International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May-2021 ISSN: 2582-3930

Voice Imitator

Dr.Shubhangi Vaikole[1], Suyash Gupte[2], Chhand Patil[3], Saloni Patil[4]

[1]*Associate Prof., Dept. of Computer Engineering, DattaMeghe College Of Engineering, Navi Mumbai, India*
[2]*Student, Dept. of Computer Engineering, DattaMeghe College Of Engineering, Navi Mumbai, India*
[3]*Student, Dept. of Computer Engineering, DattaMeghe College Of Engineering, Navi Mumbai, India*
[4]*Student, Dept. of Computer Engineering, DattaMeghe College Of Engineering, Navi Mumbai, India*

Abstract -Voice Imitator is a technology that aims to change the way we look at the online education system. The COVID-19 pandemic has changed education forever. We are now looking into the world where online education has flourished. In this current situation, there are language barriers in communication that limit the extent of our system. A Professor's reach is limited by the language they know, a german professor can only attract students who know german. A medium to reduce this barrier is needed. Machine translated voices fail to grab the attention of the listener for a long time. To overcome this, voice cloning can be used. Speech translation is the process by which conversational spoken phrases are instantly translated and spoken aloud in a second language. Speech translation technology enables speakers of different languages to communicate. Translation is necessary for the spread of information, knowledge, and ideas. It is absolutely necessary for effective and empathetic communication between different cultures. Translation is also the only medium through which people come to know different works that expand their knowledge. Language Translation has always been about inputting sources as text/audio and waiting for the system to give translated output in desired form.

***Key Words*:**Voice Cloning, Synthesize, Imitate, Librispeech.

1.INTRODUCTION

In the current scenario we realised the need for boosting our technology so that it can help us when we are in need.Language is the one thing in the world that can both enable, and at the same time, completely shut out human communication. Language is what keeps us connected to each other and the world. If it's a language known to us, we take hardly seconds to understand it. But if it's a language that we don't understand, it just cannot be understood without using dictionaries, manual parsers, translators and/or various applications available for translation. All of these solutions disrupt the flow of any conversation that someone could have with another person of a different dialect, because of the pause required to request for translation and time it takes for the actual translation process. We can only learn in the language that we know this is a disadvantage for the learner who seeks to lear something but it is not in his/her language.For this they have to learn a certain language first and then learn what they desire to learn. This takes a lot of time and effort from the learner.

Speech communication is a major medium of communication among human beings. All human beings have flexibility in changing parameters of speech like loudness, duration, pitch and intonation within their voice limits. Imitators have the ability to convince the listeners that they are listening to someone else. It is the flexibility of speech production mechanism that allows imitators to perform voice cloning in their own understandable language. Speech signals contain information about the text (speech) that is spoken, the language in which it is spoken, the speaker who uttered the text, the gender and the emotional state of the speaker. It conveys the mood of the speaker by variations in pitch, loudness, intonation, stress, pause and other such features, due to flexibility of human speech production mechanism. Every human being has a unique speech production mechanism, and hence a unique voice. But, humans also show the ability to speak in different voices, for example speaking in a soft voice if they wish to whisper or speak very loud and fast when they are anxious. Some humans have the ability to speak like someone else in a very convincing manner to the extent of fooling other humans. This act of talking like another speaker is called voice cloning. It is the flexibility of the production mechanism that allows voice cloning.

2. BACKGROUND

Speech Translation has always been about giving source text or audio input and waiting for the system to give translated output in desired form.Speech translation work has been done starting from single-speaker neural speech synthesis and moving on to multi-speech synthesis.

With regards to single-speaker speech synthesis, deep learning has been used for a variety of subcomponents, including duration prediction (Zen et al., 2016), fundamental frequency prediction (Ronanki et al., 2016), acoustic modeling (Zen and Sak, 2015), and more recently autoregressive sample-by sample audio waveform generation (e.g., Oord et al., 2016; Mehri et

al., 2016). Our contributions build upon recent work in entirely neural TTS systems, including Deep Voice 1 (Arik et al., 2017), Tacotron (Wang et al., 2017), and Char2Wav (Sotelo et al., 2017). While these works focus on building single-speaker TTS systems, our paper focuses on extending neural TTS systems to handle multiple speakers with less data per speaker. While most of us are familiar with different web-apps none of them provide translation in human voice. Our web-app will do multi-speaker translation in human voice. Users will not feel that they are talking to a machine.

The systems that are available currently have been used for a very long time and most of them we have used. The Traditional web-apps that are currently in use that translate language give us translation in machine voice ,We know that machine is talking to us. Also once the lecture is over there is nowhere to refer to what is taught.Apart From the gaps that the technology has it has still proven to be useful for now, There is also ongoing research in this field and more and more advances are coming. While the ongoing advances are done but are limited to machine language and no data for later reference.This allows us to explore more in this field as it is a technology that is used worldwide and needed worldwide as well. It gives us the liberty to make more changes and advancement in the field of online studying.

2.1 Dataset Details:

Librispeech[5] - LibriSpeech is a corpus of approximately 1000 hours of read English speech with sampling rate of 16 kHz, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.87

FeaturesDict({
'chapter_id': tf.int64,
'id': tf.string,
'speaker_id': tf.int64,
'speech':
Audio(shape=(None,),
dtype=tf.int64),
'text': Text(shape=(), dtype=tf.string),
})

Voxceleb[6] - A large scale dataset for speaker Identification. This data is collected from over 1,251 speakers, with over 150k samples in total. This release contains the audio part of the voxceleb1.1 dataset.

FeaturesDict({
'audio':
Audio(shape=(None,), dtype=tf.int64),
'label': ClassLabel(shape=(), dtype=tf.int64, num_classes=1252),
})

3. PROPOSED SOLUTION

The goal of this model is to imitate the speaker's voice in English Language.
The model has three parts:-

1. Speaker Encoder
2. Synthesizer
3. Vocoder

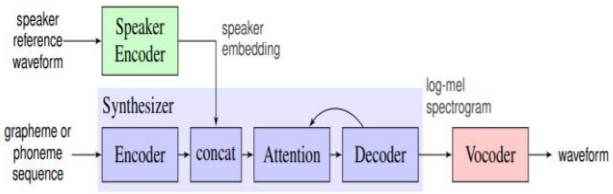


Figure 3.1 - Model overview

Speaker Encoder
The Speaker Encoder[1] is based on Generalized End-to-End Loss for speaker verification. Generalized end-to-end (GE2E)[2] training is based on processing a large number of utterances at once, in the form of a batch that contains N speakers, and M utterances from each speaker on average. Speaker verification (SV) is the process of verifying whether an utterance belongs to a specific speaker, based on that speaker's known utterances.
Speaker Verification are of two types:

1. Text Dependent - Like using "Ok Google" to activate google assistant.
2. Text Independent- Any word can be used for speaker verification.

We perform mostly Text Independent Speaker Verification in this model. Each feature vector xji ($1 \leq j \leq N$ and $1 \leq i \leq M$) represents the features extracted from speaker j utterance i. Similar sounding utterances form a cluster. In GE2E the embedding vector (d-vector) is defined as the L2 normalization of the network output. The similarity matrix Sji,k is defined as the scaled cosine similarities between each embedding vector eji to all centroids ck . $S_{ji,k} = w \cdot \cos(e_{ji}, c_k) + b$ We apply a softmax layer to Sji. In GE2E the embedding is pushed towards the centroid of the true speaker and pushes away from the rest of the clusters. This increases accuracy. Input 40-channel log-mel spectrograms are passed to a network consisting of a stack of 3 LSTM layers of 768 cells, each followed by a projection to 256 dimensions.

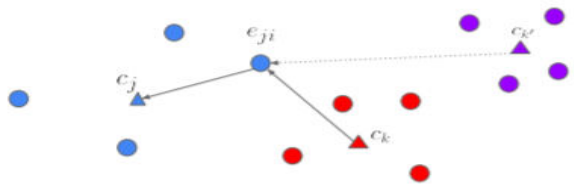


Figure 3.2 - GE2E Visualization

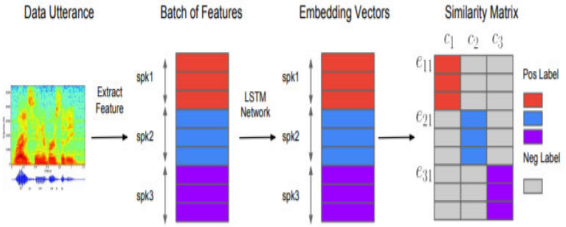


Figure 3.3 - GE2E Architecture

Synthesizer

The synthesizer[1] is trained on pairs of text transcript and target audio. The network is trained in a transfer learning configuration, using a pre-trained speaker encoder (whose parameters are frozen) to extract a speaker embedding from the target audio. It is based on the Tacatron-2[3] Architecture. The Tacotron-2 architecture is 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. Mel spectrograms are computed through a short time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop, and a Hann window function.

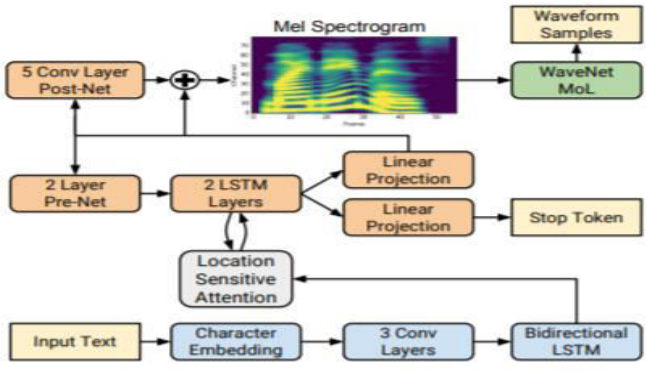


Figure 3.4 - Tacotron-2 Architecture

Studies have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies. In 1937, Stevens, Volkmann, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the melscale. A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.

Vocoder

A vocoder[1] is used to invert synthesized mel spectrogram's emitted by the synthesis network into time domain waveforms. The vocoder that we are using is the Wavenet Model..Wavenet[4] is a Deep Generative Model of Raw Audio Waveforms. The Wavenet Model is able to learn characteristics of many different voices , Male and Female. It directly models the raw waveform of an audio signal, one at a time. It is a fully convolutional neural network where the convolution layers have various dilation factors. The Wavenet Model is able to generate speech which mimics any human voice and sounds natural . The WaveNet neural network architecture directly generates a raw audio waveform, showing excellent results in text-to-speech and general audio generation . The network models the conditional probability to generate the next sample in the audio waveform, given all previous samples and possibly additional parameters. There are a number of layers in the neural network. The core of the network is a stack of casual dilated layers.

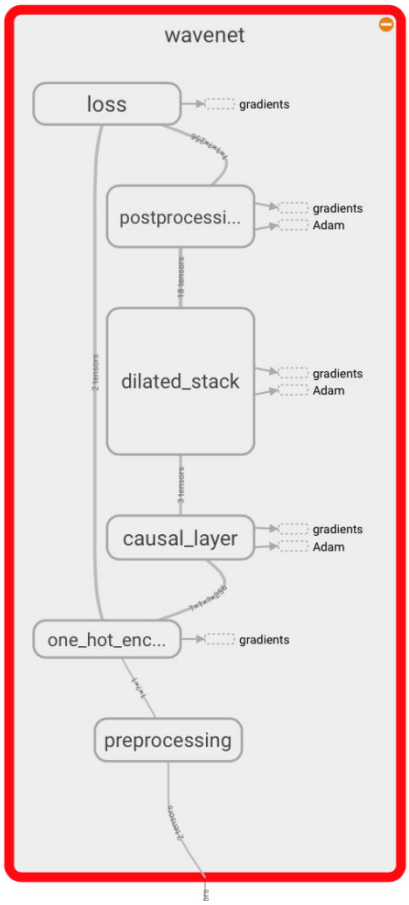


Figure 3.5 - WaveNet Working

This voice cloning model can be used to imitate the user's voice in a different language i.e English in this case. Using the Google speech to text API , the encoder will generate embeddings for the user in his preferred language , the API will generate english text for the audio and synthesizer

will generate mel spectrograms and vocoder finally will provide the output voice in english.

Video Calling
WebRTC framework will be used to implement video calling.
Steps to initialize a video call:
1. Support and request sending.
2. Processing user's response.
3. Video conference initialization.

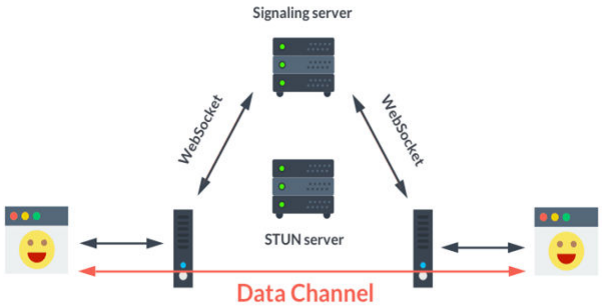


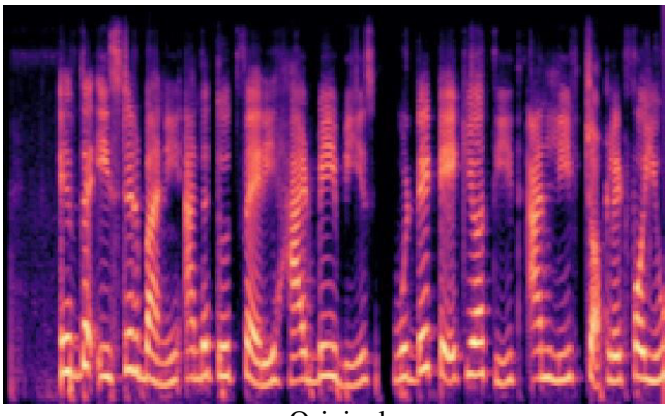
Figure 3.6 – WebRTC

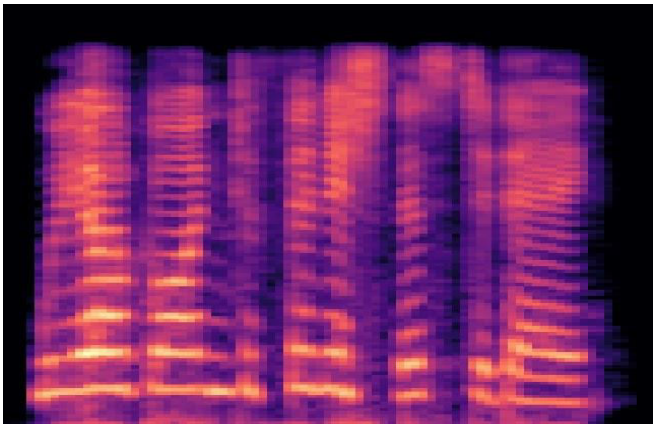
WebRTC is entirely peer to peer, so we don't have to pay for any of the bandwidth across the wire. Additionally, because WebRTC is entirely browser to browser, we get the highest performance and lowest latency possible.

4. EXPERIMENTAL RESULTS

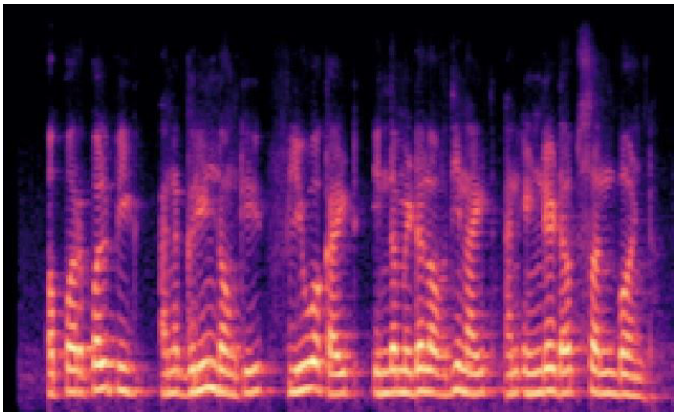
Mel Spectrograms are used for reference between the original voice and generated voice.

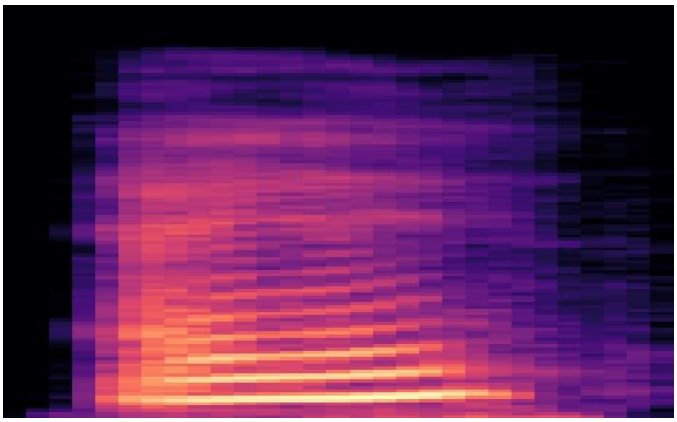
Text:-*If the Easter Bunny and the Tooth Fairy had babies would they take your teeth and leave chocolate for you?*


Original


Generated

Text:-*A purple pig and a green donkey flew a kite in the middle of the night and ended up sunburnt.*


Original


Generated

The generated voice for female voices is far better than the male counterparts. As the dataset is of English speakers from the US it fails to grab the accent of Indian speakers properly.

We asked 5 people if they recognised the generated voice belongs to which speaker
and the results are in the table below:-

	Recognised	Not Recognised
Speaker 1 (Female)	4	1
Speaker 2 (Male)	2	3

5. CONCLUSION

In this new era of online learning this web-app helps students concentrate on what the teacher wants to say and help them to learn. This web-app will also reduce all the language barriers in learning by translating languages. With the feature of script generation it also helps in future reference of the lectures.The lecturer can also see the script of the lecture and determine what is taught and what is remaining. A video calling web-app with voice cloning and translation is just the beginning it's future scope is very wide.

The use of online means for education will push the growth on more advancement in this field.The voice translation which for now is just on the lecturer side in future we can also make advancement in the technology and make the translation on the student side as well.This will make it more easy for the students to ask doubt and be clear what they have understood.Moreover, the accuracy rates will play a big part in the acceptance of this technology .Less the delay in translation more powerful and useful the technology will be.

REFERENCES

1. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.
2. Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification.
3. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, NavdeepJaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, YannisAgiomyrgiannakis, and Yonghui.Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.
4. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, NalKalchbrenner, Andrew Senior, and KorayKavukcuoglu. WaveNet: A generative model for raw audio.
5. https://www.tensorflow.org/datasets/catalog/librispeech
6. https://www.tensorflow.org/datasets/catalog/voxceleb
7. Analysis of mimicry speech based on excitation source information by D. Gomathi Alias Ramya
8. 2016 International Conference on Information Communication and Embedded Systems (ICICES), Hans Krupakar, KeerthikaRajvel, B. Bharathi
9. Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices Xin Lei, Andrew Senior, Alexander Gruenstein, Jeffrey Sorensen
10. Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Johnathan Raiman, Yanqi Zhou. Deep Voice 2:Multi-Speaker Neural Text To Speech.
11. Hans Krupakar, Angel Deborah Suseelan ,B.Bharathi, Vallidevi Krishnamurthy. A Survey of Voice Translation Methodologies - Acoustic Dialect Decoder