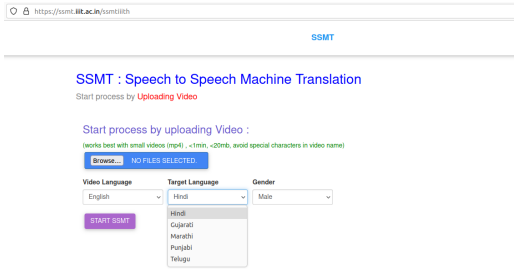

Figure 3: Speech to Speech Video Translation: Input Screen

text constitute missed words, wrongly transcribed words, spelling errors etc (Zafar et al., 2004). An editing functionality will be helpful to correct these errors for further processing.

3.2.2 MT for Translation

Translation is the process through which text content is transferred from a source language into a target language. The translation task can be carried out using machine translation systems or by human translators (Somers, 2011). For this task, we have integrated an MT system deployed by LTRC-IIIT-Hyderabad using methods presented in (Mujadia and Sharma, 2022, 2021a,b). The MT systems [5] achieved 36.33, 21.61, 18.73, 18.36, 15.89 BLEU scores (Post, 2018) on Flores Benchmarks (Goyal et al., 2022) for English-Hindi, English-Telugu, English-Gujarati, English-Punjabi, and English-Marathi language pairs, respectively. The state-of-the-art MT technology has not yet reached a level where it can directly provide publishable, usable, and accurate output in the target language. To address this, providing multiple translation options could be one possible solution. Our interface supports multiple translation options by leveraging multiple MT models for the involved language pairs. In this process, a user can choose one translation output from the available choices that can then

[4] https://github.com/Speech-Lab-IITM/English_ASR_Challenge

[5] http://ssmt.iiit.ac.in/translate

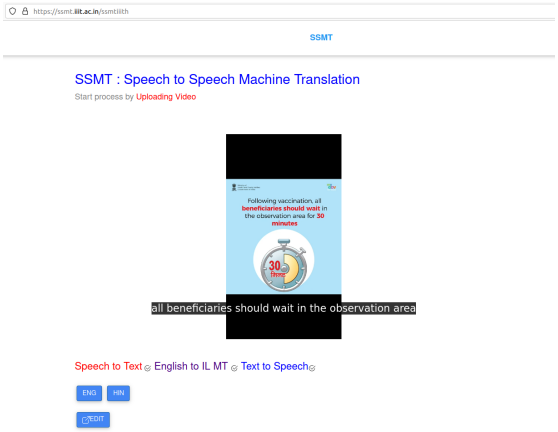

Figure 4: Speech to Speech Video Translation: Output Screen

be used for all the subsequent processing. The machine translation task becomes even more challenging when it encounters technical text. To support technical domain translation, we integrated fine-tuned machine translation systems which learn to retain already marked domain terms in the source script (Bak et al., 2021) along with domain adaptation (Ala et al., 2021; Ala and Sharma, 2020).

On top of this, human intervention in the form of post-editing is necessary to achieve fluency, adequacy, and faithfulness for the translated text. We have added the functionality for post-editing machine translation outputs that can later be used for further processing.

3.2.3 SRT - Subtitling Translated Text

Subtitling is the process of displaying spoken utterance as a text on the video screen. It is an audiovisual translation with a set of rules and guidelines [6]. The subtitle for a video is derived using the utterance speech and word alignment from ASR. We have developed an in-house mapping module which places translated text into timestamps based subtitles using source text mapping. It plays a vital role in speech to speech video translation as it helps to keep the translated text in sync with the video frame.

3.2.4 TTS for Text to Speech & Video Syncing

We integrated a Text to Speech (TTS) and video syncing system from IIT-Madras [7] (M et al., 2021; Mukherjeee et al., 2021). It uses target language subtitles and source speaker pauses to synthesize speech in the target language. To match and align the source video and synthesized audio duration, a video syncing module interpolates several frames in the middle of two adjacent frames of the original video. The integrated TTS and video syncing system has average Mean Opinion Score (MOS) (1-5) of 4 and 3.5, respectively.

3.3 Pre/Post Processing Support Tools

As discussed, to fill the gaps between core components, we have introduced pre and post processing tools. They are shown in pink color boxes in Figure-2. To bridge the gap between ASR and MT, we are using ASR post-processing tools such as punctuation marker/corrector, speech disfluency removal, repair and repeat identifier and processing. Similarly, for technical lectures, identification of domain and domain terms play an important role in translation. Therefore, we added these as a pre-processing utility to the machine translation system. Below subsections discuss each of these support tools in detail.

3.3.1 ASR Post-Processing

To prepare the raw ASR text for MT, we included 3 supporting tools as shown in Figure-2. We call these as ASR post-processing steps. First comes the **Punctuation Marker**. It is a standalone tool where it corrects, deletes, adds existing punctuation from ASR. We have integrated it for English (Mujadia et al., 2021). The second step involves **Disfluency Processing** where it removes filled pause (ahh, uhh, ah, etc) and Pet Phrases (okay, ok, so, right, etc) which are very frequent in speech. The final pre-processing step is **Repair/Repeat Processing** (Heeman, 1997), where it identifies repeated occurrences for a given ASR transcript and remove duplicate word sequences.

3.3.2 MT Domain Pre-Processing

After the ASR Post-Processing, we have integrated text based domain identification (a classification task) and text based domain term identifier as shown in Figure-2. Here, once the domain of a text is identified using the domain classifier (Sharma et al., 2020a), a domain term identifier (Sharma et al., 2020b) is used which is based on domain specific dictionaries and TextRank (Mihalcea and Tarau, 2004). These are later fed into the machine translation system which handles these identified domain terms differently in target language translation. For the purpose of this work, we integrated

[6] https://www.ted.com/participate/translate/subtitling-tips

[7] https://www.iitm.ac.in/donlab/tts/

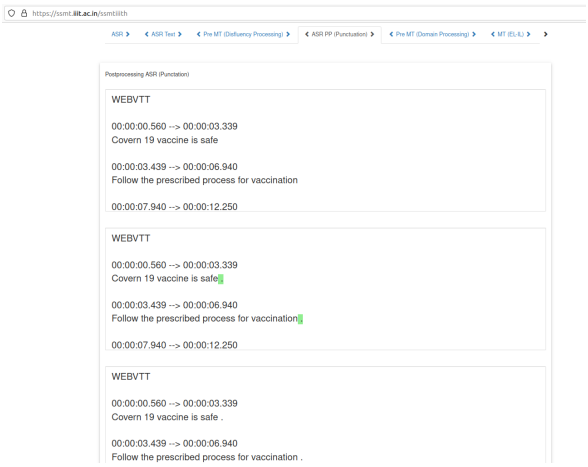

Figure 5: SSMT: ASR Post-processing for Punctuation

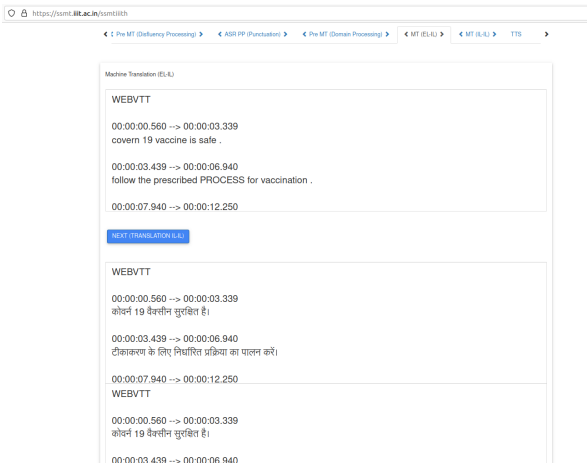

Figure 6: SSMT: MT Post-editing for Machine Translation

a domain classifier and domain dictionaries for law, computer science, biochemistry, general health awareness, and communication skill domains.

3.3.3 TTS Pre-Processing

Translated subtitled text or SRT text is a sequence of words along with a timestamp following certain language subtitling rules[8]. A valid SRT block may or may not represent a valid sentence. It can have multiple sentences or a part of a sentence as a SRT block. TTS requires valid sentences as an input to maintain target language speech flow. Therefore, the TTS pre-processing tool adjusts the subtitle timeline and keeps a valid sentence in one timeline. This tool is positioned after the SRT module in the SSMT pipeline as shown in Figure-2.

3.4 Pre/Post Editing as Human Intervention

Automatic tools are not 100% accurate. This warrants human intervention in the process. It is also required to control error propagation from one component to another. For this, we introduce a human intervention step after every automatic process. Figure-2 shows them in yellow color boxes. In our interface, one just needs to press"Edit"button to enable the editing mode after it completes processing as can be seen in Figure-4.

Figure-5 shows the editing page for punctuation marking post ASR tool. Here, it has 3 different text boxes, where 1^{st} shows the input text to the tool (here punctuation marker), 2^{nd} shows the output of the module where the differences can be viewed in green color. A user can edit in the 3^{rd} text box to make any further corrections. After this, the user needs to press the "Next" button to rerun the pipeline with the updates. Similar visual structures have been provided for editing throughout the interface. The user can navigate between different steps in the whole pipeline by clicking on the "Next" and "Previous" buttons.

Figure-6 shows the edit page for machine translation. Here, the 1^{st} text box shows the input for the MT which is received after the domain preprocessing step. Here, automatic domain terms are being shown in upper case. 2^{nd}, 3^{rd} *and* 4^{th} text boxes show translation outputs generated by different translation models. A user can pick one of them by clicking it or post-editing it. This edited/selected text box will be used for further processing in the pipeline. At each stage, the interface also gives flexibility to skip the human intervention and run the pipeline directly. As mentioned in Figure-1, we recommend that the post editing for ASR and MT output is quintessential; MT pre editing is done if required while other editing can be optional.

Core components and pre/post processing tools have been plugged based on their performance and availability. If better efficient systems are made available in future, then the pipeline has modularity to easily integrate them.

4 Performance

We evaluate the performance of the developed pipeline with two different metrics. They are: time taken to execute the pipeline and performance of major modules on their known evaluation metrics. One can access and execute presented SSMT pipeline using internet without installing specialized tools. The execution time for pipeline depends

[8] https://www.ted.com/participate/translate/subtitling-tips

V 1 CS	Options	Duration	ASR Verbatim (WER)	ASRtoMT (WER)	Eng-Hin MT (BLEU)	Eng-Hin MT + Domain (BLEU)	MOS (1-5)
1	Direct + No Punct	0:00:59	14.29%	21.74%	5.81	5.81	-
2	+ Fix len Punct	0:00:59	14.29%	23.23%	15.79	15.89	2.0
3	+ Punct by ASR	0:00:59	14.29%	22.28%	21.29	25.56	3.0
4	+ ASR PostPro	0:00:59	14.29%	18.48%	25.41	28.06	3.45
5	+ MT PreEdit	0:00:59	Gold	Gold	33.45	39.34	3.65
6	+ MT PostEdit	0:00:59	Gold	Gold	Gold	Gold	4.0
V 2 CS 1	Direct + No Punct	0:01:00	9.57%	18.48%	3.4	3.4	-
2	+ Fix len Punct	0:01:00	9.57%	19.47%	20.26	20.26	2.0
3	+ Punct by ASR	0:01:00	9.57%	16.11%	22.32	25.46	3.0
4	+ ASR PostPro	0:01:00	9.57%	12.32%	23.3	25.57	3.2
5	+ MT PreEdit	0:01:00	Gold	Gold	35.26	37.92	3.5
6	+ MT PostEdit	0:01:00	Gold	Gold	Gold	Gold	4.1

Table 1: Speech to Speech Video Translation Pipeline Evaluation at each stage; WER for ASR and ASERtoMT, BLEU for MT, Mean Opinion Score (MOS) score is for generated video in target language. Gold indicates human editing was carried out at that stage.

on the video length. On an average, it takes 1/3 of the video time for execution on a single GPU (NVIDIA-3080Ti) system. Due to resource constraints, for now we have set the input video length limit to 1 min in the interface, but this can be increased based on availability of compute infrastructure.

We created an evaluation dataset of 2 small English technical videos of **computer science domain for English-Hindi translation direction**. We hired experienced language professionals to carry out manual transcription and Hindi translation for these videos. We used WER (Favre et al., 2013) to evaluate **ASR verbatim (WER)** and **ASRtoMT Gaps (WER)** (that is verbatim + correct punctuation and without spoken disfluency). We used BLEU (Post, 2018) to evaluate English to Hindi machine translation **Eng-Hin MT (BLEU)** performance without and with domain pre-processing **Eng-Hin MT + Domain (BLEU)** to the MT. Mean Opinion Score (MOS) is used to evaluate generated Hindi speech and synced video. Table-1 shows the evaluation results for SSMT pipeline for 2 videos. 1^{st} rows "Direct + No Punct" of video 1 and video 2 show the results when ASR verbatim (without punctuation), MT, TTS, and video syncing modules are used. The 2^{nd} rows "+ Fix len Punct" of video 1 & 2 show the results when a punctuation symbol is placed after every 20 tokens on the direct ASR verbatim. The 3^{rd} rows "+ Punct by ASR" of video 1 & 2 show the results when punctuations are given by ASR along with ASR verbatim. The 4^{th} rows "+ Punct by ASR" of video 1 & 2 show the results when punctuation and disfluency processing are given by ASR post processing tools. Here, we can notice that rows from 1 to 4 for both the videos denote results for a fully automatic pipeline. For both the videos, ASR post processing tools along with domain processing for machine translation give best 18.48% and 12.32% WER scores for ASRtoMT respectively. Similarly, highest BLEU scores of 28.06 and 25.57 were achieved for machine translation with domain processing. Here, we got 3.45 and 3.2 MOS scores respectively for the Hindi audio synced video.

We have also measured the performance of the pipeline when there is human involvement in editing at major steps of the pipeline. Rows 5 of video 1 and 2 show the results after performing pre-editing for machine translation. Here, we clearly see a 12 BLEU score and 0.25 MOS score improvement in the machine translation and TTS quality, respectively when corrected texts are given to it. Rows 6 of both videos show results when post-edited machine translation output was passed to TTS and video syncing module. On the post-edited transla-

tion, we see an improvement of 0.5 MOS score. This indicates that speech to speech translation technology needs human intervention to get the best possible translated video. As speech-to-speech translation involving Indian languages is relatively a new area of research, it is difficult to compare our work with any end-to-end speech translation models.

5 Conclusion and Future Work

In this paper, we introduce an SSMT pipeline, an intelligent Speech to Speech Video Translation interface for English to Hindi, Telugu, Gujarati, Marathi, and Punjabi. This work demonstrates that speech to speech (video to video) translation is possible with a cascaded pipeline and support tools. We believe that large scale deployment of this can help lower the language barrier. The results also point out that human intervention is necessary to get high quality video translation output. In the near future, we aim to come up with benchmark corpora for speech to speech machine translation and evaluation involving English and multiple Indian Languages. We also plan to further improve the developed pipeline and its components to reduce involved human effort over a period of time. We will also plan to add multiple language directions to the pipeline in future.

Acknowledgement

We thank the reviewers for their insightful comments. We thank Pruthwik Mishra and Arafat Ahsan for their input at various stages of this work. This work is supported[9][10] by the Ministry of Electronics and Information Technology, Government of India.

References

Hema Ala, Vandan Mujadia, and Dipti Sharma. 2021. Domain adaptation for Hindi-Telugu machine translation using domain specific back translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 26–34, Held Online. INCOMA Ltd.

Hema Ala and Dipti Sharma. 2020. AdapNMT : Neural machine translation with technical domain adaptation for indic languages. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 6–10, Patna, India. NLP Association of India (NLPAI).

A Arunkumar and Srinivasan Umesh. 2022. Joint encoder-decoder self-supervised pre-training for asr. *Proc. Interspeech 2022*, pages 3418–3422.

Yunju Bak, Jimin Sun, Jay Kim, Sungwon Lyu, and Changmin Lee. 2021. Kakao enterprise's WMT21 machine translation using terminologies task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 804–812, Online. Association for Computational Linguistics.

Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, et al. 2013. Automatic human utility evaluation of asr systems: Does wer really predict performance? In *INTERSPEECH*, pages 3463–3467.

gfgc.kar.nic.in. 2014. Barriers of rural students in learning english in karnataka. https://gfgc.kar.nic.in/krpuram/FileHandler/9-4fa6882f-fb7e-4f38-b81c-cd344d1a73d9. (Accessed on 12/02/2022).

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Peter Anthony Heeman. 1997. *Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog*. University of Rochester.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Biing-Hwang Juang and Lawrence R Rabiner. 2005. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1:67.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. Janus-iii: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE.

[9]Sanction Order: 11(1)/2022-HCC(TDIL)-Part(2)/A/B/C

[10]Administrative Approval: 11(1)/2022-HCC(TDIL)-Part(2)

Mano Ranjith Kumar M, Jom Kuriakose, Karthik Pandia D S, and Hema A Murthy. 2021. Lipsyncing efforts for transcreating lecture videos in Indian languages. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 216–221.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2021. Deep contextual punctuator for NLG text (short paper). In *Proceedings of the Swiss Text Analytics Conference 2021, Winterthur, Switzerland, June 14-16, 2021 (held online due to COVID19 pandemic)*, volume 2957 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Vandan Mujadia and Dipti Sharma. 2021a. Low resource similar language neural machine translation for Tamil-Telugu. In *Proceedings of the Sixth Conference on Machine Translation*, pages 288–291, Online. Association for Computational Linguistics.

Vandan Mujadia and Dipti Misra Sharma. 2021b. English-Marathi neural machine translation for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 151–157, Virtual. Association for Machine Translation in the Americas.

Vandan Mujadia and Dipti Misra Sharma. 2022. The ltrc hindi-telugu parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3417–3424.

Bhagyashree Mukherjeee, Anusha Prakash, and Hema A. Murthy. 2021. Analysis of conversational speech with application to voice adaptation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 765–772.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Sharath Rao, Ian Lane, and Tanja Schultz. 2007. Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. In *Proceedings of Machine Translation Summit XI: Papers*.

Dipti Misra Sharma, Asif Ekbal, Karunesh Arora, Sudip Kumar Naskar, Dipankar Ganguly, Sobha L, Radhika Mamidi, Sunita Arora, Pruthwik Mishra, and Vandan Mujadia, editors. 2020a. *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TechDOfication 2020 Shared Task*. NLP Association of India (NLPAI), Patna, India.

Dipti Misra Sharma, Asif Ekbal, Karunesh Arora, Sudip Kumar Naskar, Dipankar Ganguly, Sobha L, Radhika Mamidi, Sunita Arora, Pruthwik Mishra, and Vandan Mujadia, editors. 2020b. *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TermTraction 2020 Shared Task*. NLP Association of India (NLPAI), Patna, India.

Vishwas M Shetty, S Umesh, et al. 2020. Investigation of speaker-adaptation methods in transformer based asr. *arXiv preprint arXiv:2008.03247*.

Harold Somers. 2011. Machine translation: History, development, and limitations.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. *arXiv preprint arXiv:2004.06358*.

Wolfgang Wahlster. 2000. Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final verbmobil system. *Verbmobil: Foundations of speech-to-speech translation*, pages 3–21.

Atif Zafar, Burke Mamlin, Susan Perkins, Anne M Belsito, J Marc Overhage, and Clement J McDonald. 2004. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics*, 73(9-10):719–730.