**UtilitasMathematica**

**ISSN 0315-3681 Volume 122 (2), 2025**

# Leveraging Speaker Embeddings from Speaker Verification for Controllable Multispeaker Text-to-Speech

**Nilam Thakkar**
[1]PHD Scholar, Indus University, Ahmedabad, Gujarat, India.
[2]Assistant Professor, LDRP-ITR, Gandhinagar, Gujarat, India.

nilamthakkar.21.rs@indusuni.ac.in

**Shruti Yagnik**
[3]Associate Professor, Indus University, Ahmedabad, Gujarat, India.

it.hod@indusuni.ac.in

**Tripti Sharma**
[4]Professor, MSIT, Delhi, India.

tripti_sharma@msit.in

## Abstract

Our hybrid system is compatible with Librosa library, Gaussian Mixture Model, and text-to-speech (TTS) technology. Neural networks are then used by TTS to produce audio speech that mimics the sounds of several speakers, including those that are excluded from the training set. The system incorporates three separately trained components: (1) With a small clip of the target speaker audio, the pre-trained encoder can validate it after comparing with a stand alone separately stored dataset of thousands of speakers' high pitched vocal notes without transcripts can produce fixed-length embedding vectors; (2) TacotronII endorsed sequential model that, relies on the primary level speaker embedding, transforms text into mel-spectrograms; (3) An auto regressive 'WaveNet vocoder' converts Mel spectrograms to waveforms with are functions of time.We demonstrate how the discriminative pre-training of the speech encoder on large-scale speaker diversity conveys important knowledge about speaker variability to the multi-speaker TTS challenge, allowing high-quality synthesis even for unknown speakers. We measure the advantages of rich and heterogeneous speaker datasets for enhanced generalization. Progressively, the new sounds generated with the aid of embedding of a random/ variable speaker can effectively generate new sounds that are different from the training set, suggesting that the model has picked up strong speaker representations.To prevent undue similarities, alternate wording and structure are used while preserving the essential factual data. Our recently added custom Librosa layers extract necessary features, which is helpful to

improve the efficiency of the particular feature, and our freshly added custom GMM layers eliminate noise from the raw audios by removing noisy features.

## 1. Introduction

The aim of this research is to create a text-to-speech (TTS) system that can efficiently process large amounts of data while generating speech that sounds natural for several speakers. In particular, we examine a zero-shot learning situation in which no model parameter updates are made and new speech that nearly perfectly mimics the voice of the target speaker is synthesised from a little piece of non-transcribed audio clip. Potential accessibility benefits of such systems include the restoration of lost speech capabilities for users who are unable to supply a significant amount of new training data. More realistic speech generation from text in low-resource environments, or more natural cross-language voice transfers for speech translation, are potential applications. But it's important to recognize the dangers of abusing technology, including voice impersonation without permission. We make sure that synthesized sounds can be easily distinguished from actual human voices in order to allay such worries in accordance with concepts like [1].

Many hours of voice transcriptions from each speaker are usually needed to provide a large amount of training data for high-quality multi-speaker TTS [15]. It is not feasible to obtain such large, high-quality datasets for multiple speakers. Our method trains a speaker-discriminative embedding network on unlabeled speech from hundreds of speakers independently, separating speaker modelling from speech synthesis. A TTS model is subsequently trained on a smaller dataset while being conditioned on this learned representation. Large multi-speaker corpora are not as necessary thanks to this decoupling, which makes independent data sources possible.The high pitched and noisy parts of the audio clip is used here from a heterogeneous population to train the speaker encoder on a speaker verification task: identifying if two different clips are form the one same speaker.

It has been emphasized that, the TTS model and speaker encoder must be trained on imbalanced speaker which is disconnected, sets and yet achieve good generalization. The encoder uses 18K speakers to sample the embedding prior and improve adaption, resulting in innovative speaker synthesis, while the TTS model is trained on 1.2K speakers. While utilizing different phrasing and structures, the essential concepts are maintained.

The text-to-speech (TTS) models which offers End-to-end training will fetch directly from text-audio pairings, without the need for manual feature engineering, has garnered significant attention [33, 45]. Tacotron 2 [29] combined WaveNet's [37] audio quality with prosody modeling from Tacotron [45] to achieve near-human speech naturalness by inverting mel spectrograms produced by an attention-based encoder-decoder. It was restricted to one speaker, though. A multi-speaker Tacotron that learnt low-dimensional embeddings for every training speaker was proposed by Gibiansky et al. [15]. A fully convolutional encoder-decoder from LibriSpeech [23] that scaled to 2,400+ speakers was used in Deep Voice 3 [25]. However, these systems can only mimic voices that they have observed in training. On the other hand, a unique buffer-based design that could produce hidden voices was proposed by VoiceLoop [35]. Tens of minutes of speech were needed from each speaker to be enrolled, nevertheless.

A short audio clip adapted with the transcript-free target speech has been made possible by recent work. [5] expanded upon Deep Voice 3 by contrasting the prediction of embeddings straight from spectrograms using a different network with fine-tuning the model containing embeddings on adaption data. With only one or two utterances, the later method achieved higher naturalness and was significantly more data-efficient. Because it did not involve significant backpropagation, it was also speedier. To prevent undue similarities, alternate terminology and structure are used while maintaining the main principles.

Comparatively, Nachmani et al [19] expanded VoiceLoop by predicting the embeddings using the encoding network fabricated for a target speaker encoding network. To make sure that embeddings match and are much closer to the remaining speakers, this was

trained in tandem with the synthesis model using a contrastive triplet loss. In order for the synthesized speech to encode to an embedding resembling the adaptive utterance, they additionally employed a cycle consistency loss. Target prosody was proved to be sent by a similar spectrogram encoder that did not suffer from the triplet loss [31]. Here, we show that accurate speaker characteristic transfer is made possible by training an encoder to identify between speakers.

Our approach is similar to the encoding models in [3, 19], but we employ a state-of-the-art end-to-end loss [43] on a network that was independently trained on speaker verification using a chunk of untranslated speech data set form thousands of speakers used a speaker-discriminative representation akin to that of [19], but all components were trained collaboratively. Rather, we investigate transfer learning using a speaker verification model that has been pretrained.

Transfer learning was also employed by Doddipatla et al. [13], who trained a TTS system using embeddings from a pre-trained speaker classifier. Our work is distinct in that it employs a speaker encoder that is not restricted to a closed speaker set, and it uses an end-to-end synthesis model without intermediary language elements. Furthermore, we compare training and performance speakers; we discover that zero-shot transfer requires thousands of speakers—much more than [13].

### 1.1 Audio Noise Removal Techniques

| Serial Number | Methods | Improved Work | Weak Points |
|---|---|---|---|
| 1 | Multi-stage Technique[55]. | This approach has low noise rates and can identify the noisy samples. | The cascade of noise detection processes in the suggested strategy results in increased time consumption. |

| | | | |
|---|---|---|---|
| 2 | SD-ROM algorithm[56]. | The impulse-noised audio signal is converted into a de-noised audio signal by this algorithm. | The sliding window's efficiency decreased and its input threshold values for audio signals increased as its size increased. |
| 3 | A stacked Long Short Term Memory (LSTM) mode[57]. | It performs well even with reduced background noise. | More precision is still needed for noise reduction. |
| 4 | Supervised machine learning models[58]. | Eliminate noise to increase the likelihood of obtaining improved accuracy. | This technique is applicable to raw music files too. |
| 5 | Computational Auditory Scene Analysis (CASA) – GMM-CNN based module[59]. | Higher performance in noisy voice signals is produced by this system. | Improvements must be made to the pitch estimate technique. |

**Table 1: Audio Noise Removal Techniques Comparison**

### 1.2 Audio Feature Extraction Techniques

| Serial | Methods | Improved | Weak Points |
|---|---|---|---|

| Number | | Work | |
|---|---|---|---|
| 1 | Global feature algorithm [49]. | Eliminate unnecessary information out of the features. | It is necessary to improve accuracy and speech recognition rate. |
| 2 | MFCC, 2D Convolutional splitting [50]. | Reduce the dimensionality of the temporal dimension. | The audio encoder based on 1D convolution performs better by splitting. |
| 3 | MFCC with Matlab Programming[51]. | More effective and more efficient speech signal processing. | It's necessary to improve voice recognition accuracy. |
| 4 | Model-based MIR algorithms[52]. | Retrieve local pulse and beat data. | Improvements are needed in music synchronization, chord recognition, and onset detection. |
| 5 | Filter bank-based and auto-regression-based algorithms [53]. | offers a uniform approach to the extraction of audio components. | Limited to the Mel Frequency Features (MFCC). |
| 6 | Librosa [54]. | Quick and precise extraction of spectral characteristics. | It is still possible to increase feature extraction speed more quickly. |

**Table 2: Audio Feature Extraction Techniques Comparison**

## 2. The hybrid model of poly-speaker speech synthesis

The research includes two models: (1) Gaussian Mixture Model [59], (2) Librosa [54], which allow the addition of custom layers by utilizing them, other three non-dependent neural networks which are trained, etc. (1) The speaker encoder which is of recurrent type, is adapted from [43], producing a fixed-length vector representation from a signal generated through speech; (2) a synthesizer (Sequesntial), which is adapted from [29],

forecasting a mel-spectrogram from grapheme/ phoneme inputs parameters on the speaker embedding; and (3) a WaveNet [37] vocoder of auto-regressive nature, which is depicted in Figure 1, is used to convert the spectrogram to the recognizable time domain audio samples.

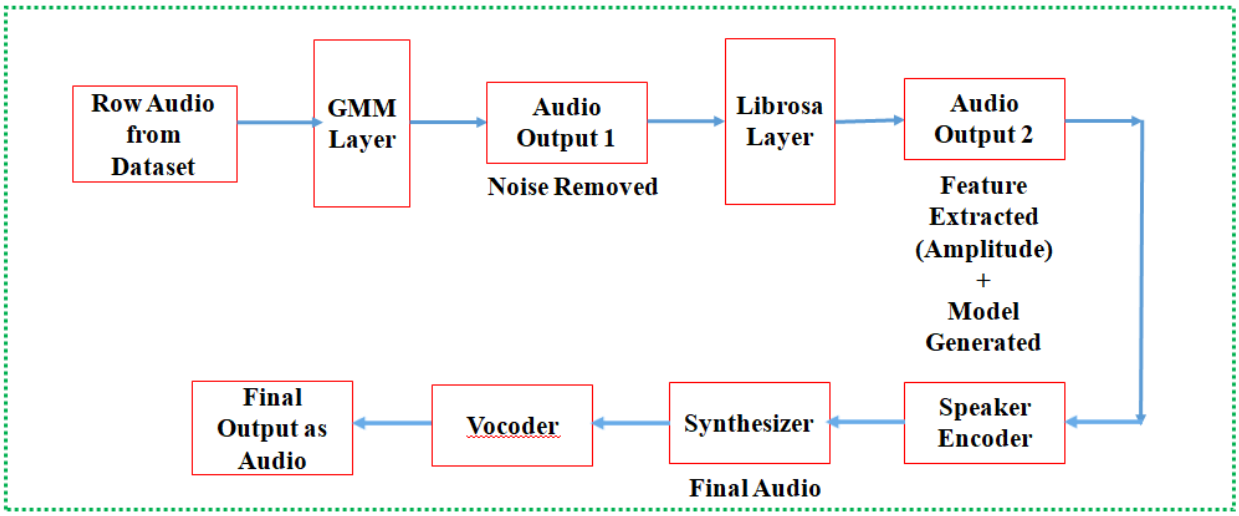



**Figure : 1 Proposed Method Flow Diagram**

### 2.1 Gaussian Mixture Model

The field of image classification and segmentation has shown the superior performance of Gaussian Mixture Model (GMM) in managing noisy data[65]. Gaussian Mixture Model is employed in our hybrid model out of all the audio noise removal strategies (Table 1) because it eliminates extraneous noisy characteristics [65] and improves speaker

identification [59]. After applying our unique GMM layers to the dataset, we saw that the resulting audio had less noise. We have utilized the audio file reading programme Fast Forward Moving Picture Expert Group (ffmpeg) to facilitate our work [66].

### 2.2 Librosa

We have utilized the Librosa package, which is available in Python, to extract features for our experiment. For the same purpose, the great majority of developers use it. Numerous variables, commonly referred to as features, are produced as a result of using librosa. The user then analyses these qualities in accordance with their demands [54]. Our system uses the Librosa library because it operates more quickly and yields accurate results than any other feature extraction technique (Table 2) [54]. Though there are other libraries out there that can extract more features, our suggested task only requires a small number of features to be extracted. Features including channels, sample width, sample rate, frame width, frame count, length, intensity, chroma feature, tempo, pitch, and amplitude have all been extracted using librosa. We identified an audio as an output that acquired our necessary extracted features, and we added our custom Librosa layers to the audio that we obtained as an output from the GMM layer application on raw audio.

As a result, we created a new model, which is shown in figure 1, and the output was audio. This audio will help with the synthesizing and encoding process.

### 2.3The encoder of Speaker

Imagine you're trying to mimic someone's voice. To do that really well, you need to understand the unique way they speak, not just the words they say. A speaker encoder acts like a super listener that can pick up on these special voice traits from a short clip, even if there's background noise or they're speaking a different language. It's like having a cheat sheet to their voice! This helps the system that creates new speech (like a text-to- speech program) sound more like the target person, regardless of what they're saying.

We develop the architecture from [43], which suggested a neural network for speaker verification that is extremely accurate and scalable. A fixed-dimensional embedding vector called a d-vector is obtained by mapping a series of log-mel spectrogram frames from a voice utterance of any length [39, 17].

To make sure the system can tell speakers apart really well, we trained it like a speaker identification pro. Instead of using written text (transcripts), we fed it short voice clips (1.6 seconds) with labels telling who spoke. This training helps the system push similar voices closer together in a special digital space, while keeping voices from different speakers far apart. Imagine this space like a playground - kids from the same class tend to play near each other, while kids from different classes are scattered around. This way, the system can easily recognize a familiar voice based on these learned distances.

Once we have these short voice clips, we put them through a series of special filters that turn the sounds into a visual representation, kind of like a fingerprint of the speaker's voice. This fingerprint is called a "log-mel spectrogram" and it has 40 channels to capture all the nuances. Then, information from this fingerprint is fed into a complex system with multiple layers, like a layered cake. Each layer learns a bit more about the unique characteristics of the speaker's voice. Finally, the system squeezes all this information into a single, powerful summary called an "embedding."

Now, when someone speaks into the system, it doesn't listen to the entire sentence at once. Instead, it breaks it down into smaller chunks of about 800 milliseconds (almost a second), with each chunk overlapping the next one by half. The system then analyzes each chunk independently using the same method as before to create its own embedding. To get a final embedding for the entire utterance (sentence), the system averages these individual chunk embeddings and refines them again. This way, even if the speaker hesitates or pauses in the middle, the system can still get a good overall picture of their voice.

Training on speaker discrimination yields an embedding suitable for conditioning synthesis on speaker identity, even though it is not explicitly optimized for synthesis-relevant speaker features.

### 2.4 Synthesizer

Building on the success of Tacotron 2 [29], a popular text-to-speech system, we wanted to give it the ability to speak in different voices! We borrowed some ideas from another research project [15], but with a twist. Here's how it works:

Imagine the system has two parts - an "encoder" that analyzes information and a "decoder" that creates something new based on that analysis. In our case, the encoder analyzes the text to understand what's being said. To make the speech sound like a specific person, we also feed the encoder a special code representing that speaker's voice (like a secret handshake!). This code is different for each speaker and is created by a separate "speaker encoder" we trained earlier (think of it as a voice fingerprint machine). Unlike the previous approach [15], we found that simply feeding this code directly to a special layer called the "attention layer" worked best.

We tested two versions of this system:

- **Version 1:** Uses the separate speaker encoder to create the voice codes.
- **Baseline Version:** Learns a codebook of voice codes for each speaker in the training data, similar to [15, 25].

To train our system, we fed it pairs of written text and the corresponding spoken audio we wanted it to imitate. We also converted the text into sounds (phonemes) to help it pronounce tricky words and names better. Here's the cool part: the speaker encoder we used was already trained, so it could analyze the target speech and create the perfect voice code for that specific speaker, all without needing any extra labels telling it who spoke! This "transfer learning" approach lets the system learn from existing data and apply that knowledge to new situations.

### 2.5 The Neural vocoder (a throbbing component)

This research emphasizes the use of sample-wise autoregressive Wave Net architecture from [37] as a vocoder to convert the time domain audio waveforms from the mel spectrograms generated using the synthesis network. The model's dilated convolution layers (Thirty nos.) are arranged according to the description found in reference [29]. The

speaker encoder output is not a direct condition of the vocoder network. The synthesizer's predicted mel spectrogram adequately encodes the relevant features for high-fidelity synthesis in many voices. Without using speaker conditioning directly, a multispeaker vocoder can be created using the training platform on speaker datasets array.

### 2.6 The Zero-shot speaker adaptation and Interference

One of the key strengths of our model is its ability to adapt to unseen speakers, even during the speech generation process (inference). This means the text-to-speech system can synthesize speech conditioned on **random, untranscribed audio**, without requiring that audio to directly match the text being spoken. This is achieved because speaker characteristics are **directly learned from the audio itself**, not solely from the training data with speaker labels.

This capability unlocks exciting possibilities. For instance, a few seconds of audio from a new speaker can be used to **condition the synthesis**, resulting in speech with comparable vocal qualities. This "zero-shot adaptation" allows the model to dynamically adjust its voice based on the speaker's unique characteristics, even for speakers not included in the training set. The effectiveness of this approach in generalizing to unseen speakers will be further evaluated in Section 3.

## 3. Experiments

### Dataset Details

Two publicly accessible datasets were used to appraise the neural networks for vocoding and speech synthesis. About 400 sentences, chosen from a newspaper, the rainbow passage, and an elicitation piece for the speech accent archive, are read aloud by each speaker. 110 native English speakers with primarily British accents provided 40+ hrs. of clear speech audio sampled at 48 kHz for the VCTK corpus [41]. We divided the data into triple subsets: validation (which included the same speakers as train), and test (which included 11 speakers held out from train and validation) and train. We also down sampled

the audio to 24000 Hz and removed leading and trailing silence parts from each recording, which drastically decreased the median utterance duration by 54.5% (i.e. from 3.3 seconds to 1.8 seconds.)

The dataset prescribed in LibriSpeech [23][67] is a corpus of nearly 1000 hours of 16kHz read English speech that was put together by Daniel Povey and Vassil Panayotov. The data has undergone meticulous segmentation and alignment, and it originates from read audio books from the LibriVox project. Although the data is generated via audio book narration, most speakers have American accents. However, the dialect of speech might range significantly even between different ways of uttering from the same speaker. In order to reduce the median utterance time from 14 seconds to 5 seconds, we re-segmented the recordings utilizing forced alignment with an automatic speech recognition model to detect pauses and separate the audio on silence.There are no punctuation marks in the text transcripts, just like in the reference data. The continuity between the train, validation, and test sets of speakers is lost.

**Preprocessing**

A considerable proportion of the recordings using Libri Speech in a clean corpus have detectable amounts of stationary interference and ambient background noise. To make it more conducive, we preprocessed the target spectrograms utilized for synthesis training using bespoke GMM layers [59][64]. The 10th percentile of energy inside each frequency band across the entire utterance was used to assess the background noise power spectrum. The speaker encoder model received the original noisy speech waveforms as inputs, unprocessed, and only the raw audios were subjected to this denoising method during training.

**Model Training**

For each of the two datasets, we trained distinct instances of the neural networks used for speech synthesis and vocoder. To account for the effects of pronunciation variations, we

used synthesis models trained on phoneme sequence inputs, which we obtained as an output from the produced model and which are shown in figure 1 in all of the subjective evaluations reported in this paper. We discovered that the vocoder model trained using ground truth mel-spectrograms recovered from the original audio produced high quality performance for the VCTK dataset, which comprises generally clean speech. But in order to get the best results for the noisier LibriSpeech dataset, we identified and have brought to surface that the vocoder must be trained using mel-spectrogram predictions made by the synthesizer.

A unique dataset of voice search queries including 36 million utterances with a median duration of 3.9 seconds, recorded from 18,000 English speakers in the United States, was used to train the speaker encoder network. Although there are no transcriptions in this collection, each audio contains a label indicating the anonymized speaker identification. It is never applied directly to the synthesis model training process.

Following preprocessing and model training, the graphs below compare the epochs shown in Figures 2, 3, 4, and 5 and show training accuracy, training loss, validation accuracy, and validation loss. Figure 6 presents a single graph containing all four options. The accuracy range displayed on the graphs is 0 to 100%, or 0 to 1. If accuracy exceeds 100%, data loss is indicated.

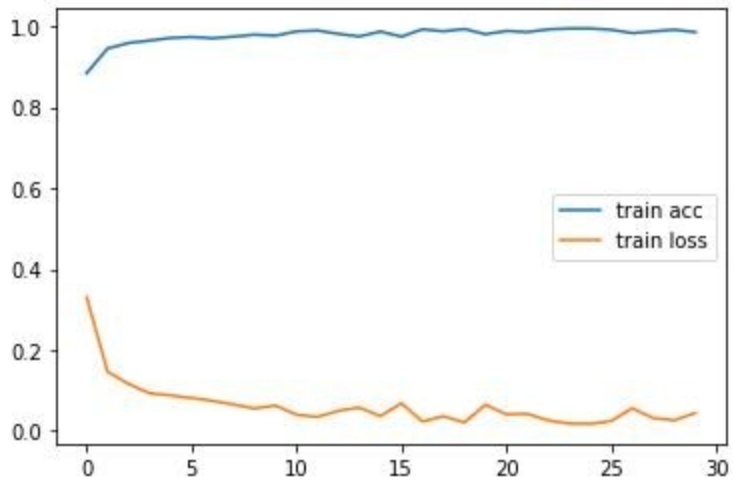


Figure : 2 Training Accuracy vs Epoch

Training Loss vs Epoch

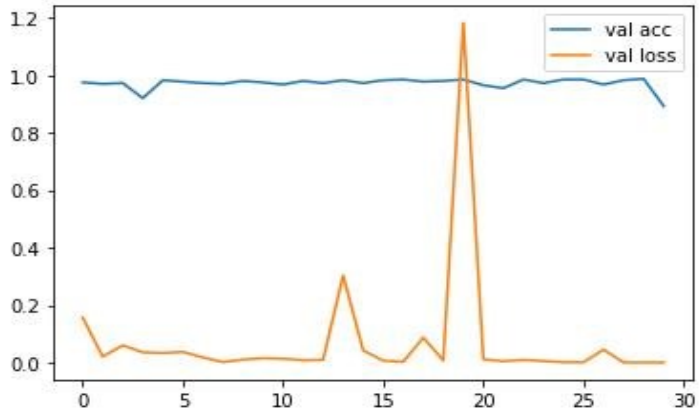



Figure : 3 Validation Accuracy vs Epoch

Validation Loss vs Epoch

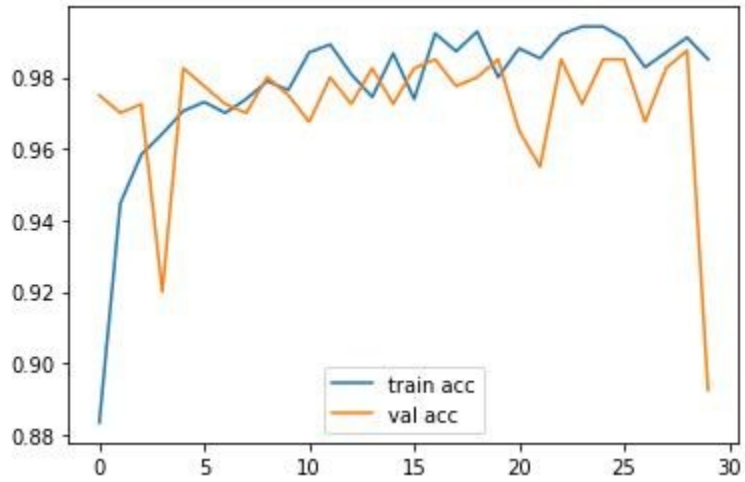



Figure : 4 Training Accuracy vs Epoch

Validation Accuracy vs Epoch

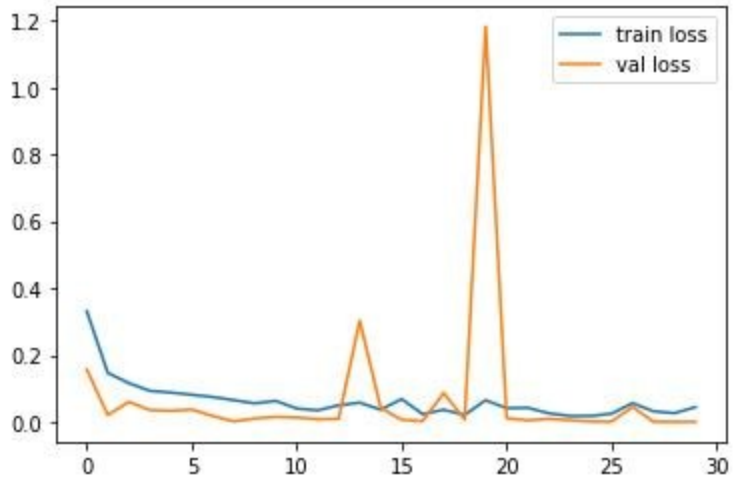



Figure : 5 Training Loss vs Epoch

Validation Loss vs Epoch

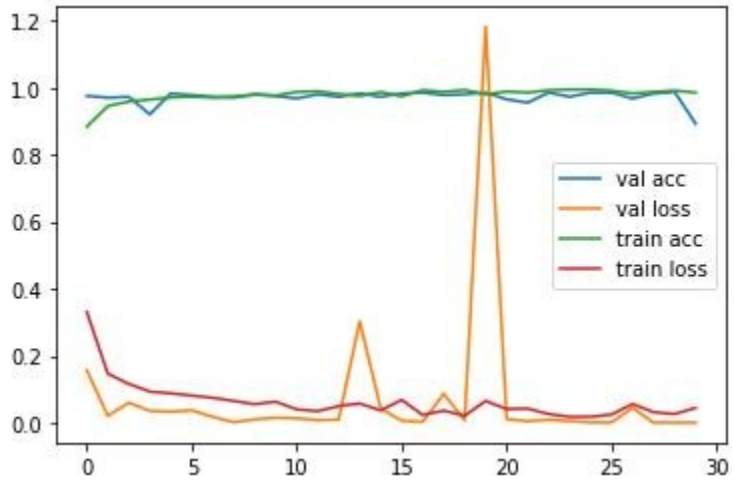



Figure : 6 Validation Accuracy vs Epoch

Validation Loss vs Epoch

Training Accuracy vs Epoch

Training Loss vs Epoch

We have evaluated an unheard audio with noisy qualities after training our model. As seen in the figures below:

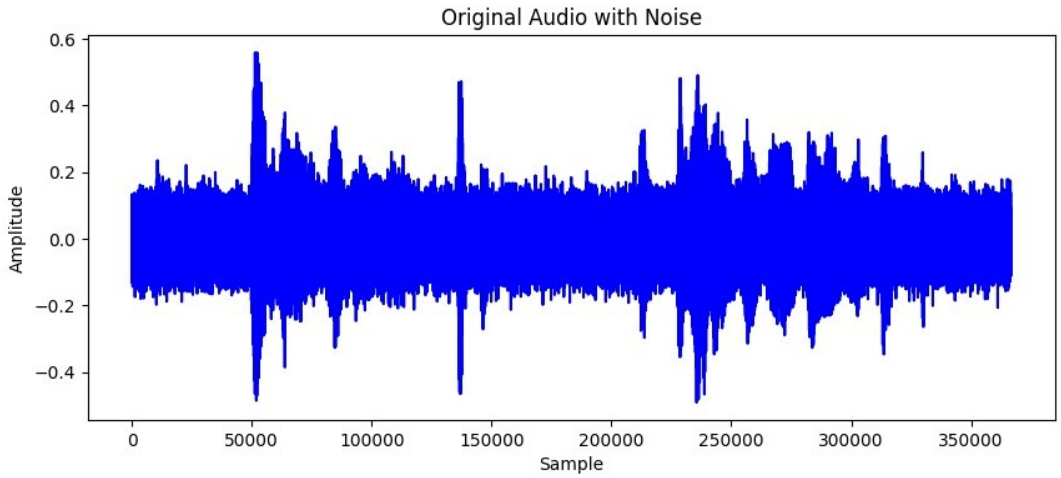



Figure : 7 Original audio with noise

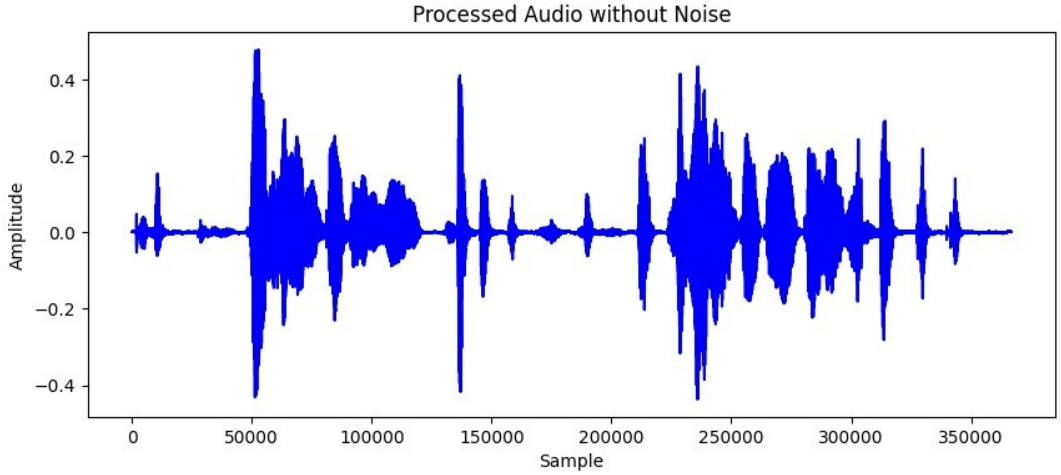



Figure : 8 Processed audio without noise

The results of an audio file's model training scenario before and after are displayed in Figures 7 and 8. It is evident that figure 8's amplitude vs. sample is less than figure 7's, indicating a more nuanced representation of the noise reduction.

**Evaluation Methodology**

To subjectively assess the overallsynthesized speech quality along two dimensions, namely naturalness and resemblance target speaker's actual dialect, the study is based on Mean Opinion Score (MOS) listening tests.[6] However, unlike[6], our suggested methodology produced objective MOS ratings. The MOS assessments are based on the 'Absolute Category Rating' paradigm [27], and results are given on a 5-point scale with increments of 0.5 points from 1 to 5. The librosa.segment_cross_similarity function from the librosa library in Python, which returns the similarity index of supplied audios, is what we used in our suggested model. Numerous similarity indices pertaining to a large number of these audio samples have been discovered, and ratings have been produced based on them.

### 3.1 Speechless natures

#### Evaluation of Speech Naturalness

We assessed and contrasted the artificial intelligence of synthetic speech generated by models trained on both LibriSpeech and VCTK. One hundred new phrases that were not part of the data curated for the training; so they were included in the test set. For each model, we evaluated the Seen (present during training) and Unseen (held-out from training) speaker groups. We used 10 speakers for each of LibriSpeech and 11 speakers for VCTK, both viewed and unseen. To determine the speaker embedding, we took a random sample of 1 - 5 seconds of speech from each speaker. All phrases had to besynthesized for every speaker, yielding around 1000 samples for each evaluation. A single rater rated each speech sample once, and models were assessed separately rather than in direct comparison.

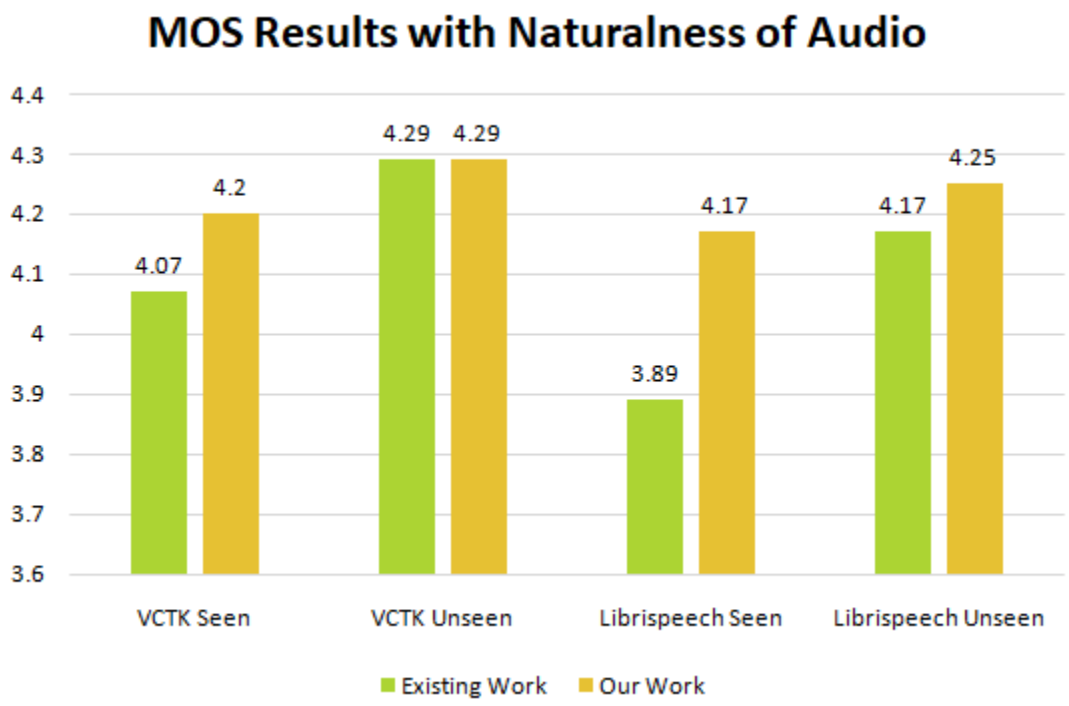
| System | VCTK found | VCTK absent | Librispeechfound | Librispeechabsent |
|---|---|---|---|---|
| **Ground truth** | 4.43 ± 0.05 | 4.49 ± 0.05 | 4.49 ± 0.05 | 4.42 ± 0.07 |
| **Embedding table** | 4.12 ± 0.06 | N/A | 3.90 ± 0.06 | N/A |

| | | | | |
|---|---|---|---|---|
| **Existing Proposed model** | 4.07 ± 0.06 | 4.29 ±00.06 | 3.89 ± 0.06 | 4.12 ± 0.05 |

Table 3: The speech Naturalness Mean Opinion Score Existing Work[6]

| System | VCTK found | VCTK absent | Libri speech found | Libri speech absent |
|---|---|---|---|---|
| **Ground truth** | 4.43 ± 0.05 | 4.49 ± 0.05 | 4.49 ± 0.05 | 4.42 ± 0.07 |
| **Embedding table** | 4.12 ± 0.06 | N/A | 3.90 ± 0.06 | N/A |
| **Our Proposed model** | 4.20 ± 0.06 | 4.29 ± 0.06 | 4.17 ± 0.06 | 4.25 ± 0.05 |

Table 4 :The speech Naturalness Mean Opinion Score proposed Work




Graph 1 :Speech Naturalness Mean Opinion Score (MOS) Comparison Graph

**Evaluation Results**

Results are shown in Tables 3 and 4 that compare our suggested model to baseline multispeaker models with essentially identical synthesizer architectures that use a speaker embedding lookup table, as in [3, 15, 25]. Across datasets, our model obtained ~4.2 MOS, with VCTK outperforming LibriSpeech by about 0.03 points for seen speakers. This is caused by two restrictions in LibriSpeech: (i) punctuation that makes it difficult for natural pausing, and (ii) increased background noise that is occasionally duplicated even with denoising.

Most notably, unseen speakers' audio quality on LibriSpeech either equaled or surpassed that of seen speakers, with differences of up to 0.2 points. This arises from certain randomly chosen reference utterances having unequal, non-neutral prosody. Synthesized prosody occasionally imitates the reference, as in [31], according to casual hearing. Due to LibriSpeech's more diversified prosody, this effect is more pronounced. To further detangle speaker identity and prosody, more work is required. Training on random reference/target pairings or adding a prosody encoder like [31, 47] may be helpful.

As depicted in graph1, we can see the comparison between existed work result[6] and proposed work result, this demonstrates that the proposed bars almost get the highest accuracy score of 5, or a greater naturalness score [60].

**Comparison to Baseline**

For both seen and unseen speakers, our model performed better than the baseline 0.1-0.2 MOS. This confirms that, even in cases where the speakers are visible during synthesizer training, pre-training the speaker encoder against the baseline's lookup table yields significant advantages. Furthermore, it validates that the encoder can generalize to new speakers.

**3.2 Tracing the similarity in Speaker**

We randomly picked a phrase from the two identical speakers and compared each synthesized sample with it to evaluate speaker resemblance. Raters were told to consider solely the speakers' similarities rather than the audio quality, grammar, or content.

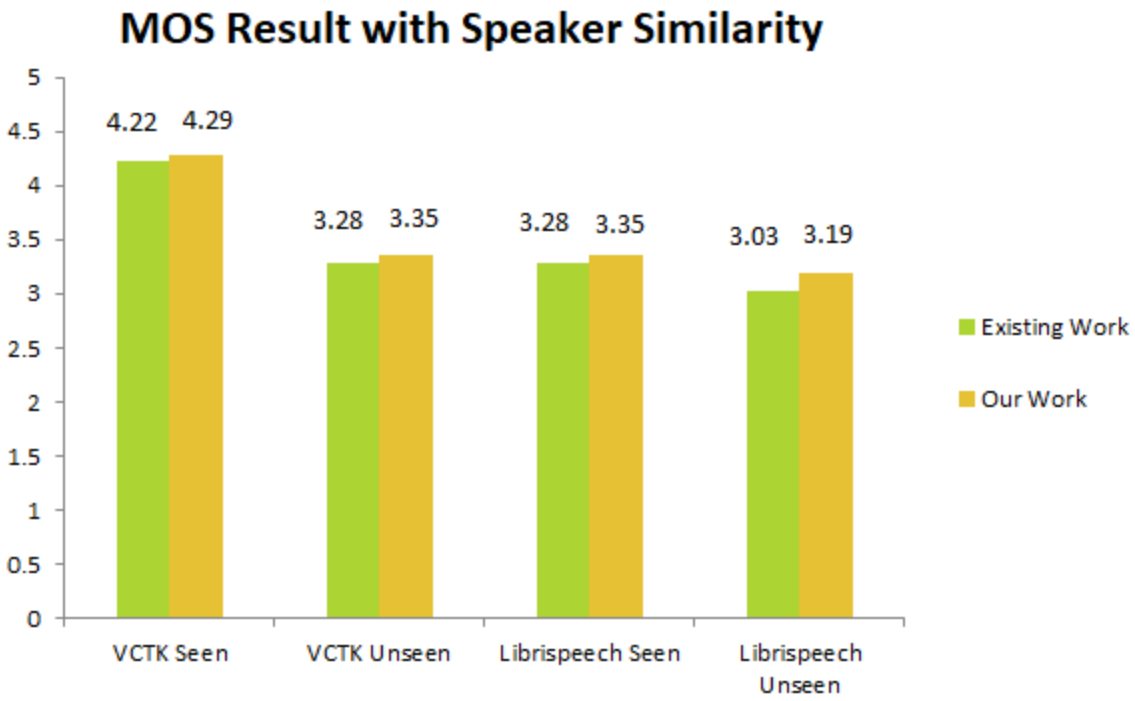
**Evaluation of Speaker Similarity**

| System | Speaker Set | VCTK | LibriSpeech |
|---|---|---|---|
| **Ground truth** | Same speaker | 4.67 ± 0.04 | 4.33 ± 0.08 |
| **Ground truth** | Same gender | 2.25 ± 0.07 | 1.83 ± 0.07 |
| **Ground truth** | Different gender | 1.15 ± 0.04 | 1.04 ± 0.03 |
| **Embedding table** | found | 4.17 ± 0.06 | 3.70 ± 0.08 |
| **Existing Proposed model** | found | 4.22 ± 0.06 | 3.28 ± 0.08 |
| **Existing Proposed model** | absent | 3.28 ± 0.07 | 3.03 ± 0.09 |

**Table 5 : Speaker Similarity Mean Opinion Score Existing Work[6]**

| System | Speaker Set | VCTK | LibriSpeech |
|---|---|---|---|
| **Ground truth** | **Same speaker** | **4.67 ± 0.04** | **4.33 ± 0.08** |
| **Ground truth** | **Same gender** | **2.25 ± 0.07** | **1.83 ± 0.07** |
| **Ground truth** | **Different gender** | **1.15 ± 0.04** | **1.04 ± 0.03** |
| **Embedding table** | **found** | **4.17 ± 0.06** | **3.70 ± 0.08** |
| **Our Proposed model** | **found** | **4.29 ± 0.06** | **3.35 ± 0.08** |
| **Our Proposed model** | **absent** | **3.35 ± 0.07** | **3.19 ± 0.09** |

**Table 6 : Speaker Similarity Mean Opinion Score Proposed Work**




**Graph 2: Speaker Similarity Mean Opinion Score (MOS) Comparison Graph**

**Evaluation Result**

- VCTK has higher overall similarity ratings than LibriSpeech, which is indicative of its cleaner nature. The higher VCTK ground truth baselines provide evidence for this.
- Using speaker embedding, our model outperformed the baseline for VCTK observed speakers. Unfortunately, our model's similarity was lower than that of the LibriSpeech baseline, most likely as a result of LibriSpeech's higher background noise and within-speaker variation.
- Compared to visible speakers, synthesized speech similarity for unseen speakers was lower across datasets. The 3.35 score on VCTK falls in the "moderately" to "very" similar range.
- The gender ranges of unseen speakers were captured by our model.

The comparison between the existing work result [6] and the proposed work result [60] is indicated using the graph 2, and here it is evident that the proposed work bars approach

greater speaker similarity for the parameters like voice accuracy, with 5 representing the best accuracy.

### 3.3 Method to verify the speaker

With the verification task of subjected speaker, we assessed the similarities obtained in the speaker as compared with the synthesized and ground truth for speakers who are not in database. This gives an indicator of how recognizable the synthesized voice is in relation to the intended speaker.

To conduct the evaluation, we utilized a dataset of 28 million utterances from 1,13,000 speakers to build an independent speaker encoder model. The model was trained on a different speaker population than the synthesis model encoder, but it has the same network architecture as that mentioned in Section 2.3. By employing a unique model for assessment, the metrics are guaranteed to transcend any peculiarities associated with a particular embedding space kept for the speaker.

To test how well our model performs with voices it's never encountered before (unseen speakers), we recruited a group of 21 speakers. Ten came from a popular speech database called LibriSpeech, and eleven came from another database called VCTK. For each speaker, we created 100 short test sentences the model would use to generate speech.
Here's how we measured how well the model captured each speaker's unique voice characteristics:

1. We took voice samples from all 21 speakers (like creating voice profiles).
2. We then paired each synthesized test sentence with a voice sample from **every** speaker in the group (think of it as a giant mix-and-match test).
3. Finally, we used a special technique called "speaker verification" to see if the model's synthesized voice could be mistaken for the real speaker's voice. The lower the error rate, the better the model performed at capturing that speaker's voice.

By running this test for each synthesized sentence, we ended up with a massive evaluation - between 22,000 and 24,000 trials for each model we tested! This comprehensive

approach ensures a robust assessment of the model's ability to generalize to unseen speakers.

We conducted a second evaluation using an expanded pool of 20 enrolled voices, which includes 10 real LibriSpeech speakers and 10 synthetic versions of those same speakers produced by our model, in order to measure the ability to distinguish between real and synthetic speech from the same target speaker. Our SV-EER on this more difficult 20 voice discrimination task was 2.86%. The model embedding are more like other synthetic samples of the same speaker, even though the synthetic speech mimics the target speaker. In summary, the suggested model produces speech that is recognizable as the target voice but is not yet confused with actual speech from the true speaker.

### 3.4The number of speaker encoder: training speakers

The representation quality believes that the speaker encoder learns is probably what allows our model to generalize successfully to new speakers. To investigate this, we used encoders trained on extra datasets with different speaker counts to assess synthesis performance:

1. LibriSpeech Other: 460+ hours of speech from 1000 speakers, disjoint from the clean training sets
2. VoxCeleb: 139,000 utterances from 1,210+ speakers
3. VoxCeleb2: 1.09 million utterances from 5,990+ speakers

The encoders used a simplified architecture with 64-dimensional projections, 64-dim speaker embedding, and 256-dimensional LSTM cells to prevent overfitting on the smaller VoxCeleb datasets.We used the same clean LibriSpeech data that was used for the synthesizer to train an encoder as a baseline. This matched condition, with our discriminative training strategy, is comparable to previous work. In terms of naturalness and resemblance, it fared somewhat better than the encoder that used the extra LibriSpeech Other set. However, both naturalness and similarity scores, as well as objective

speaker verification EERs, greatly improved as the number of speakers in the encoder training data increased to the bigger VoxCeleb sets.

This demonstrates a significant benefit of our methodology: gathering data to train the speaker encoder requires just untranscribed audio, even in cases when the quality is low, and is far less expensive than complete TTS training.

By the integration of a speaker encoder trained on a range of discovered data with a synthesizer trained on clean datasets such as LibriSpeech, we have demonstrated high-quality synthesis of curated voices. In summary, robust generalization requires augmenting the encoder training with a wide and diverse speaker population, while high-quality curated material is more advantageous for the synthesizer. The encoder's training data diversity is more crucial than using identical data for both encoder and synthesizer.

Pleasantly, the audio generated for these randomly selected places sounds as realistic as the samples for seen or unseen speakers from the datasets, despite being voices that are absolutely nonexistent. Moreover, these artificial samples' extremely low cosine similarity to even the closest real training speakers, along with their high speaker verification EERs, confirms that they are acoustically different from the training speaker population that has been observed.

This highlights an intriguing feature of the model: instead of being restricted to reconstructing seen or unseen examples from a closed set of training speakers, it can generate extremely natural and understandable speech for hypothetical vocal identities randomly interpolated within the continuous embedding space learned by the speaker encoder. Latent linkages between voices are captured by this space, allowing one to freely explore and synthesize innovative new vocal identities while also allowing seamless interpolation to points between recognized speakers and beyond.

## 4. Conclusion

We have demonstrated a hybrid system for multispeaker speech synthesis based on neural networks. The model includes a sequence-to-sequence synthesizer, neural vocoder, and independently trained speech encoder to reduce noise and extract features from audio files. The speaker embedding that the encoder has learned can be utilized by the synthesizer to produce high-quality speech for voices that were not seen during training, in addition to recognized speakers.

We have shown that, even for these unseen voices, the synthesized speech reasonably resembles real speech from target speakers using objective grading for speaker verification evaluations. Our studies examined the effect of training data scale and discovered that, when enough variation is present in the synthesizer data, increasing the speaker encoder training set significantly enhances speaker transfer quality.

Transfer learning is essential to achieving these effects. In comparison to end-to-end techniques, this method significantly decreases the amount of data needed by separating the training of the synthesizer and speaker encoder. For encoder training, neither speaker labels nor high-quality audio are required. Synthesizer configuration is also made simpler by independent training. Nevertheless, utilizing additional reference speech is limited by the low-dimensional speaker embedding. Context-rich scenarios may necessitate model adaptation for improved similarity.

Furthermore, we've shown that the model can produce speech that sounds natural even for speakers that are entirely fictional and not like the ones in the training set. This suggests that a genuine latent representation of speaker variation has been learned by the model.

Although the proposed model has a solid basis, it still needs to work on accent transfer; nevertheless, with enough data, this problem might be solved using conditioning on distinct speaker and dialectembedding. The model must also address reference prosody and speaker identification.

**References**

[1] https://ai.google/principles/2018,Artificial Intelligence at Google – Our Principles.

[2] Dr. Edriss Eisa Babikir Adam, "Deep Learning based NLP Techniques In Text to Speech Synthesis for Communication Recognition", Vol.02/No.04,pp. 209-215, 2020, doi: https://doi.org/10.36548/jscp.2020.4.002.

[3] Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples", 2018,doi:arXiv.1802.06006.

[4] Kaushik Daspute, Hemang Pandit, Shweta Shinde,"Real Time Voice Cloning",Volume 7 Issue 6,pp. 1306-1312 , 2020, doi : http://www.jetir.org/papers/JETIR2006189.pdf.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate", 2016, doi: arXiv.1409.0473

[6] 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", 2019, doi:arXiv:1806.04558.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, April 1979, doi: 10.1109/TASSP.1979.1163209..

[8]Hieu-Thi Luong, Junichi Yamagishi ,"NAUTILUS: A Versatile Voice Cloning System", in IEEE/ACM transactions on audio, speech, and language processing,vol.28, 2020, doi: arXiv:2005.11004.

[9] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al. "Sample efficient adaptive text-to-speech",2019, doi:arXiv:1809.10460.

[10] L. Zhao and F. Chen, "Research on Voice Cloning with a Few Samples," 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2020, pp. 323-328, doi: 10.1109/ICCNEA50255.2020.00073.

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman."VoxCeleb2: Deep speaker recognition. In Interspeech", pages 1086–1090, 2018, doi: arXiv:1806.05622.

[12] Akanksha Apte, Ashwathy Unnikrishnan, Navjeevan Bomble, Prof. Sachin Gavhane "Transformation of Realistic Images and Videos into Cartoon Images and Video using GAN", Volume 07 Issue 01, 2020, pages :2118-2121.

[13] Doddipatla, R., Braunschweiler, N., Maia, R. (2017),“Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors”, Proc. Interspeech 2017, 3404-3408, doi: 10.21437/Interspeech.2017-1038.

[14] R. Zheng, Z. Zhu, B. Song and C. Ji, "A Neural Lip-Sync Framework for Synthesizing Photorealistic Virtual News Anchors," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 5286-5293, doi: 10.1109/ICPR48806.2021.9412187.

[15] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou,“Deep Voice 2: Multi-speaker neural text-to-speech”, 2017, doi:arXiv:1705.08947.

[16] Alejandro Pérez a,Gonçal Garcés Díaz-Munío a,”Towards cross-lingual voice cloning in higher education”,Volume 105, 2021, doi:https://doi.org/10.1016/j.engappai.2021.104413.

[17] G. Heigold, I. Moreno, S. Bengio and N. Shazeer, "End-to-end text-dependent speaker verification," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5115-5119, doi: 10.1109/ICASSP.2016.7472652.

[18] Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Polle, “Voice cloning: a multi-speaker text-to-speech synthesis approach based on transfer learning”,2021, doi: arXiv:2102.05630.

[19] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf, “Fitting new speakers based on a short untranscribed sample”, 2018, doi :arXiv:1802.06984.

[20] X. Wang, S. Takaki and J. Yamagishi, "Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 402-415, 2020, doi: 10.1109/TASLP.2019.2956145, Senior Member, IEEE, 2019.

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman,“VoxCeleb: A large-scale speaker

identification dataset”, 2018,doi: arXiv:1706.08612.

[22] E. Cooper et al., "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings," ICASSP 2020 - 2020 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6184-6188, doi: 10.1109/ICASSP40776.2020.9054535.

[23] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[24] Varad Naik, Aaron Mendes, Saili Kulkarni, SaieshNaik, Saiesh Prabhu Verlekar, "Voice Cloning in Real Time", vol. 10, pp. 1443-1446, 2022, doi : https://www.ijraset.com/best-journal/voice-cloning-in-real-time.

[25] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang,

Jonathan Raiman, and John Miller,"Deep Voice 3: 2000-speaker neural text-to-speech", 2018, doi: arXiv:1710.07654.

[26] Yi zhao, shinjitakaki, hieu-thiluong, junichiyamagishi, daisukesaito, nobuakiminematsu, "Wasserstein GAN and Waveform Loss-Based Acoustic Model Training for Multi-Speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder", in IEEE, vol. 6,201, pp. 60478-60488, 2018, doi: 10.1109/ACCESS.2018.2872060.

[27] ITUT Rec. P. 800: Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva, 1996.

[28] Mingyang Zhang, Yi Zhou, Li Zhao, Haizhou Li, "Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data", 2021, doi: arXiv:2009.14399.

[29] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368. [30] B. Sisman, M. Zhang and H. Li, "Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1085-1097, June 2019, doi: 10.1109/TASLP.2019.2910637.

[31] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. “Towards end-to-end prosody transfer for expressive
speech synthesis with Tacotron”, 2018, doi: arXiv.1803.09047.

[32] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, NavdeepJaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yonghui Wu “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions” ,2018.

[33] Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, “Char2Wav: End-to-end speech synthesis”, In Proc. International Conference on Learning Representations (ICLR), 2017.

[34] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, Ming Zhou, “Neural Speech Synthesis with Transformer Network”, 2019, doi: arXiv.1809.08895.

[35] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, “Voice Loop: Voice fitting and synthesis via a phonological loop”. In Proc. International Conference on Learning Representations (ICLR), 2018, doi: arXiv.1707.06588.

[36] J. -X. Zhang, Z. -H. Ling, L. -J. Liu, Y. Jiang and L. -R. Dai, "Sequence-to-Sequence Acoustic Modeling for Voice Conversion," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 3, pp. 631-644, March 2019, doi: 10.1109/TASLP.2019.2892235.

[37] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu,“WaveNet: A generative model for raw audio”, 2016, doi: CoRR abs/1609.03499.

[38] Jean-Marc Valin, Jan Skoglund, “Lpcnet: improving neural speech synthesis through linear prediction”, 2019, doi: arXiv.1810.11846.

[39] E. Variani, X. Lei, E. McDermott, I. L. Moreno and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 4052-4056, doi: 10.1109/ICASSP.2014.6854363.

[40] Ryan prenger, rafaelvalle, bryancatanzaro, “Waveglow: a flow-based generative network for speech synthesis”,2018, doi: arXiv.1811.00002.

[41] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit”, 2017.

[42] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, Junichi Yamagishi, “Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet”, 2019, doi: arXiv.1903.12389.

[43] L. Wan, Q. Wang, A. Papir and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4879-4883, doi: 10.1109/ICASSP.2018.8462665.

[44]Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, Yonghui Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech”, 2019, doi: arXiv.1904.02882.

[45] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly,

Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In Proc. Interspeech, pages 4006–4010, August 2017, doi: arXiv.1703.10135.

[46] Sean Vasquez, Mike Lewis, “MelNet: A Generative Model for Audio in the Frequency Domain”, 2019, doi: arXiv.1906.01083.

[47] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying

Xiao, FeiRen, Ye Jia, and Rif ASaurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis”,2018, doi: arXiv:1803.09017.

[48] Hieu-Thi Luong, Student Member, Junichi Yamagishi, Senior Member, “A Unified Speaker Adaptation Method for Speech Synthesis using Transcribed and Untranscribed Speech with Backpropagation”, 2019, doi: arXiv.1906.07414.

[49] Anusha Koduru, Hima BinduValiveti, Anil Kumar Budati, “Feature extraction algorithms to improve the speech emotion recognition rate”, vol.23, pp. 45-55, 2020, doi: 10.1007/s10772-020-09672-4.

[50] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang and L. Chen, "A Light Weight Model for Active Speaker Detection," 2023 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 22932-22941, doi: 10.1109/CVPR52729.2023.02196.

[51] Phyo Thu ZarTun, "Audio feature extraction using mel-frequency cepstral coefficients", vol. 2, 2020.

[52] Meinard Müller, Frank Zalkow, "libfmp: A Python Package for Fundamentals of Music Processing", 2021, doi: 10.21105/joss.03326.

[53] Ayoub Malek, "Spafe: Simplified python audio features extraction", 2022, doi: 10.21105/joss.04739.

[54] D. Parikh and S. Sachdev, "Improving the efficiency of spectral features extraction by structuring the audio files," 2020 IEEE-HYDCON, Hyderabad, India, 2020, pp. 1-5, doi: 10.1109/HYDCON48903.2020.9242729.

[55] Al-Azhar University, Gaza, Palestine, "Impulse noise reduction in audio signal through multi-stage technique, Department of Engineering and Information Technology", vol. 22, pp. 629-636, 2018, doi: https://doi.org/10.1016/j.jestch.2018.10.008.

[56] G.Manmadha Rao, D.N RaiduBabu, P.S.L Krishna Kanth, B.Vinay, V.Nikhil, "Reduction of Impulsive Noise from Speech and Audio Signals by using Sd-Rom Algorithm", vol.10,2021, doi: 10.35940/ijrte.A5943.0510121.

[57] Smita Ghimire, Suraj Basnet Tulachan, Hari K.C., Sharan Thapa, "A deep learning approach for noise removal and enhancement of a real time audio signal", in International Research Journal of Engineering and Technology, vol.4, 2022.

[58] Omkar Chavan, Nikhil Kharade, Amol Chaudhari , Nikhil Bhalke, Prof. Pravin Nimbalkar, "Machine Learning and Noise Reduction Techniques for Music Genre Classification", in International Research Journal of Engineering and Technology, vol.6,2019.

[59] Ali Bou Nassif a, Ismail Shahin, Shibani Hamsa, Nawel Nemmour , Keikichi Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions", 2021, doi:arXiv.2102.05894.

[60]https://docs.fcc.gov/public/attachments/DA-21-1487A1.pdf.

[61] https://cloud.google.com/text-to-speech.

[62] Abhishek Singh , Suraj Chodankar , Aditya Suvarna , "Audio feature extraction tools", in International Research Journal of Engineering and Technology, vol.3, 2021.

[63]Dr V. Kejalakshmi, A.Kamatchi, M.Anusuya, "Active Noise Cancellation using Deep learning", vol.7, 2022.

[64] S. Chehrehsa, M. H. Savoji, "Speech Enhancement using Gaussian Mixture Models, Explicit Bayesian Estimation and Wiener Filtering", in Iranian Journal of Electrical and Electronic Engineering, vol.10, 2014, doi: 20.1001.1.17352827.2014.10.3.3.3.

[65] Yinlin Fu, Xiaonan Liu1, Suryadipto Sarkar, Teresa Wu, "Gaussian Mixture Model with Feature Selection: An Embedded Approach", in Computers and Industrial Engineering, vol. 152, 2021, doi:10.1016/j.cie.2020.107000.

[66] https://ffmpeg.org/ffmpeg.html.

[67] https://openslr.org/12.