Improved SV2TTS Implementation for High-Quality Text-to-Speech Synthesis

G005 (s1862323, s1826377)

Abstract

The goal of the project was to provide left-behind children with high-quality audio accompaniment, and this system aims to deliver a high-quality audio output and performance improvement on the SV2TTS system. The proposed solution consists of three main components: 1. noise denoising module using U-Net 20; 2. SV2TTS system (models: WaveRNN, Tacotron, and GE2E); 3. Prosody transfer module. By incorporating the DCUnet-20 denoise model and the prosody transfer module, the study aims to enhance TTS system performance in noisy environments. The method involves purifying speech signals, synthesizing target speech from denoised speech, and adjusting the speech rate to match the target speaker's prosody. The improved system demonstrates evidently higher performance in terms of DTW distance metric. Our ultimate goal in the future is to deploy the proposed solution via smartphones or personal computers for emotional support to left-behind children who have lost loved ones with high-quality synthesis speeches.

1. Introduction

1.1. Motivation

Voices are vital for socialising and bonding. Due to low local employment, many children in India, China, the Philippines, and Pakistan grow up without parents. Lack of parental direction and company affects children's education and access to key resources, perpetuating poverty. So, creative solutions are needed to help these youth overcome daily educational and health challenges.

Our Text-to-Speech (TTS) solution supports children in these circumstances with high-quality audio. Our approach produces better audio by improving the baseline model. We want to help various people, especially the socially isolated, disabled, and grieving. We also aim to make this technology easy to utilise for those who need it.

1.2. Baseline system

With the aim of accurately simulating the genuine human voice, text-to-speech (TTS) systems are used to translate written text into spoken language. With potential advantages in areas including accessibility, entertainment, and education, TTS is a quickly expanding pitch. TTS systems can be used to produce interactive and interesting educational content, such as audiobooks or language learning materials, that is suited to the needs and interests of children who are left behind in the educational and emotional support systems. TTS technology can also be included in chatbots or virtual friends to offer emotional support and lessen loneliness and isolation sensations. In our project, we put a lot of emphasis on implementing real scenarios and providing emotional support.

The paper Transfer Learning from Speaker Verification to Multi-speaker Text-To-Speech Synthesis (SV2TTS) is the foundation for our TTS system's concept, and basic operation (Jia et al., 2019). Corentin Jemine created the current TTS system. The authors of the SV2TTS study suggest a method for developing a multi-speaker Text-to-Speech (TTS) system that can synthesise a target speaker's voice using only a few seconds of their speech as input. The technique uses transfer learning from a speaker verification model to extract speaker embeddings that aid in the synthesis of natural-sounding speech. WaveRNN is the system's vocoder, Tacotron is the synthesiser, and GE2E is the encoder.

The SV2TTS paper offers a novel method for multi-speaker TTS by combining few-shot learning and transfer learning. The system is highly adaptable and useful for a variety of applications since it can learn to synthesise a speaker's voice using just a few seconds of their speech. The system is able to extract speaker-specific features without the need for considerable training data for each speaker by drawing on knowledge from a previously trained speaker verification model. Real-time speech synthesis that closely resembles the target speaker's voice is possible with this technique.

1.3. Research question

Our research question is whether we can improve the performance of the baseline system in noisy environments by incorporating noise reduction and prosody transfer techniques.

1.4. Objective

After utilizing the existing TTS system developed by Corentin Jemine, we discovered that the system relies on clean studio audio recordings for training, and the subsequent testing results exhibit clear audio without any environmental noise. This situation presents a discrepancy between the goals and practical applications of our project since we cannot assume users will have access to studio conditions for recording. Therefore, we aim to incorporate a noise reduction model into the TTS system, enabling it to adaptively reduce noise according to various ambient noise levels. This integration will provide cleaner input audio compared to the original noisy recordings, ultimately enhancing the quality of the final audio output.

The primary objective of this study is to integrate a denoising module into an existing TTS system to enhance its performance and robustness in the presence of background noise.

2. Data set and task

2.1. Data set of the baseline system implementation

The implementation of the SV2TTS baseline system is based on the LibriSpeech corpus. The LibriSpeech dataset is a vast, diversified, and high-quality resource for research on English speech recognition. Its primary advantages are its size, speaker variety, and audio clarity. Its shortcomings, however, include an emphasis on reading the speech from audiobooks rather than real-world scenarios and a paucity of loud audio samples. Although the LibriSpeech dataset has several benefits, it does not match all of our specifications since it lacks background noise, which is crucial for mimicking real-world circumstances. We need a dataset with ambient noise in the recordings in order to handle our real-world issue more effectively..

2.2. Data set of the proposed solution

Text-to-Speech (TTS) systems are utilised to transform written information into spoken language while attempting to mimic the natural human voice. TTS is a fast-expanding discipline with potential applications in fields including accessibility, entertainment, and education. In the areas of education and emotional support, TTS systems can be utilised to provide dynamic and interesting educational products, such as audiobooks or language-learning materials, that are personalised to the needs and interests of children left behind. In addition, TTS technology can be incorporated into chatbots or virtual companions to offer emotional support and alleviate feelings of loneliness and isolation. In our initiative, we emphasise emotional support and the realisation of actual scenarios.

Based on the work Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis (SV2TTS), our TTS system's concept and guiding principle were developed (Jia et al., 2019). Corentin Jemine designed the current TTS system. The authors of the SV2TTS paper offer a method for developing a multispeaker Text-to-Speech (TTS) system that can synthesise the voice of a target speaker using only a few seconds of their speech as input. The method employs transfer learning from a speaker verification model to derive speaker embeddings that facilitate the synthesis of speech that sounds natural. WaveRNN as the vocoder, Tacotron as the synthesiser, and GE2E as the encoder are the three primary components of the system.

The SV2TTS study (Jia et al., 2019) proposes a unique approach to multispeaker TTS using transfer learning and few-shot learning. The system can learn to synthesise a speaker's voice from as little as a few seconds of their speech, making it very adaptable and applicable to a wide range of applications. By using knowledge from a speech verification model that has already been trained, the system may extract speaker-specific features without requiring considerable training data for each speaker. The system is capable of synthesising speech in real-time that sounds natural and closely resembles the target speaker's voice.

2.3. Evaluation metric

We will evaluate the efficacy of our suggested approaches using Dynamic Time Warping (DTW) distance, which is generally acknowledged and broadly utilised in speech detection and processing, as the evaluation measure. This measure allows for a thorough comparison of synthesised and target speech, taking into account changes in speaking speeds, pitch, and other prosodic characteristics. This paper's "Experiment" section will have a thorough overview of the DTW distance and how it was utilised in our inquiry.

3. Methodology

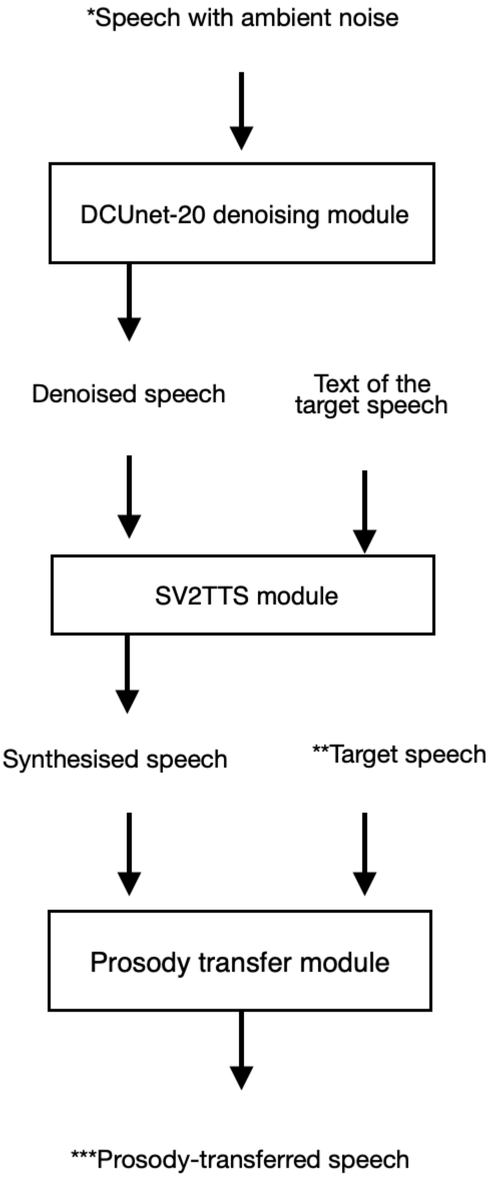
*Speech with ambient noise
DCUnet-20 denoising module
Denoised speech
Text of the target speech
SV2TTS module
Synthesised speech
**Target speech
Prosody transfer module
***Prosody-transferred speech

Figure 1. Structure of the proposed solution

1 2 3

3.1. Overview of the improved system

In this section, we provide an overview of the proposed solution that improved the SV2TTS system, highlighting the key structure and components based on the baseline implementation. Our project aims to enhance the quality of the synthesised speech by incorporating a denoising method and a prosody transfer method into the baseline system. The proposed solution consists of three main components: the denoising module, the original SV2TTS implementation by Corentin Jemine ("SV2TTS system"), and the prosody transfer module.

The summarised overall workflow of the proposed solution:

Denoising Module: The input speech signal is passed through the denoising module, which employs a DCUNET20 model to purify the quality of the input speech. DCUNET20 model enhances the clarity of the input speech by removing unwanted noise. The enhanced speech will improve the baseline system to extract the speaker's characteristics better and eventually embed voice-cloning in the SV2TTS system.

SV2TTS system (the baseline system): The denoised speech signal is fed into the original SV2TTS system, which synthesises the target speech. The system generates the output speech by utilising the denoised speech with better speaker characteristics and embeddings obtained from the denoising module.

Prosody Transfer Module: The synthesised speech is further processed by the prosody transfer module, which modifies the rate of speech by adjusting the playback speed. This technique ensures that the synthesised speech has more similar prosody, closely cloning the speaker's original speaking style.

In the following sections, we will further discuss the denoising and prosody transfer methods in detail and explain their integration with the SV2TTS system.

3.2. Denoising method

3.2.1. Algorithm and implementation

The denoising method employed in the proposed solution is based on the Noise2Noise approach for speech denoising, as introduced by the paper "Speech Denoising Without Clean Training Data: A Noise2Noise Approach" (Kashyap et al., 2021). We adopt the open-source implementation provided by the authors, available at "https://github.com/madhavmk/Noise2Noise-audio/_denoising_without_clean_training_data". This implementation consists of a DCUnet-20 architecture, which we will refer to as the "Noise2Noise Denoising Module" for the remainder of this paper.

The model learns to denoise the input signal by minimising the reconstruction error between two different noisy instances of the same clean speech signal. The DCUnet-20 architecture consists of an encoder-decoder design with 20 convolutional layers. The encoder part extracts important features from the noisy speech input by reducing spatial dimensions and increasing feature channels through convolutional and downsampling layers. The decoder part then rebuilds the denoised speech from these features by increasing spatial dimensions and decreasing feature channels using upsampling and convolutional layers. Skip connections between corresponding layers in the encoder and decoder help the network effectively use both low-level and high-level features.

[1] *Noisy speech

[2] **Clean speech by the same speaker with same content as *Noisy speech

[3] *** Modified rate of speech according to the **Clean speech

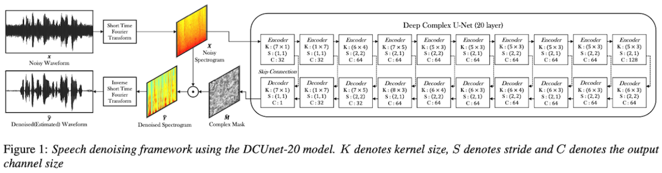
Short Time Fourier Transform
Noisy Waveform
Noisy Spectrogram
Deep Complex U-Net (20 layer)
Encoder K : (7 × 1) S : (1, 1) C : 32
Encoder K : (1 × 7) S : (1, 1) C : 32
Encoder K : (6 × 4) S : (2, 2) C : 64
Encoder K : (7 × 5) S : (2, 1) C : 64
Encoder K : (5 × 3) S : (2, 2) C : 64
Encoder K : (5 × 3) S : (2, 1) C : 64
Encoder K : (5 × 3) S : (2, 2) C : 64
Encoder K : (5 × 3) S : (2, 1) C : 64
Encoder K : (5 × 3) S : (2, 2) C : 64
Encoder K : (5 × 3) S : (2, 1) C : 128
Skip Connection
Decoder K : (7 × 1) S : (1, 1) C : 1
Decoder K : (1 × 7) S : (1, 1) C : 32
Decoder K : (7 × 5) S : (2, 2) C : 32
Decoder K : (8 × 3) S : (2, 1) C : 64
Decoder K : (6 × 4) S : (2, 2) C : 64
Decoder K : (6 × 3) S : (2, 1) C : 64
Decoder K : (6 × 4) S : (2, 2) C : 64
Decoder K : (6 × 3) S : (2, 1) C : 64
Decoder K : (6 × 3) S : (2, 2) C : 64
Decoder K : (6 × 3) S : (2, 2) C : 64
Inverse Short Time Fourier Transform
Denoised(Estimated) Waveform
Denoised Spectrogram
Complex Mask
Figure 1: Speech denoising framework using the DCUnet-20 model. K denotes kernel size, S denotes stride and C denotes the output channel size

Figure 2. DCUnet-20 Model Architecture (Kashyap et al., 2021)

The pre-trained denoising model is to be applied to the input speech signal to obtain a denoised version. The denoised speech signal is then used for speaker embedding extraction and synthesis in the SV2TTS system, resulting in improved voice quality and clarity.

3.3. TTS system based on the SV2TTS framework

This TTS system focuses on generating speech in real-time while mimicking the characteristics of a given speaker. The main components of this implementation are as follows:

1. Speaker Encoder: The speaker encoder is a neural network that learns to extract speaker embeddings from a short reference audio clip. These embeddings capture unique speaker characteristics and are used to condition the speech synthesis process. The model is trained on a speaker verification task, learning to differentiate between different speakers.

2. Synthesizer: The synthesizer is a sequence-to-sequence model that takes text input and speaker embeddings as input and generates mel-spectrograms as output. This component is responsible for learning the relationship between linguistic features, speaker embeddings, and acoustic properties. The Tacotron 2 architecture is used as the synthesizer in this implementation, employing an attention mechanism to align the input text with the generated mel-spectrograms, ensuring that the speech is intelligible and follows the intended prosody.

3. Vocoder: The vocoder is responsible for converting the mel-spectrograms generated by the synthesizer into raw audio waveforms. In the Real-Time Voice Cloning implementation, the WaveRNN vocoder is used. WaveRNN is a lightweight, autoregressive generative model that is designed to work in real-time, enabling the generation of high-quality speech with low computational overhead.

In summary, this TTS system is based on the SV2TTS framework and consists of a speaker encoder for extracting speaker embeddings, a Tacotron 2 synthesizer for generating mel-spectrograms based on the input text and speaker embeddings, and a WaveRNN vocoder for converting mel-spectrograms into raw audio waveforms. This system focuses on generating high-quality speech that mimics a given speaker's voice while maintaining real-time performance.

3.3.1. Evaluation the denoising method with the baseline System

To evaluate the effectiveness of the Noise2Noise denoising method in conjunction with the SV2TTS system, we performed a set of experiments where we manually processed the input speech signals using the denoising model before passing them to the SV2TTS system for synthesis. As discussed in the experiment part, we assess the impact of the denoising method on the quality of the synthesised speech.

In addition, to make the best evaluation of comparing the target speech and the synthesised speech with prosody transfer applied, the output of the SV2TTS system with the denoising method is used as the input speech for evaluating the prosody transfer method (introduced in 3.3).

3.4. Prosody Transfer Method

To create the more realistic voice output with aligned prosody, we will incorporate the rate of speech transfer method into the baseline SV2TTS system.

3.4.1. Algorithm and implementation

The prosody transfer method we employed in our evaluation is based on adjusting the playback speed of the synthesised speech to modify the rate of speech while preserving the pitch. This effective technique ensures that the synthesised speech has transferred the target speaker's prosody, making the generated speech closely resemble the speaker's original speaking style. The implementation of the prosody transfer method consists of two main components: 1. Calculating the speaker's rate of speech using the librosa library in Python. 2. Modifying the playback speed of the synthesised speech according to the target rate of speech while preserving the pitch using the pyrubberband library, which employs the Rubber Band time-stretching algorithm.

3.4.2. Implementation

a. Rate of speech feature extraction: The librosa library is used to load the audio file, and the speech_recognition library is employed to recognise the speech and calculate the rate of speech for both the target speech and the synthesised speech. The provided implementation calculates the speech rate as words per second.

b. Prosody Transfer: Apply the extracted rate of speech to the synthesised speech generated by the SV2TTS system. This is be achieved by adjusting the playback speed of the speech using the pyrubberband library, which employs the Rubber Band time-stretching algorithm that preserves the pitch of the speech while modifying its playback speed.

c. Evaluation: Assess the similarity between the prosody transferred speech and target speech using normalised similarity (between 0 and 1) based on Dynamic Time Warping (DTW) distance.

The above implementation demonstrates how the proposed solution extracts the speech rate and employs the prosody transfer method by adjusting the playback speed of the synthesised speech while preserving the pitch. The implementation is a proposed solution of incorporating the prosody transfer method into the SV2TTS system, allowing for a more accurate evaluation of the generated speech.

3.4.3. Evaluation the prosody transfer method with the baseline system

To evaluate the effectiveness of the prosody transfer method in conjunction with the SV2TTS system, we performed a set of experiments where we manually processed the synthesised speech signals using the playback speed adjustment technique before comparing them to the target speech samples. This approach allowed us to assess the impact of the prosody transfer method on the quality of the synthesised speech without making any changes to the baseline system.

In addition, to make the best evaluation of comparing the target speech and the synthesised speech with prosody transfer applied, the synthesised speech is the output of the SV2TTS system with the denoising method (introduced in 3.2).

4. Experiments

4.1. Experimental setup

4.1.1. Dataset and preprocessing

The dataset used for the experiments was the VCTK dataset, which was utilized to evaluate the efficacy of the noise-denoising model. The data resulting from the VCTK noise-denoising served as the input data for the TTS model. A total of 12 pairs of (noisy-clean) samples were used for both the Denoising method and Prosody Transfer method experiments.

4.1.2. Evaluation metrics

The evaluation metric used to assess the performance of the models was the Dynamic Time Warping (DTW) distance, which was calculated between the target speech and the synthesised speech. The similarity score was normalised between 0 and 1.

The Dynamic Time Warping (DTW) distance was used as the evaluation metric in this project because it is an effective and widely accepted method for comparing the similarity between two time series or, in this case, two wav audio files. DTW has been extensively used in the field of speech recognition and processing for various tasks, including speaker identification, speech synthesis, and speech emotion recognition (Müller & Ellis, 2007) (Dhingra et al., 2013). In the context of this study, the goal was to assess the performance of the proposed denoising and prosody transfer methods by comparing the synthesised speech with the target speech. Since the main objective was to generate synthesised speeches that closely resemble the target speech in terms of prosody and overall quality, it was essential to use an evaluation metric that could effectively capture these aspects. DTW distance has two major advantages for this purpose (Müller & Ellis, 2007): Robustness to time-warping: DTW can align and compare two time series with varying lengths or different time scales, which is a common occurrence in speech data due to differences in speaking rates. Ability to handle non-linear relationships: DTW can compare two time series with non-linear relationships, which may occur in speech signals due to variations in pitch, intonation, and other prosodic features. The normalised results (between 0 and 1)can be more easily interpreted and compared across different methods and experiments. A higher similarity score indicates that the synthesised speech is more similar to the target speech, which is the desired outcome in this study.

4.2. Results

In conclusion, our experiment effectively demonstrates that the improved TTS system, featuring a denoising model and prosody transfer, significantly outperforms the baseline TTS system when processing noisy audio files. By comparing the output audio generated by both systems, we found that the enhanced TTS system produced more realistic and natural-sounding speech and yielded better Dynamic Time Warping (DTW) distance values, indicating a closer resemblance to the clean reference speech.

The integration of the denoising model in the improved TTS system proved crucial in mitigating the impact of background noise on the synthesized speech, resulting in more intelligible and higher-quality audio. Furthermore, the incorporation of prosody transfer contributed to the generation of more expressive and engaging speech, enhancing the overall listening experience for users.

This experiment underscores the importance of integrating noise reduction techniques and prosody transfer mechanisms in TTS systems, particularly when addressing real-world scenarios where noisy environments are prevalent. The improved TTS system offers considerable promise for a variety of applications, such as audiobook narration, virtual assistants, and educational tools, where high-quality speech synthesis is essential for user engagement and comprehension.

Table 1. Experiment Results

Method	**DTW Similarity Score (avg)**
Baseline SV2TTS System	8.925×10^{-5}
Denoising Method	9.378×10^{-5}
Denoising + Prosody Transfer	9.576×10^{-5}

4.2.1. Denoising method performance

The average DTW similarity score for the baseline evaluation was $8.925x10^{-}5$. The similarity score after applying the denoising method was $9.378x10^{-5}$. The results shows that the denoising method has improved the similarity between the target speech and the synthesised speech compared to the baseline.

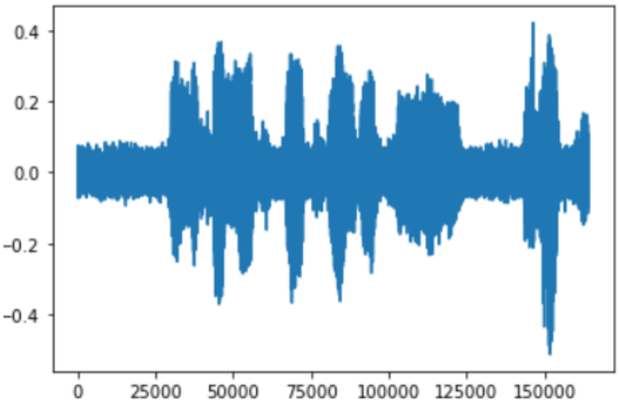

Figure 3. Noisy waveform image

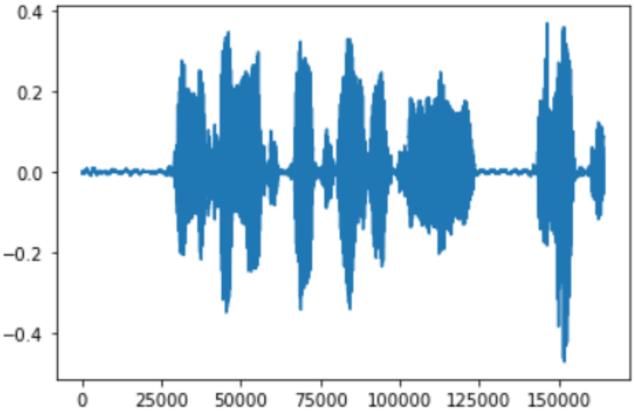

Figure 4. waveform image

4.2.2. Prosody transfer method performance

The average DTW similarity score after applying both the denoising and prosody transfer methods was $9.576x10^{-}5$. The results indicate that the prosody transfer method has improved the similarity between the target speech and the synthesised speech compared to the denoising method alone.

4.2.3. Overall system performance

In conclusion, the denoising method and the prosody transfer method showed promising results in improving the synthesised speech's quality.

4.2.4. Comparison with the baseline SV2TTS system

The baseline SV2TTS system achieved a similarity score of $8.925x10^{-}5$. The proposed denoising method alone improved the score to $9.378x10^{-}5$. The combination of the denoising method and the prosody transfer method resulted in the best performance, with a similarity score of $9.576x10^{-}5$. This indicates that the proposed solution is more effective than the baseline SV2TTS system in generating synthesised speech with improved similarity to the target speech.

5. Related works

This section reviews the relevant existing works and previous studies that relate to our work, placing our research in a broader view. We will discuss works related to each major component of the proposed solution.

5.1. Denoise methods

Various noise reduction methods have been proposed in the literature to improve the quality of speech signals in noisy environments. Some popular methods include spectral subtraction (Boll, 1979), Wiener filtering (Scalart & Filho, 1996), and deep learning-based approaches, such as deep auto-encoders with recurrent neural networks (RNNs) (Weninger et al., 2014). These techniques have demonstrated positive results in various applications.

5.2. Text-to-Speech (TTS) Synthesis

TTS synthesis is a popular research topic that has had significant advancements recently. The development of deep learning-based methods, such as WaveNet (van den Oord et al., 2016), Tacotron (Wang et al., 2017), and DeepVoice (Arık et al., 2017), has led to improved naturalness and intelligibility in synthesized speech. The SV2TTS system (Jia et al., 2019) is a widely discussed TTS system that has been used in various applications, and its implementation by Corentin Jemine serves as the baseline for our work.

5.3. Prosody Transfer in Text-to-Speech Synthesis

Prosody transfer for modifying the prosody of synthesised speech to match the target speech is an important aspect of the TTS application. Methods like prosody embedding (Chen et al., 2021) (Lee & Kim, 2019) and prosody modification through adjusting the pitch of speeches (Lee, 2021) have been proposed to achieve better prosody control in TTS systems. Our work builds upon these methods to improve the naturalness and expressiveness of the synthesised speech.

5.4. Integrating Denoise algorithms to ST2TTS

Rahman et al. (Rahman et al., 2022) explored the application of noise reduction techniques to improve speaker verification in multi-speaker text-to-speech input. The authors proposed adding a noise reduction system to the recorder of a speaker verification to a multi-speaker text-to-speech

(SV2TTS) system, aiming to enhance the quality of synthesized speech. They compared six noise reduction algorithms and applied the best-performing one to the SV2TTS system. Though their work shares the goal of improving TTS systems with us, it approaches the problem from a different angle, focusing on the preprocessing of input speech data. In contrast, our work put more effort into improving the similarity between target speech and synthesized speech through denoising and prosody transfer methods. The research by Rahman et al. indicates the potential benefits of preprocessing input speech data, and their findings could be combined with our approach for even better results in TTS systems.

5.5. Future work

In the scope of further research, our work presents a foundation for the area of noise reduction and prosody transfer for TTS systems. Some potential directions for future work include exploring additional deep learning-based denoising methods, incorporating more advanced prosody control techniques, and extending the evaluation to include other datasets and languages. Also, the investigation of the impact of different types of noise and environmental factors on the performance of the proposed system could provide valuable insights for practical applications.

In view of real-world applications, integrating the proposed noise reduction and prosody transfer techniques into existing TTS systems with voice-cloning could significantly improve user experience. For instance, this accessible TTS system provides an accessible and affordable way for people who do not need high-quality audio and still can enjoy the TTS system to improve their lives, especially for the left-behind children who can get more audio accompaniment and emotional support from the TTS system. Moreover, the most important one of our future goals is to transform it into a mobile phone or other smart devices.

6. Conclusions

In conclusion, our proposed noise reduction and prosody transfer method improved the SV2TTS system's speech synthesis and resulted in clearer speech with higher similarity to the ground truth. Combining these two methods enhances the overall performance, which performs the baseline system by producing synthesised speech that closely clones the target speech. The proposed solution will help create accessible TTS systems with voice-cloning for left-behind children.

Future research should study more deep learning-based denoising methods and prosody control techniques, evaluate the system's performance with different datasets and languages, and assess its response to various kinds of noises in the real-world environment.

Overall, the proposed solution's improvements will foster a more supportive and connected world, enabling high-quality TTS with voice cloning to reach those in need.

References

Arık, Sercan Ö., Chrzanowski, Mike, Coates, Adam, Diamos, Gregory, Gibiansky, Andrew, Kang, Yongguo, Li, Xian, Miller, John, Ng, Andrew, Raiman, Jonathan, Sengupta, Shubho, and Shoeybi, Mohammad. Deep voice: Real-time neural text-to-speech. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 195–204. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arik17a.html.

Boll, Steven F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979. doi: 10.1109/TASSP.1979.1163209.

Chen, Liping, Deng, Yan, Wang, Xi, Soong, Frank K., and He, Lei. Speech bert embedding for improving prosody in neural tts, 2021.

Dhingra, Shivanker Dev, Nijhawan, Geeta, and Pandit, Poonam. Isolated speech recognition using mfcc and dtw. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(8):4085–4091, 2013.

Jia, Ye, Zhang, Yu, Weiss, Ron J., Wang, Quan, Shen, Jonathan, Ren, Fei, Chen, Zhifeng, Nguyen, Patrick, Pang, Ruoming, Moreno, Ignacio Lopez, and Wu, Yonghui. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019.

Kashyap, Madhav Mahesh, Tambwekar, Anuj, Manohara, Krishnamoorthy, and Natarajan, S. Speech denoising without clean training data: A noise2noise approach. In *Interspeech 2021*. ISCA, aug 2021. doi: 10.21437/interspeech.2021-1130. URL "https://doi.org/10.21437%2Finterspeech.2021-1130".

Lee, Jaeryoung. Generating robotic speech prosody for human robot interaction: A preliminary study. *Applied Sciences*, 11(8), 2021. ISSN 2076-3417. doi: 10.3390/app11083468. URL https://www.mdpi.com/2076-3417/11/8/3468.

Lee, Younggun and Kim, Taesu. Robust and fine-grained prosody control of end-to-end speech synthesis, 2019.

Müller, Meinard and Ellis, Daniel P. W. Dynamic time warping. In *Information Retrieval for Music and Motion*, pp. 69–89. Springer, Berlin, Heidelberg, 2007. ISBN 978-3-540-74047-6. doi: 10.1007/978-3-540-74048-3_4.

Rahman, Md. Masudur, Pranto, Sk. Arifuzzaman, Ema, Romana Rahman, Anfal, Farheen, and Islam, Tajul. Application of noise reduction techniques to improve speaker verification to multi-speaker text-to-speech input. In Khanna, Ashish, Gupta, Deepak, Bhattacharyya, Siddhartha, Hassanien, Aboul Ella, Anand, Sameer, and Jaiswal, Ajay (eds.), *International Conference on Innovative Computing and Communications*, pp. 43–56, Singapore, 2022. Springer Singapore. ISBN 978-981-16-2594-7.

Scalart, Pascal and Filho, Jozue Vieira. Speech enhancement based on a priori signal to noise estimation. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 2:629–632 vol. 2, 1996.

van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio, 2016.

Wang, Yuxuan, Skerry-Ryan, RJ, Stanton, Daisy, Wu, Yonghui, Weiss, Ron J., Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, Le, Quoc, Agiomyrgiannakis, Yannis, Clark, Rob, and Saurous, Rif A. Tacotron: Towards end-to-end speech synthesis, 2017.

Weninger, Felix, Watanabe, Shinji, Tachioka, Yuuki, and Schuller, Björn. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4623–4627, 2014.