International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 132-135 www.ijaem.net ISSN: 2395-5252

Voice Cloning

Saiesh Prabhu Verlekar[1], Saili Kulkarni[2], Varad Naik[2], Aaron Mendes[2], Saiesh Naik[2]

[1]*ProjectGuide,DepartmentofInformationTechnology,ShreeRayeshwarInstituteofEngineeringand Information Technology Goa,India*
[2]*Students, Department of Information Technology, Shree Rayeshwar Institute of Engineering and Information Technology Goa,India*

Submitted: 25-01-2022 Revised: 05-02-2022 Accepted: 08-02-2022

ABSTRACT
Deep learning models for natural voice cloning methods begin in 2016, since then the main focus of the researchers is to make the voice more natural and get the output voice in real-time.Previouslyitusedtotakemanyhoursofvoicesamplestocloneafewseconds.Afterusing deeplearningmodelsnow,itisreducedtoafewseconds. Inthispaper,wewillstudydifferent techniques used for voice cloning. Some of these methods are multi-speaker generativemodel, speaker adoption, speaker encoding, vector quantization,etc.
Index Terms - Voice cloning, text-to-speech, voice conversation, speaker adaption, speaker encoding, vector quantization, neural network.

I. INTRODUCTION

The definition of VC (voice conversion) is to change the voice of the source speaker to the target speaker's voice without changing the language features. To imitate the target speaker, a VC system should modify the tone, accent, and pronunciation of the voice of the source speaker. The process of producing natural speech from a written input remains a challenging task despite decades of research[2]. These days there are many text-to-speech (TTS) methods able to get great results in terms of synthesis of natural voices very close to human once. Unfortunately, many of these systems learn to synthesize text only with a single voice. The goal of this project is to create a system that can effectively generate a data efficient natural speech for many different speakers. In practice, this technology can be used in many applicationssuchasentertainment,creativityandthistechnologyalsoallowsrestoringofvoice or customizing digital assistants such asSiri[5].

Voice conversion consists of supervised and unsupervised scenarios. In the earlier times, researches on one-to-one or many to-one conversion systems such as GMM-based and regression-based models, which convert one or many speakers voice into specific target speaker, is supervised learning. It learns to map the data into specific target distribution and usually achieves higher voice similarity and speech quality. But they need frame-level alignment on training data which may lack flexibility[10].

UnsupervisedVC,whichdoesnotneedparalleldata,becameapromisingresearchdirectionto tackle the many-to-many voice conversion problem, and most of the works achieved good performance on seen speakers. Generative adversarial network (GAN) is one of the successful voice conversion methods owing to its distribution guarantee between the generated data and true data, and the difference between speakers is the most obvious one. Some worksutilize vector quantization technique to extract content information and feed into WaveNetcombined with one-hot speaker embedding to synthesis the voice. This output generated is remarkably closertonaturalhumanvoice.However,theystillsuffer oneproblemwhichisthattheycannot synthesis the voice which does not exist in the trainingdata[2].
One-shot technique can solve the unseen speaker problem. One-shot model only needs one utterance from the source speaker and the target speaker. In the training phase, they should learn to differentiate the content information and speaker information, and most of them are not trained on speaker label to avoid overfitting on the training data. Proposed IN-based one- shot VC, which is a technique widely used in the computer vision, and it shows that this approach can learn the meaningful speaker embedding.
The important elements are:

Multi speaker generative model- In speech synthesis, generative models (speaker generative model) can be conditioned on text and speaker

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 132-135 www.ijaem.net ISSN: 2395-5252

identity. Although the text contains language proficiency and controls the content of the produced speech, the speaker identity captures features such as tone, speech rate, and pronunciation[1].

Speakeradaption-Speakeradaptationisbasedonfine-tuningamulti-speakergenerativemodel with a few cloning samples[1].

Speaker encoding - Derives an embedding from the short utterance of a single speaker. The embedding is a meaningful representation of the voice of the speaker, such that similar voices are close in latent space [1].

Vector quantization - Vector quantization is a method of constraining an input from a continuous or otherwise large set of values by using a shared codebook [2].

II. LITERATURESURVEY

AccordingtoShijunWangandDamianBortht heytriedtoperformvoiceconversionthatcould transformvoicefromaudiotosourcewithoutlosingthei mportantcontentsinalanguage.They performed methods such as voice conversion, vector quantization, contrastive predictive coding, augmentation. The results that were found out were with contrastive predictivecoding and Noise Augmentation the content encoder achieves very good results in capturing the content information. Meanwhile they also concluded that speaker encoder was successfully able to extract speaker characteristics[7].

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee proposed a model that could effectively perform one shot voice conversion with a single utterance from the source and the speaker, before this voice conversion with parallel data was not successful. They used disentangled representationsapproach,instancenormalizationandg enerativemodel,usingthismethodsand modelstheywereabletoperformvoiceConversiononu nseenspeakersaswellasthesubjective and objective progression showed positive results for the target speaker[3].

Hieu-Thi Luong and Junichii Yamagishi Nautilus: A Versatile Voice Cloning System It can actasaTTSorVCsystemwithhighconsistencyinterms ofspeakercharacterswhenswitched between two[8].

KaushikDaspute,HemangPanditRealTime VoiceCloningInthistheytriedtoreplicatethree phase pipe line model that allows you to clone unseen voice with few sec voice sample[9].

SercanÖ.Arık,JitongChen,KainanPeng,We iPing,YanqiZhou.NeuralVoiceCloningwith a Few Samples In this they discussed about speaker encoding and speaker adoption, speaker encoding gives fast output but the voice feels little synthetic as compared to speaker adoption [1].

Merlijn Blaauw Jordi Bonada and Ryunosuke Daido DATA EFFICIENT VOICE CLONING FOR NEURAL SINGING SYNTHESIS From small amount of voice sample we can produce voice fitting or speaker adoption [6].

Paarth Neekhara Shehzeen Hussain Shlomo Dubnov Expressive Neural Voice Cloning Evaluating 3 aspects of extracting and transferring audio reference to speech [4].

Da-Yi Wu, Hung-yi Lee proposed a vector quantization based one-shot VC approach without anysupervisiononspeakerlable.Theirexperimentsho wedthattheVQVClearnsameaningful embedding space without any supervision. Then they eve performed VC to unseen speakers with only one utterance, and subjective evaluations showed good results in terms of similar to target speakers[2].

III. ALGORITHM

1.1 Multi-Speaker Generative Modeling to VoiceCloning

Consider multi-speaker generative model, $f(t_{i,j}, s_i; W, e_{si})$, which takes text input $t_{i,j}$ and a speaker samples s_i . Trainable parameters are parameterized by W, and e_{si}. The latter s_i corresponding to the trainable speaker embedding. W and e_{si} are optimized by minimizing L (loss function) that penalizes the difference between generated and ground-truth audios (eg.: a regression loss for spectrograms):

$$\min_{W,\mathbf{e}} \mathbb{E}_{\substack{s_i \sim \mathcal{S}, \\ (\mathbf{t}_{i,j},\mathbf{a}_{i,j}) \sim \mathcal{T}_{s_i}}} \{L\left(f(\mathbf{t}_{i,j}, s_i; W, \mathbf{e}_{s_i}), \mathbf{a}_{i,j}\right)\}$$

When S means a set of speakers, T_{si} means a training set of text-audio pairs for speaker s_i, $a_{i,j}$ means the ground-truth audio for $t_{i,j}$ of speaker s_i . The expectation is guessed over text-audio pairsof each training speakers. □□ and □□ are used to designate the trained parameters□mbeddings. The speaker embeddings have been shown to effectively capture speaker characteristics with low-dimensional vectors. Generative loss, discriminatory features (e.g., gender and verbal communication) are seen in the speaker embeddingspace.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 132-135 www.ijaem.net ISSN: 2395-5252

For cloning voice, we take out the speaker characteristics of an unknown speaker s_k from a set of cloning audios A_{sk}, and generate an audio given any text for that speaker. Speech naturalness and speaker similarity are the two-performance metrics for the generated audio[1].

1.2 Speaker adaptation

The idea is to fine-tune a trained multi-speaker model for an unseen speaker using audio-text pairs. It can be applied to either the speaker embedding or the entire model. For embedding-only adaptation, we have the following:

$$\min_{\mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j},\mathbf{a}_{k,j})\sim\mathcal{T}_{s_k}} \left\{ L\left(f(\mathbf{t}_{k,j}, s_k; \widehat{W}, \mathbf{e}_{s_k}\right.\right.$$

Where T_{sk} is a set of text-audio pairs for the target speaker s_k. For whole model adaptation, we have the following:

The entire model provides more degrees of freedom for speaker adaptation but, its optimization is difficult for small amount of cloning data. To avoid overfitting early stopping is required[1].

1.3 Speaker encoding

In this case they suggested a speaker encoding method to directly measure the embedding of the speaker from the audio samples of the unseen speaker. Such type of model doesn't require any fine-tuning during voice cloning. so, the same model can be used for all unseen speakers. The speaker encoder, $g(A_{sk}; \Theta)$, takes a set of cloning audio samples A_{sk} and measures e_{sk} for speaker s_k. The model is parametrized by Θ [1]:

$$\min_{W,\boldsymbol{\Theta}} \mathbb{E}_{\substack{s_i\sim\mathcal{S},\\ (\mathbf{t}_{i,j},\mathbf{a}_{i,j})\sim\mathcal{T}_{s_i}}} \left\{ L\left(f(\mathbf{t}_{i,j}, s_i; W, g(\mathcal{A}_{s_i}; \boldsymbol{\Theta})), \mathbf{a}_{i,j}\right)\right\}.$$

$$\min_{W,\mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j},\mathbf{a}_{k,j})\sim\mathcal{T}_{s_k}} \left\{ L\left(f(\mathbf{t}_{k,j}, s_k; W, \mathbf{e}_{s_k}), \mathbf{a}_{k,j}\right)\right\}.$$

1.4 Vector Quantization

Vector quantization is a method of constraining an input from a continuous or otherwise large set of values by using a shared codebook. It also helps the decoder to do reconstruction with the quantized vectors. In signal domain let X = { $x_1, x_2, \ldots x_T$} where T denotes the sequence of acoustic features. An encoder is applied to transform X into a sequence of representation E = { $e_1, e_2, \ldots e_T$ }. Each value of the E set is quantized into a sequence of codes Q = { $q_1, q_2, \ldots q_T$ }.

The quantization function can be written as :

$Q_t = \arg\min (\|e_t - q\|_4)\ q \in QV$

the function takes the encoded e_t and selects the closest q from the codebook based on similar distances. A decoder is employed to reconstruct the correct input features which might have been lot with input Q. The decoder gives the output in the form of $\hat{\mathbf{X}}=\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_t\}$[2].

1.5 Contrastive predictive coding

Contrastive predictive coding learns self-supervised representations by predicting the future in the latent space. This model encourages the model to capture time-variant information while discarding time-invariant features. Contrastive predictive coding loss is contrastive. With the contrastive loss, the prediction is achieved by minimizing the dot product provided by an encoder. In CPC the sequence from the encoder is passed to current network to produce context.

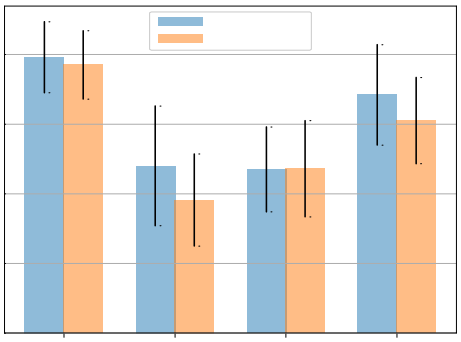
International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 2 Feb 2022, pp: 132-135 www.ijaem.net ISSN: 2395-5252

C = { c1,c2,…..c4 }.The main function of the model is to predict the future patterns given the current input [3], [5].

IV. OBSERVATION

Similarity test were conducted with Mean Opinion Score. In this test subjects were presented with pairs of utterances. Subjects were asked to assign a score ranging from 1-5. 5) Same, absolutely sure 4) Slightly sure 3) Not sure 4) Slightly different 5) Different absolutely sure. They randomly selected 10 speakers from seen speakers and unseen speakers. MOS scoresare presented in the belowfigure.

Blue (seen speaker), Pitch (unseen speaker)



Thus, we can show that both speaker adaption and speaker encoding can achieve an MOS similartothebaseline.Wealsoobservethatdrawbacksoftrainingamulti-speakermodelusing a dataset of low-quality audios and limited speaker diversity. This test also showed us to complete VC to unseen speakers in one shot try [3],[5].

V. CONCLUSION

We have studied the different approaches for voice cloning like Speaker adaptation, speaker encoding, vector quantization, multi-speaker generative model, etc. From the learnings we found that both speaker adaptation and speaker encoding approaches can get good cloning quality even with a few audio sample, for naturalness both speaker adaptation & speaker encoding can achieve good results. On other hand, Speaker adaptation takes more time togive resultsbutthequalityofclonedaudioisbetterthanspeaker encoding.

REFERENCE

[1] Sercan Ö. Arık, Jitong Chen, Kainan Peng, Wei Ping, Yanqi Zhou. Neural Voice Cloning with a Few Samples. Baidu Research 1195 Bordeaux Dr. Sunnyvale, CA 94089.

[2] Da-Yi Wu, Hung-yi Lee. One-shot Voice Conversion by Vector Quantization.College of Electrical Engineering and Computer Science, National Taiwan University (x07922119,hungyilee)@ntu.edu.tv.

[3] Ju-chieh Chou, Cheng-ChiehYeh. To implement one shot voice cloning by separating speaker from sourse and targetspeaker.

[4] Paarth neekhara, Shehzeen Hussain. Expressive neural voicecloning.

[5] Giuseppe Ruggiero, Enrico Zovato. Voice cloning: A multi speaker TTS synthesis approach based on transferlearning.

[6] MerlinBlaauw,JordiBonada,andRyunosukeDaido.DataEfficientVoiceCloningFor Neural Singing Synthesis.

[7] Shilun Wang, Damian Borth. Towards High Quality Zero-Shot VoiceConversion.

[8] Hieu-ThiLuongandJunichiiYamagishi.Nautilus:A VersatileVoiceCloningSystem.

[9] Kaushik Daspute, Hemang Pandit. Real Time VoiceCloning.

[10] Yen-Hao Chen, Da-vi wu. Again-VC A one shot voice conversion using activation guidance and AdalN.