Survey: Storytelling Text-to-Speech Synthesis

T Pavan Kalyan
IIT Bombay, Mumbai
190020124@iitb.ac.in

Abstract

This paper presents a survey of the importance of incorporating storytelling speaking style in text-to-speech (TTS) technology. The paper highlights the significance of prosodic features, such as intonation and rhythm, in conveying meaning and emotion in spoken language. It discusses the challenges of capturing human narrators' vocal characteristics and speaking style and the ways to overcome them using advanced machine learning algorithms and neural network architectures. The paper extensively covers state-of-the-art expressive TTS models and different TTS datasets. The potential of TTS technology in enhancing spoken language quality and impact in various domains, from entertainment to education, is also emphasized.

1 Introduction

Text-to-speech (TTS) synthesis has made significant progress in recent years, with systems capable of generating speech with diverse prosody and speaking styles. One interesting application of TTS is in creating a story-telling machine that can take a story for children in text format as input and output a well-narrated story in speech format. The final expected outcome is a TTS system that can narrate stories to children, rich in prosody and exaggerating certain emotions and expressions to make it more interesting for children. In this survey, we explore the state-of-the-art in storytelling speaking style TTS systems. We begin by discussing the challenges involved in this task, such as the need for expressive prosody, the importance of understanding how to narrate a story expressively, and the difficulty in training models on data labeled with prosodic features. We then review the different approaches taken by researchers to address these challenges, including using multi-speaker TTS, single-speaker/multi-role TTS, and incorporating linguistic and paralinguistic information.

1.1 Problem statement

The aim is to build a story-telling machine that takes a story for children in text format as input and outputs a well-narrated story in speech format. This problem is part of the bigger problem called expressive text-to-speech synthesis system that can generate speech with diverse prosody and speaking styles. The final expected outcome is a TTS system that can narrate stories to children. The output should be rich in prosody. In fact, the speech should exaggerate certain emotions and expressions to make it more interesting for children. An even more interesting problem is to produce such speech as output without explicitly training the model on data labeled with any prosodic features. Hence, we expect the TTS system to not only speak the story but also understand how to narrate a story expressively to children aged 7-12 years.

1.2 Motivation

Modern neural text-to-speech (TTS) systems have achieved human-like quality in terms of naturalness and intelligibility. However, most TTS systems are trained on a standard 24-hour LJ Speech dataset, which consists of non-fiction audiobooks read by professional actors. To effectively model all expressions of speech, a more expressive speech dataset is required. Children's stories, with their exaggerated emotions, provide a suitable alternative. Despite the high quality of current TTS systems, they lack an understanding of the spoken text, resulting in a lack of human prosody and expressive speech. Motivated by the need to create a TTS system that can narrate stories to children in an interactive and expressive way, this survey explores the state-of-the-art in storytelling speaking style TTS systems. We discuss the challenges involved in this task, such as the need for expressive prosody, the importance of understanding how to narrate a story expressively, and the difficulty in training models on data labeled with prosodic features.

2 Background

The art of storytelling is found culturally everywhere in the world. In fact, most stories children hear in India are either from their parents or grandparents. The advent of urbanization and technology has allowed people to forget this tradition of telling stories to children and instead YouTube has taken its place. Though this is easier for parents, it does not help children interact and learn actively from stories. The proposed TTS system may tell the story the parents want and even mock parents' voices using zero-shot voice cloning. Creating such a system will open a plethora of opportunities and will help the research of TTS systems further in terms of expressiveness. Storytelling speaking mainly comprises two primary research areas: speech production and emotions in the produced speech. The following sections provide clear explanations of these two parts.

2.1 Speech

Speech production is the process in which humans produce meaningful speech that can be perceived by others. Speech is produced as a by-product of human respiration. CO2 is let out from the lungs during exhalation, which passes through the vocal tract. The rest is controlled by the brain and the vocal tract to produce meaningful speech. Sounds are classified broadly into vowels and consonants, where vowels are produced by unrestricted airflow through the vocal tract, and consonants are produced by forming a constriction at some place in the vocal tract. Most common sounds are because of the vibration of vocal cords, some sounds are produced by a narrow constriction in the oral cavity. Some sounds like /t/ are produced because of a sudden release of air called plosion and such sounds are called plosives. All vowels are voiced as air flows through vocal cords which vibrate and create voiced sounds. Vocal cords do not vibrate while producing voiceless sounds.

The amount of air exhaled by the lungs and the muscular strain on the articulators that produce the sound are the key determinants of a speech sound's volume or intensity. For instance, speech in rage typically has more volume than regular or calm speech. The volume or intensity is a prosodic parameter that is related to emotions and sentence type. For example, interrogative sentences tend to end with a higher intensity as compared to neutral statements. The fundamental frequency of voiced sounds is the frequency at which the vocal cords vibrate (F0 or pitch). One of the most significant prosodic factors is the fundamental frequency, which is dependent on the strain placed on the vocal cords and the amount of airflow generated by the lungs. The fundamental frequency may be modified to give the phrase a certain intonation. Emotions and phrase patterns are significantly influenced by the fundamental frequency. The signal's spectral envelope is a highly helpful tool from signal processing for speech analysis. This spectral envelope often displays a few maxima at the vocal tract's resonance frequencies, or formants, which are traits of the various phonemes. In fact, the formants of the various vowels can be used to differentiate them. With the aid of voicing (and fundamental frequency for tonal languages like Chinese), the spectral envelope is capable of differentiating between speakers and distinct phonemes of a language. The durations of the phonemes are determined by the coordinated movement of the speech production system across time. Duration is regarded as a prosodic feature that provides useful data for identifying phonemes and speakers.

2.2 Emotion

There are recognized theories of emotions from the majority of the great classical thinkers. Defining emotions is a 124-year-old unsolved mystery. Since Darwin, researchers have been studying emotions, and many psychological schools have developed several theories that reflect various approaches to comprehending emotional state. The three basic kinds of theories of emotion are physiological, neurological, and cognitive. According to physiological theories, emotions are caused by internal processes in the human body. According to neuroscientific ideas, emotional responses are the result of brain activity. According to cognitive theories, ideas and other mental activities are crucial in the development of emotions. The categorical model and the dimensional model are the two distinct methods for representing emotions. The representation in the dimensional model is built on a number of quantitative metrics scaled on many dimensions. Both models offer perceptions on how emotions are represented and perceived by the human mind and each one serves to express a certain aspect of human emotion. These models evaluate a person's actual emotional states. According to Oxford dictionary emotion is "A strong feeling deriving from

one's circumstances, mood, or relationships with others." Emotion was introduced into academic discussion as a catch-all term to passions , sentiments and affections (Dixon, 2003). Plutchik was one of the psychologists working at the frontiers of emotions. He proposed that are eight primary emotions : sadness, fear, disgust, anger, trust, anticipation, surprise and joy. He also proposed a wheel of emotions to depict the relationship between different emotions. He used color theory to depict the combination of emotions and the result of this combination as another emotion.

2.3 Expression of emotions in speech

Humans have the innate ability to comprehend the underlying emotional state and linguistic substance of spoken communication. Typically, humans notice the emotions of a stranger through departures from their typical condition. This suggests that a reference (neutral/normal) exists and that departures from the reference are perceived.
The word voice quality refers to the distinctive marking of a person's speaking. Typically, each speaker has a unique voice quality characteristic. They express essential information, such as intentions, emotions, and attitudes, by utilizing a variety of voice characteristics. Some of the characteristics of the many emotions share comparable traits. A voice signal's spectrum is sound-specific and comprises characteristics such as F0, durations, loudness, and spectral parameters. Several studies have demonstrated that the amplitude and shift of formants during emotional states vary between vowels. The concept of seeing emotions as points in a continuous spatial dimension was initially proposed in (Schlosberg, 1941). Principally, emotions are understood as mixtures of three dimensions: valence, arousal, and dominance. There are many levels of feature representation, including frame level, segment level, and utterance level. Voice characteristics include shimmer, jitter, and NAQ, which are connected to glottal excitation traits.

2.4 Storytelling

Storytelling, a sub theme of fiction literature, is built on discourse modes, which commonly include narrative, descriptive, and conversational styles. The primary purpose of narrative storytelling is to enlighten the audience about the events and individuals influencing the plot. In contrast, the descriptive mode provided the listener with specific information about a character or incident so that they could form a clear mental image of what was presented. Lastly, dialogue storytelling is when the narrator transforms his or her voice into a character's, generating an exaggerated register of expressions and full-blown emotions. In the majority of storytelling speaking styles, children's stories and folk tales are the preferred narrative kinds.

Corpus	Hours	Speakers	Sampling rate (kHz)
ARCTIC	7	7	16
VCTK	44	109	48
Blizzard-2011	16.6	1	16
Blizzard-2013	319	1	44.1
LJSpeech	25	1	22.05
LibriSpeech	982	2484	16
LibriTTS	586	2456	24
VCC 2018	1	12	22.05
HiFi-TTS	300	11	44.1
CALLHOME	60	120	8
RyanSpeech	10	1	44.1

Table 1: Various English TTS corpora compiled in Table 17 in (Tan et al., 2021)

3 Datasets

The LJSpeech dataset ((Ito and Johnson, 2017)) is the benchmark dataset used by State-of-the-Art English TTS systems. It is a US accent dataset with approximately 24 hours of audio of 7 non-fiction books. Other TTS corpora like LibriTTS ((Zen et al., 2019)) and VCTK ((Yamagishi et al., 2019)) are also famous for multi-speaker training. The libriTTS dataset consists of 585 hours of speech data sampled at 24kHz recorded by 2456 speakers. Since none of these datasets contain audio for children, the presented dataset is more expressive than the currently available TTS corpora. Table 1 is a list of English TTS corpora and their related properties, like the number of hours of speech data, the number of speakers, and the sampling rate. Most modern production quality TTS use 22.05kHz, 32kHz, 44.1kHz, or 48 kHz sampling rate. Higher sampling frequency allows the acoustic model to learn the audio's detailed acoustic information and reproduce the same from the text.

4 Neural Text-to-speech systems

A neural text-to-speech synthesis system can be modular or end-to-end. A typical TTS system consists of three components: 1. Text-processor,

2. Acoustic model, 3. Vocoder. In an end-to-end model, all these components are modeled together as a single neural network architecture. Here end-to-end means the input to the model is text and the output is a speech waveform. A text-processing module converts textual input i.e. characters into linguistic features using a neural architecture. These linguistic features are input to the acoustic model which outputs an acoustic representation. These acoustic features are fed into Vocoder which produces the output speech waveform.

4.1 Text-processing module

This module is also called the front end in conventional text-to-speech systems. A typical text-processing module consists of the following steps:

- Text-normalization: This involves converting numbers like 1989, abbreviations like Mr., and other non-standard words from raw-text format to spoken form like "nineteen eighty-nine" and "Mister". This module is important when there are multiple ways of verbalizing non-standard words. For example, 3 Lb can be spoken as "three lb" or "three pounds" depending upon the context ((Zhang et al., 2019)). Another such instance is for numerical addresses. Consider "345 Tilak Marg", as this can have two verbalizations. One where the number is expanded completely as "Three hundred and forty-five Tilak Marg" but this option is not the most suitable for the case of navigation systems where the better output is "Three forty-five Tilak Marg". All such words are called semiotic words that differ in the way they are written and verbalized. Some of these words include dates, times, numbers, and monetary amounts.

- Part-of-Speech Tagging: This module assigns a part-of-speech tag to each word in the text. This will help the TTS system to convert the graphemes to phonemes easily as a word may have different phonetic transcription based on the POS tag. Though this module is very impactful for statistical TTS systems, neural architectures almost always skip this step.

- Prosody Prediction: Prosody plays an important role in human speech and the inclusion of prosody in TTS-generated speech makes the speech natural. Prosody includes rhythm, stress, and intonation of speech which are modeled by the duration, pitch, and loudness of the phonemes. Neural architectures have separate modules to learn the elements of prosody like pitch, duration, and intensity.

- Grapheme-to-phoneme conversion: The most important step is to convert the graphemes to phonemes. This can be done using a grapheme-to-phoneme dictionary available for the language. But for an out-of-vocabulary word, the lexical and pronunciation dictionary available for that particular language is used to give the phonemic representation of the word. In all our experiments E-speak Phonemizer has been used for converting the graphemes to phonemes.

Note: Neural network-based Text-to-Speech systems almost all the time use characters or phonemes as input features. So, a separate neural network to extract linguistic features from the characters or words is not required for the TTS system.

4.2 Acoustic Model

Acoustic models convert linguistic features into acoustic features. These acoustic features can be Mel Cepstral Coefficients (MCC), Line Spectral Pairs (LPS), Mel Generalized Coefficients (MGC), Pitch, Fundamental Frequency, and Mel-Spectrograms. But out of all these features, Mel-Spectrograms are widely utilized as the output of neural acoustic models. Different architectures have been used to build these acoustic models. Some popular architectures used for building these acoustic models are elaborated below:

1. RNN-based models :
   The Tacotron series is based on the RNN framework, i.e., an encoder-attention-decoder framework that takes characters as input and outputs Mel-spectrograms.

2. CNN-based models :
   DeepVoice ((Arik et al., 2017)) is a system that uses convolutional neural networks to obtain linguistic features, which are then used to generate waveforms. DeepVoice 2 ((Gibiansky et al., 2017)) is an improved version of DeepVoice that uses a more complex network structure and is able to model multiple speakers. DeepVoice 3 ((Ping et al., 2017)) is the most recent version of DeepVoice, and it uses a fully convolutional network to generate

mel-spectrograms from characters. ClariNet ((Ping et al., 2019a)) is a system that generates waveforms from the text in a fully end-to-end way. ParaNet ((Peng et al., 2019)) is a system that is similar to ClariNet but is faster and has better speech quality. DCTTS ((ho Kang et al., 2021)) is a system that uses a fully convolutional network to generate mel-spectrograms from character sequences.

3. Transformer-based Models :
   Tacotron 2 ((Shen et al., 2018)) model (which uses an RNN-based encoder and decoder) has two issues: 1) it can't be trained or run in parallel, which makes it inefficient, and 2) it's not good at modeling long dependencies. The TransformerTTS ((Li et al., 2018)) model (which uses a Transformer-based encoder and decoder) is similar to Tacotron 2 but doesn't have these issues. However, the Transformer-based model has its own issue of not being robust due to parallel computation. Some works have proposed ways to improve the robustness of the Transformer-based model. TransformerTTS, Tacotron, and DeepVoice series are auto-regressive in nature and hence have two major problems: 1) Slow inference speed as autoregressive generation of mel-spectrogram is slow. 2) Robustness, i.e., These autoregressive models have problems like word skipping and repetition due to inaccurate attention alignments between text and mel-spectrograms.Hence a non-autoregressive model called FastSpeech ((Ren et al., 2019)) is introduced which is a feed-forward Transformer network that generates mel-spectrograms in parallel. This parallel generation greatly speeds up inference. FastSpeech also removes the attention mechanism between text and speech to avoid word skipping and repeating issues and instead uses a length regulator to bridge the length mismatch between the phoneme and mel-spectrogram sequences. The length regulator uses a duration predictor to predict the duration of each phoneme and expands the phoneme hidden sequence according to the phoneme duration. This expanded phoneme hidden sequence can match the length of the mel-spectrogram sequence and facilitate parallel generation.

Apart from these architectures, other models are also there that are generating flow-based, VAE-based, GAN-based, and Diffusion-based models. In later sections, VITS TTS ((Kim et al., 2021)) will be discussed which is an end-to-end model that uses both Normalizing flows and VAE for acoustic modeling and performs adversarial learning for waveform generation.

4.3 Vocoder

This module takes the output of the acoustic model and converts it into a speech waveform. The input can be acoustic features or mel-spectrogram depending upon the acoustic model. Autoregressive generation of a waveform from mel-spectrograms is slow and therefore other methods like GAN, flow, and Diffusion-based models are used for waveform generation. These representative models are described below:

1. Autoregressive models:
   Wavenet ((van den Oord et al., 2016)) is the first neural-based vocoder, which leverages dilated convolution to generate waveform points autoregressively. WaveNet can be easily modified to condition on linear-spectrograms and mel-spectrograms, although the original WaveNet and certain subsequent efforts that use WaveNet as a vocoder generate speech waveform conditioned on linguistic features. The urge for a fast and lightweight vocoder arose as the Wavenet has a slow inference speed though the output speech quality was good. LPCNet ((Valin and Skoglund, 2018)) introduces conventional digital signal processing into neural networks and uses linear prediction coefficients to calculate the next waveform point while leveraging a lightweight RNN to compute the residual.

2. Flow-based:
   A generative model that transforms a probability density into standard/normal probability distribution using invertible transforms is called normalizing flow. Neural flow-based TTS can be classified based on autoregressive and bi-partite transforms. Examples of flow-based autoregressive vocoders include WaveNet ((van den Oord et al., 2016)) and bi-partite vocoders consisting of FloWaveNet ((Kim et al., 2018)) and WaveGlow ((Prenger et al., 2018)). WaveFlow ((Ping et al., 2019b)) offers the benefits of both autoregressive and bipartite transforms.

3. GAN-based:
   Generative adversarial networks (GANs) have been widely used in data generation tasks, such as image generation and text processing. A lot of vocoders leverage GAN to ensure audio generation quality, including WaveGAN ((Donahue et al., 2018)), MelGAN ((Kumar et al., 2019)), and HiFi-GAN ((Kong et al., 2020a)). The research efforts focus on how to design models to capture the characteristics of the waveform, in order to provide a better guiding signal for the generator. Multiple-scale discriminators, proposed in MelGAN ((Kumar et al., 2019)), use multiple discriminators to judge audio in different scales (different downsampling ratios compared with original audio). Multi-period discriminators can capture different implicit structures by looking at different parts of an input signal in different periods.Hierarchical discriminators are leveraged in VocGAN ((Yang et al., 2020)) to judge the waveform in different resolutions from coarse-grained to fine-grained. Other specific losses such as STFT loss and feature matching loss are also leveraged to improve performance.

4. Diffusion-based:
   Recently, various vocoder works, including DiffWave ((Kong et al., 2020b)), WaveGrad ((Chen et al., 2020)), and PriorGrad ((Lee et al., 2021)), have used denoising diffusion probabilistic models (DDPM or Diffusion). The basic idea is to use diffusion and reverse processes to formulate the mapping between data and latent distributions: in diffusion, a waveform data sample is gradually mixed with random noises until it becomes Gaussian noise; in reverse, random Gaussian noise is gradually denoised into a waveform data sample. Due to their lengthy iteration process, diffusion-based vocoders may produce speech with extremely high voice quality, but they struggle with sluggish inference speed. As a result, many studies on diffusion models focus on finding ways to shorten inference times without sacrificing generation quality.

4.4 Fully end-to-end TTS model

A fully end-to-end TTS system takes input as characters and directly generates the corresponding speech waveform. The advantages of this method are that it requires less human annotation and feature development, and can avoid error propagation. However, the main challenge of this method is the different modalities between text and speech waveform, as well as the huge length mismatch between character/phoneme sequence and waveform sequence. The experiments presented in this report are performed using VITS TTS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text to Speech) model mentioned in (Kim et al., 2021). Given below is a detailed description of the model architecture:
The suggested model's overall architecture is comprised of a posterior encoder, a prior encoder, a decoder, a discriminator, and a stochastic duration predictor. The posterior encoder and discriminator are only employed for training and never for inference. The normal posterior distribution's mean and variance are generated by the linear projection layer over the blocks.
The prior encoder is comprised of a text encoder that processes the input phonemes and a normalizing flow that increases the prior distribution's flexibility. We may derive the hidden representation from input phonemes by using the text encoder and a linear projection layer above the text encoder that generates the prior distribution's mean and variance. For the sake of simplicity, the normalizing flow is designed as a volume-preserving transformation with a determinant of one.
Essentially, the decoder is a HiFi-GAN generator from (Kong et al., 2020a). It consists of a stack of transposed convolutions that are each followed by a multi-receptive field fusion module (MRF).
The stochastic duration predictor calculates the phoneme duration distribution based on the conditional input i.e. phonemes. Residual blocks are stacked with dilated and depth-separable convolutional layers for the efficient parameterization of the stochastic duration predictor.

5 Expressive TTS systems

The goal of TTS systems is to produce natural and intelligible speech. The expressiveness of the generated speech determines the naturalness of the speech. Elements like prosody, content, timbre, and style are essential to synthesizing an expressive speech. Expressive speech synthesis is a one-to-many mapping problem, i.e., given a text, there are many factors the expressiveness of the synthesized speech depends on, like pitch, duration, loudness,

style, and emotion. Giving this variation information as input to the TTS model and effectively modeling them will allow synthesizing expressive speech comparable to human quality. According to (Tan et al., 2021), this variation information can be classified into four types:

- Text information
- Speaker or timbre information
- Prosody, style, and emotion information
- recording devices or noise environments information.

For storytelling, the text and prosody information is more relevant than information on the speaker, timbre, or noise environments.

5.1 Text Information as a variation

Text information is the main content that tells the TTS system "What to say?". Characters or phonemes are passed as text information to the TTS system. Some works have explored the ideas of representation of text using word embeddings and pre-training to improve the expressiveness of speech systems. In (Fang et al., 2019) the authors passed the text to tacotron2-encoder and Bert ((Devlin et al., 2019)) simultaneously and concatenated the resulting representations to feed it into the decoder of Tacotron2. The authors have done a preliminary study and found that while the quality of the synthesized audio is not significantly improved, the model does converge faster during training and produces less babbling at the end of the synthesis. (Hayashi et al., 2019) proposed a model for text-to-speech synthesis that uses information from pre-trained text embeddings. The model is designed to improve the quality of synthesized speech. The text embeddings contain information about the meaning of the text and the importance of each word, which should help the text-to-speech system produce more natural prosody and pronunciation. The proposed model is evaluated using the LJSpeech corpus, and it is found to improve the quality of the synthesized speech. The paper proposes two models- a subword-level model and a phrase-level model. Both models have three neural networks- a text-context extraction network, a spectrogram prediction network, and a vocoder network. The text-context extraction network extracts contextual embeddings from an input text. The spectrogram prediction network is a seq2seq model that generates log mel-spectrogram features from the inputs of a sequence of characters and the output of the text-context extraction network. The vocoder network is a deep convolutional neural network that generates a waveform from given acoustic features. According to (Xiao et al., 2020) there are two key factors for expressive Chinese speech synthesis, the first type of factor is a linguistic feature. Linguistic features can provide a lot of contextual information that can be helpful in improving a Chinese text-to-speech system. These linguistic features are grouped into two categories: phoneme-related and prosody-related. Phoneme-related features include the ID of the current phoneme and the two phonemes before and after it. Prosody-related features include the break tag, predicted by a CRF model, and the Chinese tones. (Jia et al., 2021) introduces PnG BERT, an augmented BERT model that can be used as a drop-in replacement for the encoder in typical neural TTS models, which takes advantage of both phoneme and grapheme representation, as well as by self-supervised pre-training on large text corpora to better understand natural language in its input. This has been used by Shyaam in his dual degree thesis on Text-to-Speech synthesis of Indian languages at CFILT lab. PnG BERT takes in two segments, a phoneme sequence, and a grapheme sequence, and outputs a hidden state for each token in the phoneme sequence. The model is trained using a self-supervised method, where the model is given a corpus of text and is then tasked with predicting masked tokens in the text. The model can be fine-tuned for use in a neural TTS model by freezing the weights of the lower layers and only training the higher layers.

5.2 Prosody, style, and emotion as variation information

This information answers the question "How to say a text?". This information is represented primarily by intonation, stress, and rhythm of the speech. Some works deal with prosody transfer using a reference clip or speaking style modeling and transfer. Since the current problem is of storytelling speaking style which is a specific type of style of speaking, we concentrate more on modeling prosody from text instead of style transfer or modeling. (Skerry-Ryan et al., 2018) proposes an extension to the Tacotron speech synthesis architecture that allows it to better match the prosody

of a reference signal, even when the reference and synthesis speakers are different. The extension involves learning a latent embedding space of prosody from a reference acoustic representation. A bank of embeddings called "global style tokens" ((Wang et al., 2018)) is jointly learned in Tacotron. Although they are not explicitly labeled during training, embeddings learn to reflect a wide variety of acoustic expressiveness. Independent of the text content, the soft interpretable "labels" they produce can be utilized to govern synthesis in unique ways, like pace and speaking style. The Text-Predicted Global Style Token (TP-GST) architecture, which interprets GST combination weights or style embeddings as "virtual" speaking style labels within Tacotron, is introduced in (Stanton et al., 2018). The ability of TP-GSTs to synthesize speech with background noise eliminated is further demonstrated, and findings on human-rated listener preference for audiobook tasks that support these analyses are encouraging.

5.3 Storytelling speaking style in TTS systems

Research in the area of storytelling speaking style in TTS systems is very sparse. A larger problem, namely style transfer, and modeling are actively being pursued but this report talks about a simpler problem of narrating a story. Concatenative synthesis has been used by some researchers to create story-telling speaking style systems. (Sarkar et al., 2014) presents a rule-based system that provides prosody rules to convert neutral speech into a story-telling speaking style. The neutral TTS system is a syllable-based TTS system trained on neutral speech corpus. Various prosodic characteristics are modified to generate a story-telling speaking style from the original neutral speech. These parameters include pitch contour, duration patterns, intensity patterns, pause patterns, and tempo. Three Indian languages—Bengali, Hindi, and Telugu—have unique rule sets created for each of these prosodic criteria. The rule sets are created by comparing the perception of synthetic neutral speech utterances to the corresponding naturally spoken utterances, as told by a storyteller. In (Sarkar and Rao, 2015) The authors try to model and evaluate the pause pattern in order to collect story semantic data. The goal of the essay is to specify a first step in creating a Story TTS based on types of discourse. For each style of discourse — narrative, descriptive, and dialogue — the author based their study on the pauses in Hindi children's stories. The story-semantic information is then gathered by analyzing the pause pattern after the sentences have been grouped into modes. For each mode, a three-stage data-driven approach is given to forecasting where and how long pauses would last.

6 Summary

The survey paper provides an in-depth analysis of the role of storytelling speaking style in text-to-speech (TTS) technology. The paper highlights the importance of incorporating prosodic features, such as intonation, rhythm, and emphasis, in order to convey meaning and emotion in spoken language. The paper goes on to discuss the challenges associated with capturing the unique vocal characteristics and speaking style of human narrators, as well as how these challenges can be addressed through the use of advanced machine learning algorithms and neural network architectures. Moreover, the paper extensively discusses several state-of-the-art expressive TTS models and different TTS datasets. The paper emphasizes the exciting possibilities of TTS technology for enhancing the quality and impact of spoken language in a wide range of applications, from entertainment to education and beyond. The use of storytelling speaking style TTS has the potential to create highly personalized and engaging audio content that meets the diverse needs of modern audiences, while also navigating ethical considerations and delivering a seamless listening experience.

References

Sercan Ö. Arik, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. 2017. Deep voice: Real-time neural text-to-speech. *ArXiv*, abs/1702.07825.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Thomas Dixon. 2003. *From Passions to Emotions: The Creation of a Secular Psychological Category*. Cambridge University Press.

Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis.

Wei Fang, Yu-An Chung, and James R. Glass. 2019. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. *ArXiv*, abs/1906.07307.

Andrew Gibiansky, Sercan Ö. Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In *NIPS*.

Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu. 2019. Pre-trained text embeddings for enhanced text-to-speech synthesis. pages 4430–4434.

Min ho Kang, Jihyun Lee, Simin Kim, and Injung Kim. 2021. Fast dctts: Efficient deep convolutional text-to-speech. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7043–7047.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. Png bert: Augmented bert on phonemes and graphemes for neural tts. In *Interspeech*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *CoRR*, abs/2106.06103.

Sungwon Kim, Sang-gil Lee, Jongyoon Song, and Sungroh Yoon. 2018. Flowavenet : A generative flow for raw audio. *CoRR*, abs/1811.02155.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *CoRR*, abs/2010.05646.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020b. Diffwave: A versatile diffusion model for audio synthesis.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. 2021. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and M. Zhou. 2018. Close to human quality tts with transformer. *ArXiv*, abs/1809.08895.

Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. 2019. Parallel neural text-to-speech. *ArXiv*, abs/1905.08459.

Wei Ping, Kainan Peng, and Jitong Chen. 2019a. Clarinet: Parallel wave generation in end-to-end text-to-speech. *ArXiv*, abs/1807.07281.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ö. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: 2000-speaker neural text-to-speech. *ArXiv*, abs/1710.07654.

Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song. 2019b. Waveflow: A compact flow-based model for raw audio. *CoRR*, abs/1912.01219.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2018. Waveglow: A flow-based generative network for speech synthesis.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *ArXiv*, abs/1905.09263.

Parakrant Sarkar, Arijul Haque, Arup Kumar Dutta, Gurunath M. Reddy, M. D. Harikrishna, Prasenjit Dhara, Rashmi Verma, P. N. Narendra, B. Kr S. Sunil, Jainath Yadav, and K. Sreenivasa Rao. 2014. Designing prosody rule-set for converting neutral tts speech to storytelling style speech for indian languages: Bengali, hindi and telugu. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 473–477.

Parakrant Sarkar and K. Sreenivasa Rao. 2015. Analysis and modeling pauses for synthesis of storytelling speech based on discourse modes. In *2015 Eighth International Conference on Contemporary Computing (IC3)*, pages 225–230.

Harold Schlosberg. 1941. A scale for the judgement of facial expressions. *Journal of Experimental Psychology*, 29:229–237.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Robert A. J. Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *ArXiv*, abs/1803.09047.

Daisy Stanton, Yuxuan Wang, and R. J. Skerry-Ryan. 2018. Predicting expressive speaking style from text in end-to-end speech synthesis. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 595–602.

Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *ArXiv*, abs/2106.15561.

Jean-Marc Valin and Jan Skoglund. 2018. Lpcnet: Improving neural speech synthesis through linear prediction.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*.

Yujia Xiao, Lei He, Huaiping Ming, and Frank K. Soong. 2020. Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Jinhyeok Yang, Junmo Lee, Youngik Kim, Hoonyoung Cho, and Injung Kim. 2020. Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network.

Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*.

Hao Zhang, Richard Sproat, Axel H. Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Comput. Linguist.*, 45(2):293–337.