International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

A study on the techniques for speech to speech translation

Shrishti Sandeep Gupta[1], Vaishali Ramakant Shirodkar[2]

[1]Student, Information Technology Department, Goa College of Engineering, Farmagudi – Goa.
[2]Assistant professor, Information Technology Department, Goa College of Engineering, Farmagudi – Goa.

Abstract - *As globalism continues to advance, language barriers obstruct free communication. Speech translation targets at converting speech from one language into speech or text in another language. This technology helps overcome communication barriers between people communicating via different languages and can allow access to digital content in different languages. One of the central applications of automatic speech translation is to translate documents such as presentations, lectures, broadcast news, etc. However, direct speech-to-speech translation is a very complicated task that involves recognizing and automatically translating speech in real time. This paper outlines some of the work done in this field.*

Key Words: Speech-to-Speech, Translation, Transformer, Encoder, Decoder, Attention.

1. INTRODUCTION

Speech translation refers to the task of transcribing a spoken utterance in a source language into the target language. These systems are typically categorized into the cascade and End-to-End systems. The traditional speech translation system follows a step-by-step process which can be broken down into three components:

- automatic speech recognition (ASR)
- text-to-text machine translation (MT)
- text-to-speech (TTS) synthesis.

The ASR, MT, and TTS systems are trained and tuned independently. The ASR processes and transforms the speech into text in the source language, MT then transforms this text into the corresponding text in the target language. Finally, TTS converts the target language text into speech utterances.

Researchers are now analyzing direct speech-to-speech translation (S2ST) models that translate speech without relying on text generation as an intermediate step. Direct S2ST comprises fewer decoding steps. Therefore such systems have lower computational costs and low inference latency.

2. ANALYSIS OF VARIOUS METHODS USED FOR SPEECH TO SPEECH TRANSLATION

2.1 Cascaded speech translation model

This paper [1] implements a speech-to-speech translation robot in the domain of medical care that helps English speaking patients describe their symptoms to Korean doctors or nurses. The system consists of three main parts - speech recognition, English-Korean translation, and Korean speech generation. English-Korean translation in this system is based on the rule-based translation. This system consists of five main modules: tokenization, part-of-speech tagging, sentence components grouping, Korean grammar application, and word-by-word translation

It utilizes CMU Sphinx-4 as a speech recognition tool which is an open source program of Java speech recognition library. Once the recognition is successful, it passes the transcribed text to the translation system. Then the translation algorithm divides a sentence into basic sentence components, such as subject, verb, object, and prepositional phrase. It rearranges the parsed components by applying syntactic rules of Korean.

As a last step, the DARwIn-OP speaks the result of a translated sentence in Korean. As there was no appropriate Korean TTS program that can be applied to this program, pre-recorded MP3 files were used. Each word is matched to each Korean voice recording by looking up the hash table.

2.2 Listen, Attend, and Spell (LAS) model

Chiu et al. [2] presents the Listen, Attend, and Spell (LAS) model for direct speech to speech translation. The LAS model is a single neural network that includes an attention-based encoder-decoder. The LAS model consists of 3 modules. The encoder takes the input features, x, and maps them to a higher-level feature representation, h^{enc}. The output of the encoder is passed to an attender, which determines which encoder features in h^{enc} should be attended to in order to predict the next output symbol, y_i. The output from the attention module is passed to the decoder, which takes the attention context c_i, generated from the attender, and an embedding from the previous prediction, y_{i-1}, to produce a probability distribution, P(yi|yi−1, . . . , y0, x), given the previous units, {yi−1, . . . , y0}, and input, x.

In order to optimize the structure, Chiu et al explores word piece models (WPM) and multihead attention. WPM uses sub-word units, ranging from graphemes to entire words. The WPM is trained to maximize the language model over the training set. Words are segmented passively and independent of context using a greedy algorithm. Multi-head attention (MHA) extends the traditional attention mechanism to have multiple heads, where each head generates a different attention distribution. Thus each head has a different role on attending to the encoder output, which makes it easier for the decoder to fetch information from the encoder. Further Weiss et al focuses on the minimum expected word error rate (MWER) training. The main objective here is to minimize the expected number of word errors. Scheduled sampling is used to train the decoder. This helps reduce the gap between training and inference. Label smoothing is used as a regularization mechanism to prevent the model from making over-confident predictions.

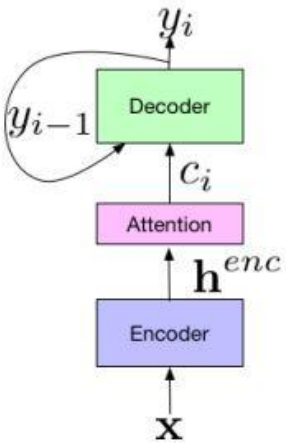

Fig -1: Components of the LAS model

2.3 *Translatotron Model*

Jia et al. [3] demonstrates a Translatotron which is an attention-based sequence-to-sequence neural network model trained end-to-end. This model does not require any text representation during inference. However, the model does not perform as well as the cascaded system.

The proposed Translatotron model architecture is composed of several separately trained components:

1) An attention-based sequence-to-sequence network (blue) which generates target spectrograms

2) A vocoder (red) which converts target spectrograms to time domain waveforms.

3) A pre-trained speaker encoder (green) which can be used to retain the identity of the source speaker, enabling inter-language voice conversion along with translation.

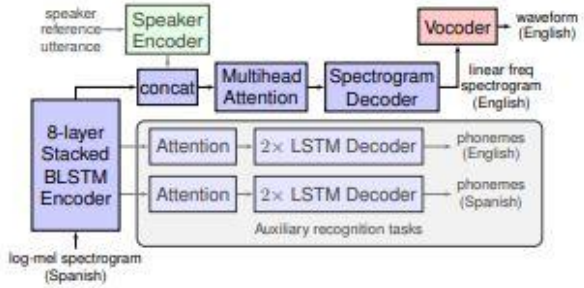

Fig -2: Proposed Translatotron model architecture

The encoder is composed of a stack of 8 bidirectional LSTM layers. The intermediate activations are passed to auxiliary decoders predicting phoneme sequences whereas the final layer output is passed to the primary decoder. The sequence-to-sequence encoder stack maps log-mel spectrogram input features into hidden states which are passed to an attention-based alignment mechanism that conditions the decoder. The decoder predicts log spectrogram frames in the target speech. Two auxiliary decoders, each having their own attention components, predict source and target phoneme sequences. This model uses multi-head additive attention with 4 heads. The primary decoder uses 4 or 6 LSTM layers while the auxiliary decoders use 2-layer LSTMs with single-head additive attention. The model also incorporates a speaker encoder network in order to control the output speaker identity. BLEU scores are used to measure the performance of the translation system. The model achieves high translation quality on two Spanish-to-English datasets although performance is not as good as the cascade models. The Translatotron sometimes mispronounces words, especially proper nouns.

2.4 Vector quantized variational autoencoder (VQ-VAE) Model

Tjandra et al. [4] proposes a method that can directly generate target speech without any pre-training steps that use source or target transcription. Vector quantized variational autoencoder (VQ-VAE) is used to extract the discrete symbols and capture the context without any supervision. The proposed model is trained using three different modules: VQ-VAE, a speech-to-codebook seq2seq, and a codebook inverter. The input to the encoder are speech features such as mel-spectrograms and input x's speaker identity. The encoder generates discrete latent variables.

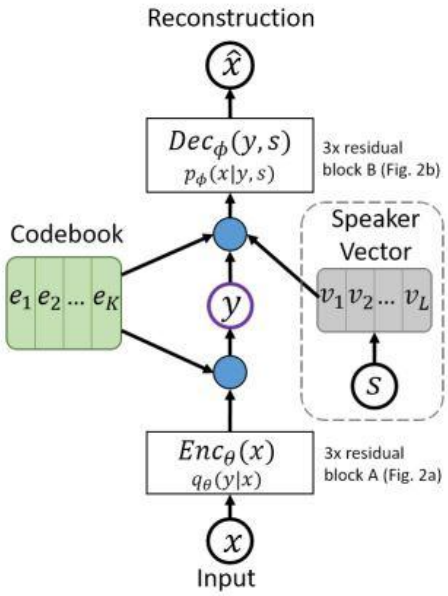

Fig -3: VQ-VAE model

The speech-to-speech translation model is built on an attention based sequence-to-sequence (seq2seq) framework. The seq2seq model is trained from the source language speech to the target language codebook. The inference stage uses Griffin-Lim to reconstruct the phase from the spectrogram. An inverse short-term Fourier transform (STFT) is applied to invert it into a speech waveform. Tjandra et al works on two data sets: French-to-English and Japanese-to-English. The performance is evaluated using BLEU scores and METEOR. This model can be applied to any language, with or without a written form because the target speech representations are trained and generated unsupervised.

2.5 Transformer-based Model

Kano et al. [5] proposes a Transformer-based speech translation system using Transcoder. The attention-based encoder-decoder component is trained step-by-step with curriculum learning from easy to complex tasks while changing the model structures. The learning scheme changes from single-task to multi-task learning while training the decoders and transcoders. The architecture consists of a single encoder that encodes the source language speech, three decoders that predict source language text transcriptions, target language text transcriptions, target language speech and two Transcoders that transfers the attention context information.

First, the pre-trained ASR, MT, and TTS models are prepared. The pre-trained ASR encoder is used for the source language speech encoding, the ASR, MT, and TTS decoders are used for the source language text, target language text and target language speech generation, respectively. We then fine-tune the overall framework by connecting these components with two Transcoders. The major difference is that previous multitask systems used only one attention module to align the input speech to the target speech. Whereas the proposed system relies on three attention modules.

Kano et al. [6] carries out experiments on English-toSpanish and Japanese-to-Korean datasets as syntactically similar language pairs and English-to-Japanese and Japanese-toEnglish datasets as syntactically distant language pairs. BLEU and METEOR scores are used to evaluate the model's performance. Experimental results show that end-to-end models outperformed the cascade models. The transcoder model also outperformed Google's multi-task-based speech translation model.

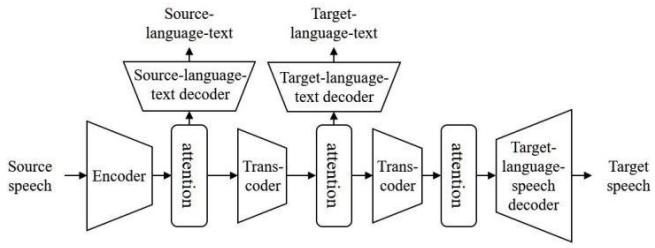

Fig -4: Proposed framework of the end-to-end speech-to-speech translation architecture

2.6 Direct speech-to-speech translation with discrete units

Lee et al. [7] presents a model that predicts self-supervised discrete representations of the target speech instead of mel spectrogram features.

Four targets for the auxiliary tasks are translation with discrete units is useful in scenarios where the source and target transcripts may or may not be available. The length mismatch issue between the speech and text output during decoding is resolved using connectionist temporal classification (CTC). In addition to the auxiliary tasks proposed, this paper designs a new text decoding task conditioned on the intermediate representation of the decoder.

The proposed system is a transformer-based sequence to sequence model with a speech encoder and a discrete unit decoder and incorporates auxiliary tasks. A vocoder is separately trained to convert discrete units into waveforms. The output of the auxiliary tasks can be either phonemes, characters, subword units or any discrete representations of the source or target utterances. These auxiliary tasks are used only during training and not in inference. Teacher-forcing with the target discrete units is used during training. Discrete unit decoding and CTC

decoding is used during inference. We train the enhanced vocoder by minimizing the mean square error (MSE) between the module prediction and the ground truth. The experiments conducted use the Fisher Spanish- English speech translation corpus. The dataset consists of 139k sentences from telephone conversations in Spanish, the corresponding Spanish text transcriptions and their English text translation.

For source phonemes (sp), a multihead attention module with 4 heads and a decoder with 2 transformer layers is used. For target phonemes (tp), we attach the attention and the decoder to the encoder. Eight bidirectional LSTM layers for the encoder and four LSTM layers for the decoder are used.

An open-sourced ASR model is used for evaluation. The BLEU scores are also listed and the difference between BLEU scores and ASR model results are evaluated. The results show that source units are effective in guiding the model to learn the attention effectively and the discrete representations learned in a self-supervised manner can capture basic pronunciations.

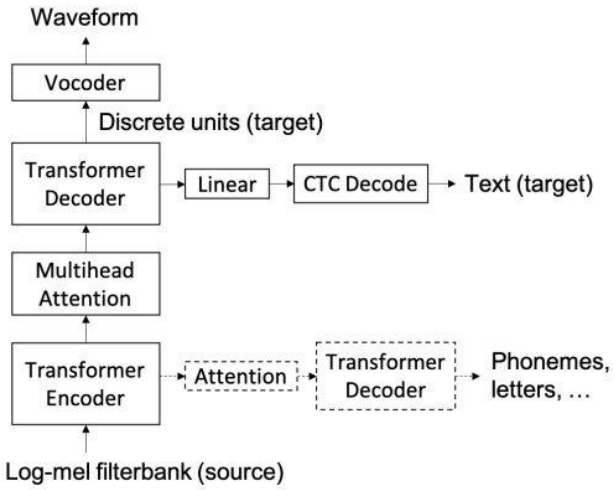

Fig -5: Direct S2ST model with discrete units.

2.7. Listen-Attend-Spell model with TacoTron

Guo et al. [8] proposes a method for end-to-end spoken language translation. The input to the model is an acoustic sentence in one language and the output is an acoustic representation of the same sentence translated into a different language. The model operates on spectrograms and combines the Listen-Attend-Spell model [9] with the TacoTron [10] architecture. The backbone of the model is a sequence-to-sequence architecture with attention. The model consists of a convolutional network, a bidirectional encoder to capture acoustic features and a decoder with attention.

In order to address the issue of learning from a large number of timesteps, a pyramidal RNN is used. A pyramidal RNN is the same as a standard multi-layer RNN but instead of each layer simply accepting the input from the previous layer, successively higher layers in the network only compute during particular timesteps. Pyramidal RNN also reduces the inference complexity.

To aid Learning, an attention based LSTM transducer is used. Here, attention is defined as the alignment between the current decoder frame i and a frame j from the encoder input. At each timestep, the transducer produces a probability distribution over the next character conditioned on all the previously seen inputs. The method proposed by Tacotron reduces training time and shows better spectrogram reconstruction performance. Since the model predicts only the spectrogram magnitudes, Griffin-Lim phase recovery is used to generate the final waveform of the translated sentence.

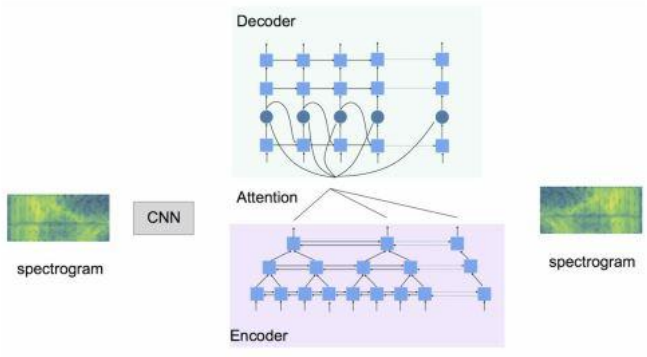

Fig - 6: Model overview

Experiments are conducted on a new spoken language translation dataset, titled Pearl. The dataset consists of grammatically correct input and output sentences spoken by the same speaker. A total of twelve speakers contributed to the dataset. The languages include Hindi, Mandarin, English, and Spanish. Using a newly collected dataset of multiple speakers in multiple languages, the model is able to learn acoustic and language features.

3. CONCLUSIONS

Most of the works carried out in direct speech to speech translation use an RNN model for modeling sequential data. However, we conclude that using transformer architecture in speech translation is more suitable than RNN. Since the transformer uses a self-attention function, it can learn long context information. The transformer can also bring down the calculation time, specifically when the data sequence is long.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Nilesh B. Fal Dessai from Goa College of Engineering for his valuable comments and suggestions.

REFERENCES

[1] S. Shin, E. T. Matson, Jinok Park, Bowon Yang, Juhee Lee and Jin-Woo Jung, "Speech-to-speech translation humanoid robot in doctor's office," *2015 6th* International Conference on Automation, Robotics and Applications (ICARA), 2015, pp. 484-489, doi: 10.1109/ICARA.2015.7081196.

[2] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li 0028, Jan Chorowski, Michiel Bacchiani. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. pages 4774-4778, IEEE, 2018.

[3] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu, "Direct speech-to-speech translation with a sequence-to sequence model," Proc. Interspeech 2019, pp. 1123– 1127, 2019.

[4] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Speech-to-speech translation between untranscribed unknown languages," in 2019 IEEE Automatic Speech Recognition and UnderstandingWorkshop (ASRU). IEEE, 2019, pp. 593–600.

[5] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, "End-to-end speech translation with transcoding by multi-task learning for distant language pairs," IEEE ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 1342–1355, 2020.

[6] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 958–965.

[7] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct Speech-to-Speech Translation With Discrete Units. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

[8] M. Guo, A. Haque, and P. Verma, "End-to-end spoken language translation," arXiv preprint arXiv:1904.10760, 2019.

[9] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

[10] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., … Saurous, R. A. (2017). Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. *CoRR*, *abs/1703.10135*. http://arxiv.org/abs/1703.10135

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[12] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, "Structured-based curriculum learning for end to-end english-japanese speech translation," in Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, 2017, pp. 2630–2634.

[13] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, "Attention-passing models for robust and data-efficient end-to-end speech translation," Transactions of the Association for Computational Linguistics, vol. 7, pp. 313–325, 2019.

[14] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in Proc. Interspeech, 2019.