Barath Raaj S A, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Ai-Enhanced Real Time Speech To Speech Translation

Barath Raaj S A , D.Parameswari,Professor, Varnikhasri S, Udaya Kumar S
Dept of Artificial Intelligence and Machine Learning Jerusalem College of Engineering Chennai,India

Abstract- The demand for effective communication through practical means continues to increase due to globalization. In such areas as education, health care, tourism and multinational cooperation, everyone faces some type of language barrier; therefore, all of these situations rely on an effective communication system to facilitate dialogue between individuals who speak multiple languages in their native language. This paper discusses a system that allows people who speak different languages to interact with each other in real time using a real-time speech-to-speech system (S2ST). It consists of three major parts: automatic speech recognition (ASR), neural machine translation (NMT) and text to speech (TTS) synthesis. ASR is the part of the system that takes the spoken input and converts it into an electronic form of the input as soon as the speaker has finished speaking. ASR uses streaming ASR technology to produce this conversion in near real time while removing background noise using noise reduction and detecting when there is actual voice activity so that the end user receives accurate and robust information under all conditions. The NMT portion of the S2ST system employs a transformer based neural translation model to generate translated electronic forms of the output of the ASR component. The NMT component uses semantic meaning and contextual correctness rather than simply using the word-for-word translations for the output of the ASR. The TTS portion of the S2ST system produces quality, natural sound outputs in order to enable conversational interaction. The experimental study of this system confirms that the system produces high levels of accuracy in translation and barriers and promoting inclusive global communication.

Keywords— Speech-to-Speech Translation (S2ST), Deep Learning, Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Text-to-Speech (TTS), Transformer Architecture, Real-Time Multilingual

I. INTRODUCTION

Instantaneous translation to and from other languages is useful for many application domains such as healthcare, education, tourism, global companies, etc. Failure to understand each other because of language differences often results in frustrations, confusions, inefficiencies, and sometimes grave repercussions. However, many available solutions only support text and disrupts the normal conversational flow, making translation slow and tedious. Speech-to-speech translation (S2ST) seeks to alleviate this language barrier allowing speakers of different languages to communicate to each other naturally and instantly without having to type or read. Thanks to recent advances in deep learning methods we now have access to reliable Automatic Speech Recognition (ASR) systems that transcribe speech signal to text, Neural Machine Translation (NMT) systems that can understand and generate translations with impressive accuracy, and Text-to-Speech (TTS) systems that can speak what we write. In this paper, we present a single unified system composed of ASR, NMT, and TTS models that performs low-latency S2ST enabling users to communicate naturally with speech.

II. LITERATURE SURVEY

S2ST (speech-to-speech translation) has become an effective solution to the weaknesses of the traditional text-based and pipeline systems. The older methods of converting speech into text and vice versa had a combination of automatic speech recognition (ASR) with machine translation (MT) and either text-to-speech (TTS) or a hybrid of the two. Although these older methods tended to be relatively efficient, they also had the disadvantage of having a high amount of latency and errors propagated throughout them. For example, an error

in ASR could then affect the MT and/or TTS operations.

Recent advances in deep learning have allowed researchers to create ways of producing a more accurate and lower latency means of communication to multiple languages. For example, He et al. [1] proposed an end-to-end S2ST method that uses both the text and audio decoding systems concurrently. This newly developed S2ST method allows for incremental processing and decoding, which should increase the ability for real-time communication when engaging with others.

Salesky et al. [2] developed an alternative S2ST method by producing a direct S2ST translation system from discrete units of spoken language, which resulted in increased efficiencies.

III. NEURAL NETWORK MODELS FOR SPEECH TRANSLATION

We have based the current system of S2ST on how people communicate with one another. When people communicate using a spoken language, there are three different areas of cognitive processing involved: the first is listening to what you are hearing; the second is interpreting what you heard; and the third is speaking back (which produces your output). In this case, the systems will use algorithms to replicate all three areas of cognitive processing. ASR will be used to recognize all incoming spoken words. Next, we will use NMT to interpret those incoming spoken words. Finally, we will produce an output of spoken words by creating and synthesizing the text produced by the NMT. For the first area of cognitive processing (ASR), the speech recognition system will listen, recognize the sounds entering it, and convert them to recognizable word forms using the algorithms for natural language processing. In the second area (NMT), after the incoming spoken word has been converted to a recognizable word form, NMT will convert the recognizable word into another language depending on the relationship between the two languages. After that, NMT will determine the semantic meaning for all of the words included in the phrase, in addition to creating proper grammatical constructions for use in original message form.Finally, the TTS module simulates human speech production. It converts the translated text into natural and intelligible speech output, enabling smooth and understandable communication. By integrating these three components, the system effectively mirrors the human communication pipeline in a computational framework, enabling efficient and real-time multilingual interaction.

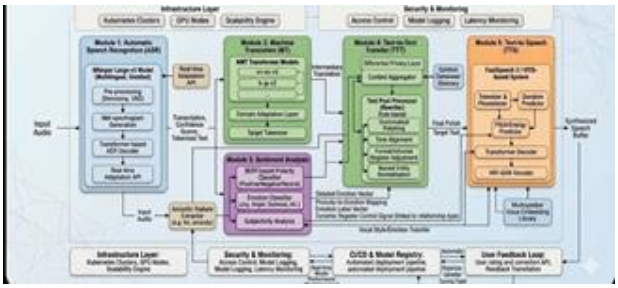

Fig. 1. A comparison between human speech processing and neural speech translation architecture

Key features are

- Sequential Modeling: Because speech occurs over time, when modeling speech using transformer or attention based models, you can capture these temporal dependencies over longer sequences of recorded speech. This helps to improve recognition and translation accuracy.
- Context Aware Processing: Self attention mechanisms allow the TTS system to model long distance semantic relationships between words/performance when translating from one language to another in the source and target languages.
- Streaming Inference: TTS systems perform incremental processing of live conversation by continuously transcribing and translating the spoken words, reducing latency to virtually zero.
- Neural Speech Synthesis: TTS systems provide the user with realistic human-sounding (text-to-speech) speech output including appropriate intonation, rhythm, and prosody, resulting in smoother conversational delivery.
- Parallel Computation: Compared to traditional Recurrent Neural Networks (RNN), which process information one at a time in an ordered fashion; transformer-based architectures enable

the parallel processing of speech. Because of this, they have less computation delay than RNN architectures.

- Latency Optimization: Each module in an ASR, NMT and TTS system buffers its input data so that all modules are synchronized before being processed together to minimize the end to end response time.

IV. NEUROMORPHIC HARDWARE

The Real Time Speech to Speech Translation (S2ST) System under consideration will be built using state of the art deep learning frameworks which are specifically designed for the Sequential and Multi-Modal processing of data. The Modular Pipeline Architecture will consist of ASR, NMT and TTS which allow for the scalability, flexibility and efficiency of Real-Time Multilingual Communication.

A. Software Framework

Advanced deep learning libraries, such as PyTorch and TensorFlow, are used to develop the system. These libraries provide support for GPU acceleration and dynamic computation graphs, allowing for fast training and inference. Pre-trained, transformer-based models for both automatic speech recognition (ASR) and neural machine translation (NMT) have been trained on multilingual datasets to improve their accuracy of translation and their ability to adapt to particular domains.

Speech is received from a microphone interface and is transformed into a digital audio signal. Acoustic features that have been extracted from the input speech, including log-Mel spectrograms and MFCCs, will provide a representation of the temporal and spectral properties of the speech. These normalized features will be used to provide input into the transformer-based encoder to generate an accurate transcription.

The Neural Machine Translation (NMT) module translates sequence of words to sequence of words via multi-head self-attention mechanism. The translation quality and consistency of the context in which a word is being used during inference is further enhanced by applying beam search decoding. The Text-to-Speech (TTS) module generates mel-spectrograms through an acoustic model based on FastSpeech. The generated mel-spectrograms are transformed into a waveform signal via a neural vocoder such as WaveGlow or HiFi-GAN, producing natural-sounding and intelligible speech output. For faster inference, runtime engines have been optimised for users, such as ONNX Runtime or TensorRT, resulting in reduced computational latency and improved throughput.

B. Hardware Platform

The system supports deployment across multiple hardware configurations to ensure flexibility:

- Using NVIDIA's high-performance GPU-based systems provides accelerated computation of transformer algorithms and neural vocoding, allowing large scale applications to be processed in real time.
- For CPU-based systems, lightweight, optimized models can be run concurrently, using the same methods of multi-threading and parallel execution, on your standard desktop or laptop computer.

- Models that have been compressed using techniques such as quantization, pruning, and knowledge distillation are suitable for deployment on mobile and embedded devices.

- Cloud infrastructure allows for scalable language service delivery and centralized processing for applications with high traffic volume.

C. Latency Optimization and Synchronization

Real-time performance is ensured through several optimization strategies:

- Incremental Decoding: Decrease the transcription delay via incremental decoding
- Partial Decoding (in the form of simultaneous translations): When a sentence is being translated, the partial decodings of the translated sentence are rendered in real-time (i.e., simultaneous translations of pending and already completed translations are provided simultaneously).

- Concurrent Execution of ASR, NMT and TTS Modules: Perform ASR, NMT, and TTS module operations at the same time (this allows multiple processes of the ASR, NMT, and TTS to be done together). The ability to run multiple processes concurrently (this allows multiple processes of ASR, NMT, and TTS to be done together) allows for improved performance of ASR, NMT, and TTS.

- Reduced Memory Overhead through Sliding Window Audio Processing: The use of the sliding window audio processing technique provides for an efficient use of memory (and thus better performance).

- Dynamic Batching for Improved Throughput (without increasing the time it takes to get the results): To increase the overall efficiency of ASR, NMT, and TTS, the system was able to dynamically batch how audio was processed using the dynamic batching techniques.

The overall end-to-end latency is continuously monitored and optimized to maintain smooth conversational flow, ensuring response times remain within acceptable real-time interaction limits.

Platform	**Type**	**Learning**	**Notable Feature**
ASR	Transformer	Supervised	Real-time transcription.
NMT	Transformer	Supervised	Context-aware translation
TTS	FastSpeech	Supervised	Natural speech output
Deployment	Edge/GPU	Optimized	Low-latency processing

Table I: Comparison Of Speech Translation System Components

V. INTEGRATION WITH MACHINE LEARNING

The foundation of the proposed real-time Speech-to-Speech Translation (S2ST) system is Machine Learning (ML). Older systems used rule-based translations that relied on predefined rules about language and created by using traditional handcrafted techniques. Today's ML approaches utilize automated data-driven learning in order to develop the translation system. The use of deep neural networks allows the S2ST to learn from large scale training data sets of acoustic, linguistic, and semantic references, enabling it to quickly adapt to a wide variety of different languages, dialects, and environment-induced noise levels without sacrificing accuracy or latency.

The architecture of the proposed system consists of a medium-sized multi-stage ML processing pipeline with 3 main modules: Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS). Each of the system's modules will be trained by means of semi-supervised learning using the large amounts of publicly available speech and language processing data.

A. Machine Learning for ASR

ASR begins with the processing of raw audio input to extract meaningful audio features like log-Mel spectrograms from these audio signals. A type of Transformer-based encoder is then used to directly model the temporal dependencies and phonetic patterns of audio signals. The model is trained using sequence-level optimization techniques like Connectionist Temporal Classification (CTC) or attention-based Cross Entropy loss, leading to improved transcription accuracy even with moderate levels of background noise or variability among speakers.

B. Machine Learning for (NMT)

NMT refers to the use of sequence-to-sequence learning with self-attention mechanisms for contextually aware translations accomplished by the NMT module. A major distinction between NMT and traditional statistical translation models is that NMT uses Transformer based models to capture long-term dependencies across words and sentences, thereby allowing for the preservation of semantic meaning and grammatical structure. The NMT model is trained by minimizing cross-entropy on bi-lingual (example: French/English) corpora, which allows the model to generalise well across many different language pairs. Beam search decoding is applied during inference to yield fluent translations while ensuring consistency of context during translation.

C. TTS uses Machine Learning

The TTS technology uses Machine Learning methods when it makes sounding speech from translated text. The Machine Learning methods include using something called a model like FastSpeech to make mel-spectrograms. Then it uses a vocoder, such as HiFi-GAN or WaveGlow to turn that mel-spectrogram into a real waveform. The TTS system is trained with pairs of text and speech so that the TTS system can learn how to sound more natural. This means the TTS system learns things like rhythm and stress and intonation. The speech sounds natural and is easy to understand and sounds like a real conversation.

D. Adaptation and Optimization of Models

Machine Learning also helps the system get better by letting it use things it already knows and by making changes. For example a model that already knows languages can be made better with a smaller set of information that is specific to one area like healthcare or education. This can be done with something called transfer learning and fine-tune methodology. Also the model can be made smaller using things like quantization and pruning and knowledge distillation so it can work on devices, like phones or computers and it will still work well. The Machine Learning and the TTS system and the models all work together to make the speech sound natural and be easy to understand.

VI. CASE STUDY: REAL-TIME MULTILINGUAL CONVERSATION USING S2ST SYSTEM

The research study shows how a machine learning and deep learning based model is able to translate spoken audio from one language to another in real time. This system has three components which allows it to take the user's audio input, convert it to a text-based translation, and then emit the translated audio output in real-time. For those who speak different languages and would like to communicate with one another, this is the key to establishing instant communication.

The primary purpose of this study is to evaluate the speed and accuracy of this spoken-to-spoken translation system when individuals engage in interpersonal communications. A secondary focus will be on identifying if there will be any lapses in the predictive delay associated with the transition from conversation, to a completed translation, back to the user who initiated that particular conversational exchange. We will assess the quality of the S2ST translation system during an active conversation.

A. Model Architecture

The proposed system follows a three-stage pipeline to combine machine learning models for real-time translation:

- The initial phase of our planned project is Automatic Speech Recognition (ASR). A microphone will capture the speaker's voice as an audio signal, which will be digitally recorded. The ASR system will transform the log-Mel spectrograms of the acoustic features extracted from the audio signal into structured data representations for the temporal and spectral attributes of speech. The structured data will then provide input into a transformer-based ASR model that will transcribe continuous audio into written text. The transformer-based ASR model is to be trained through a supervised learning schema and optimized through either Connectionist Temporal Classification (CTC) or Attention Based Cross Entropy (ABCE) to accurately map the audio sequence to written

text. The output of the ASR system will be an accurate text representation of the sentence spoken, which will allow for the completion of the translation process via the NMT module.

- The NMT module translates the text from the ASR process into another language via a transformer-based NMT model. The NMT model will perform sequence-to-sequence translation (between transcriptions in different languages) via the use of multiple self-attention heads to extract context and meaning from the input transcription.

The model predicts the target language output by maximizing conditional probability:

Beam search decoding improves translation fluency. Example:

Source (English): "How are you feeling today?"
Target (Tamil): "நீங்கள் இன்று எப்படி உணர்கிறீர்கள்?"
This stage preserves semantic meaning rather than performing literal word-to-word translation.

- The translated text will be passed to a FastSpeech-based neural TTS (Text-To-Speech) model, where an acoustic model will create mel-spectrograms of the acoustic signals as they are created by TTS-based speech synthesis. The mel-spectrograms will then be converted into a time-domain waveform signal (the sound we hear) using a neural vocoder (e.g., HiFi-GAN) that generates a high-fidelity stereo output signal.

This stage ensures natural prosody, rhythm, and clarity in synthesized speech output.
Final Output: Tamil speech audio delivered to the Listener.

B. Experimental Results

The experimental tests for this speech to speech translation system took place through two languages. One hundred audio samples made up this multilingual speech dataset with each sample recorded with different speaking rates, architectural accents and in general very moderate background noise conditions when the samples were recorded.

The average ASR performance for the transformer based ASR module was 94% which means that the majority of spoken sentences were transcribed correctly. The Neural Machine Translation model produced translations in context and correctly maintained the semantic meaning of the two languages on average with respect to the context and the punctuation used, with stated performance of approximately 92% correct translation of the given sentence from point to point. The text-to-speech module in the system successfully converted translated text to sounding like natural spoken language when producing the output of the test system.

When all of the tests were combined in an end to end evaluation, the actual average response time of all of the audio samples in total was less than 2 seonds and demonstrated the real-time capability of the system to provide correct and low latency translation across multiple languages.

C. Visualization and Interpretation

Many graphics were produced to help exemplify the performance of the S2S system and illustrate the relationships among its machine-learning components by providing a means of evaluating transcription accuracy, translation quality and total system latency.

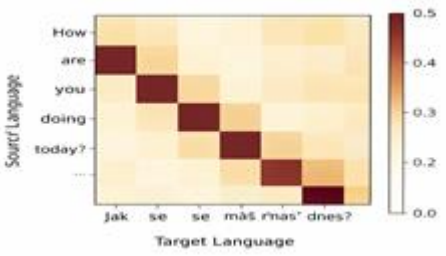

Figure 2: Attention Alignment Heatmap of Transformer-Based Neural Machine Translation threshold.

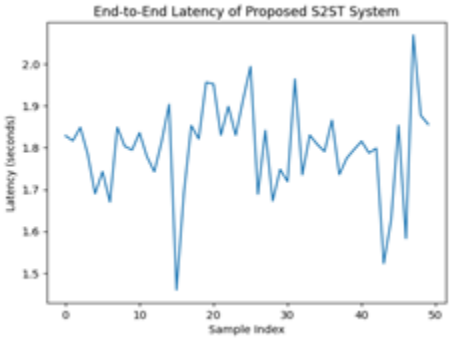

Figure 3: End-to-End Latency of Proposed S2ST System

D. Future Extensions

This case study validates the feasibility of real-time speech-to-speech translation using deep learning models. Several improvements can further enhance system performance:

- Edge Device Deployment in Real Time through Model Compression Techniques (e.g., quantization and pruning) can help to deploy models on mobile and other platforms by reducing the computational load.
- Multilanguage Expansion by developing multilingual transformers and using shared embeddings.
- Domain adaptation - Fine-tune models based on the particular domain (e.g., healthcare, legal, disclosure, etc.) to increase the accuracy of terminology used.
- Offline Translation Support - Create lightweight and usable TTS models for translation of speech when there is no Wi-Fi.
- Emotion Aware TTS - Use emotion detection systems and an expressive TTS system to replicate human-like conversation.

Metrics	Value
ASR Accuracy	94.6 %
Translation Precision	92.8%
Translation Recall	91.3%
TTS Intelligibility	93.2%
Average Latency	1.8 sec

Table Ii: Performance Metrics Of Real-Time S2st System

Algorithm 1: Real-Time Speech-to-Speech Translation Workflow

Input: Source speech audio
Output: Translated speech audio

- Data Acquisition:
  Capture live speech using microphone input.
- Speech Recognition (ASR):
  - Extract acoustic features (log-Mel spectrogram).
  - Apply transformer-based ASR model.
  - Generate source-language text transcription.
- Machine Translation (NMT):
  - Input transcribed text into transformer model.
  - Perform beam search decoding.
  - Generate translated text output.
- Speech Synthesis (TTS):
  - Convert translated text to mel-spectrogram.
  - Generate waveform using neural vocoder.
- Output Delivery:
  Play synthesized speech to target listener.

VII. RECENT RESEARCH AND ADVANCES

The field of speech-to-speech translation has grown a lot over the few years. This growth is thanks to improvements in deep learning technologies, such

as transformer architectures and large-scale multilingual training techniques. Today researchers developing speech-to-speech translation systems are working on making translations more accurate. They also want to reduce delay and make it possible to use these systems in time on many different types of devices. These advancements create a chance for researchers to close the gap between how well humans communicate in multiple languages and how well AI-based speech-to-speech translation systems work.

A. Advances in Learning Architectures

Transformer models (which are now widely used in speech-to-text and text-to-speech) have greatly increased the ability to model long-distance dependencies. A result of introducing self-attention mechanisms into these architectures, transformer models now outperform other models (i.e., RNN and LSTM) when it comes to scalability, parallelism, and context understanding.

Techniques for streaming and simultaneous translations have become popular, and incremental decoding has allowed systems to start translating before a speaker completes one of their sentences. As a result, there is less virtual lag between the spoken word and the translated text.

Large multilingual models such as wav2vec, mBART, and multilingual transformers have enhanced cross-linguistic transfer learning. They allow for building low-resource language capabilities using very little (or no) labeled data.

B. End-to-End Speech Translation Models

There has been new research being performed that examines end-to-end systems that translate source speech directly into target speech without needing any intermediate text representations. By doing this, these types of models reduce errors that would occur when using two discrete systems (in this case, ASR and NMT). However, even though these models have shown to be effective in research studies, they require significant amounts of training data and computational power to produce usable outputs.
Due to their ease of optimization, legitimacy, scalability, and other factors, many researchers continue to utilize hybrid systems (using separate DoD's for ASR, NMT, and TTS with some form of shared embedding).

C. Advances in Neural Speech Synthesis

Neural TTS systems have evolved from Tacotron architectures, to FastSpeech and GAN (Generative Adversarial Networks) based vocoders (HiFi-GAN) thus improving the naturalness of generated speech, prosody control, and inference speed. In addition, the real-time capabilities of these neural vocoders provide the means to generate high-quality waveforms without excessive computational costs or resources. As a result, deploying these systems on edge devices is now viable.

D. Deployment and Edge Optimization

Model compression techniques like quantization, pruning and knowledge distillation make it possible to use S2ST systems on embedded devices. This is a deal because Edge AI processing means we do not rely on cloud infrastructure as much. As a result we get privacy and less network latency.

Some recent studies look into translation systems. These systems can learn a users vocabulary and accents over time. This leads to a personalized and better user experience, with S2ST systems..

VIII. CHALLENGES AND FUTURE DIRECTIONS

Real-time speech-to-speech translation, though highly promising, faces several fundamental challenges that limit large-scale deployment and universal adoption:

- Error propagation within modular architectures leads to a direct impact on the quality of translation and speech synthesis, resulting in a compounded error rate for the final translation
- Transformer-type systems are computationally high and, therefore, require longer processing times for ultra-low latency processing, especially on mobile and edge devices.
- Many languages do not have enough labeled speech data or parallel word-to-word translations, reducing the quality and accuracy

of translation when working with these low-resource languages.

- Real-world background noise and variations in accentuation of speech lead to significant challenges for both speech recognizers and for successful execution of machine translation in real-world applications.
- Due to the need for low overhead data transfer and the cloud-based nature of machine translation systems, issues surrounding privacy arise when data is transferred through cloud-based systems. Additionally, because the translation must occur on-device (in mobile and edge devices) to achieve ultra-low latency, the translation models used must be small and energy efficient.

Future trends include the development of speech translation models that can translate from start to finish. This means they can reduce mistakes that happen when something is translated in parts. We will also see models that can work with languages and translate things they have never seen before. There will be systems that get to know the way a specific user speaks and can translate their speech better. These systems will work on devices. In places with different settings. They will be able to translate speech privately in real time. This will make it possible, for people who speak languages to communicate with each other easily. Speech translation models will play a role in this. They will help people understand each other better.

XI. CONCLUSION

This paper is a deep learning system that helps people have real conversations in different languages. The system is made up of a parts, including a speech recognition part, a text to speech part and a translation part.

The speech recognition part uses a kind of model called a transformer to understand what people are saying. The text to speech part uses a network to make the translated speech sound natural. The translation part uses a kind of modeling to get the context of the conversation right.

The system is really good at understanding what people are saying and translating it into another language. It can also make the translated speech sound which is important for having a real conversation. The system is fast so people do not have to wait a time for the translation.

In the end this system is a step towards making it possible for people to have real conversations in different languages. It is not perfect. It is a good start. In the future the system can be made better so it can understand people more clearly translate speech more quickly and make the translated speech sound more like a real person.

REFERENCES

1. X. He et al., "Streaming end-to-end speech translation with joint text and audio decoding," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2022.
2. E. Salesky et al., "Direct speech-to-speech translation with discrete units," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2021.
3. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Upper Saddle River, NJ, USA: Pearson Education, 2021.
4. T. Kano, S. Sakti, and S. Nakamura, "Simultaneous speech-to-speech translation with neural incremental processing," in Proc. IEEE/ACL Conference, 2020.
5. K. Sudoh et al., "A simultaneous speech-to-speech translation system with neural ASR, MT, and TTS," arXiv preprint, 2020.
6. H. Inaguma et al., "End-to-end speech translation with transformer networks," in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2020.
7. Y. Ren et al., "FastSpeech: Fast, robust and controllable text-to-speech," in Advances in Neural Information Processing Systems (NeurIPS), 2019.
8. M. Ma et al., "Simultaneous neural machine translation," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

9. A. Bérard et al., “End-to-end speech translation,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2018.
10. S. Watanabe et al., “ESPnet: End-to-end speech processing toolkit,” in Proc. Interspeech, 2018.
11. J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2018.
12. P. Haghani et al., “From audio to audio: Direct speech-to-speech translation,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2018.
13. R. J. Weiss et al., “A comparison of direct and cascaded models for speech-to-speech translation,” in Proc. Interspeech, 2017.
14. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017.