Neural Synthesis of Expressive and Emotional Speech: a Survey

Pavan Kalyan and **Pushpak Bhattacharyya**
Indian Institute of Technology Bombay, India
190020124@iitb.ac.in, pb@cse.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 various neural network architectures. The paper extensively covers state-of-the-art expressive TTS models and different TTS datasets. In the context of emotional speech synthesis, this paper summarizes controllable emotional speech synthesis. We summarize the idea of task arithmetic that has been shown to be useful in steering the behaviour of neural models for NLP and vision tasks. 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 and emotional 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. Then we discuss about emotional speech synthesis and the application of task arithmetic for editing models at inference. We explore different ways of controlling the emotions of generated speech using controllable emotional TTS systems.

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. Another interesting problem this survey talks about is including and controlling the emotions expressed in speech for each sentence.

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

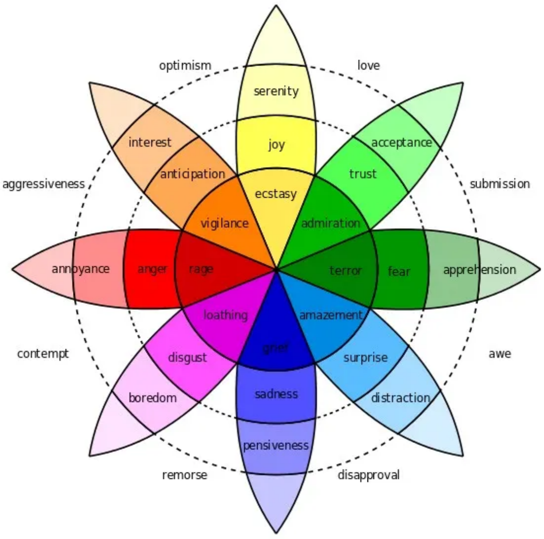

Figure 1: Plutchik's wheel of emotions

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.

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

Corpus	Domain	#Hours	#Spk	f_s(kHz)
ARCTIC	Read speech	7	7	16
VCTK	Read Speech	44	109	48
Blizzard-2011	Audiobook	16.6	1	16
Blizzard-2013	Audiobook	319	1	44.1
LJSpeech	Audiobook	25	1	22.05
LibriSpeech	Audiobook	982	2484	16
LibriTTS	Audiobook	586	2456	24
VCC 2018	Read speech	1	12	22.05
HiFi-TTS	Audiobook	300	11	44.1
CALLHOME	Conversational	60	120	8
RyanSpeech	Conversational	10	1	44.1

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

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., 2019a)). 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., 2021a)) 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.

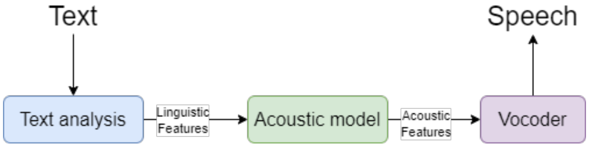

Figure 2: The three key components of TTS

4. Diffusion-based:
   Recently, various vocoder works, including DiffWave ((Kong et al., 2020b)), WaveGrad ((Chen et al., 2020b)), 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

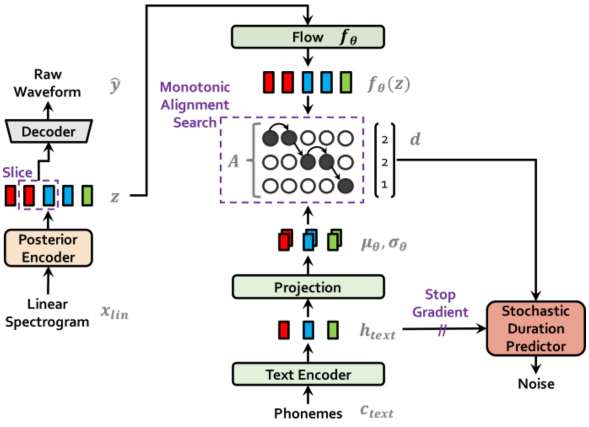

Figure 3: Training pipeline for VITS model

Speech) model mentioned in (Kim et al., 2021a). 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.

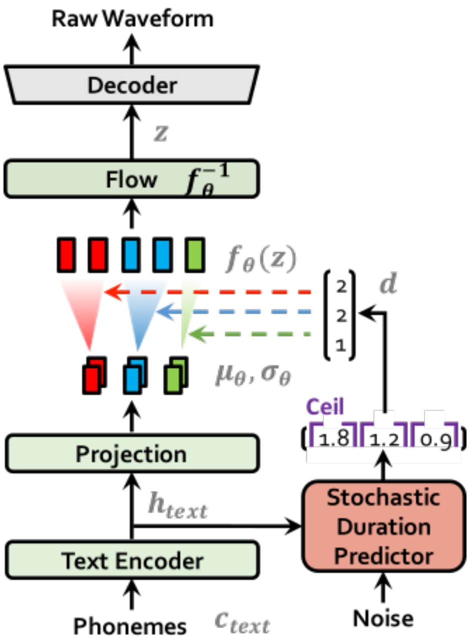

Figure 4: VITS inference piepline

5 Expressive TTS

The investigation of expressiveness in Text-to-Speech (TTS) encompasses a wide range of subjects such as modeling, disentanglement, control, and transfer of elements like content, timbre, prosody, style, and emotion, among others. One of the fundamental aspects in achieving expressive speech synthesis lies in effectively addressing the issue of one-to-many mapping. This concept pertains to the existence of numerous speech variations associated with a single text, encompassing factors such as duration, pitch, sound volume, speaker style, and emotion. Attempting to model the one-to-many mapping using the standard L1 loss (Gazor and Zhang, 2003) without adequate input data can result in excessive smoothing of mel-spectrogram predictions (Takamichi et al., 2016). For instance, this may involve predicting the average mel-spectrograms across the dataset instead of capturing the nuanced expressiveness of each individual speech utterance. Consequently, this approach leads to the production of low-quality, less expressive speech output. Hence, it is crucial to include these variations as input data and enhance the modeling of such variations to address this issue and enhance the expressiveness of the synthesized speech. Moreover, through the inclusion of variation information as input, it becomes possible to disentangle, regulate, and shift the variation information. Firstly, through the modification of these variation details (including specific speaker characteristics like timbre, style, accent, speaking speed, etc.) during inference, we gain the ability to regulate the produced speech. Secondly, by supplying the variation information corresponding to a different style, we can transform the voice into that particular style. Lastly, for the purpose of attaining precise voice regulation and transformation, it is essential to disentangle various types of variation information, such as content and prosody, timbre and noise, among others.

The information needed to synthesize a voice can be categorized into following types:

1. Text information: This can be characters or phonemes that represents the content of the synthesized speech (i.e., what to say). Some works improve the representation learning of text through enhanced word embeddings or text pre-training (Xiao et al., 2020; Jia et al., 2021), aiming to improve the quality and expressiveness of synthesized speech.

2. Speaker information or timbre data embodies the unique characteristics of speakers (i.e. how they sound). Some multi-speaker text-to-speech (TTS) systems use methods such as speaker lookup tables or speaker encoders to explicitly capture these properties (Moss et al., 2020; Chen et al., 2020a).

3. Prosody, style, and emotion information, which covers the intonation, stress, and rhythm of speech and represents how to say the text (Wagner and Watson, 2010). Prosody/style/emotion is the key information to improve the expressiveness of speech and the vast majority of works on expressive TTS focus on improving the prosody/style/emotion of speech (Um et al., 2019; Sun et al., 2020a).

4. Recording devices or noise environments, which are the channels to convey speech, and are not related to the content/speaker/prosody of speech, but will affect speech quality. Research works in this area focus on disentangling, controlling, and denoising for clean speech synthesis

6 Voice modulation in storytelling TTS

Though there has been a lot of work on neural speech synthesis, expressive neural speech synthesis is constrained by data scarcity. Since this work concentrates on a very specific type of expressive speech called storytelling speech synthesis, only very few works have appeared in the literature. This section contains brief descriptions of all such recent works for storytelling speech synthesis.

Previous works, such as Greene et al. (2012) make an attempt to predict and apply the suitable character voice to a text-to-speech (TTS) system for storytelling. However, these works solely rely on objective tests to evaluate the match between the character voice and the retrieved voice, without conducting any subjective tests on the TTS outputs. Using distinct and appropriate synthetic voices for characters in a children's story can enhance engagement and comprehension. This paper presents a data-driven approach for predicting suitable voices based on character attributes, using Mechanical Turk for labelling and Naive Bayes for modelling. The system performs well in the objective evaluation of speaker voice prediction, showing the effectiveness of the approach. In contrast, another study (Xin et al., 2023) focuses on synthesizing speech with enhanced prosody for audiobooks. This study takes into account both the acoustic and textual contexts. However, it should be noted that the dataset used in this study is a multi-speaker Japanese audiobook TTS dataset (Takamichi et al., 2022), which differs from single-speaker storytelling speech.

Moving on, Nakata et al. (2022) explores character acting in Japanese audiobooks. The authors predict character-appropriate voices by utilizing character embeddings derived from the character's name, conversational sentences, and surrounding characters. This paper presents a speech-synthesis model for predicting appropriate voice styles based on character-annotated text for audiobook speech synthesis. The goal is to produce distinctive voices for different characters in an audiobook. The proposed model involves character-acting-style extraction and style prediction from quotation-annotated text, enabling the automated creation of audiobooks with character-distinctive voices. Subjective evaluations indicate that the proposed model generates more distinctive character voices while maintaining the naturalness of synthetic speech. However, the sample audio does not fully capture the ground truth in terms of expressiveness, even though they make an effort to mimic the character's voice. Furthermore, the authors do not provide any results on character voice consistency, which is a crucial aspect of storytelling speech.

Another work, known as Kato et al. (2020) concentrates on synthesizing Rakugo speech, a form of comic storytelling that only includes character dialogues and not narrator sentences. Using Tacotron 2 and enhancements, the authors aimed to model rakugo speech and measure its quality compared to professional performances. While the synthesized speech did not reach the professional level, the study highlighted the importance of not only naturalness but also character distinguishability and content understandability for audience entertainment. The authors develop a database and annotate the character descriptions based on the conversation. However, in the case of storytelling speech, the character descriptions are derived from the stories themselves. Additionally, storytelling speech requires controllability in expressiveness, particularly when it comes to the narrator's text compared to the character's text. Lastly, there is a work referred to as the Kalyan et al. (2023) that presents a single-speaker English Storytelling TTS dataset, allowing for the transition of voice from the narrator to the character. In our work, we present a more expressive Hindi TTS dataset where the narrator modulates an average of 3-4 character voices in addition to the narration.

End-to-end TTS models, such as VAE (Zhang et al., 2019b) and GAN-based models (ShuangMa et al., 2019), have demonstrated the ability to generate high-quality speech using phoneme sequences and audio as input. While many TTS models can produce speech comparable to human speech, models utilizing GAN and Normalizing Flows (Aggarwal et al., 2020) have shown improved expressiveness (Ren et al., 2022). The paper (Kumar et al., 2023) performs an analysis of various kinds of neural TTS for Indian languages. Due to its competitive performance for Indian languages, we use VITS TTS (Kim et al., 2021b) in a multi-speaker setting. VITS, a non-auto-regressive TTS model, utilizes the Variational Auto-encoder architecture along with normalizing flows to model the prior distribution and employs a GAN pipeline to enhance voice quality.

7 Emotional TTS

Speech consists of both lexical (the words we use) and non-lexical (the way we say them) elements. While both convey emotions (Schuller and Schuller, 2020), research indicates that prosody—comprising aspects like tone, rhythm, and voice quality—is particularly crucial in expressing emotions through speech (Cowen et al., 2019). Emotional TTS methods can be broadly classified into three different categories based on the nature of their conditioning data. These categories include models that use:

1. Categorical labels to represent one or more emotions
2. Referenced speech with the desired emotional state
3. Textual descriptions of the emotional target state as a form of conditioning data.

The first approach is traditionally used when working with a labeled data set, as it facilitates the implementation of conditioning simply by introducing an embedded lookup table.

Modeling emotion in a text-to-speech task typically entails the development of a conditional model that emulates emotional speech based on text and emotion representation. The manner in which emotions are depicted plays a crucial role in determining the specific attributes of emotions that can be replicated by the model. In the study by (Lee et al., 2017), emotions are portrayed as distinct labels (e.g., joy, sadness). While this approach enables the model to replicate primary emotions by explicitly defining the input emotion label, it poses limitations on simulating more intricate characteristics such as emotion intensity and a blend of emotions. To address this challenge, some recent studies incorporate the aforementioned framework of emotions into their models. (Zhou et al., 2022) stands out as one of the pioneers in introducing the capability to replicate emotion intensity and secondary emotions through a rank-based emotion attribute vector. In the study conducted by (Tang et al., 2023), emotion is depicted as a vector embedding derived from a preexisting speech emotion recognition model.

This approach facilitates the replication of various characteristics through the integration of the hidden state of the embedding.

Emotion can be most effectively characterized through the utilization of explicit emotion labels (Lee et al., 2017; Tits et al., 2019), in which the model is trained to establish connections between labels and styles of emotion. (Lee et al., 2017) demonstrates the utilization of an emotion label vector by the attention-based decoder in order to generate the intended emotion. Similarly, in (Tits et al., 2019), a model adaptation approach is employed to construct a low-resourced emotional text-to-speech system with the incorporation of a limited number of emotion labels. Apart from the explicit labels related to distinct emotion categories, endeavors have been made to condition the decoder with continuous variables (Rabiee et al., 2019).

An alternative methodology involves utilizing a style encoder for the purpose of replicating and transferring the reference style (Skerry-Ryan et al., 2018). Global style token (GST) (Wang et al., 2017) serves as an illustration of acquiring style embeddings from the reference audio in an unsupervised fashion. Various research endeavors incorporate supplementary components such as emotion recognition loss (Cai et al., 2020), perceptual loss (Li et al., 2020), or adversarial training (Ma et al., 2018) to enhance the emotional expression. Subsequent studies (Cornille et al., 2022; Klimkov et al., 2019; Li et al., 2021; Zhang et al., 2020) opt to substitute the global style embedding with phoneme or segmental level prosody embedding to encompass emotion variations across multiple scales. Analogous methodologies have been extended to the domain of emotional voice conversion research. In (Zhou et al., 2021), the style encoder additionally functions as the emotion encoder to glean genuine emotion details via a two-phase training process. In (Choi and Hahn, 2021), a speaker encoder is introduced to safeguard the speaker attributes.

Controllable Emotional Speech Synthesis

Speech emotion is frequently expressed through various aspects of prosody (Maupomé and Isyutina, 2013). The manipulation of different prosodic cues can influence the expression of emotion. Current research (Lee and Kim, 2018; Tan and Lee, 2020) predominantly focuses on formulating the prosody embedding as a control vector derived from a framework of representation learning. For instance, style tokens (Wang et al., 2018) are specifically created to encode high-level styles such as speaker characteristics, pitch variability, and speech tempo. Emotion expression can be regulated by selecting particular tokens. Recent efforts (Sun et al., 2020a,b) explore incorporating a detailed, hierarchical prosody representation into the style token-based framework (Wang et al., 2018). Additionally, some studies utilize variational autoencoders (VAE) (Kingma and Welling, 2013) to regulate speech style through the acquisition, adjustment, or fusion of disentangled representations.

8 Model editing at inference - Task arithmetic

Although neural networks are inherently non-linear in nature, prior research has demonstrated through empirical studies that the process of interpolating between the weight configurations of two distinct neural networks can result in the preservation of their high level of accuracy. This phenomenon occurs when these two neural networks have overlapping segments within their optimization trajectory, indicating a convergence or similarity in the direction of optimization (Ilharco et al., 2022b; Wortsman et al., 2022; Fort et al., 2020).

An increasing body of research is currently delving into the exploration of utilizing interpolations between the weights of models and task arithmetic in order to manipulate and enhance the capabilities of pre-trained models. Specifically, numerous research studies have indicated that the process of interpolating between the fine-tuned weights of a model and its pre-trained initialization has the potential to result in enhanced performance on individual tasks, sometimes even surpassing the accuracies achieved through fine-tuning alone (Ram'e et al., 2022; Ramé et al., 2022; Wortsman et al., 2021). Within the context of multi-task scenarios, a proposed approach involves averaging the parameters of numerous fine-tuned models, aiming to create superior multi-task models (Wortsman et al., 2022; Li et al., 2022; Ilharco et al., 2022a) that are able to prevent catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999) and may even offer a more advantageous starting point for subsequent fine-tuning endeavors (Don-Yehiya et al., 2022; Choshen et al., 2022). Notably, the advantages associated with weight ensembles and interpolations are not limited to pre-trained models but also extend to models that are trained from scratch, provided that they are appropriately aligned prior to merging (Singh and Jaggi, 2019; Ainsworth et al., 2022).

(Achille et al., 2019) as well as (Vu et al., 2021) delved into various strategies for representing tasks through continuous embeddings, aiming to forecast task similarities and transferability, or establish taxonomic relations. Although the task vectors we construct could serve such purposes, our primary objective is to utilize them as instruments for guiding the behavior of pre-trained models. Furthermore, (Lampinen and McClelland, 2020) introduce a framework for adjusting models based on the interconnections between tasks.

In the context of fine-tuning, the precision demonstrates a consistent rise as the parameters of a pre-existing model are gradually adjusted towards its fine-tuned equivalent (Wortsman et al., 2021; Ilharco et al., 2022b). (Ilharco et al., 2022b) discovered that beyond a singular task, enhancing accuracy on fine-tuning tasks can be achieved by fine-tuning multiple models with different tasks but the same initialization and then averaging their weights. Similarly, (Li et al., 2022) observed analogous outcomes by averaging the parameters of language models fine-tuned across diverse domains. Through their research, (Choshen et al., 2022) demonstrated that amalgamating the weights of fine-tuned models through averaging can establish a more optimal starting point for fine-tuning on a subsequent task. Moreover, (Wortsman et al., 2021) revealed that aggregating the weights of models fine-tuned on various tasks leads to an enhanced accuracy when applied to a new downstream task, obviating the need for additional training.

In view of the fact that re-training models is typically cost-prohibitive, numerous scholars have investigated more resourceful approaches to adjusting a model's behavior through interventions post pre-training, denoting this procedure with various terms such as patching (Murty et al., 2022; Ilharco et al., 2022b), editing (Mitchell et al., 2022, 2021), aligning (Glaese et al., 2022; Askell et al., 2021), or debugging (Geva et al., 2022). Diverging from earlier scholarly works, our study introduces a distinctive method of modifying models, allowing for the addition or removal of capabilities in an efficient and modular fashion by leveraging fine-tuned models. A related study by (Subramani et al., 2022) delves into steering language models by incorporating vectors into their hidden states; however, our research involves the application of vectors in the weight space of pre-trained models, without altering the standard fine-tuning process.

9 Summary

This survey article introduces two facets of speech synthesis - Expressive and Emotional speech synthesis within the narrative storytelling framework. The study encompasses an overview of different Text-to-Speech (TTS) datasets and architectures designed for both English and Hindi languages. The exploration of emotive speech synthesis and the controllability in speech synthesis are detailed within this research. The summary of model editing during the inference phase is concisely presented to facilitate the application of task arithmetic for speech synthesis. In addition to providing insights into emotional and expressive TTS systems, this survey also briefly elucidates fundamental concepts such as emotions and expressions in speech.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6429–6438.

Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. 2020. Using Vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. 2022. Git re-basin: Merging models modulo permutation symmetries. *ArXiv*, abs/2209.04836.

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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova Dassarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, John Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861.

Xiong Cai, Dongyang Dai, Zhiyong Wu, Xiang Li, Jingbei Li, and Helen M. Meng. 2020. Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5734–5738.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, and Tao Qin. 2020a. Multispeech: Multi-speaker text to speech with transformer. *ArXiv*, abs/2006.04664.

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

Heejin Choi and Minsoo Hahn. 2021. Sequence-to-sequence emotional voice conversion with strength control. *IEEE Access*, 9:42674–42687.

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining. *ArXiv*, abs/2204.03044.

Tobias Cornille, Fengna Wang, and Jessa Bekker. 2022. Interactive multi-level prosody control for expressive speech synthesis. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8312–8316.

Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3:369 – 382.

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

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2022. Cold fusion: Collaborative descent for distributed multitask finetuning. *ArXiv*, abs/2212.01378.

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

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. 2020. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *ArXiv*, abs/2010.15110.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135.

Saeed Gazor and Wei Zhang. 2003. Speech probability distribution. *IEEE Signal Processing Letters*, 10:204–207.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. *ArXiv*, abs/2204.12130.

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*.

Amelia Glaese, Nathan McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, A. See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sovna Mokr'a, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William S. Isaac, John F. J. Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *ArXiv*, abs/2209.14375.

Erica Greene, Taniya Mishra, Patrick Haffner, and Alistair Conkie. 2012. Predicting character-appropriate voices for a tts-based storyteller system. In *Thirteenth Annual Conference of the International Speech Communication Association*.

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.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022a. Editing models with task arithmetic. *ArXiv*, abs/2212.04089.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022b. Patching open-vocabulary models by interpolating weights. *ArXiv*, abs/2208.05592.

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*.

T Pavan Kalyan, Preeti Rao, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. Narrator or character: Voice modulation in an expressive multi-speaker tts. *Proc. INTERSPEECH 2023*, pages 4808–4812.

Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, and Junichi Yamagishi. 2020. Modeling of rakugo speech and its limitations: Toward speech synthesis that entertains audiences. *IEEE Access*, 8:138149–138161.

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

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021b. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

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

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Viacheslav Klimkov, S. Ronanki, Jonas Rohnke, and Thomas Drugman. 2019. Fine-grained robust prosody transfer for single-speaker neural text-to-speech. *ArXiv*, abs/1907.02479.

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.

Gokul Karthik Kumar, SV Praveen, Pratyush Kumar, Mitesh M Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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.

Andrew Kyle Lampinen and James L. McClelland. 2020. Transforming task representations to perform novel tasks. *Proceedings of the National Academy of Sciences*, 117:32970 – 32981.

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.

Younggun Lee and Taesu Kim. 2018. Robust and fine-grained prosody control of end-to-end speech synthesis. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. 2017. Emotional end-to-end neural speech synthesizer. *ArXiv*, abs/1711.05447.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. *ArXiv*, abs/2208.03306.

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.

Tao Li, Shan Yang, Liumeng Xue, and Lei Xie. 2020. Controllable emotion transfer for end-to-end speech synthesis. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5.

Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia, and Helen M. Meng. 2021. Towards multi-scale style control for expressive speech synthesis. *ArXiv*, abs/2104.03521.

Shuang Ma, Daniel J. McDuff, and Yale Song. 2018. Neural tts stylization with adversarial and collaborative games. In *International Conference on Learning Representations*.

Gerardo Maupomé and Olga Isyutina. 2013. Dental students' and faculty members' concepts and emotions associated with a caries risk assessment program. *Journal of dental education*, 77:1477–87.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2021. Fast model editing at scale. *ArXiv*, abs/2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. *ArXiv*, abs/2206.06520.

Henry B. Moss, Vatsal Aggarwal, Nishant Prateek, Javier I. González, and Roberto Barra-Chicote. 2020. Boffin tts: Few-shot speaker adaptation by bayesian optimization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643.

Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Tulio Ribeiro. 2022. Fixing model bugs with natural language patches. In *Conference on Empirical Methods in Natural Language Processing*.

Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, Yuki Saito, Yusuke Ijima, Ryo Masumura, and Hiroshi Saruwatari. 2022. Predicting vqvae-based character acting style from quotation-annotated text for audiobook speech synthesis. In *Proc. Interspeech*, pages 4551–4555.

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.

Azam Rabiee, Tae-Ho Kim, and Soo-Young Lee. 2019. Adjusting pleasure-arousal-dominance for continuous emotional text-to-speech synthesizer. In *Interspeech*.

Alexandre Ram'e, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. 2022. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*.

Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. 2022. Diverse weight averaging for out-of-distribution generalization. *ArXiv*, abs/2205.09739.

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.

Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2022. Revisiting over-smoothness in text to speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8197–8213, Dublin, Ireland. Association for Computational Linguistics.

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

Dagmar M. Schuller and Björn Schuller. 2020. A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emotion Review*, 13:44 – 50.

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.

ShuangMa, Daniel McDuff, and Yale Song. 2019. Neural TTS stylization with adversarial and collaborative games. In *International Conference on Learning Representations (ICLR)*.

Sidak Pal Singh and Martin Jaggi. 2019. Model fusion via optimal transport. *ArXiv*, abs/1910.05653.

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.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. *ArXiv*, abs/2205.05124.

Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu. 2020a. Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6699–6703.

Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuanbin Cao, Heiga Zen, and Yonghui Wu. 2020b. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6264–6268.

Shinnosuke Takamichi, Wataru Nakata, Naoko Tanji, and Hiroshi Saruwatari. 2022. J-mac: Japanese multi-speaker audiobook corpus for speech synthesis. *arXiv preprint arXiv:2201.10896*.

Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. 2016. Postfilters to modify the modulation spectrum for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:755–767.

Daxin Tan and Tan Lee. 2020. Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement. In *Interspeech*.

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

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. *ArXiv*, abs/2306.00648.

Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2019. Exploring transfer learning for low resource emotional tts. *ArXiv*, abs/1901.04276.

Seyun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, Chung Hyun Ahn, and Hong-Goo Kang. 2019. Emotional speech synthesis with rich and granularized control. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7254–7258.

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.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Matthew Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *ArXiv*, abs/2110.07904.

Michael Wagner and Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25:905 – 945.

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

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*.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *ArXiv*, abs/2203.05482.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7949–7961.

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. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708.

Detai Xin, Sharath Adavanne, Federico Ang, Ashish Kulkarni, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2023. Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Guangyan Zhang, Ying Qin, and Tan Lee. 2020. Learning syllable-level discrete prosodic representation for expressive speech generation. In *Interspeech*.

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

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019b. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.

Kun Zhou, Berrak Sisman, and Haizhou Li. 2021. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. *ArXiv*, abs/2103.16809.

Kun Zhou, Berrak Sisman, Rajib Kumar Rana, B.W.Schuller, and Haizhou Li. 2022. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14:3120–3134.