Generating Podcast Episodes Based On Text Inputs

Uswa Khan[1][0000−1111−2222−3333]

FAST Nuces, Islamabad nu.edu.pk

Abstract. With content creation increasing, many people are using AI to generate digital content. Many people create podcasts because of the diverse niche to choose from and the potential privacy it offers. To cater to this, we have to automate the generation of podcasts. A potential framework that can be used is WaveNets. It uses dilated causal convolutions to generate high-fidelity audio from text input. It produces natural-sounding speech and by leveraging conditional WaveNets, features like speaker identity can be added. Another method that can be used is GAN architecture called MelGAN for conditional audio synthesis. The generator is a fully convolutional feed-forward network that uses the mel-spectrogram representation of the input text as input and produces raw audio as output. The discriminator architecture would be multiscale with multiple discriminators at different scales. A hinge loss version of the GAN objective is used to train the model, along with feature matching to reduce the L1 distance. The evaluation metrics used are WER (Word Rate Error) which measures the accuracy by comparing the generated audio with audio text input. Another metric used is PESQ (Perceptual Evaluation of Speech Quality) which evaluates how closely the audio matches the original audio in terms of perceived quality. To train the model, the LibriSpeech dataset will be used

Keywords: Automatic Podcast Creation · Text-to-Speech · Speech Synthesis· Generate High Quality Podcast

1 Introduction

We are living in an age where content creation has experienced an exponential surge fuelled by the availability of versatile digital platforms. Among the different types of digital content, podcasts have emerged as a popular medium. They offer creators a plethora of niches to choose from and a diverse range of audiences to engage with. The appeal of podcasts isn't just limited to the diversity they offer but rather to the privacy they offer to the creator.

As the pool of creators of podcasts grows so does the need for effective methods of content creation. Creators who want increased privacy would particularly find using someone else's voice appealing. Many creators are therefore turning to AI technologies to automate the process of content creation. This calls for the use of AI to provide effective podcast generation. Using AI has numerous advantages, including increased scalability, reduced time of production, and the ability to translate your podcast into multiple languages to expand to a diverse range of audiences.

This paper aims to explore the effectiveness of using multiple text-synthesis models to automate the generation of podcasts. This entails exploring the architecture of all the models and their mechanisms for text-to-speech synthesis and even conditional generation capabilities. Through comparative studies and objective analysis, we aim to quantify the performance of these models in terms of the quality of audio, accuracy and perceived quality. The metrics used will include Word Error Rate (WER) and Perceptual Evaluation Of Speech Quality (PESQ).

Our secondary objective is to propose a framework for automatic podcast generation. This will encompass data pre-processing, fine-tuning a pre-trained model, and evaluating the model. The framework will serve as a guide for developers aiming to use AI technologies for content creation.

2 Related Work

A considerable amount of research has been done to explore the various aspects of automated speech synthesis, laying the foundations for advancements in automated podcast production. Prior studies have shown valuable insights into the challenges and opportunities in this field.

2.1 Wavenet: A Generative Model For Raw Audio

One such work is that by multiple authors titled "Wavenet: A Generative Model For Raw Audio". It introduces WaveNets, a deep-learning architecture that is popular for its ability to generate high-quality audio from text input. WaveNet uses dilated casual convolution to model the temporal dependencies in audio data. This enables the generation of speech that sounds natural and is of high quality. By using conditional WaveNets, additional features such as the identity of the speaker and be easily integrated into the generation process. This means that the user will be able to personalize the content according to their liking.

2.2 MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis

Another significant work is "MelGAN:Generative Adversarial Networks for Conditional Waveform Synthesis" by multiple contributors. MelGANs can be tailored for conditional audio synthesis. The Generator network takes as input the mel spectrogram representations of text input and produces as output raw audio. The discriminator architecture in MelGAN comprises multiple discriminators and all these discriminators operate at different resolutions to provide a comprehensive assessment of the generated audio. A hinge-loss version version of the GAN objective is used to train the model. To minimize the L1 distance. Figure

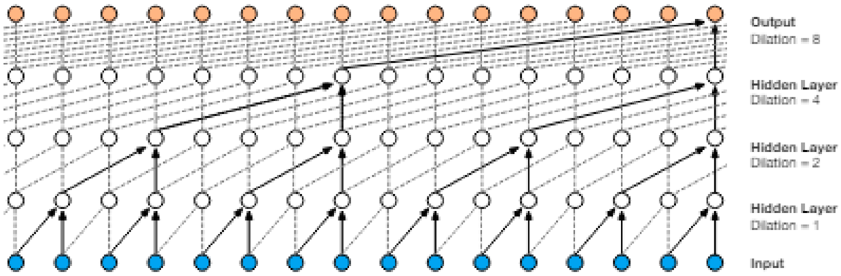


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

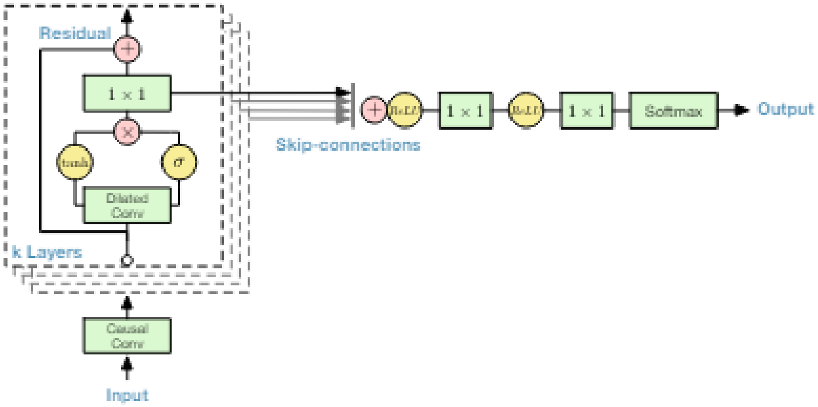
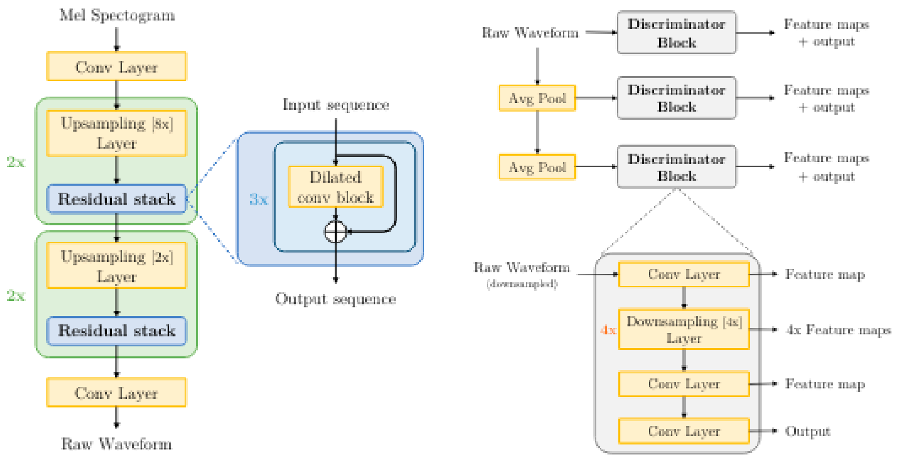


Figure 4: Overview of the residual block and the entire architecture.

5 shows the architecture of MelGAN and Figure 6 shows the the pipeline of MelGAN architecture.



(a) Generator (b) Discriminator

Fig. 1. Figure 5: Architecture of MelGAN

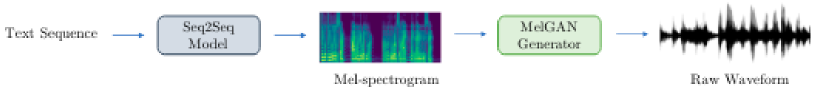


Figure 2: Text-to-speech pipeline.

Fig. 2. Figure 6: Pipeline of MelGAN

2.3 Automatic Podcast Generation

This paper discusses modern advances in natural language processing (NLP) as it relates to podcasts specifically, where improvements in machine-generated podcasts will be the focus. It starts by distinguishing Google Duplex as a podcast-caliber AI device that facilitates humans and computers to converse naturally, and then, continues with the developments in human-computer interaction.

The intrinsic part of the paper's objectives stays the task of including contextual awareness in the machine learning models that bring about summaries whose overall quality is high. Research findings provide evidence that one can experience sensory perception through VR technologies.

The author tends to present his goal of writing a system that can pass Turing's speech test in an attempt to create content for the masses of people. The written content of a particular medium can be transformed into an audio-based medium by making use of the current advancements in NLP with a particular focus on text summarization and speech synthesis. The aim is to close the gap between text and audio content consumption by the users. Figure 2 shows the cloner architecture used in this approach.

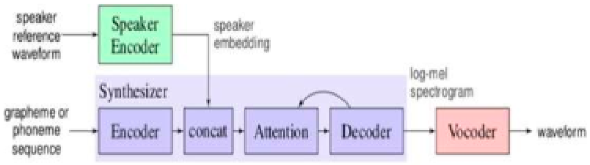


Figure 2. Voice Cloner Architecture

2.4 SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing

The paper presents SpeechT5, a unified framework for learning joint contextual representations from both speech and text data. Its model architecture comprises an encoder-decoder module alongside six modal-specific pre/post-nets. These pre-nets convert input speech or text data into a unified space of hidden representations, which are then processed by the shared encoder-decoder for sequence-to-sequence conversion. Post-nets then generate output in either speech or text modality based on the decoder output.

In terms of input/output representations, the paper formulates tasks as "speech/text to speech/text," handling input speech/text to generate corresponding output within the same modality.

For speech encoding and decoding, the paper utilizes the convolutional feature extractor of wav2vec 2.0 and a neural network for text encoding and decoding. During pre-training, the model learns from large-scale unlabeled speech and text corpus through bidirectional masked prediction and sequence-to-sequence generation tasks for speech data, and text reconstruction tasks for text data.

Post-pre-training, the encoder-decoder backbone undergoes fine-tuning for downstream tasks such as ASR, TTS, speaker diarization (SD), and speaker identification (SID). However we are only interested interested in the models ability to perform the task of TTS for the scope of this project.

2.5 TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

This paper discusses the implementation of modern TTS systems. It is a multi-routine process where there are such phases as feature generalization, duration forecasting, and vocoding. Feature engineering in an integrated end-to-end TTS system is now a thing of the past, new conditioning is possible, and the proposed systems may be more robust due to the all-in-one nature of the systems. Instead of looking for the most suitable melody and then, for example, scaling it to the specifications mentioned above, Tacotron (a proposed end-to-end generative TTS model) solves this problem by getting directly to the unprocessed spectrograms. The Tacotron 2 model, without the need for phoneme-level alignment or extensive human annotation, achieved a MOS for the naturalness of 82 for US English evaluation, exceeding the characteristics of the current parametric systems.

2.6 Deep Voice: Real-time Neural Text-to-Speech

This paper stands out in its approach compared to the other papers in the sense that it uses a TTS system constructed from the deep neural network. This presents a novel end-to-end approach to neural speech synthesis. The system comprises five components: a segmentation model, grapheme-to-phoneme conversion model, phoneme duration prediction model, fundamental prediction model, and lastly audio synthesis model. The model is simpler than the system of the traditional TTS pipeline since each component uses deep neural networks. Overall this paper demonstrates that the system can perform real time inference and achieve up to 400x speedups over existing implementation. This makes the model very suitable for various applications.

2.7 The AT T Next Gen TTS System

This paper proposes the AT T Next Gen TTS System which combines elements from FlexTalk TTD, the Festival System, and the CHATR to enhance English Text Synthesis. From Flextalk and Festival's modular architecture, it incorporates text normalization, letter to sound conversion, and prosody generation. It employs CHATR's unit selection algorithms for high intelligibility. To enhance speech synthesis it uses the Harmonic plus Noise Model (NHM) backend. There have been challenges in database pruning and post lexical processing but this system shows potential for further improvements. Further Research focuses on new voices and language varieties.

2.8 TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks.

In this paper, a novel approach to TTS synthesis is introduced. The approach uses a Bidirectional LSTM model-based Recurrent Neural Network RNN. Traditional TTS systems encounter challenges in capturing long-term contextual effects and generating smooth speech trajectories. The proposed architecture solves these problems by capturing correlations between any two instants in a speech input. Experimental results show that using a hybrid system that combines DNN and BLSTM-RNN is better than using conventional HMM-based and DNN-based systems. The hybrid approach is superior both objectively and subjectively. The proposed model is promising when it comes to improving TTS synthesis by leveraging deep representations of long-span features. However, it remains a challenge to optimize model architecture and training procedures for larger corpora.

2.9 ESPNET2-TTS: EXTENDING THE EDGE OF TTS RESEARCH

This paper introduces ESPnet2-TTS which is an advanced end to end text to speech toolkit It builds upon ESPnet-TTS with new features and models. It simplifies the training pipeline by offering flexible on the fly preprocessing, joint training with neural vocoders and state of the art TTS models. It not only offers a unified Python interface but also pre-trained models in Models Zoo. When evaluated on English and Japanese datasets it achieves a quality that is comparable to ground-truth speech. Future work aims to train using noisy data sets and offer speech to speech translation

2.10 Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech

This paper introduces a TTS model which leverages denoising diffusion probabilistic models. It uses a score based decoder to generate mel-spectograms for noise predicted by an encoder. Control over the trade-off between sound quality and inference speed is established using stochastic calculus. The encoder converts the text into features and the decoder transforms noise into mel-spectograms. The experiments performed on this model demonstrate its ability to generate high-quality speech with variable inference speed and therefore it outperforms existing TTS Models

2.11 Enhancing BARK Text-to-Speech Model: Addressing Limitations through Meta's Encodec and Pretrained HuBert

This paper uses an already existing model called Bark and introduces enhancements to that model. By extracting cookbooks and semantic tokens, the suggested method aims to improve Barks generative capabilities. The modified Bark model demonstrates improved performance in various applications including the domain of multilingual support abd music generation. This paper works to provide advancements in the process of text-to-audio generation and provides access to pretrained model checkpoints for commercial use.

2.12 Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search

This paper introduces a parallel TTS model that is capable of learning its own alignment without the help of external aligners. Glow-TTD achieves robust TTS by combining flow-based generative modeling and dynamic programming. Overall this paper presents a novel alignment search algorithm and model architecture that demonstrates it effectiveness through experiments on single and multiple-speaker datasets.

3 Methodology

The methodology used is simple and wasy to understand. Three different models were experimented with to generate speech from text. The first model used is SpeechT5, the second used is Bark and the third used is Massively Multilingual Speech.

3.1 Purpose For Use

The very first model SpeechT5 is used because of its ability to generate high fidelity speech. Results when analyzed subjectively are good and can work to generate podcasts using the text as an input. However, there are two main problems with this model. First it generates speech in limited languages and for our model we would require speech to be generated in multiple languages. Second, it is trained on the speech of one speaker only which means that we can generate speech in the voice of a female only.

The second model solves the problem of having a single speaker. It allows us to choose from a library of multiple speaker presets having different accents and available both in the voice of female and male. This allows a lot of flexibility for the user. Since we are generating podcasts we would want our audio to have speech patterns like laughing, coughing, and clearing throat. This works in our favor since generating speech with audio patterns adds a human element to your speech.

The third model used can generate speech in multiple languages. For this paper, we have used to generate speech in Hindi since there doesn't exist a version of a model which can generate speech in Whereas this model allows for us to use different presets Bark has a edge over it in the sense that it doesn't allow us to incorporate speech patterns.

3.2 Architecture and Algorithm

SpeechT5 Model is first pretrained using large scaled speech and text data which is unlabeled. After pretraining the entire decoder and encoder structure backbone is fine tuned for each original task. Only thise prenets and postnets are used which are relevant to the specific task that is employe. This way we can obtain multiple models that are all fine tuned for different speech tasks.

Following are the pre and post nets that SpeechT5 uses for the TTS task specifically: Text encoder pre net: A text embedding layer responsible for mapping text tokens to the hidden representation. These hidden representations are those that the encoder expects. Speech decoder prenet: This takes as input a log mel spectrogram and employs a sequence of linear layers which compress the spectogram into hidden representation. Speech decoder post net: This is responsible for predicting a residual to add to the output spectrogram. It is used also to refine the results.

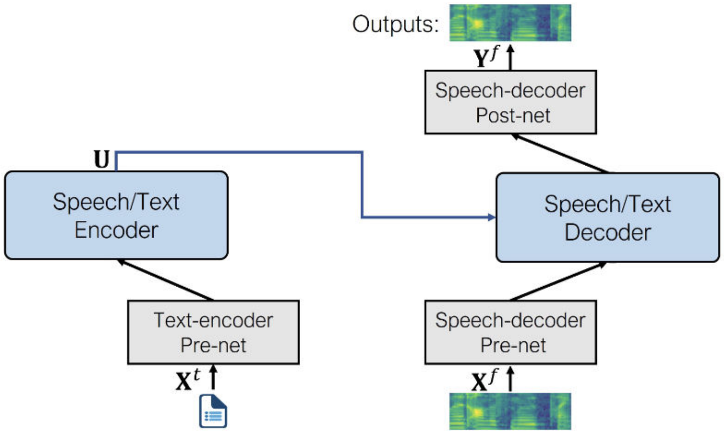


Fig. 3. SpeechT5 Architecture

The second model Bark is different than SpeechT5 since it generates speech waveforms directly. This eliminates the need for a separate vocoder during inference since its already encoder.

With encodec we can also compress audio into a lightweight format to reduce memory usage and subsequently decompress it back to the original audio. To facilitate the process of compression the model has 8 codebooks and each of them consists of integer vectors These codebooks can be thought of as representations or embeddings of the audio integer form. What is interesting about these codebooks is that each improves the output of the previous codebook. Since codebooks integer vectors model transformers can learn them. The Bark consists of four models: BarkSemanticModel, BarkCoarseModel, BarkFineModel and the Encodec model.

MMS (Massive Multilingual Speech) is an incredible model that can synthesize speech in over 1100 languages. Its based on VITs, a state of the art TTS approach. VITS is a speech synthesis network responsible for converting text into raw speech waveforms. It functions like a conditional variational autoencoder and estimates audio features from the input text. The first step is the generation of acoustic features in the representations of spectrograms. The second step is the decoding of the waveforms. For this, transposed convolutional layers adapted from HiFi-GAN are used. During the process of inference, our text encodings are upsampled. They are also transformed into waveforms and for this purpose flow module and HiFi-GAN decoder are employed. As was the case with Bark, there isn't any need for a vocoder since the waveforms are generated directly. The following image sshow the variational inference and the reconstruction loss of the model.

3.3 Experimental Setup and Results

Since we are using pretrained models, the dataset used is default dataset. This has several advantages. First is the reduced time. The time spent on loading the dataset can now be spent on fine-tuning the model and qualitatively analysing is. Next is importing the pretrained models. We have imported three models. First for generating high fidelity audio, second for incorporating speech patterns in the audio and third for generating the text in Hindi.

As for evaluating the model, it is incredibly difficult for us to provide a quantitative analysis since the analysis of speech is subjective.

The visualization of dataset is shown below. However visualizing the output was challenging since the output was audio.

3.4 Innovation and Analysis

When it comes to podcast generation we have seen the best results with Bark model. It allows for us to use multiple voice presets, incorporates different speech patterns and supports multiple languages although not as many languages as the MMS model. However it offers a considerable advantage over MMS since it is able to generate long form audio and is able to differ between the speakers which makes it invaluable for podcast generation.

3.5 Conclusion and Future Work

The work can be extended to generate even longer form of podcasts. Bark can be improved to support more languages. Apart from this the model can be optimized so that it can process text in parallel and in significantly less time.

$$L_{recon} = \|x_{mel} - \hat{x}_{mel}\|_1$$

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p_\theta(z|c)}\right]$$

References

1. University of Shanghai for Science and Technology. (2021). Automatic Podcast Generation: Extractive Summarization of Text Data and Synthesis of High-Quality Audio. *Journal of University of Shanghai for Science and Technology*, **23**(10), 22–26.
2. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499.*
3. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., & Courville, A. (2019, December 9). MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. *arXiv:1910.06711v3 [eess.AS].*
4. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. *arXiv preprint arXiv:1703.10135.*
5. Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (n.d.). The AT&T Next-Gen TTS System. *AT&T Labs – Research.* Retrieved from http://www.research.att.com/projects/tts.
6. Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K. Soong. "TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks." *Proceedings of Interspeech 2014*
7. Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, Shinji Watanabe. "ESPNET2-TTS: EXTENDING THE EDGE OF TTS RESEARCH." *Proceedings of Interspeech 2022*
8. Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov. "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech." *Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.*
9. Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, Furu Wei. "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing." *Electrical Engineering and Systems Science > Audio and Speech Processing*, arXiv:2110.07229v3 [cs.SD], 24 May 2022.
10. Arık, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., and Shoeybi, M. (2017). Deep Voice: Real-time Neural Text-to-Speech. *Proceedings of the 34th International Conference on Machine Learning*, pages 195–204.
11. Devin Schumacher and Francis LaBounty Jr. "Enhancing BARK Text-to-Speech Model: Addressing Limitations through Meta's Encodec and Pretrained HuBert." *arXiv preprint arXiv:4443815*
12. Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search." *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*