International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Advances in Real-Time Speech-to-Speech Translation and Voice Cloning: A Comprehensive Survey

Varad Gandhi
SKNCOE BE Computer Department (of SPPU)
Pune , INDIA
gandhivarad02@gmail.com

Bithika Pradhan
SKNCOE BE Computer Department (of SPPU)
Pune , INDIA
pradhan168bithika@gmail.com

Anushka Chillal
SKNCOE BE Computer Department (of SPPU)
Pune , INDIA
anushkachillal23@gmail.com

Shyamsundar Gadde
SKNCOE BE Computer Department (of SPPU)
Pune , INDIA
shyamgadde1602@gmail.com

Guided by
Prof. Dhanashree Nevase
SKNCOE Computer Department (of SPPU)
Pune , INDIA

***Abstract*—In today's interconnected world, effective communication across language barriers is paramount. This project presents an innovative solution for seamless cross-lingual communication through real-time speech-to-speech translation combined with advanced voice cloning techniques. The goal is to enable individuals to converse in their native language while still being understood by speakers of a different language, all while preserving the authenticity of their original voice. The proposed system utilizes automatic speech recognition (ASR) coupled with neural machine translation (NMT) models to transcribe and translate spoken content in real time. These models are trained on diverse multilingual datasets to ensure accurate and contextually appropriate translations. Voice cloning technology is seamlessly integrated into the translation pipeline to preserve the individual's unique voice characteristics. Leveraging generative adversarial networks (GANs) and other voice synthesis techniques, the system generates translated speech that closely mimics the original speaker's voice, including intonation, rhythm, and emotional nuances.**

***Keywords*—Automatic Speech Recognition, Voice cloning, Neural Machine Translation**

INTRODUCTION

Overcoming language barriers is more important than ever in today's increasingly interconnected world, when globalization and multiculturalism are at the forefront of our culture. Real-time speech-to-speech translation is at the vanguard of this linguistic revolution, aided by the extraordinary possibilities of voice cloning technology. This dynamic team is changing the way we interact by breaking down barriers between languages and, more importantly, humanizing the entire process.

Real-time speech-to-speech translation, aided by voice cloning technology, is a game-changer in the field of communication and language translation. This cutting-edge technology leverages the power of machine learning to enable seamless communication between people who speak various languages. It allows two individuals to converse and understand each other regardless of their native languages, thanks to the power of instantaneous translation. While this is a big step forward in overcoming linguistic gaps, voice cloning technology takes it a step further by adding a very personal and humanizing touch.

Voice cloning technology is a significant breakthrough that enables machines to reproduce and simulate human voices. It can accurately duplicate the vocal qualities, subtleties, and intonations of certain individuals. This means that, in addition to communicating words in a foreign language, the translation can be presented in the user's voice. The key to the translation process's effectiveness is its humanization. It helps the interaction feel more personal, engaging, and natural in many ways. It adds warmth and familiarity to the encounter, changing translation from a sterile exercise into a genuine exchange between two people.

A user can talk in their native language in a real-time speech-to-speech translation scenario improved by voice cloning, and the system not only accurately translates the words but also delivers the translation in a voice that mimics their own. This customization can dramatically improve the user's ability to connect with their conversation partner, effectively developing rapport and trust. This technology has several applications, ranging from international business and diplomacy to education, healthcare, and social connections. It removes the language barriers that have frequently hampered cross-cultural understanding and cooperation, encouraging better empathy and connection among people from various linguistic origins.

Historical Overview of Speech Translation

Automatic speech translation is an enticing digital niche. The incorporation of technology into the process of speech translation dates back to 1947. Warren Weaver, then-director of the Rockefeller Foundation's Division of Natural Sciences, examined the outcomes of wartime machine decoding.

He felt that there was a flaw in encryption that was causing problems with translation accuracy. While the world's earliest machine translation systems were government-funded ventures employed during World War I, there was no true computing market anywhere in the globe at the time.

Machines were the size of an entire automobile back then, and the best translation they could accomplish was rule-based. This shortcoming in the function of decoding machines gave rise to the idea that machines should be able to translate independently utilizing grammar and language rules.
The IBM 701 was introduced in 1954. It could translate roughly 49 sentences of chemistry from Russian to English. This was a significant accomplishment and the first effective step towards developing non-numerical computer applications. This generated a lot of initial excitement, until flaws in its utilization were uncovered, such as incorrect syntax and grammatical problems.
The Canadian METEO system was then launched in 1976 as a prototype. It may translate weather forecasts from French to English. Environment Canada mostly used it to convert forecasts into said languages.
Machine translation services went live for the first time in 1992, delivering subscribers translations from English to German. AtlaVista's "Babel Fish" was born in 1997. The name was inspired by the best-selling novel series Adams, D. (1979) The Hitchhiker's Guide to The Galaxy. Babel Fish could translate between German, English, Dutch, French, Italian, and Spanish. Despite being a significant advancement in the field of early speech translation, Babel Fish was not dependable when it came to selecting the correct translation for words with numerous meanings in a specific language. It also provided translations that were semantically challenging. Kwabena Boahen, an electrical engineer, improved on Alan Turing's (British scientist who attempted to make a computer mimic human thought in the year) work. This also shed some insight on the idea that teaching computers human language was a process that could only be accomplished if the computers spoke the raw numerical language fed by data successfully.

RELATED WORK:

1. Jong Wook Kim, Greg Brockman, Tao Xu, Alec Radford, – "Robust Speech Recognition via Large-Scale Weak Supervision" [1]
The work discusses a study on the capabilities of speech processing systems trained to predict transcripts of audio data from the internet. The study involves scaling the training to a large dataset and aims to achieve competitive results without the need for fine-tuning or self-supervision techniques. Instead of relying on traditional supervised techniques, the approach involves predicting transcriptions for a vast amount of internet audio data. The study has successfully scaled up to 680,000 hours of multilingual and multitask supervision. [1] Even in instances in which there is zero-shot transfer, the generated models demonstrate strong generalization on standard benchmarks. Even without considerable fine-tuning, models trained using this method generalize well to established benchmarks. When compared to fully supervised models, models trained using this method frequently obtain competitive performance.
While the research claims that this approach simplifies speech recognition training, it may still necessitate complicated preprocessing, model design, and data management in order to perform effectively at such a vast scale. Access to a large amount of online audio data may not be readily available to all academics or organizations, limiting the utility of this approach.

2. Pratyush Kumar, Praveen S V, Mitesh M. Khapra, Karthik Nandakumar, Gokul Karthik Kumar, - "Towards buildingtext-to-speech systems for the next billion users" [2]
They have demonstrated the use of deep learning techniques to design and evaluate text-to-speech (TTS) systems for Indian languages. It focuses on the challenges and developments in TTS, particularly for Indian languages, and offers the findings of research undertaken to determine the ideal configuration for developing TTS models across multiple criteria.
The study intends to fill a gap in the development of TTS systems for Indian languages. This is beneficial since it caters to a broad language and cultural population. It also identified the FastPitch and HiFiGAN V1 combo, trained with male and female speakers, as the optimum setup.
When compared to other frequently spoken languages, the availability of resources for Indian languages may be limited, making it challenging to collect the necessary data and construct high-quality TTS models. While the research focuses on enhancing TTS quality, several topics are discussed but not completely researched, such as voice cloning and unheard speaker generalization. This shows potential research gaps.

3. Vivek Raghavan, Jay Gala, Pranjal A. Chitale, Janki Nawale, Raghavan AK, Sumanth Doddapaneni, Anupama Sujatha, Ratish Puduppully, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, Anoop Kunchukuttan- "Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages" [3]
This paper examines the significance of Machine Translation (MT) systems for India's linguistically varied country, focusing on the 22 languages listed in the Indian Constitution. It emphasizes the lack of accessible and high-quality MT systems for these languages and describes a research initiative aimed at closing the gap.
The project prioritizes four areas for improvement: curating and building larger training datasets, developing varied benchmarks, training multilingual models, and distributing open-access models.
The research work intends to fill a void in MT for Indian languages by developing previously unavailable materials and models. The IndicTrans2 translation model, which is intended to handle all 22 languages, improves on existing models by improving the quality of MT in these languages.
While the article emphasizes the improvements, it does not go into detail on the translation quality of the MT models or the difficulties they may have when dealing with the linguistic intricacies of each language.

4. Shashank Subramanya, Jan Niehues – "Multilingual Simultaneous Speech Translation" [4]
This research paper explores the development of applications for real-time speech translation during events like conferences or meetings, aiming to strike a balance between translation quality and latency for an optimal user experience. The paper investigates the use of multilingual models and different architectural approaches (end-to-end and cascade)

for online speech translation, utilizing a technique to adapt monolingual models to this context. The experiments are conducted on the multilingual TEDx corpus, showing that the approach generalizes across different architectures, resulting in a notable reduction in latency (40% relative) across languages and models. Interestingly, the end-to-end architecture exhibits smaller translation quality losses after adapting to the online model.
It provides a substantial reduction in latency (40% relative) across different languages and models. This is a crucial advantage, especially in real-time scenarios, as lower latency leads to faster and more responsive translations. It also investigates various architectural approaches, including end-to-end and cascade models. This exploration helps researchers and developers understand which architecture is more suitable for online speech translation applications.

5. Andros Tjandra, Bowen Shi, Alexis Conneau, Paden Tomasello Arun Babu Sayani Kundu, Ali Elkahky, Vineel Pratap, Wei-Ning Hsu, Michael Auli – "Scaling Speech Technology to 1,000+ Languages" [5]
This research addresses the Massively Multilingual Speech (MMS) initiative, which intends to enhance the language coverage of speech technology. It emphasizes the limitations of current speech technology, which is available for only a small fraction of the world's languages, and describes the MMS project's accomplishments, which include the creation of pre-trained models, automatic speech recognition, speech synthesis models, and language identification models for a significantly larger number of languages.
The primary benefit of the MMS project is a significant increase in language coverage. The project provides access to information and technology for a more diverse linguistic community by increasing the number of supported languages by 10-40 times. Scaling to more languages and dialects increases the danger of introducing undesired biases in model performance. Inadequate representation in training data, particularly for dialects and low-resource languages, might make attaining unbiased results difficult.

6. Liang Ding, Hexuan Deng, Xuebo Liu, Dacheng Tao, Meishan Zhang, Min Zhang – "Improving Simultaneous Machine Translation with Monolingual Data" [6]
The paper discusses the use of simultaneous machine translation (SiMT), a technique that aims to provide real-time translations by starting the translation process before the source sentence is complete. However, SiMT's partial source sentence conditioning can lead to difficulties in capturing the complete semantics, particularly for distant language pairs like English and Japanese. To address this issue, the paper proposes a two-step approach for training SiMT models. It leverages sequence-level knowledge distillation (Seq-KD) as the first step, with a full-sentence neural machine translation (NMT) model serving as the teacher. This Seq-KD step helps generate coherent knowledge and reduces data complexity. Monolingual and bilingual data are generally complementary, and using monolingual data alongside bilingual data transfers knowledge from both sources, maintaining the benefits of Seq-KD. Monolingual data is significantly more abundant than bilingual data, offering the potential for substantial improvements in SiMT performance.
The paper also addresses the challenge of long-distance reordering in SiMT, which makes the pseudo-targets generated by the full-sentence NMT teacher model unsuitable. Inspired by human simultaneous interpretation strategies, the paper introduces novel techniques for sampling monolingual data suitable for SiMT, taking into account chunk lengths and monotonicity.

7. Weiping Tu, Jingyi Li, Li Xiao – "FreeVC: Towards high-quality text-free one-shot voice conversion" [7]
The paper discusses voice conversion (VC), a technique used to change the voice of a source speaker to match the style of a target speaker while preserving the linguistic content. [7]
The authors identify several challenges with existing VC approaches, including the contamination of content information with speaker information, the need for a large amount of annotated data for training, and the potential degradation in the quality of the converted waveform due to mismatches between conversion models and vocoders.
In response to these challenges, the paper introduces an end-to-end framework called FreeVC for high-quality waveform reconstruction in voice conversion.
The authors propose a method to separate content information from speaker information without relying on text annotation. They do this by imposing an information bottleneck on WavLM features, a type of self-supervised learning (SSL) feature extracted from the waveform. A data augmentation technique called spectrogram-resize (SR) is introduced to enhance the disentanglement of content and speaker information. SR distorts speaker information while preserving content information. The focus of the paper is on one-shot voice conversion, where only a single reference utterance of the target speaker is available for conversion. The proposed FreeVC framework leverages the advantages of VITS for high-quality waveform reconstruction while addressing the content disentanglement problem. The authors use SSL features from WavLM for content extraction, and they employ a speaker encoder for extracting speaker information.

8. Kyung Hyun Cho, Dzmitry Bahdanau, Yoshua Bengio – "Neural Machine Translation by Jointly Learning to Align and Translate" [8]
They suggested that the neural translation machines aim towards constructing a single neural network which will be jointly tuned to increase the performance between translation. With this method it can gain a rapid translation which can be compared to the existing state of art phrase-based system. It performs well regardless of sentence length. It provides better alignment mechanism which correctly associates each target word with the relevant words or annotations in the source sentence.
They were unable to deal with out-of-vocabulary words or those with limited occurrences in the training data.

9. Sakriani Sakti, Andros Tjandra, Satoshi Nakamura- "Speech-to-speech translation between untranscribed unknown language" [9]
They proposed a method for developing speech-to-speech translation without any kind of linguistic supervision.
It consists of 2 steps –
a. Monitor and generate representation with unsupervised

discovery with discrete auto-encoder.
b. Execute a sequence-to-sequence model that directly matches the core language.
No Need for Transcription or Linguistic Supervision: It can perform speech-to-speech translation tasks without relying on transcription.
Performance Variation: performance of the method may vary across different language pairs and data conditions. Achieving consistent high--quality translations, especially for less-resourced languages, remains a challenge.

10. Laurent Besacier, Alexandre Berard, Christophe Servan, Olivier Pietquin, – "Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation" [10]
This paper presents an initial effort to develop an end-to-end speech-to-text translation system that does not rely on transcription of the source language for decoding or learning. End-to-End approach: System represents true end-to-end solution for speech to text translation.
Elimination of transcription requirement: Completely eliminating the need for source language text transcription. Synthetic Nature of Data: This might lead to overfitting and may not fully reflect the complexities of real-world speech.

11. Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird and Trevor Cohn – "An Attentional Model for Speech Translation Without Transcription" [11]
They experiment with the neural, attentional model applied to this data. Without using transcriptions or a vocabulary, their model performs almost as well as GIZA++ on gold transcriptions.
In many low resource languages, it is easier to find spoken translation than transcription.
While spoken translations are more available than transcriptions, producing them at scale can still be costly.

12. Zhaoyu Liu, Brian Mak- "Cross lingual Multi- speaker Text-to-Speech synthesis for voice cloning without using parallel corpus for unseen speaker" [12]
This study explores an innovative method for producing high-quality native or accented voice for native speakers of Mandarin and English using cross-lingual multi-speaker text-to-speech synthesis. The system consists of three separately trained components: an x-vector speaker encoder, a Tacotron-based synthesizer and a WaveNet vocoder.
It can effectively clone foreign speakers' voices and generate accented speech
Capable of generating speech in multiple languages (English and Mandarin) and accommodating various speakers. Limited Accented Speech Quality: Quality of accented speech is noted as being lower, which may limit its utility in certain applications.

13. Chengyi Wang, Yu Wu, Zhengyang Chen, Shujie Liu, Zhuo Chen, Jinyu Li, Sanyuan Chen, Naoyuki Kanda, Takuya Yoshioka - "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing" [13]
The paper outlines an innovative technique to voice processing that employs Self-supervised Learning (SSL). It introduces WavLM, a pre-trained model built to handle numerous speech processing tasks. During pre-training, the model simultaneously learns masked speech prediction and denoising, which improves its ability to handle speech content and non-ASR (Automatic Speech Recognition) tasks. The training dataset for the model has been increased from 60k to 94k hours and includes data from multiple sources such as Libri-Light, GigaSpeech, VoxPopuli, podcasts, YouTube, and European Parliament event recordings. WavLM Large outperforms the competition on the SUPERB benchmark and shows considerable gains across a variety of speech processing tasks. WavLM's masked speech prediction framework aids in the improvement of ASR information modelling, perhaps leading to improved performance in ASR challenges. The addition of several components, such as masked speech prediction, speech denoising, and gated relative position bias, may increase the model's complexity, making it more difficult to design and maintain.

System Architecture

The process begins when a speaker talks into a microphone. The microphone captures the speaker's voice and converts it into digital audio data. This audio data is then sent to a component called Whisper ASR, which stands for Automatic Speech Recognition. This component listens to the audio data and transcribes it into written text in real-time. The transcribed text is stored in a stream buffer, which is a temporary storage area.
As soon as new text is added to the stream buffer, another component called Sentence Tokenization comes into play. This component breaks down the transcribed text into individual sentences. Each sentence is then sent to a Translation Model, which translates the sentence from English to the target language (e.g., Hindi, Bengali). The translated text is then sent to a Text-to-Speech (TTS) model, which converts the written text back into spoken words. The synthesized speech is then passed through a Retrieval-Based Voice Conversion (RVC) model. This model adjusts the voice in the synthesized speech to match the original speaker's voice. Finally, the converted speech is streamed to the listener's device via WebRTC, which stands for Web Real-Time Communication.

MODEL ARCHICTURE

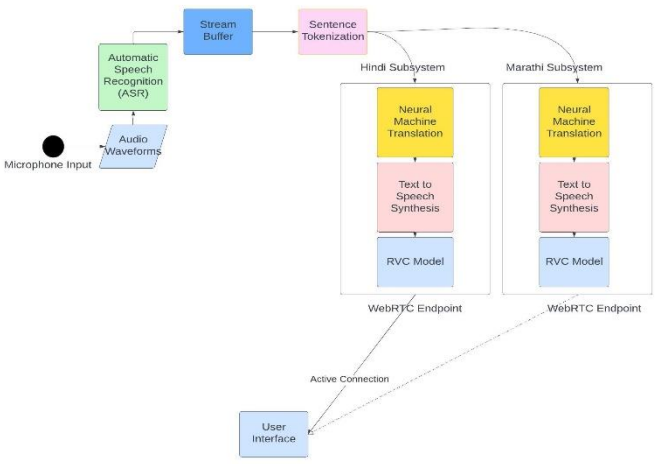

Fig 1: The above figure shows the parallel multi-lingual machine translation coupled with the RVC model.

Conclusion

In conclusion, the combination of real-time speech-to-speech translation and advanced voice cloning technology provides a potent solution for overcoming language barriers and improving cross-linguistic communication. This novel approach not only allows people to communicate in their native languages while being understood by speakers of other languages, but it also adds a personal and humanising touch by keeping the authenticity of their original voices. The historical overview demonstrates the progress of machine translation from its early rule-based systems to today's powerful neural machine translation techniques.

The related work area demonstrates ongoing attempts to increase translation accuracy, efficiency, and usefulness through research and development in the field of speech translation. To improve speech-to-speech translation and voice cloning technologies, researchers have investigated numerous methodologies such as neural networks, unsupervised learning, and end-to-end systems.

References

[1] Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." In International Conference on Machine Learning, pp. 28492-28518. PMLR, 2023.

[2] Kumar, Gokul Karthik, S. V. Praveen, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. "Towards Building Text-to-Speech Systems for the Next Billion Users." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

[3] Gala, Jay, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale et al. "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages." arXiv preprint arXiv:2305.16307 (2023).

[4] Subramanya, Shashank, and Jan Niehues. "Multilingual Simultaneous Speech Translation." arXiv preprint arXiv:2203.14835 (2022).

[5] Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky et al. "Scaling speech technology to 1,000+ languages." arXiv preprint arXiv:2305.13516 (2023).

[6] Deng, Hexuan, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. "Improving simultaneous machine translation with monolingual data." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12728-12736. 2023.

[7] Li, Jingyi, Weiping Tu, and Li Xiao. "Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

[8] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[9] Tjandra, Andros, Sakriani Sakti, and Satoshi Nakamura. "Speech-to-speech translation between untranscribed unknown languages." In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 593-600. IEEE, 2019.

[10] Bérard, Alexandre, Olivier Pietquin, Christophe Servan, and Laurent Besacier. "Listen and translate: A proof of concept for end-to-end speech-to-text translation." arXiv preprint arXiv:1612.01744 (2016).

[11] Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. "An attentional model for speech translation without transcription." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 949-959. 2016.

[12] Liu, Zhaoyu, and Brian Mak. "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers." arXiv preprint arXiv:1911.11601 (2019).

[13] Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16, no. 6 (2022): 1505-1518.

[14] Weiss, Ron J., Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. "Sequence-to-sequence models can directly translate foreign speech." arXiv preprint arXiv:1703.08581 (2017).

[15] Zen, Heiga, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. "Libritts: A corpus derived from librispeech for text-to-speech." arXiv preprint arXiv:1904.02882 (2019).

[16] Zheng, Yibin, Xi Wang, Lei He, Shifeng Pan, Frank K. Soong, Zhengqi Wen, and Jianhua Tao. "Forward-backward decoding for regularizing end-to-end TTS." arXiv preprint arXiv:1907.09006 (2019).

[17] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.