Direct Speech to Speech Translation: Model Improvements

Bryant McArthur, Stephen D. Richardson

Brigham Young University
bmcarth4@byu.edu

Abstract

We propose a sequence-to-sequence LSTM attention-based neural network to perform machine end-to-end speech translation from English to German. This network is trained on English Mel-spectrograms to predictively generate corresponding German Mel-spectrograms, which can then easily be transformed back to audio waveform. We propose modifications to current SOA models to improve ease of implementation, training stability, and generalizability for future end-to-end speech translation models. In practice, this architecture was specifically built for performing speech translation of German audio into English, but may be easily trained on any language pair.

1 Introduction

Recent MT research has provided evidence that the quality of an End-to-End Speech Translation system could potentially match or outperform that of the traditional "cascaded" system. There is noise and increased error introduced to a cascaded system, especially in the ASR step that grows when passed through the following models. In order to eliminate unnecessary error a direct speech to speech model may be used.

Google's "Translatotron" (Jia et al. 2019) was one of the first attempts to improve the cascaded approach. It consists of a direct Speech-to-Speech model using a generative sequence-to-sequence convolutional neural network (CNN) similar to the architecture of "Tacotron" (Wang et al. 2017). The model takes as input the Mel-spectrogram of the source language audio and generates a predicted Mel-spectrogram in the target language.

Google's "Translatotron2" (Jia et al. 2021) closely follows the architecture of Translatotron (Jia et al. 2019) except that it also incorporates a speaker encoder to capture features of the spoken language in order to concatenate them to the embeddings and transfer speech naturalness through the translation.

Google's Translatotron and Translatotron2 are considered current SOA models, however some simple adjustments can be made to make the model more robust for implementation and training and improve quality of generalizability.

We will attempt to build an end-to-end speech translation system modifying the architecture of Google's Translatotron (Jia et al. 2019) in order to make it simpler and easier to use, resolve a common issue of exploding gradients in neural networks, and implement some more recent practices for improving any CNN.

2 Related Works

Google's Tacotron (Wang et al. 2017) and Translatotron (Jia et al. 2019) are CNNs using a generative sequence-to-sequence model. They both incorporate audio and Mel-spectrograms. In order to alternate between audio, spectrograms, and back to audio they wrote in-house functions including the short-time Fourier transform to go from audio to spectrogram, and the Griffin-Lim (Griffin and Lim 1984) algorithm to go back from spectrogram to audio.

These in-house functions are difficult to manage, work with, interpret, and scale for various audio file types. TorchAudio (Yang, et al. 2022) is an easy-to-use, versatile python package with various speech processing tools. Some of these tools include the Mel-spectrogram and inverse spectrogram transformations as well as the Griffin-

Lim (Griffin and Lim 1984) algorithm. All of TorchAudio is compatible with PyTorch Tensors for training and developing deep neural networks.

Furthermore, Tacotron (Wang et al. 2017) and Translatotron (Jia et al. 2019) use Dropout and Batch Normalization as regularization techniques in their CNN. Research done by Xiang Li, et al. (2019) shows that Dropout and Batch Norm should generally not be used simultaneously, especially in CNNs because they create a "disharmony by variance shift" from training to inference time that may cause "erroneous predictions". Their findings suggest batch normalization generally has a better overall effect than dropout.

We will incorporate both of these resources into our model adaptations by replacing delicate in-house functions with the more robust TorchAudio methods, as well as removing dropout layers from the CNN to solely rely on batch normalization as a regularization technique.

3 Data

3.1 LibriS2S

We used the LibriS2S dataset by Pedro Jeuris built for training direct speech to speech machine translation models (Jeuris and Niehues 2022). This dataset contains approximately 40 hours of source and target aligned speech in German and English by several different speakers. This dataset was built off of the LibriVoxDeEn corpus for speech translation and speech recognition (Beilharz et al. 2020). LibriVoxDeEn contains aligned triplets of English audio, English text, and German text. The LibriS2S dataset contains corresponding German audio from LibriVox for approximately half of the original LibriVoxDeEn dataset to create the quadruplets now found in LibriS2S.

Nobody has used and cited this dataset yet to train a direct Speech-to-Speech translation model. As opposed to Google's datasets used for Translatotron (Jia et al. 2019) and Translatotron2, (Jia et al. 2021) we will not be using synthesized speech for training. All the German and English recordings in LibriS2S are authentic speakers to help retain speech naturalness during training time.

The LibriS2S (Jeuris and Niehues 2022) dataset contains approximately 40 hours of speech for both German and English for a total of 80 hours. In comparison, Translatotron (Jia, et al. 2019) used approximately 100 hours of speech for both source and target languages.

4 Methodology

Below is an outline of the baseline model and our proposed adjustments as well as our evaluation approach.

4.1 Translatotron Baseline

First, we built our baseline model to closely resemble the architecture of Google's Translatotron (Jia et al. 2019). We used the code publicly available from Tacotron (Wang et al. 2017) and made the proper adjustments noted by the authors of Translatotron to ensure everything matched.

Early testing showed that the given preprocessing functions to convert audio to Mel-spectrograms for the DataLoader were incompatible with my audio files. At this point, we made the switch to TorchAudio (Yang, et al. 2022) for all preprocessing. It is easy to convert audio to Mel-spectrogram and back to audio with any file type and it provides built-in methods for easy plotting and visualization.

However, even after this switch for the data preprocessing, training the model was quickly cut short because of the issue with exploding gradients. We were never able to train a model for any substantive or promising results using the Tacotron (Wang, et al. 2017) code and proposed Translatotron (Jia, et al. 2019) alterations.

4.2 Single Training-Instance Model

We made several modifications to the architecture, parameters, and hyperparameters of the baseline model above and then used a single training-instance technique to show at the very least that our proposed model "works" and is able to "memorize" and reproduce the single instance it was trained on.

The first modification done was to the preprocessing as stated above using TorchAudio (Yang, et al. 2022) for easy implementation.

The second problem fixed was that of exploding gradients. We first adjusted the data normalization in the DataLoader. Originally, the data normalizer scaled the values of spectrograms, which ranged from [0, 60000], by using a logarithmic function log(x)/C to map the values to [-1, 1]. The negative values and the density of values around zero added to exploding gradients and other issues from activation functions. We simply changed the function to be log(x+1)/C to logarithmically map the values to [0, 1].

Next, we changed every occurrence of tanh activation functions to LeakyReLU. It is well known that the tanh activation function may cause problems with exploding and vanishing gradients that ReLU and LeakyReLU can help fix in many cases.

Third, we simply implemented gradient clipping. Instead of clipping the gradient norm like Tacotron (Wang, et al. 2017) and assumingly Translatotron (Jia, et al. 2019), we clipped the actual value of the gradient to avoid any possibility of explosion. While this method ensures we will not encounter exploding gradients in training time, adjusting the data normalization and switching the activation functions help ensure very large gradients are less likely to actually occur during our optimization step.

Finally, we made some other slight modifications to help during training time and increase generalizability of the model, i.e. increase the accuracy of testing for a broader range of data input. We did this by adjusting the regularization techniques and the learning rate.

As for the regularization, we follow Xiang Li, et al. (2019) and eliminate the use of dropout when both batch normalization and dropout were being used on convolution layers.

For our learning rate we use a cyclical learning rate between 1e-3 and 1e-6 as per Leslie Smith (2017). The method of using an exponentially decreasing cyclical learning rate allows us to descend our loss function quickly at the beginning of training time and find a more precise minimum in later training time while still being able to jump out of local minima.

With all of these adjustments we trained our model with a single sentence to ensure the model would be able to memorize and reproduce that given sentence in testing.

4.3 Training Full Dataset

Lastly, we started the training of our model using our adjusted architecture and hyperparameters on the full dataset. Even though we distributed the training onto six Nvidia A100 GPUs it is unlikely the model will converge in a reasonable time with the resources we have. However, we will evaluate preliminary results to better understand the promises of our architecture.

4.4 Evaluation

In order to evaluate the quality of the translation from the single-training instance model we use the standard ASRBleu. We simply run an automatic speech recognition (ASR) on the predicted audio file and then run the Bleu score (Papineni, et al. 2002) on the resulting text to compare it to the reference.

We will rely on the training and validation loss plots, the gradient norm plots, as well as a human evaluation on the progress of predicted Mel-spectrograms during validation steps of the training of our full dataset to show how promising our proposed model is in practice.

5 Results

No results are mentioned for the baseline model. Our baseline model following Translatotron (Jia, et al. 2019) would only prematurely stop training due to exploding gradients. For this reason, we proposed certain adjustments, and we evaluate the training process of our proposed model.

5.1 Single Training-Instance Model

With our proposed methodology we were able to train a model that successfully memorized a single training instance and reproduce that instance with minimal loss.

As you can see in Figure 1, the two Mel-spectrograms closely align with just a little noise added into the prediction.

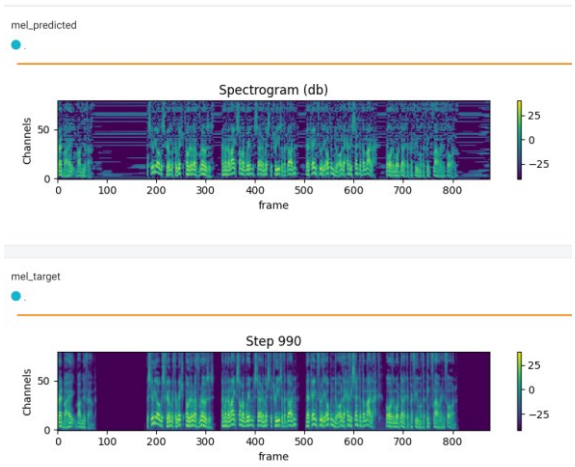

Figure 1: English Mel-spectrograms for predicted (top) and target (bottom) audio.

When we converted our prediction to audio form the speech itself exactly matched the reference audio. There was additional white noise as if the

speaker was speaking through a fan, but all the words were distinguishable to the human ear.

Unfortunately, when we ran ASR on the output of the text it greatly underrepresented the accuracy of the 'translation'.

Our reference English text was, "THE studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden, there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pinkflowering thorn." The output of the ASR was, "The studio was filled with the rich owner of roses I went to Light Summer Windstar damage the Trees of the garden there came through the open door the heavy scent of the Lilac or the more delicate perfume."

The ASR mixed up 'odour' with 'owner', 'and when the' with 'I went to', and 'wind stirred amidst' with 'windstar damage'. It also completely missed the end of the sentence, "of the pinkflowering thorn".

The predicted audio only received an ASRBleu score of 40.68 because of these errors due to the ASR and not our proposed model or methodology.

This result alone makes us very skeptical of any ASRBleu scores, however it must be used because it is currently the norm for evaluating speech to speech translation models. Ideally when evaluating a full training set we will also use BLASER (Chen, et al. 2022) a "text-free" evaluation metric specifically built for speech-to-speech translation, and human evaluations.

5.2 Training Full Dataset

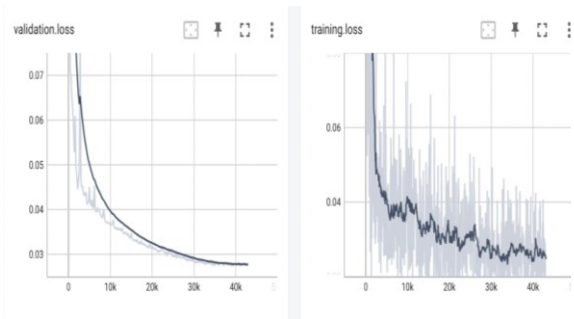

Figure 2: Training loss (right) and validation loss (left) of distributed training on full dataset.

Due to resources, we were not able to finish training our model on the full dataset. However, we are able to get a glimpse at its promising ability by the training and validation loss plots, gradient norm plot as well as the progress of predicted Mel-spectrograms in the validation steps.

As we can see from the loss plots in Figure 2, when training was terminated neither of the loss plots had yet converged and were still continually decreasing. This denotes that further progress would be made if training were to continue.

We are able to clearly see from Figure 3 that the issue with exploding gradients approaching infinity was resolved by our techniques. We used a gradient clip value of 5 which helped keep the gradients small at the very beginning of the training. However, for the majority of our training steps the gradient norm was kept below .05 and our gradient

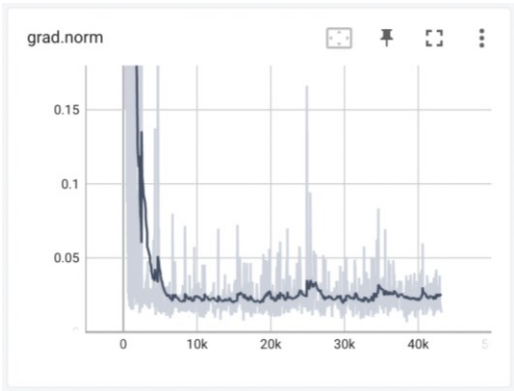

Figure 3: Gradient Norm plot of distributed training on full dataset.

clipping was never used. The real advantage to our methodology was adjusting the data normalization, activation functions, and learning rate.

From Figure 4 we can see the progression of Mel-spectrograms. It does not require a trained eye to see that the two bottom predictions (3rd row) for steps 31,800 and 43,000 much more closely align with the target Mel-spectrograms (4th row) than the top two predictions (1st row) for steps 4,200 and 6,600 with their target Mel-spectrograms (2nd row).

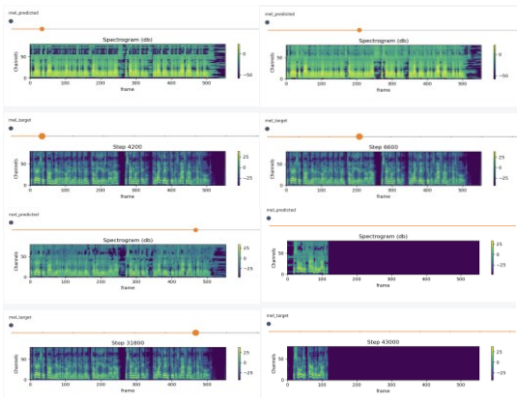

Figure 4: Predicted and Target Mel-spectrograms for various validation steps. Top Left: Step 4,200. Top Right: Step 6,600. Bottom Left: Step 31,800. Bottom Right: 43,000.

This shows promising preliminary results that the model is successfully learning how to generate accurate Mel-spectrograms.

The Mel-spectrograms at step 43,000 are still not developed enough to convert to an audible waveform, however these results give us confidence that with further training the model would converge just as it did with the single-instance training.

6 Conclusion

In conclusion, we have improved the SOA direct Speech-to-Speech translation model Translatotron (Jia, et al. 2019). Our proposed preprocessing with TorchAudio (Yang, et al. 2022) lends to more versatile use cases for various file types and easy implementation. We have shown certain techniques may be used to handle and mitigate the probability of exploding gradients in training time. And finally, we use a Cyclic learning rate (Smith 2017), LeakyReLU activation function and more current methods for regularization in a CNN by using only batch normalization and removing dropout (Li, et al. 2019) to improve training and generalizability at inference time.

7 Future Work

Our top priority for future work is to acquire the resources to complete the training of our full dataset. This will require much more than six A100 GPUs. After complete training we will be able to make proper evaluations on the quality of our translation system compared to current SOA systems.

We have shown ASRBleu is unreliable as an evaluation metric and hope to move to BLASER (Chen et al. 2022) and human evaluations. BLASER as a "text-free" evaluation metric specifically built for Speech-to-Speech translation shows promise and appears to be a much more reliable measure of quality for Speech-to-Speech models.

Furthermore, Automatic Dubbing is a MT constrained problem where the model attempts to, above all, retain timing constraints (isochrony) without sacrificing the translation quality. There is current research engineering cascaded models to meet these constraints, but there lacks development for the automatic dubbing problem in a direct Speech-to-Speech model.

Brannon, et al. (2022) analyzed Amazon Prime videos to determine how humans actually perform dubbing to obtain insights needed to improve automatic dubbing. They found that translation quality is of paramount importance, and although there are high rates of isochrony, speech tempo is of higher priority. This means, while human dubbing tries to meet isochronic constraints they are not willing to vary speech tempo to achieve isochrony, nor are they willing to significantly alter the meaning of the sentence and sacrifice translation quality for a shorter or longer target sentence to match the source. Brannon, et al. (2022) also found "strong evidence for several levels of non-textual transfer" including emotional features such as pitch, energy and vocal profiles to replicate in synthesizing natural target speech.

In a follow-up work with Amazon Prime videos, Chronopoulou, et al. (2023) built a model that largely maintained isochrony without much loss of translation quality. Although the model was a step in the right direction for isochronic MT, they used the conventional cascaded approach of ASR, MT, and TTS rather than a direct method.

We propose follow-up work to this paper to not only continue training this model for direct German to English Speech translation with proper resources, but also to incorporate isochrony, prosody, and other emotional features into the model by simple adjustments to the loss function to penalize over- and under-generation and adding a speaker encoder similar to the one found in Translatotron 2 (Jia et al. 2021).

Acknowledgments

We acknowledge Ammon Schur and Lawry Sorenson for their assistance.

References

Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel M. Lakew, and Marcello Federico. 2023. *Jointly Optimizing Translations and Speech Timing to Improve Isochrony in Automatic Dubbing*. arXiv preprint arXiv:2302.12979.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. *LibriVoxDeEn: A Corpus for German-to-English Speech Translation and Speech Recognition.* LREC. Marseille, France. https://arxiv.org/pdf/1910.07924.pdf

Daniel Griffin and Jae Lim. 1984. "Signal estimation from modified short-time fourier transform". Acoustics, Speech and Signal Processing, IEEE

Transactions on, vol. 32, no. 2, pp. 236-243. (Pubitemid 14608418)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318.

Leslie N. Smith. 2017. "Cyclical learning rates for training neural networks." In 2017 IEEE winter conference on applications of computer vision (WACV), pp. 464-472. IEEE.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2022. *BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric*. arXiv preprint arXiv:2212.08486.

Pedro Jeuris and Jan Niehues. 2022. *LibriS2S: A German-English Speech-to-Speech Translation Corpus.* arXiv preprint arXiv:2204.10593.

William Brannon, Yogesh Virkar, and Brian Thompson. 2022. *Dubbing in practice: A large scale study of human localization with insights for automatic dubbing.* arXiv preprint arXiv:2212.12137.

Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. 2019. "Understanding the disharmony between dropout and batch normalization by variance shift." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2682-2690.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack et al. 2022. "Torchaudio: Building blocks for audio and speech processing." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6982-6986.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. *Translatotron 2: High-quality direct speech-to-speech translation with voice preservation.* arXiv preprint arXiv:2107.08661.

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.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. 2017. *Tacotron: Towards end-to-end speech synthesis*. arXiv preprint arXiv:1703.10135.