Medicon Engineering Themes
Volume 1 Issue 1 September 2021
Review Article

DeepGen Network based Voice Conversion

Sheena Christabel Pravin[1*], M Palanivelan[2], S Saravanan[3]

[1]*Assistant Professor, Department of ECE, Research Scholar (Anna University), Rajalakshmi Engineering College, Chennai, India*

[2]*Professor, Rajalakshmi Engineering College, Chennai*

[3]*Student, Rajalakshmi Engineering College, Chennai*

***Corresponding Author:** Sheena Christabel Pravin, Assistant Professor, Department of ECE, Research Scholar (Anna University), Rajalakshmi Engineering College, Chennai, India.

Received: August 18, 2021; **Published:** September 30, 2021

Abstract

A DeepGen network is proposed for voice conversion, which is the process of modification of a speech utterance by a source orator to that of a target orator, preserving the linguistic contents. Automatic dubbing is an application of speech processing which facilitates the modelling of the large variability of pitch. The proposed DeepGen network automizes voice conversion by taking blocks of consecutive frame-wise linguistic and fundamental frequency features. This block-wise approach models temporal dependencies within the features of the input block. Voice conversion, which is also called as voice cloning has always required significant amount of recorded speech but the proposed DeepGen network intends to convert one voice to another using a relatively smaller set of speech samples by bootstrapping samples from a larger speech dataset. The proposed generative model is a variant of the convolutional generative model with encoder and decoder blocks that bring into line the hidden structures of the feature spaces from the source and target voices on a two-stage training process. An attractive and efficient voice conversion is thus obtained in the real-world scenario using the proposed DeepGen network. The novelty of the work lies in the induction of a deep learning generative network for cloning voice in Tamil language.

Keywords: DeepGen network; Voice Conversion; Generative Network; Convolutional generative model; Feature spaces

Introduction

Voice is one of the natural predispositions of man which gets distorted with age. Voice conversion is an area of speech processing that deals with the conversion of the perceived speaker identity. In other words, the speech signal uttered by the source speaker, is modified to sound as if it was spoken by a second speaker, referred to as the target speaker. Voices have various patterns and are used in many fields by using them. Especially, in entertainment programs voice modulation is used as a tool to give fun and laughter. And in the news and current programs, voice modulation is used as an identification tool of the emergent person. It is also widely used for diagnosis and diagnosis, such as pronunciation analysis and language disorder diagnosis. As such, voice signals are usefully used in various fields. However, crime reflects society. Voice conversion is gaining prevalence across various fields in applications ranging from audio book narration in a familiar voice for children to a personal digital assistant that can render speech in a desired voice. The most obvious use of voice conversion is text-to-speech synthesis where voice conversion techniques can be used for creating new and personalized voices in a cost-efficient manner. Other potential applications include security related usage, voice pathology, voice restoration, games and other entertainment applications as well. Yet other possible applications could be speech-to-speech translation and dubbing of television programs.

Speech conveys a variety of information that can be categorized, for example, into linguistic and nonlinguistic information. Before diving deeper into different aspects of voice conversion, it is essential to understand the factors that determine the perceived speaker identity. Linguistic information has not traditionally been considered in the existing voice conversion systems but is of high interest

in the field of speech recognition. Even though some hints of speaker identity exist on the linguistic level, non-linguistic information is more clearly linked to speaker individuality.

With the recent improvements in computation power and high scale datasets, many interesting studies have been presented based on discriminative models such as Convolutional Neural Network (CNN) [22] and Recurrent Neural Network (RNN) [27] architectures for various classification problems. These models have achieved current state-of-the-art results in almost all applications of computer vision but not efficient sampling out-of-data, understanding of data distribution. By pioneers of the deep learning community, generative adversarial training is defined as the most exciting topic of computer vision field nowadays. With the influence of these views and potential usages of generative models, many kinds of researches were conducted using generative models especially Generative Adversarial Network (GAN) and Autoencoder (AE) based models with an increasing trend [24]. Apart from CNN, RNN based discriminative models, which take consideration of conditional probability and work well for representation learning but not efficient to predict out-of-samples, out-of-samples can be generated to reject or sample with generative models. For predicting out-of-samples, generative models based on the joint probability of input pairs will be meaningful. Nowadays, adversarial networks attract the attention of researchers through the discovery of adversarial examples and their effects on neural networks. Adversarial examples are obtained by manipulating original images via perturbations. These manipulations could not be seen easily on images but cause different predictions. Adversarial examples may not be commonly seen in practice but adversarial trained networks will be more robust and will be performing well at the same time. Until now, unsupervised autoencoder and adversarial learning based generative models are designed for generating synthetic contents. Deep generative models are handled to clarify recent trend in the deep learning society. The generative models are categorized into unsupervised fundamental models, AE based models, autoregressive models, GAN based models and AE-GAN hybrid models to associate generative models easily. Conventional approaches using Gaussian Mixture Model [2] worked well but it was outperformed by Artificial Neural Networks [3]. The usage of Redundant Convolutional neural network [4] model for voice conversion helps to achieve the exact voice of the target speaker. Generative modelling has finally controlled the idiosyncrasies of the human voice.

The paper is organized as follows: section II briefs on the related work, section 3 discusses on the speech corpus used in the current study followed by data pre-processing while section IV introduces the proposed generative network. Section V elaborates on the results of experimentation with the proposed model while section VI concludes the paper with a note on the possible future work in voice conversion.

Related Work

Voice conversion was endeavoured by quite a few researchers in the recent days. Yist Y. Lin et al. aimed to convert voice from any to any speaker even unseen during training which is more attractive in real world scenarios. They use FragmentVC, a parallel data free Artificial Neural Network based approach for any-to-any voice conversion with an encoder-decoder architecture. FragmentVC consists of a source encoder, a target encoder and a decoder and uses the latent phonetic structure of the utterance of the source speaker. This model has one utterance from a source speaker and 10 utterances from a target speaker. Any-to-Any voice conversion is achieved by extracting and fusing voice fragments to construct the desired utterances with attention. FragmentVC is more flexible and easier to implement.

Ye Jia *et al.* proposed a deep learning network that consisted of three independently trained components viz. a speaker encoder network, a sequence-to-sequence synthesis network and an auto-regressive Wave-Net based vocoder network. This model is able to generate realistic speech from fictitious speakers that are dissimilar from the training set, implying that the model has learned to utilize a realistic representation of the space of speaker variation. This network is trained on non-transcribed speech containing reverberation and background noise from a large number of speakers, trained directly from text-audio pairs without depending on hand-crafted intermediate representations. The network is reported to be highly scalable and promisingly accurate for speaker veri-

fication. The drawback of the models is its inability to transfer accents and also not able to completely isolate the speaker voice from the prosody of the reference audio.

Ali Bou Nassif *et al.* proposed a model founded on Convolutional neural network (CNN) and Recurrent Neural Networks (RNN). A Convolutional Neural Network (CNN) is used mainly for image processing, classification, segmentation and also for other auto correlated data. RNNs are the state-of-the-art models for sequential data. These networks are considered as a discriminative architecture where every convolutional layer and a pooling layer are stacked on top of each other. The weight sharing together with properly chosen pooling schemes results in various properties of CNN that has shown its best in image recognition and computer vision tasks. Yutian Chen *et al.* proposed a meta-learning approach where the model has two types of parameters such as task-dependent parameters and task-independent parameters. This has a large model with shared parameters to capture the generic process of mapping text to speech to deploy the wavenet model. Joon Son Chung *et al.* proposed a system that make use of two key contributions namely a very large-scale audio-visual speaker recognition dataset collected from open-source media and a convolutional neural network model. In this system, they present a deep CNN based neural speaker embedding system named VGGVox, trained to map voice spectrograms to a compact Euclidean space where distances directly correspond to a measure of speaker utility. ResNet-34 and ResNet-50 architectures were used for the spectrogram input. The Voxceleb2 dataset was used which is several times larger than any speaker recognition dataset. Srinivas Desai et al. proposed a voice conversion model using Artificial Neural Networks (ANN) and Gaussian Mixture Model (GMM); ANN to transform the source speaker formants to target speaker formants and the GMM for mapping of source to target speaker space.The system was trained on CMU ARCTIC databases consisting of 7 speakers and each speaker has recorded a set of 1132 phonetically balanced utterances. Artificial Neural Networks (ANN) based spectral transformation yields better results both in objective and subjective evaluation than that of Gaussian Mixture Model (GMM) with Maximum Likelihood Parameter Generation (MLPG).

Speech Corpus and Data Preprocessing

In this work, the Librispeech audiobook corpus was used as the training speech corpus. LibriSpeech dataset consists of 2302 speakers from the train speaker subsets and approximately 500 hours of utterances, sampled at a frequency of 16 kHz. Proprietary speech corpus consists of 10 American English speakers and approximately 300 hours of utterances at frequency 16 kHz. The utterances were generated with 5 and 10 minutes of training data per speaker from LibriSpeech and VCTK. Also, the crowdsourced high-quality Tamil multi-speaker speech dataset provided by Google Research Dataset was utilized. The dataset contains approximately 10,500 speech samples each with a 4 second recording duration. The speech samples were re-sampled to 16 KHz after which their spectrograms were plotted using the librosa library in Python 3.6. For the output, approximately 8250 samples have been recorded among which 4327 has been created using audio data augmentation methods includes harmonic-percussive source separation function, random shifting, shuffling, addition to low-power noise.

The speech samples downloaded from the Google Dataset [2] was sampled to 16000 Hz. Each sample has approximately 4 to 7 sec of audio. The data that was used has a male: female ratio of 85:15. First, the audio samples is loaded by using librosa library. After that, the spectrogram was created by sampling audio using the parameters given in Table 1.

Parameter	***Value***
Length of Fast Fourier Transform	512
Spectrogram dimension	260
Hop length	128

Table 1: Parameters Passed to the Model.

Proposed Generative Network

Generative models can be used to change the speaking style of the source speaker to that of a target speaker. The proposed model's architecture is described in the following section.

Proposed Model Architecture

The proposed DeepGen architecture is given in Fig.1 and Fig. 2. The model consists of a source encoder, a target encoder and a decoder. Librosa library was used to convert an audio file into data files and the same can be used for creating spectrogram of dimension 260x260 by two linear layers with ReLU activation, used as the input to the decoder. The mel-spectrogram generated from the audio of the target speaker is fed into the target encoder, which has six ReLU-activated Id-convolution layers, for extracting the voice. The decoder consists of extractors and smoothers along with linear projection and a Tacotron-2-styled PostNet [18], to predict the mel-spectrogram for the desired output. The head and hidden size of both extractors and smoothers is 512. The convolutional network in each Transformer layer is placed in feed-forward layers for high correlation among adjacent features in speech.

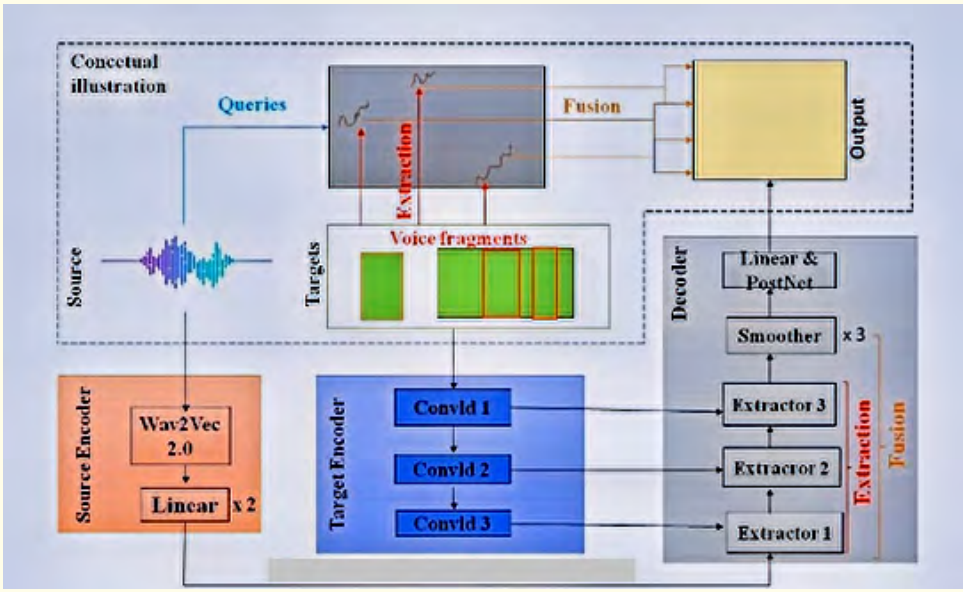

Figure 1: The overall architecture of the proposed Model.

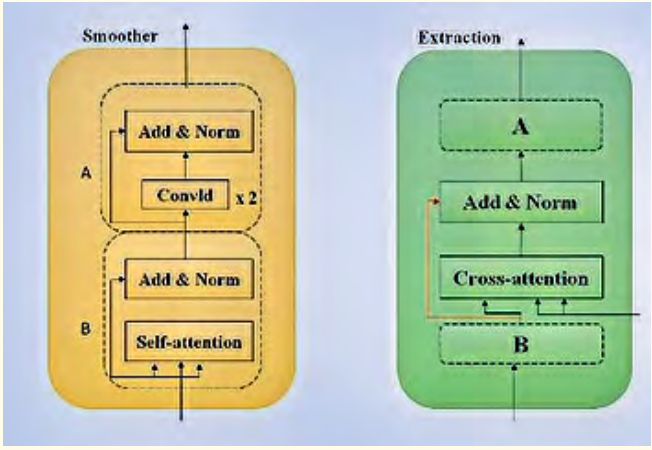

Figure 2: The model architecture of the extractor and the smoother.

The extractors are based on the structure of source speaker utterance by cross-attention to extract fine-grained voice fragments from target speaker utterances. Extractor 1 is used to construct highest-level phonetic structure of the output utterance based on the source representations. The Conv1D of the target encoder produces the most abstractive spectral information. So, the extractor depends on the output of the third convolutional layer Conv1D. Extractor 3 is to offer only slight modifications or minor adjustments in the spectrograms obtained from Conv1D 1'st layer of the target encoder. The smoothers take the output of the extractor stack to further smooth the output utterances.

Training Phase

A two-stage training scheme was adopted to train the proposed model. In the first stage, a single utterance from source speaker is fed as input to source and target encoder and set the goal as to reconstruct the Mel-spectrogram of the utterance. The model has to align the hidden structures between the source encoder and the target encoder by extracting and fusing the voice fragments. The linear layers in the source encoder provides conversion between the two different feature spaces. It was found that if the target utterance was different from the source utterance, the model was able to extract voice fragments properly and the output voice was similar to the target speaker at this stage.

In the second stage, the spectrograms of utterances were concatenated and into the target encoder. The goal is to reconstruct the spectrogram of the source utterance. In order to preserve the already well-trained attention modules, the learning rates of the source encoder was reduced. The learning rate of the source encoder, target encoder and the extractors are reduced 100 times in this stage, while the other parameters remained unchanged.

Results and Discussions

The entire model was training with 10273 samples of source speaker and 8670 samples of target speaker. All the utterances are sampled to 16KHz with the parameters mentioned in Section III. The model was optimized with the Adam Optimizer with the learning rate of 10^{-4}, β_2 =0.999, = 10^{-8} and weight decay =0.01 at 8000 epochs. The original spectrogram and the DeepGen network generated spectrogram is displayed in Fig.3 (a) and (b) respectively.

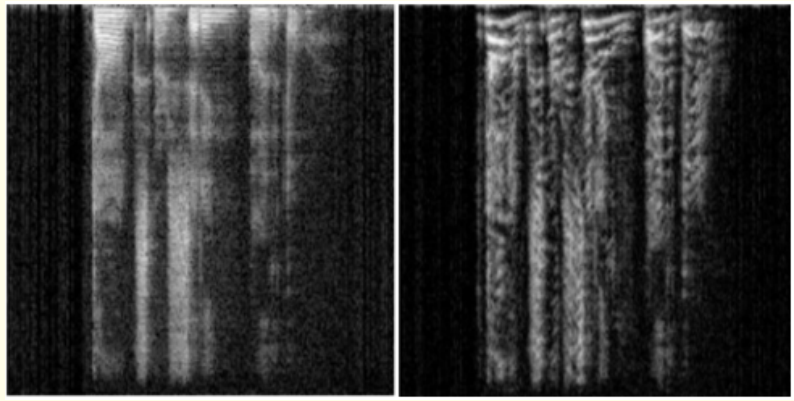

Figure 3: (a) Original audio spectrogram (b) Generated audio spectrogram.

The training and modulation accuracy of the proposed model is presented in Table 2. In this model accuracy column, the metrics used for finding the similarity between two voices is Euclidean Distance between the generated voice and target voice. It has been found that number of steps below 8000 are seems to be underfitting and epochs above 10000 are seems to be overfitting. So, either 8000 and 10000 steps may be chosen for the training purpose. Among these two, 10000 steps have been chosen as the optimum step for training the model.

S. No	*Steps*	*Training time (in hour)*	*Modulation Accuracy*
1	2000	1	30.40%
2	4000	2.5	45.80%
3	8000	5	68.70%
4	10000	7.4	87.20%
5	16000	16	98.30%

Table 2: Training and Modulation Accuracy.

The similarity index and absolute category rating on a scale of 1-4 with the following representation: 1- not good, 2-good, 3- perceptible, 4- more perceptible were recorded for testing the model on speech from various speaker to a particular target voice. The gender, age and similarity index of the target voice with the source voice in presented in Table 3.

Gender	*Age*	*Similarity to target speaker*	*Absolute Category Rating*
Male	32	85%	4
Male	20	80%	4
Female	47	79%	3
Female	17	86%	3

Table 3: Absolute Category rating of our model.

The similarity percentage has been calculated based on the loss parameter used in this model. The similarity index was observed to be high for male voices which were transformed with high accuracy preserving the speaker characteristics. The overall evaluation of the model proves it to transform the source speech to target speech with perceptible change.

Conclusion and Future Work

In this paper, Redundant Neural Network with 6 blocks of RES-Net is proposed for the style transfer of source speaker's utterance to that of the target speaker's voice. The model performed well on subjective evaluation measurement of absolute category rating. Also, various spectrogram dimensions with different hop length were experimentally tried by which the 4 seconds short period samples were observed to work well with 260x260-dimension spectrogram.

In future, the proposed generative model is aspired to be trained on multi-language speech samples for language independent voice conversion. Also, different optimizers will be studied for their efficacy in our future endeavours.

References

1. Emily L Denton., et al. "Deep generative image models using a laplacian pyramid of adversarial networks". In Advances in neural information processing systems 1 (2015): 1486-1494.
2. Y Stylianou., et al. "Continuous probabilistic transform for voice conversion". IEEE Transactions on Speech and Audio Processing 6.2 (1998): 131-142.
3. S Desai., et al. "Voice conversion using artificial neural networks". IEEE International Conference on Acoustics, Speech and Signal Processing (2009): 3893-3896.
4. Ju chieh Chou., et al. "multi-target voice conversion without parallel data by adversarially learning disentangled audio representations". in Proc. Interspeech (2018): 501-505.
5. G Hinton., et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". Signal Processing Magazine, IEEE 29.6 (2012): 82-97.
6. S Desai., et al. "Spectral mapping using artificial neural networks for voice conversion". Audio, Speech, and Language Processing, IEEE Transactions on 18.5 (2010):954-964.
7. A van den Oord S., et al. "Wavenet: A generative model for raw audio". in SSW 1 (2016).
8. S Desai., et al. "Voice conversion using artificial neural net-works". in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (2009): 3893-3896.
9. Y Xie., et al. ''Deep learning for natural language processing". Handbook of Statistics. Amsterdam, The Netherlands: Elsevier (2018).
10. J Shen., et al. "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions". in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018): 4779-4783
11. Wen-Chin Huang., et al. "Voice transformer net- work: Sequence-to-sequence voice conversion using trans-former with text-to-speech pretraining" (2020).
12. L Deng., et al., ''Recent advances in deep learning for speech research at Microsoft''. in Proc. IEEE Int. Conf. Acoust., Speech Signal Process (2013): 8604-8608.
13. T Salimans., et al. "Improved techniques for training gans". Advances in neural information processing systems (2016): 2234-2242.
14. C Donahue., et al. "Adversarial audio synthesis". International Conference on Learning Representations (2019).
15. Georg Heigold., et al. "End-to-end text-dependent speaker verification". In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference (2016): 5115-5119.
16. Sercan O Arik., et al. "Neural voice cloning with a few samples". arXiv preprint arXiv:1802.06006, 2018.

17. Ju chieh Chou and Hung-Yi Lee. "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization". in Proc. Interspeech (2019): 664-668.
18. G Hinton., et al. ''Deep neural networks for acoustic modeling in speechrecognition: The shared views of four research groups''. IEEE SignalProcess. Mag 29.6 (2012): 82-97.
19. Yutian Chen., et al. "Sample efficient adaptive text-to-speech". arXiv preprint (2018).
20. IMA Shahin. ''Gender-dependent emotion recognition based on HMMs and SPHMMs". Int. J. Speech Technol 16.2 (2013): 133-141.
21. Jose Sotelo., et al. Char2Wav: "End-to-end speech synthesis". In Proc. International Conference on Learning Representations, ICLR (2017).
22. Tomi Kinnunen and Haizhou Li. "An overview of text-independent speaker recognition: From features to supervectors". Speech communication 52.1 (2010): 12-40.
23. Ehsan Variani., et al. "Deep neural networks for small footprint text-dependent speaker verification". In Acoustics, Speech and Signal Processing (2014).
24. Kyunghyun Cho and Yoshua Bengio. "Attention-based models for speech recognition". In Advances in Neural Information.

Volume 1 Issue 1 September 2021