Literature Review

ABHIJEET ANAND, Indraprastha Institute of Information Technology Delhi
ANKIT GAUTAM, Indraprastha Institute of Information Technology Delhi
KARTIKEY DHAKA, Indraprastha Institute of Information Technology Delhi
SAMRIDH GIRDHAR, Indraprastha Institute of Information Technology Delhi
SOUMYA MOHAPATRA, Indraprastha Institute of Information Technology Delhi
SUSHANE DULLOO, Indraprastha Institute of Information Technology Delhi

ACM Reference Format:
Abhijeet Anand, Ankit Gautam, Kartikey Dhaka, Samridh Girdhar, Soumya Mohapatra, and Sushane Dulloo. 2018. **Literature Review**. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

This project aims at studying the capabilities of large Language models (LLMs) for personalising and generating Audiobooks Narrations. This is done with advanced text-to-speech (TTS) generation that empowers unique narration for users to select preferred voices, accents tones and emotions set specifically per genre mood and audience type. The project will also allow for dynamic adjustments, providing a unique and personalised listening experience to users that is personalised to various literary contexts. LLM narrations are audio solutions for video game worlds or fake emotions performance in a fictional piece or formal tone on educational content. It uncovers a way to create the audiobooks for entertainment as well as access and education, opening up new possibilities as this sort of project can deliver additional material to have listeners involved in reading themselves into an enriched storytelling experience.

2 Methods

Considering our problem statement we could not find any direct research papers that caters to our problem statement so we decided to divide the problem statements into sub-parts that include:

Emotional text-to-speech for better user experience and Multi speaker text to speech for incorporating multiple characters in story. We reviewed the research papers on each of the sub-topic as they are key components of our project. Following are the key findings from each paper we reviewed:

Authors' Contact Information: Abhijeet Anand, abhijeet21509@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi ; Ankit Gautam, ankit21518@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi ; Kartikey Dhaka, kartikey21534@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi ; Samridh Girdhar, samridh21282@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi ; Soumya Mohapatra, soumya21103@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi ; Sushane Dulloo, sushane21292@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi .

Manuscript submitted to ACM

3 Key Findings

3.1 MultiSpeech – Multi-Speaker Transformer TTS

This paper proposed MultiSpeech model for multiple speaker TTS systems. They have utilised the ability of parallel computation during training to improve the training of the model they have proposed. They have highlighted that although RNN based encoder decoder models have existed for multiple speaker text-to-speech but they face problem with sequential training. Therefore they have utilised transformer approach. They have also highlighted some issues with transformer based training. They have mentioned that transformer based training works well for single speaker TTS systems but works poorly for multi speaker systems. The key reasons they have mentioned include learning the alignment between text and speech during parallel training as the dataset of multi speaker is noisy and diverse. They have proposed several techniques to improve text-to-speech alignment.

- As the multispeaker data is noisy and diverse largely. The voice of each speaker distincts in speech and acoustic properties therefore really hard to for the model to align text and speech. To mitigate this issue the paper has proposed to add constraints on the diagonal of the attention matrix to force the model to learn correct alignments.

- The second issue is that when they convert the graphemes to phonemes embeddings the values of the embeddings can vary a lot. On the other hand the range of positional embeddings is almost fixed and have less variance. This affects the alignment of text and speech. To solve this the paper proposes to add a layer of normalisation before adding the positional encodings.

- Thirdly they have introduced a bottle-neck layer in the decoder pre-net by reducing the size of the hidden layer. This is done so that the model does not copy the information from the previous speech frames and focuses on the current text input. This helps for a smooth switch between speakers.

Datasets Used

• They have utilised the VCTK dataset (108 speakers with different accents) and LibriTTS dataset (2000 speakers).

Results:

• **Improved Speech Quality:** The mean opinion score improved (MOS) indicating the sound is much pleasant for the speaker.

• **Better text and speech alignments:** The diagonal attention rate improved for their model which shows better text and speech alignments.

Since we want to include multiple characters in book narration with different accents, linguistic properties and customization this paper can help us to improve our model in that direction.

3.2 StyleSpeech: Parameter-efficient Fine Tuning for Pre-trained Controllable Text-to-Speech

This research paper introduced a new model StyleSpeech a text to speech model for natural speech generation. This model has three key innovations

1) Unique style Decorator

2) LoRA technique in text -to-speech generation task

3) Novel automatic evaluation metric LLM-Guided Mean Opinion Score (LLM-MOS) or efficient evaluation.

- **Style Decorator:** They have utilised the transformer based approach to train the distinct phonemes features and style features in parallel thus saving time for training.

- **LoRA** They have utilised LoRA technique to reduce the parameters in TTS model. This will help the model to train with less computational resources.

- **LLM-Guided Mean Opinion Score** Since human based evaluations are cost inefficient as it is labour intensive. They have introduced a novel approach for automatic evaluation in TTS task.

Components of StyleSpeech:

1. **Acoustic Pattern Encoder (APE):**

It uses a FFN layer to encode the phoneme and the style features of the input text and it processes the text and converts the Grapheme to phoneme and style features using G2P conversion and then convert to Hp and Hs

2. **Phoneme Duration Adapter:**

This checks for any misalignment in the lengths of phoneme embeddings and the target Mel-Spectrogram. It has two components:

Components:

- **Duration Predictor:** Estimates the duration of each phoneme and style features.

- **Length Regulator:** This adjusts the embeddings to match the predicted durations, resulting in adaptive embeddings (H_L).

3. **Style Decorator**

Its function is to fuse the phoneme and style embeddings Hp and Hs into a combined embedding H without altering the phoneme features. This helps to prevent the blending of style features and phoneme features that are seen in traditional TTS systems.

4. **Mel-Spectrogram Encoder and Vocoder:**

- **Mel-Spectrogram Encoder:** It processes the fused embeddings through additional FFT layers to generate Mel-Spectral embeddings (Hm), which are then transformed into a Mel-Spectrogram (Y').

- **Vocoder:** It uses the Griffin-Lim algorithm to reconstruct the speech audio from the Mel-Spectrogram by estimating and refining the phase information iteratively.

Dataset:

It uses the Baker dataset containing 10, 000 high quality recordings of a professional actress. The language of the dataset is Chinese.

Evaluation:

For evaluation they have utilized the following metrics:

1. **Quantitative metrics:**

- WER (Word error rate)

- Mel Cepstral Distortion (MCD)

- Perceptual Evaluation of Speech Quality (PESQ)

2. **LLM-Guided MOS (LLM-MOS):**

- Rates speech quality on a scale of 1 to 5 using large language models.

- Combines quantitative metrics to provide an overall quality rating.

- Offers an objective and efficient alternative to traditional human-based MOS evaluations.

Thus this paper shows us the effective way to train our TTS model using style decorators and LoRA adaptation for training the phoneme and style features simultaneously without losing the distinct features of the phoneme. This will be useful to utilise a transformer based model that too in a cost and computationally effective manner.

3.3 Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability

This paper proposes an i-ETTS model to improve the emotion discriminability of synthesised speech by interacting with a Speech Emotion Recognition (SER) model using Reinforcement Learning. This paper tries to solve various problems with the existing emotional text-to-speech models like:

Firstly the style embeddings that are generated from text input lacks meaning hence various models that are trained on these embeddings lack interpretability and often have emotion confusion. To improve the emotional rendering issue the various models have introduced emotion recognition loss and perception loss but this has made the style embeddings misaligned with the output acoustic features and heavily focused on emotions. The MSE loss on output helps to train for the acoustic features for the sound. So there is a trade off between the emotion rendering and acoustic features in the output. To mitigate this several research papers have introduced method of interacting the speech emotion recognition model with our TTS system for better emotion predictability in speech. The issue here is a large amount emotional labelled speech dataset is required . To solve this issue the paper proposes a method. In this method the TTS system will serve as an agent and the SER model will serve as a reward model and will interact with the TTS system to improve its performance through a feedback loop. This will reduce the amount of speech labelled dataset. This methodology formulates the TTS training as an RL problem with a reward function based on SER accuracy. Since just using the SER model to train will make the speech embeddings highly emotionally classified and will deviate from our prime objective of generating high quality audio, the model proposes an iterative training strategy where first the model is pre trained with general MSE loss for some time , then after pretraining the model is iteratively trained with MSE loss and Rl based loss. This helps to balance between our prime objective of speech generation and auxiliary objective of aligning emotions with speech.

Model workflow

1. Model generates speech from text

2. The SER model predicts the emotion perceived from the speech.

3. SER model prediction is compared to the actual emotions and a feedback is sent to the TTS agent.

4. The TTS system utilises RL based training to optimise the parameters based on the reward generated from the reward model to get higher rewards.

5. This loop is done for many iterations with RL based training and MSE loss training done alternatively.

Dataset

They utilised an ESD dataset containing five different type of emotions: happy, sad, angry, neutral and surprise.

Evaluation: They performed two types of evaluation:

1. Objective: They used SER accuracy to predict the accuracy of the model in emotion synthesis in speech.

2. Subjective: They used MOS score on each emotion category and A/B testing to judge user experience with the audio. They found the i-ETTS model performs better than baseline models MTL-ETTS and CET-TTS.

Thus this paper effectively utilised RL based approach and integrating TTS systems with SER model to improve emotional recognition and reduce confusion, at the same time preserved the acoustic quality of the audio through iterative training. This paper is useful to fine tune our model for incorporating emotions in the speech while narrating the AudioBook.

3.4 EMOTIONAL VOICE CONVERSION USING MULTITASK LEARNING WITH TEXT-TO-SPEECH

This paper proposes a study of effectiveness of multi task learning for Voice Conversion and TTS task. The paper speaks about challenges faced in various traditional seq2seq models for voice conversion from input speech to output speech . It says that when we use a seq2 seq model for an input speech and try to generate a new output speech with different linguistic properties like accent, gender, etc there is a loss in the linguistic information of the input speech. This lead to mispronunciation and training instability. To mitigate this issue people have tried to add textual supervision was added to each step at the decoder output. But this required alignment of the text with the speech explicitly which is a complicated task. In their proposed method they decided to train the VC model with the embeddings space of the TTS model instead of the embedding space of VC model as input. This is because the embedding space of TTS model is similar to the phonetic information of the VC model. In this way they are performing multitask learning of the VC and TTS model simultaneously. Using TTS system will help to preserve the linguistic information in the speech. The model architecture basically consist of a style encoder that takes the log Mel spectrogram of style-reference speech, text encoder that takes the one hot represented text, content encoder that takes the log Mel spectrogram of speech-carrying linguistic content. The text, content, style encoders generate the text, content, style embeddings respectively. Now since the VC and TTS are trained simultaneously they hay created a switch like system. This switch will ensure the correct set of inputs flow for the different tasks. For VC we require the content and the style encodings only to train. On the other hand we require the text and style encodings for TTS systems. Then we have the attention RNN layer followed by a DECODER layer to generate the output at each time steps. In this way by using multitask learning we can efficiently train both the models simultaneously. This model has a lot of interesting features. Firstly we dont require to explicitly align the text in the model using Dynamic Time Warping. Secondly there is no requirement of emotional labels to train the model.

Dataset: They have utilised the mkETTS dataset which contains audio recording in Korean Language by a 30 year old male speaker. It includes 3000 sentences in each of the seven emotions: anger, sad, happy, neutral, surprise, fear, disgust. Thus resulting in 21000 utterances. They claim the dataset to be of high quality.

Evaluation:

- The objective evaluation had been done using WER prediction

- The subjective evaluation was done using MOS and A/B preference testing.

- The results clearly showed that the model that multitask learning is a very effective way to create VC and TTS models simultaneously and this technique can be easily extended for emotional text-to-speech.

- They tested there models on the emotional speeches with input as neutral and converted it to 6 different emotions and found that the style encoder in the model was efficiently generating embeddings for the emotions. The embeddings belong to similar emotions had cosine similarity much similar. Like style embeddings of sad and fear were more similar than happy and sad.

- When the mel spectrogram was plotted it was found that the overall shape of the spectrogram is the same but there is a temporal shift and change in frequency.

This paper is useful to us as we can explore the capabilities of VC models using multi task learning to convert the audio generated using a traditional TTS model and change it to different accents, dialects and characters. This model has only used Korean dataset that too with only male speaker. We can explore it in a multi speaker and explore its capabilities and try to fit this approach in our project to build customizable AudioBook.

4 Conclusion

In this Survey paper we have tried to covered sub-topics like multi-speaker TTS, emotional TTS using RL, multi task learning of VC and TTS and PEFT finetuning of TTS systems which would be of key use when we will be building our project. The following papers align with our problem statement to build a customizable audiobook that would contain emotions, different characters belonging to different genders, generations etc. Thus utilising the key findings and improving the weaknesses in the above papers can help us to meet the requirements of our problem statement.

5 References

- https://arxiv. org/abs/2408. 14713 - Parameter-efficient Fine Tuning for Pre-trained Controllable Text-to-Speech

- https://arxiv. org/abs/2006. 04664 - Multi-Speaker Text to Speech with Transformer

- https://ieeexplore. ieee. org/abstract/document/9053255?casa$_{t}oken$ = 73$tBVplBj8oAAAAA$: $BJF53GYF4MWGHa-zRNDrN7MjjYpLXbiTZQc5H2YoJg9OceG7E3gdjJc4aaLZwhpztztpNv_{m}MMb2788-EmotionalVoiceConversionUsingMultitaskLearnin To - Speech$

- https://arxiv. org/abs/2104. 01408 - Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability