Journal of Computer Engineering & Information Technology

Research Article A SCITECHNOL JOURNAL

Translation of Speech To Speech Using Cloud-Based Services and Deep-Learning Models

Ajit R Patil[1*], Kamlesh Patil[2] and Sonal Patil[3]

[1]Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Lavale, Pune, India

[2]Department of Information Technology, Bharati Vidyapeeth College of Engineering, Pune, India

[3]Department of E and TC, Bharati Vidyapeeth College of Engineering, Pune, India

***Corresponding author:** Ajit R Patil, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Lavale, Pune, India, Tel: 9975115055; E-mail: Patilajit667@gmail.com

Received date: 10 January, 2023, Manuscript No. JCEIT-23-86497;
Editor assigned date: 13 January, 2023, PreQC No. JCEIT-23-86497 (PQ);
Reviewed date: 27 January, 2023, QC No. JCEIT-23-86497;
Revised date: 18 April, 2023, Manuscript No. JCEIT-23-86497 (R);
Published date: 25 April, 2023, DOI: 10.4172/2324-9307.1000272

Abstract

In the last few decades, with the advent of the internet and social media, a global community has formed and its direct application for an underdeveloped community. Cognitive learning is one of the most sought after research fields which seeks to improve human computer interaction. This paper presents an application for speech to speech translation using a combination of cloud services and state of the art machine learning models. For speech to speech translation, a three phase architecture is investigated which contains cloud based speech to text, speech translation, token extraction, speech synthesis model based on deep neural networks, and a vocoder also based on deep neural networks. Our main focus in this study is creating a robust system for speech to speech language translation and the application of this system.

Keywords: Neural networks; Cognitive learning; Tacotron 2; HifiGAN; Speech to speech translation introduction

Introduction

Speech to speech translation is one of the most phenomenal advances in technology that has the capacity to make a great impact. It can influence global communication in various parts of the world since it is humanly impossible to learn all the existing languages. The total number of languages known to mankind is around 7000, and learning all of them stands out of the question. Thanks to the advancement in technologies, especially the internet, today's world is more 'global' than it ever was. This has made it much more urgent and necessary for us to have automated systems at the fingertips of every human being on this planet, probably through a web or a mobile application. Research in this field has mostly been focused on translation between one or more local languages and English. This is due to various factors firstly, due to the unavailability of the dataset in multiple languages. Secondly, since the computational power for just one language voice can require training the model for days to weeks, it does not seem cost effective. For many years the main goal of researchers has been not just to improve the efficiency of the models but to make it sound almost like human speech. Towards this end, many models like wavenet, HifiGAN, etc. have started synthesizing convincing human like speech, but progress is still needed. In this paper, we aim to implement a robust system that can address the implementation of this whole system. Our solution is to use cloud services for translation between languages in the speech to text format then use a deep learning model named tacotron 2 along with HifiGAN to synthesize the speech from the translated text. The advantage of using the tacotron 2 model is that with proper training we can mimic a human voice with very high accuracy and in many supported languages including Hindi, German, French, etc.

Materials and Methods

This section gives information about the language translation services that are offered. These services use a part presents network similar to the one used in this project for translation but may not always translate speech to speech. There are four translation services are explained here and its help to speech to speech translation.

Speech To Text (STT)

Speech to text is the first stage for speech to speech translation and has been the first topic under research. This is usually done by using an Automatic Speech Recognition (ASR) method which takes in audio files as input and converts them into phonemic and textual representations. In human communication, on the other hand, an open minded approach with sound responses from the speaker's mouth to his ear is essential. In their paper, they have taken a step forward and developed a closed speech chain model based on deep learning [1].

While ASR writes non-labeled speech features, TTS attempts to create the original speech format based on text from ASR [2]. On the other hand, the ASR is also trying to recreate the original text provided in the integrated speech. Efforts have been taken towards training models for local languages like Marathi, Gujarati, Bengali, etc. These models and datasets are very important when training for advanced systems and architectures.

Text To Speech (TTS)

Text To Speech (TTS) is a model for the synthesis of speech that takes textual data as input and produces human like speech. These models are mostly phoneme based which means that they consider the way in which letters of the words are spoken by a human [3]. Here, phonemes are any of the perceptually distinct units of sound in a specified language that distinguishes one word from another. This field of research has been under consideration for decades mostly focusing on phoneme based models and spectrogram analysis through acoustic methods [4]. However, the results seem to have become much more reliable since the introduction of neural networks. The most challenging task in this research field has been to make the synthesized sound indistinguishable from human speech.

Since understanding and critically comparing different models requires some kind of evaluation system is essential. Towards this end, the most common way to measure the performance of TTS models is subjective human evaluation using MOS (Mean Opinion Score), including some of the most novel quantitative metrics which are the

Frechet deep speech distance and Kernel deep speech distance which are well correlated with MOS [5].

Machine translation

In machine translation, we use an automatic system that converts one natural language into another without altering the actual meaning of the text while producing high quality translations [6]. Machine translation has been under study for at least 3 decades now. Starting from rule based systems to data driven systems these systems have shown a tremendous increase in output accuracy. In the rule based approach, words are matched with their transliterated words from a dictionary like database. A corpus based or data driven approach analyzes large document collections comparable or parallel corpus to construct a statistical translation model [7].

However, with the advent of faster and easier global communication, it still poses various challenges like;

- The absence of equivalent words between different languages.
- Some languages have words with multiple meanings.
- Languages having different grammatical structures. These problems are quite difficult to come around and systems, therefore, are required to predict the results by making inferences from the training data [8]. Another huge problem of machine translation is the need for tremendous amounts of data and being computationally expensive. Another challenge that is unique to multilingual communities like India where people mix two or more languages in one or more sentences. Such a situation is a nightmare for machine learning or NLP based systems [9].

Cloud cognitive services

With the advent of high performance computing resources and high speed and high bandwidth internet services available, outsourcing of computing resources through cloud services has become quite common. Various organizations like Microsoft's Azure, Google Cloud, Amazon's AWS, etc. have become commonplace in various applications as it provides security and saves money and time as storage and computation locally gets very costly. Azure provides various cognitive services through its cloud APIs (Application Programming Interface) which solves the issue of reinventing the wheel every time [10].

Results and Discussion

System architecture

Our proposed system consists of three phases, shown in Figure 1.

- A speech to text translator and a text to text language translator based on cloud.
- A recurrent sequence to sequence feature prediction network with attention which predicts a sequence of Mel spectrogram frames from an input character sequence.
- A HifiGAN model generates time domain waveform samples conditioned on the predicted Mel spectrogram frames, these waveforms are sent to the website as output for speech to speech translation some common mistakes.

In the first phase of our system, we have used cloud services for both speech to text and language translation using APIs. However, these are just placeholders at this point and should be replaced with better performing models. The second component uses the slightly modified tacotron 2 model, replacing wavenet vocoder with HifiGAN and adding a new component called DDC (Double Decoder Consistency). The Mel-Spectrogram generated by the tacotron 2 model is passed through the HifiGAN vocoder which generates. Wav files which can be sent to the third phase. The third phase is simply the linking of the model to a web application. In the following sections, we will be explaining in brief what those components are and why they can be called state of the art methods.

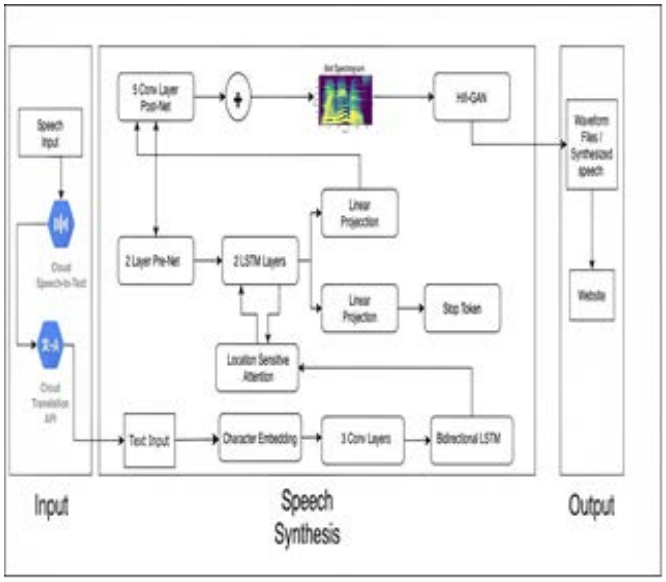

Figure 1: Sysytem architechure.

Tacotron model

Tacotron 2 can be said to be a combination of the best features of Google's wavenet, a deep generative model of raw audio waveforms, and tacotron proposed in, and its earlier speech recognition project [11]. The sequence to sequence model that generates Mel spectrograms has been borrowed from tacotron, while the generative model synthesizing time domain waveforms from the generated spectrograms has been borrowed from wavenet.

Tacotron 2 basically works on the principle of superposition of two deep neural networks [12]. One that converts text into a spectrogram, which is a visual representation of a spectrum of sound frequencies, and the other that converts the elements of the spectrogram to corresponding sounds (Figure 2). According to the paper, the system consists of two components;

- A recurrent sequence-to-sequence feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence.
- A modified version of wavenet which generates time-domain waveform samples conditioned on the predicted mel spectrogram frames.

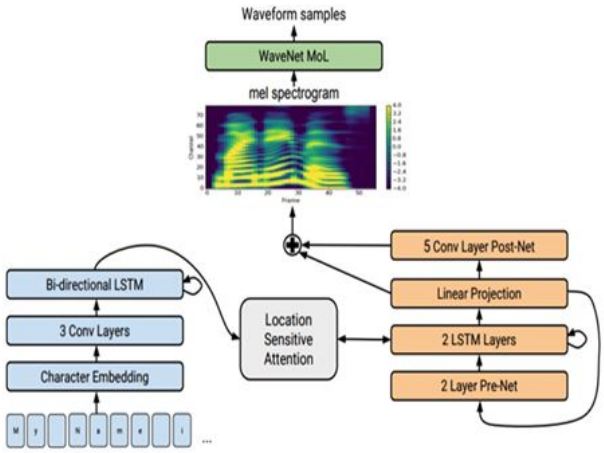

Figure 2: Tacotron architecture.

DDC

Despite being successful in solving the text-to-speech synthesis problem to a significant extent, most models suffer from attention alignment problems at inference time. They occur especially with long-text inputs or out-of-domain character sequences [13]. Basically, DDC consists of two decoders that learn synchronously with different reduction factors. It uses the level of consistency of these decoders to attain better attention performance. Good examples of attention-based TTS models are tacotron and tacotron 2.

The architecture of our work is based around tacotron 2 and a cocktail of some different models like HifiGAN and DDC which improve the performance of tacotron 2. These models constitute architecture, which has an encoder, an attention module, a decoder, and an additional stack of layers called postnet, all of which is stacked in a sequence-to-sequence architecture. The encoder generates a hidden representation from which the decoder generates predictions for the target acoustic feature frames. A context-based attention mechanism is used to align the input text with the predictions. Finally, decoder predictions are sent across the postnet, which predicts residual information to help the model perform better during reconstruction. Mel-spectrograms are acoustic features that are utilised to represent audio information in a perceptually relevant way with a lower temporal resolution.

Wavenet

Tacotron 2 inverts the mel spectrogram feature representation into time-domain waveform samples using a modified version of the wavenet architecture from. There are 30 dilated convolution layers, organised into three dilation cycles, as in the original architecture, with layer k (k=0...29) having a dilation rate of 2 k (mod 10). Only two up sampling layers, rather than three, are employed in the conditioning stack to work with the spectrogram frames' 12.5 ms frame hop. The research published by and demonstrates this [14].

The Mean Opinion Score (MOS) for this model is 4.53, which is similar to the MOS of 4.58 for professionally recorded speech. Offers ablation analyses of major system components and assesses the effects of employing mel spectrograms as the wavenet conditioning input rather than language, duration, and F0 parameters. It also demonstrates that by employing this tiny acoustic intermediate representation, the wavenet architecture may be significantly reduced in size.

And unlike original tacotron, tacotron 2 uses simpler building elements in the encoder and decoder, including the vanilla LSTM (Long Short Term Memory) and convolutional layers instead of CBHG stacks and GRU recurrent layers. In the tacotron text-to-speech model, CBHG is a building block. It consists of a bank of 1-D convolutional filters, highway networks, and a Bidirectional Gated Recurrent Unit, in that order (BiGRU). A Gated Recurrent Unit, or (GRU), on the other hand, is a form of recurrent neural network. It's similar to an LSTM, except that it just has two gates: A reset gate and an update gate, and no output gate. Each decoder step in tacotron 2 corresponds to a single spectrogram frame, hence there is no “reduction factor.” Instead of additive attention, location-sensitive attention is used (Table 1).

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
Wavenet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2	4.526 ± 0.066

Table 1: MOS values of different models.

HifiGAN

Recently, generative models have gained in prominence, notably after the development of generative adversarial models [15]. In recent years, generative adversarial networks have advanced rapidly, resulting in significant advances in picture generative modelling. However, its use in the audio domain has gotten less attention, and autoregressive models like wavenet are still the gold standard for generative modelling of audio signals like human speech.

HifiGAN presents GAN-TTS, a Generative Adversarial Network for Text-to-Speech, to overcome this deficiency. This system consists of a conditional feed-forward generator that generates raw speech audio and an ensemble of discriminators that act on random windows of various widths. The discriminators assess the audio in terms of general realism as well as how well it matches the utterance that should be said. GAN-TTS can produce high-fidelity speech with

naturalness equivalent to state-of-the-art models, and unlike autoregressive models, it can be highly parallelized thanks to an efficient feed-forward generator.

HifiGAN distinguishes itself by employing an ensemble of Random Window Discriminators (RWDs) that work on randomly sub-sampled segments of actual or produced samples rather than a single discriminator. The ensemble is created by taking a Cartesian product of two-parameter spaces.

- The size of the random window input into the discriminator.
- Whether a discriminator is conditioned on language and pitch attributes. For example, in our best-performing model, we take into account five window widths (240, 480, 960, 1920, and 3600 samples), yielding a total of ten discriminators. The random windows for these discriminators are sampled at full 24 kHz frequency without limitations, increasing the amount of training data even further.

Conclusion

We have implemented the above architecture with satisfying results of speech-to-speech language translation supporting more than 20 languages including Hindi, Bengali, and Marathi. This application can be implemented in places where we have a multilingual workforce without a common language. We have used cloud services in phase 1 to avoid reinventing the wheel as cloud STT services are quite satisfactory. More research is needed into this to make sure that the results are accurate even in our society where code-mixing is a commonplace.

Future Work

This technology is being implemented for a desktop application, but it can also be used for a mobile phone in the future. As a result, customers can more efficiently use this system by simply pressing a button on their mobile device rather than relying on a desktop for language conversion. By adding machine learning models for the translation, document reading, or over-the-phone use, it can be transformed into a genuine offline application. Translation procedure after transcribing. The document translation can initially support simple documents with extensions like. txt, docx, etc.; but, by introducing image detection models, it can also be utilized for translating pdf documents. It will encourage more research because it will cut down on the cost of translating research papers and free up lots of time for reading or making citations.

Acknowledgment

We would like to acknowledge the help provided by the technical and support staff in the computer engineering department of the Bharati Vidyapeeth's college of Engineering, Lavale, Pune which is affiliated with the Savitribai Phule Pune university and L and D Infotech Pvt. Ltd. under whose constant guidance we have accomplished this humongous task. We would also like to show our deep appreciation to our Prof. Ajit Patil who supported us throughout this process.

References

1. Tjandra A, Sakti S, Nakamura S (2017) Listening while speaking: Speech chain by deep learning. ASRU 301-308.
2. Sultana R, Palit R (2014) A survey on Bengali speech-to-text recognition techniques. Int J Technol Manag 26-29.
3. Kastner K, Santos JF, Bengio Y, Courville A (2019) Representation mixing for tts synthesis. Speech Signal Process 5906-5910.
4. Klatt DH (1987) Review of text-to-speech conversion for English. J Acoust Soc Am 82:737-793.
5. Streijl RC, Winkler S, Hands DS (2016) Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. Multimed Syst 22:213-227.
6. Hutchins J (2005) Example-based machine translation: A review and commentary. Mach Transl 19:197-211.
7. Thara S, Poornachandran P (2018) Code-mixing: A brief survey. In Int Conf Adv Comput Commun Inform (ICACCI) 2382-2388.
8. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, et al. (2017) Tacotron: Towards end-to-end speech synthesis. ArXiv org 1703-1735.
9. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proc IEEE Int Conf Acoust Speech Signal Process 4779-4783.
10. Wu YC, Hayashi T, Tobing PL, Kobayashi K, Toda T (2021) Quasi-periodic wavenet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. IEEE/ACM Trans. Audio Speech Language Process 29:1134-1148.
11. Oord AVD, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A (2016) Wavenet: A generative model for raw audio. ArXiv org 1609-3499.
12. Su J, Jin Z, Finkelstein A (2020) HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. ArXiv preprint arXiv 2006-5694.
13. Saini S, Sahula V (2015) A survey of machine translation techniques and systems for Indian languages. Int Conf Comput Intell Commun Technol 676-681.
14. Jia Y, Johnson M, Macherey W, Weiss RJ, Cao Y, et al. (2019) Leveraging weakly supervised data to improve end-to-end speech-to-text translation. Proc IEEE Int Conf Acoust Speech Signal Process 7180-7184.
15. Kuang S, Li J, Branco A, Lout W, Xiong D (2017) Attention focusing for neural machine translation by bridging source and target embeddings. ArXiv preprint arXiv 1711-5380.