Towards building controllable Text to Speech systems

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

K Saiteja
2020701004
saiteja.k@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Towards building controllable Text to Speech systems” by K Saiteja, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Vineet Gandhi

To my family, my guide and my friends

Acknowledgments

First, I would like to thank my advisor Prof. Vineet Gandhi, for the wonderful experience throughout my Masters, working towards my thesis. He provided me with freedom and flexibility to explore ideas worth spending time and helped me in formulating concrete ideas out of them. At every stage of the research, his ideas , discussions, constant encouragement motivated me to pursue, experiment, analyse some projects. Not only has he been my guide, he helped in shaping my research career by being a friend supporting and helping in shaping my career.

I want to thank my colleague Sarath Sivaprasad, for being a constant support in understanding and approaching any idea. Many hour long crazy discussions made this thesis fruitful. I would like to thank Dr Anil Nelakanti, for his guidance on the idea exploration. His ideas were the base of this thesis. I would also like to thank Niranjan Pedanekar, for helping me run some experiments which are crucial for our work. And lastly, I thank my fellow colleagues at CVIT, Jeet Vora and Ritam Basu for the great work environment they have created. All the discussions about research in general made me well verse with the new works and technologies happening around.

This work would not have been possible with the coursework at IIIT Hyderabad. I would like to thank Dr C.V.Jawahar, Dr Anoop Namboodari, Dr Pawan Kumar, Dr Manish Srivastava for the valuable lectures, assignments and projects which helped me build a base on my research profile. I would also like to thank Ashok Urlana, Neil Kumar shah, Neha Sherin, Vishal T, Amit Pandey, Chandradeep Pokhariya, Dhawal Sirikonda, Akash KT, Astitva Srivastava for making my time in Masters fun and memorable.

Finally, more important acknowledgement goes to my family. My Dad, Ma, Anna, Baabi and chinni for being a constant support throughout my professional and personal life. This thesis wouldn't have happened without their immense support when I decided to resign from a well established job and pursue Masters.

Abstract

Text-to-speech systems convert any given text to speech. They play a vital role in making Human-computer interaction (HCI) possible. As humans, we don't just rely on text (language) to communicate; we use many other mechanisms like voice, gestures, expressions, etc., to communicate efficiently. In natural language processing, vocabulary and grammar tend to take center stage, but those elements of speech only tell half the story. Affective prosody of speech provides larger context and gives meaning to words, and keeps listeners engaged. Current HCI systems largely communicate in text, and they lack a lot of prosodic information, which is crucial in a conversation. To make the HCI systems communicate in speech, text to speech systems should be able to synthesize speech that is expressive and controllable. But the existing text to speech systems learn the average variation in the dataset it's trained on, which synthesizes samples in a neutral way without many prosodic variations. To this end, we develop a text-to-speech system that can synthesize the given emotion where the emotion is represented as a tuple of Arousal, Valance and Dominance (AVD) values.

Text to speech systems have a lot of complexities. Training such a system requires the data to be very clear, noiseless, and collecting such data is difficult. If the data is noisy, it will reflect unnecessary artifacts in the synthesized samples. Training emotion based text to speech models is considerably more difficult and not strait forward. The fact that obtaining emotion annotated data for the desired speaker is costly and very subjective makes it a cumbersome task. Current emotion based systems can synthesize emotions with some limitations. (1) Emotion controllability comes at the cost of loss in quality, (2) Have discreet emotions which lack the finer control, and (3) cannot be generalized to new speakers without the annotated emotion data.

We propose a system that overcomes the above-mentioned problems by leveraging the largely available corpus of noisy speech annotated with emotions. Even though the data is noisy, our technique trains an emotion based text to speech system that can synthesize desired emotion without any loss of quality in the output. We present a method to control the emotional prosody of Text to Speech (TTS) systems by using phoneme-level intermediate variances/features (pitch, energy, and duration) as levers. We learn how the variances change with respect to emotion. We bring the finer control in the synthesized speech by using AVD values, which can represent emotions in a 3D space. Our proposed method also doesn't require emotion annotated data for the target speaker. Once trained on the emotion annotated data, it can be applied to any system which has the prediction of the variances as an intermediate step.

With thorough experimental studies, we show that the proposed method improves over the prior art in accurately emulating the desired emotions while retaining the naturalness of speech. We extend the traditional evaluation of using individual sentences for a complete evaluation of HCI systems. We present a novel experimental setup by replacing an actor with a TTS system in offline and live conversations. The emotion to be rendered is either predicted or manually assigned. The results show that the proposed method is strongly preferred over the state-of-the-art TTS system and adds the much-coveted ”human touch” in machine dialogue.

Contents

Chapter Page

List of Figures

Figure Page

List of Tables

Chapter 1

Introduction

Speech synthesis is the artificial production of speech. And the process of synthesizing speech from text is known as Text to speech (TTS). Synthesized speech is used in many applications of daily life, from railway announcements to automated telephone responses. Artificial speech synthesis has been an area of interest for a long time. Any science fiction movie is incomplete without a computer talking. For instance, Jarvis in Avenger series, TARS, and CASE in Interstellar are the perfect examples of speech synthesis. Often these movies synthesize the fake speech using an actor, and some processing is done to make the voice computerized. Speech synthesis has wider applications, it is often used as a communication mechanism to aid disabled persons, it can be widely used as a reading system for the blind, where a system would read a book and convert it into speech. Professor Stephen Hawking has been one of the main contributors to direct exposure to Speech synthesis. It is also used as a virtual assistant, like Alexa, Siri, Google home, Cortana, etc., which can perform some actions with voice commands and gives us a response back. Although TTS systems are popular in recent times, feedback on TTS was not good in the past due to its quality. Apart from blind people using it for accessibility, other applications rarely use synthesized speech. With recent advancements in the quality of the synthesis, TTS systems are deployed in many applications. The main application of a TTS system is probably in call center automation, where there are predefined automated responses like paying mobile bills, delivery status, or conducting the entire transaction through an automatic dialogue system.

1.1 Communication and Language

Communication is crucial for the survival of any being. Even animals communicate by making sounds, nodding, or making gestures in a way that other animals can interpret the signs. Communication can be thought of as the process of creating information and sharing such that other beings can interpret, understand and respond accordingly. Communication can be done in many ways. Broadly it can be classified into three types, a. **Affective Communication**: expressing with external means like pain or crying for an injury. b. **Iconic Communication**: using icons that are meant to convey the intended meaning, like pictures, gestures, or sounds. And lastly, c. **Symbolic Communication**: where we predefine some symbols to communicate and learn how to interpret them before we communicate. For instance, language is one symbolic communication. Language consists of rules, grammar, and semantics. To communicate, you need to learn the language. Human communication mainly happens in Symbolic Communication. From childhood, we learn these symbols, their meaning, and how to use the combination of these symbols to effectively communicate. Spoken communication can happen in two ways, verbal communication, and prosody communication. Verbal communication is a way of expressing using words or a combination of words in larger sentences. In contrast, prosodic communication is the way of expressing emotion, intent, emphasis, style, and so on. Prosody doesn't contain any symbols and thus cannot be represented using discrete units. But prosody is crucial for human communication. Prosodic information doesn't have to be dependent on verbal communication; for example, if someone says "It's wonderful" in a gloomy way, although the verbal information gives a positive sense, it could be inferred that the person is being sarcastic.

1.2 What is a Text to speech (TTS)?

A Text To Speech (TTS) is a system that converts text to synthesized speech. This technology can be used as an assistive mechanism to communicate with users for whom reading something on a screen is either inconvenient or not possible. A computer system can be implemented in hardware or software to create a speech for normal language represented in the text. This system can be used to read out the content on the screen and make many applications more accessible to people who can't read. A TTS system is generally composed of two parts, front-end, and back-end. Front-end is referred to the tasks of preprocessing the text in such a way the computer understands, and Back-end is referred to the speech synthesis given the text. Developing a TTS system involves understanding languages and human speech production and involves multiple disciplines, including linguistics [15], acoustics [50], digital signal processing [126], and machine learning [6].

1.2.1 Challenges

Over the years, developing a quality text to speech synthesizer has been a challenging task. A good TTS should synthesize speech that is both natural and intelligible. Naturality refers to how close the voice is to the human voice, and intelligibility refers to if the voice is easily understandable. Representing text in itself is a challenging problem. Text normalization and converting text to units that any model can understand are not straightforward. The challenges are detailed below.

Text Normalization: Any TTS system takes natural language as input. But normalizing natural language, as in representing language in common form, is very difficult. For example, consider two sentences, (a) Give me a *minute* (b) Don't miss the *minute* details of the project. The word *minute*, even though has the same letter sequence, in two sentences it has different meanings and different pronunciations. A more difficult problem is natural language content often not only contains text but numbers, dates, times, currencies, abbreviations, etc. Converting all types of text into standard forms is difficult. But often, this is possible by considering based on the context. In other scenarios, there is no standard way of rendering some words; for example, an email address john242@gmail.com can be spoken as ”john two four two at the rate gmail dot com” or ”john two forty-two at the rate gmail dot com.”

Naturalness: The ultimate goal of any TTS is to make the system sound as natural as possible, which means closer to a human voice and can speak like all the human variations. Building such a TTS system has been the utmost challenge over the years. Many TTS systems were able to create robotic voices without many artifacts, but many studies proved that people prefer speech closer to humans than robotic speech. Also, getting rid of some noise artifacts like pops, clicks, buzzes, or any other mechanical sounds is also not straightforward if we are working on data that have these artifacts. Other than the noise artifacts, variations in the speech can also contribute to the naturalness of the speech. A particular phoneme has a particular variability depending on the phonemic context, position in the sentence, and supra-segmental influence; getting this variability is probably one of the key tasks. Considerable efforts have been made to increase the naturalness of the synthesized speech.

Intelligibility: Intelligibility can be defined as the ability of the listener to decode and understand the message from the speech. Often intelligibility is measured by the capability to decode the message since measuring the understanding ability is subjective and difficult. This can be considered as the easiest challenge of all the difficulties. In fact, the synthesizers developed in older days like [1] has reasonable intelligibility compared to modern TTS systems, given that proper text normalization is done. Even though the intelligibility of these systems is good, they sound quite unnatural. So, achieving intelligibility along with naturalness can be a challenging problem to fix.

Evaluation Challenges: The evaluation of the speech synthesis systems is not easy due to the lack of universally agreed objective evaluation criteria. A few subjective techniques like Mean Opinion Score (MOS), and Mushra are used to assess the naturalness of the synthesized outputs as compared to the original recordings. To access intelligibility, people have used the Word error rate (WER) and Phoneme error rate (PER) from a common ASR (Automatic Speech Translation) model on the original and the synthesized samples. Although these evaluation techniques are used widely, there are no commonly accepted criteria because people work on recordings from various speakers, accents, languages, and regions.

1.2.2 Terminology

In this section, we describe common terminology used in this manuscript.

- **Characters:** A Character is a semiotic sign or symbol, typically a letter, numerical digit, punctuation, or ideogram. Characters can represent the natural language.

- **Phonemes:** Principle unit of sound is called a phoneme. It is the smallest unit of sound. A phoneme can distinguish one word from another in a particular language. They include vowels

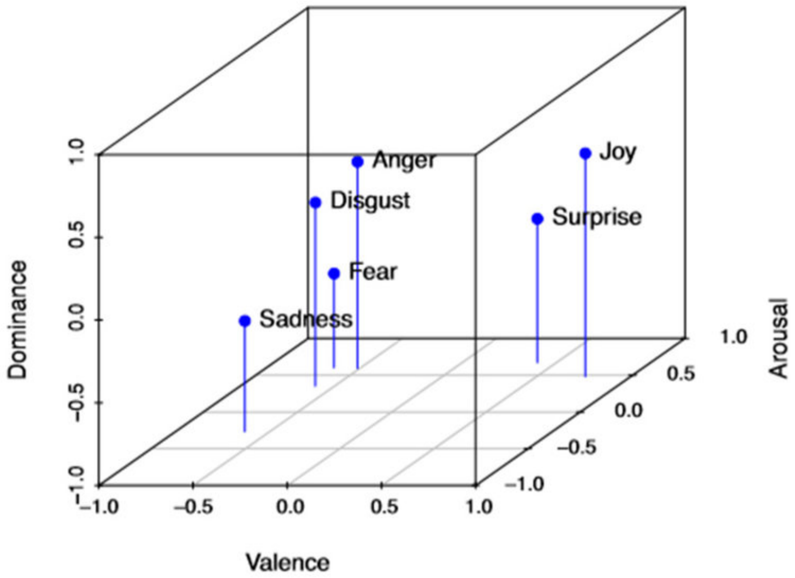


Figure 1.1: VAD space representing emotions in 3D space [4]

and consonants. Phonemes can work in a contrastive system of sounds, in which a fixed number of phonemes can combine to form a large number of words. In turn, these words can combine to form longer sentences. For example, the English language has around 44 phonemes but a longer word dictionary. Different languages have different sets of phonemes.

- **Syllables:** Syllable is a basic functional unit of speech. Syllables are structural sound units used to group phonemes. In general, syllables are a cluster of phonemes with at least one vowel. Syllables form words. The number of syllables in a word is unrelated to the number of phonemes.

- **Spectrogram:** A spectrogram is a visual representation of a spectrum of frequencies for any time-varying signal. As shown in Figure 1.2.2, the vertical axis of the spectrogram represents frequency, the horizontal axis represents time, and the color values represent the amplitude of the signal at that frequency. By viewing the spectrogram of a spoken sentence, professionals can roughly infer what phonemes are spoken. They also contain prosodic information.

- **Mel-spectrogram:** A mel spectrogram is a spectrogram where the frequencies are converted to mel scale. Studies have shown that humans cannot perceive frequencies on a linear scale. We are better at distinguishing variations in lower frequencies than higher frequencies. A mel scale is a unit of pitch such that equal distances in pitch sound are equally distant to the listener. A simple mathematical formula can be used to convert from linear to mel scale.

- **F0:** F0 or the fundamental frequency is the frequency at which the vocal cords vibrate in voiced sounds.

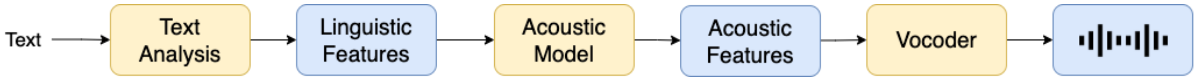


Figure 1.2: Three important components of Text to speech systems.

- **Duration:** In this manuscript duration of a phoneme is referred to as the length of the corresponding phoneme spoken in the speech.

- **Prosody:** Prosody is the component of speech that are not individual phonetic segments but are properties of syllables, words, or sentences like intonation, stress, rhythm, emotion, pause, tempo, etc. Prosody tells us the features of the speaker or the utterance, like if it's a question or command, its emotion, sarcasm or irony, emphasis.

- **Arousal:** Emotional arousal is the intensity of the emotion. It tells us how strong or weak is particular emotion is. For example, anger has positive arousal, and sadness has negative arousal. Arousal is represented on Y-axis in the figure 1.1.

- **Valence:** Emotional valence is the polarity of emotion. Positive or negative emotion. For example, happy and excited has a positive valence, whereas sad and angry have a negative valence. As shown in figure 1.1, valence is represented in the x-axis.

- **Dominance:** Dominance is the degree of control exerted by the stimulus. It represents controlling and dominant vs. controlled or submissive. For example, anger and fear are negative emotions with high arousal, whereas anger is dominant emotion while fear is submissive emotion. Dominance is shown on Z-axis in figure 1.1.

- **Vocoder:** A vocoder is a system that converts acoustic information of the audio signal to the audio waveform. Often, a TTS system is developed as a model from text to acoustic features (Mel spectrogram). Later vocoder is used to convert the mel spectrogram to the audio waveform.

1.3 Key components in TTS

As shown in Figure 1.2, TTS consists of three main components, a text analysis model, which extracts linguistic features from raw text, and an acoustic model, which converts linguistic features to acoustic features. And a vocoder that converts acoustic features to raw audio.

1.3.1 Text Analysis

Text analysis, often called ”Fronted” in TTS, converts input text to linguistic features which contain rich information like pronunciation and prosody of the text easing the speech synthesis process. The same words can often be spoken in many ways. Determining how the pronunciation should be is not strait forward. Many techniques have been developed by linguists to convert text to phonemes or grapheme to phonemes. Text analysis generally contains several functionalities like text normalization, word segmentation, parts of speech tagging, pronunciation prediction, and prosody extraction.

- **Text normalization** is the process of converting the raw text to a common format, for example, digits to words, and is the hardest problem in text analysis. For example, the year ”2022” has to be converted to ”two thousand and twenty-two”. Some works use rule-based text normalization technique [110]. Later neural network-based systems have been used to model text normalization as a sequence-to-sequence task [111, 135]. Recent works which use both rule-based and neural network-based text analysis have also been proposed to further improve the text normalization performance. Once the standard word format is obtained from character input using text normalization, grapheme-to-phoneme conversion is applied to convert the text to a sequence of phonemes.

- **Word Segmentation** is the process of finding word boundaries. For some character-based languages like Chinese and Japanese, where there are no word boundaries, the model needs to understand where the words are ending, which is also necessary for grapheme-to-phoneme conversion, prosody prediction, etc.

- **POS tagging** is the process of tagging parts of speech for each word like noun, verb, adjective, etc. This is important for grapheme-to-phoneme conversion and prosody predictor, as in, depending on the parts of speech tag, the pronunciation of the word may differ.

- **Prosody prediction** model predicts the prosody information like pitch, duration, and loudness of speech, which captures the rhythm, stress, and intonation of speech. The prosody is important in speech production because the wrong prosody can lead to totally unnatural speech.

- **Grapheme to Phoneme** models convert characters to phonemes which can ease speech synthesis. After the text normalization, the standard word format is phonemes. A manually collected vocabulary called lexicon is usually used for conversion, although in English, there will be many out-of-vocabulary words for which the grapheme-to-phoneme conversion is mainly built.

1.3.2 Acoustic Models

An acoustic model converts the linguistic features to acoustic features, which are responsible for speech generation. Various acoustic features like mel-cepstrum coefficients (MCC), mel spectrogram, mel-generalized coefficients (MGC), fundamental frequency (f0), band aperiodicity (BAP), bark frequency cepstral coefficients (BFCC), etc., are used in different models. Out of these, mel-spectrograms are used widely in modern neural TTS systems. The acoustic models are first developed with an HMM-based parametric model converting to acoustic features from linguistic features. Later, the neural TTS acoustic model is developed, a sequence-to-sequence learning model, where the source sequence is text (character/phoneme) and the target sequence is mel spectrogram frames.

This sequence-to-sequence mapping can be learned using various architectures. An encoder-attention-decoder was proven to work in most of scenarios. A basic LSTM [122, 106], CNN [2, 85], self attention [63] and the recent feed-forward networks with CNN's and self attention [98, 97] are used for mel spectrogram generation either from characters or phonemes.

1.3.3 Vocoders

Vocoders are systems that convert either intermediate features to the audio waveform. These intermediate features can be either linguistic features or acoustic features. In parametric synthesis, acoustic features like mel-cepstral coefficients, band aperiodicity, and f0 are extracted from the speech and used to train a model which generates the audio waveform back given these features [46]. In neural-based TTS systems, mel-spectrogram is used to condition the generation of the audio waveform. Early works like Wavenet [81], Char2Wav [109], WaveRNN [42] directly take linguistic features and generate the audio waveform. Later, a few methods were proposed to generate audio waveform conditioning on mel spectrogram, which has most of the information required for audio generation. WaveGlow [92], MelGan [60], Hifi-GAN [55] and several other flow-based, GAN-based, DDPM-based models are proposed which achieves a state of the art quality for unseen speakers as well.

1.4 History of Speech Synthesizers

In this section, we talk about how the problem of speech synthesis is perceived. Various types of technologies are used to develop systems using Mechanical devices, Electromechanical devices, Electrical and electronic devices, and modern digital devices.

1.4.1 Mechanical Devices

Much before electronic signal processing was pursued, people tried developing mechanical devices which could replicate the human vocal tract and synthesize speech. Synthesizing human speech from machines dates back to the 12th century [124]. The existence of some legends in the early modern period, like ”brazen heads” [125], suggests people have tried to build mechanical devices which can talk automatically. The real interest in the scientific community developed in the later part of the 18th century when a competition was announced by the Imperial Academy of Sciences and Arts. Inspired by the competition, German scientist Christian Gottlieb Kratzenstein developed a mechanical talking device that resembles the human vocal tract and could produce five long vowel sounds [58]. The model was designed with five organ-pipe-like resonators, which are used to produce the vowels (/a,e,i,o,u/), when excited with different vibrations of reed. Leonhard Euler, who is famously known in the areas of Mathematics, proposed this competition had a huge regard for the research of developing talking machines. He wrote in 1761 that ”The construction of a machine capable of expressing sounds, with all the articulations, would no doubt be a very important discovery" [21]. During the same period, another researcher Wolfgang von Kempelen came up with talking "Acoustic-mechanical speech machine" which could elicit more than five vowels. It consisted of bellows as a respiratory source of air pressure, a wooden wind box as a trachea, a rubber funnel as a vocal tract, and a reed system as the source of generating voice. The speech is generated by controlling the device by bellows, some other ports and levers, and manipulating the rubber vocal tract along the time. The device can synthesize consonants along with vowels and can clearly make up the words, although it was difficult for it to generate sentences [117]. Almost a century later, a German scientist named Joseph Faber, inspired by the von Kempelen idea of the talking machine, came up with a more advanced way of controlling the mechanical simulation which can produce sentences instead of just words.

Although all these systems are mechanical and rely on trial and error in finding the right controllers, no one has ever explored to understand the acoustic theory. Robert Willis, A professor at Cambridge University, tried to approach the speech production problem based on physical acoustics. He built reed-driven organ pipes, which by changing the length of the pipes, produced different sounds. A telescope working idea is used to control the lengths of the tubes, and changing the length of the tube is able to produce different resonant frequencies. Research on speech communication technologies has moved on to understanding the spectral components of speech signals. Similarly, people have moved away from developing speech synthesizers inspired by human organs. A German scientist named Hermann Helmholtz developed a system to control the amplitude by maintaining the vibration of eight or more tuning forks, each coupled to resonating chamber [32]. With careful choice of frequencies and amplitude, the system was able to synthesize vowels.

1.4.2 Electro Mechanical Devices

Research on sound acoustics has led to learning the spectrum bands of sound frequencies. In 1859, Koenig developed a system to record and visualize sound signals. They developed a system called phonautograph which has a receiving cone, a diaphragm, and a stylus to convert sound waveform to pressure waveform etched on smoke paper rotating around a cylinder. A few years later, he also proposed a system where sound is used to flicker the flame, and the movements of flame are captured on a rotating mirror producing a visualization of sound [56] as a waveform. Dayton Miller later developed a system called "phonodeik" to visualize and study waveforms of sounds generated by musical instruments and human vowels.

A shift from mechanical and electromechanical devices started in 1922 by John Stewart. He worked on "An Electrical Analogue of the Vocal Organs" [113], an electrical circuit with two resonant branches containing resistors, inductors, and capacitors. He proposed that by adjusting these circuit elements in the resonant branches, various vowel sounds can be synthesized. Since manipulating these circuit elements can be done just by turning knobs or moving sliders, he was able to synthesize two vowels by manipulating the circuit elements and shifting resonance frequency after the synthesis of the first vowel.

Although the circuit is easily manipulative, they are not the electrical analogue of voice organs; instead, they replicate the acoustic resonances produced by human voices, which don't sound natural.

In the 1930s, Bell Labs came up with a system that could analyze and synthesize speech waveforms called Vocoder (VOice CODER). Homer Dudley developed the system where some set of parameters are analyzed from the original speech waveform stored or transferred to other locations, which can be used later to synthesize back the original speech signal [96]. Dudley designed a circuit that could extract low-frequency spectral information from an acoustic speech signal via a bank of filters, transmit that information along the low-bandwidth cable, and use it to modulate a locally supplied carrier signal on the receiving end to reconstruct the speech. Vocoder was shown to be also working for instrumental music or any other sounds like a train locomotive. Although the analyzed signals were synthesized back, extracting fundamental frequency from the original signal was not possible at that time, and so the reconstruction of natural-sounding speech was not possible. Dudley later modified this system to have manual controls at the analysis stage called VODER (Voice Operation DEmonstratoR) [19]. A keyboard, wrist bar, and foot pedal are used as controls. The foot pedal controls the pitch of the resonant oscillator, and the keyboard is used to control the amplitude of periodic signals. Playing with keys and modulating the foot pedal, an operator of the device could learn to generate speech. Voder was not the same as the telephone; the telephone did not talk. A speech signal is just transmitted over a distance in the telephone without any manipulation, whereas the Voder did talk by manipulating the controls.

1.4.3 Electrical and Electronic Devices

During the same period of Bell Labs developing Voder, Ralph Potter developed ”sound spectrograph” [91], a visualization tool that can represent the time-varying record of the spectrum of sound instead of the waveform. The visualization is called a ”spectrogram”, which has frequency on the y-axis, time on the x-axis, and intensity represented as gray bars. Potter also suggested that spectrogram can also be used as a potential application to aid persons who are hearing impaired. The idea was if we train a user to understand the spectrogram, they can read the content by seeing the auditory information as visual content. Later many works followed on design, analyzing, and understanding of spectrograph ([53], [54], [57] and [112]).

Frank Cooper and Alvin Liberman, researchers at Haskins Laboratories, started research on analyzing spectrogram and modifying the spectral representation of speech, transforming back to sound so that the modification made in the spectrogram is perceived. In 1951, they developed a device called ”Pattern Playback,” which allows users to draw the spectrogram on a transparency film and convert it to sound waves using a system including a light source, tone wheel, photocell, and amplifier. The tone wheel contained 50 circular soundtracks that, when turned by a motor at 1800 rpm, would modulate light to generate harmonic frequencies from 120-6000 Hz, roughly covering the speech spectrum. The photocell would receive only the portions of the spectrum corresponding to the pattern that had been drawn on the film and convert them to an electrical signal which could be amplified and played through a loudspeaker. Pattern playback is the first speech synthesizer used widely for experimentation in understanding the structure of speech. One such study was used to understand how the formant transitions happen at the onset and offset of the consonants ([16], [7]). People also formalized rules ([38],[65]) for generating utterances using pattern playback. Although the only drawback of the system was the user had to hand draw the spectrogram on the transparent film. During the same period, various other works, including Parametric Artificial Talker [61] and ”OVE II” [24] are developed, which also use spectrograms and an electric circuit with resonant branches which produces the frequencies according to the spectrogram.

Simultaneously research on storing speech waveforms is largely explored, and it was possible to store audio on magnetic tapes. This led to pursuing the speech synthesizer as a concatenating small speech segment of prerecorded natural speech. Harris ([31],[30]) developed a system that isolates the segments of tape that contain many instances of consonants and vowels. Synthetic speech is created by piecing together isolated segments which have matching harmonics and formant frequencies. The speech was intelligible but quite unnatural because of discontinuities at the boundaries of isolated segments. Later [83] developed an alternate segmentation technique called ”dyad”, which is the segment in time from a steady state location of one phoneme to another. Along with phone and phone dyad, other segments like syllable nuclei, syllable, half syllable, syllable dyads, and word segments can be used from a small set of units to stitch together to form larger utterances. This approach was responsible for modern TTS with unit selection techniques.

1.4.4 Digital Systems

The emergence of digital computing in the 1960s has led to the advent of modern speech synthesizers. The first digital speech synthesizer was developed at Bell Labs by John Kelly and Lochbaum [47]. They used an IBM 704 computer to generate the speech by typing on the keyboard. This was considered one of the most prominent events in Bell labs [124]. Their system is a computer algorithm that generates acoustic waves in an analog of the vocal tract configurations. The keyboard is used to control the parameters of the analog circuit, synthesizing speech according to the controls. This allowed the control of the synthesis using a computer interface, while the process of generating waveform is in an analog circuit. [34] proposed a set of rules which converts the controls to time series parameters of the resonance synthesizer.

More research on using computer controls for vocal tract synthesis started during this time. Observing and understanding the spatial and temporal movements of the human vocal tract was also possible due to advancements in the field of X-ray technology. This led to a new type of synthesis called ”Articulatory synthesis.” This system produces speech by simulating the behavior of human voice organs such as lips, tongue, glottis, and moving vocal tract. [14] developed a system based on a computation model of speech articulators. Controlling the articulator behaviors is quite difficult and also involves large amounts of collecting data and understanding the simulation behavior. In the meantime, people also used the computer to build controls for analog circuit synthesis called formant synthesizers. Various systems called ”Klattalk” [51], ”MITalk” [1], ”DecTalk”, and ”KLSYN88” have particularly become quite popular as a text to speech synthesizers. [104] and [52] developed systems called which can control the resonant parameters like frequency and amplitude of the waveforms. The speech synthesized was just the sounds of the frequencies generated from the circuits, so they are not natural.

Owing to the naturalness of the synthesized quality approaches based on collecting and manipulating human voices are probably the way to go in speech synthesis. Many types of speech synthesis systems have been developed in the digital era focusing primarily on the quality of the speech for automated messaging and digital assistants. These systems leverage a database containing many hours of recordings collected from voice actors. One such method is called ”Concatenative Synthesis” similar to [30], where parts of audio are stitched together to form complete audio, only the algorithms are much more efficient in finding the smaller sound segments and stitching them together longer utterances. [79], [75] and [102] are some works uses concatenative approaches to synthesize. The quality of the synthesis is limited by the size of the original database. Also, stitching the small segments causes discontinuities in the prosody, emotion, and pitch patterns in the synthesis. A different technique was proposed to overcome the challenges of the concatenative approach, ”Statistical Parametric Speech Synthesis (SSPS)”. In 1999 Yoshimura [129] proposed an HMM-based speech synthesizer that relies on generating spectral features of the audio wave and then converting them to audio wave using some techniques ([37],[36], and [46]). This idea has clear advantages over the previous approaches, with good quality, flexibility in controlling the features for synthesis, and low cost of data collection. SSPS has become quite popular, and many techniques ([116],[128], and [134])followed the idea of developing and improving the quality and naturalness of the synthesized samples. However, the generated voice is still robotic and has some artifacts in the output.

In 2010, with the advent of Deep Neural Networks (DNNs), speech synthesizers were also developed using DNNs. [131] and [93] incorporated DNNs in Parametric synthesizers. Few works, including [23], [132] and [133] worked on using Recurrent Neural Networks (RNNs) to generate acoustic features and later used a vocoder to convert these acoustic features to the audio waveform. [121] proposed an end-to-end speech synthesis system directly predicting acoustic features from phoneme sequence. Wavenet [81] is one of the first neural methods which can synthesize direct audio from a sequence of linguistic features. These methods have largely improved the quality of the synthesized speech. Later various methods were proposed using deep learning techniques and various architectures like CNN’s, RNN’s, LSTM’s, and Transformers. Most notable ones include Tacotron [122], Tacotron2 [106], DeepVoice 1[2], DeepVoice 2 [2], Fastspeech [98], Fastspeech2 [97] and many other methods which directly predicts Mel-spectrogram. And various vocoders like Waveglow [92], MelGAN [60], and HiFi-GAN [55] are proposed, which convert mel spectrogram to audio signals. Some other methods like ClariNet [84], FastSpeech2s [97], and EATS [18] predict and generate audio directly instead of spectrogram prediction.

1.5 Speech Synthesis Techniques

The most important qualities of the TTS system are naturalness and intelligibility. Naturalness refers to how close the speech is to humans' voices, and intelligibility refers to the ease of understanding the output. Building TTS systems with high quality was approached in various ways. Each has its limitations and advantages. They are broadly categorized and detailed below.

1.5.1 Articulatory Synthesis

Articulatory synthesis produces speech by simulating the behavior of human voice organs such as lips, tongue, glottis, and moving vocal tract. This synthesis could be very natural because it is closer to how humans speak. Since the synthesis is closer to the human vocal tract, the synthesized speech is fairly natural and intelligible, and many models were developed on proper control of articulators using computer algorithms ([66],[71], [100] and [105]). Although these systems are closer to human vocal synthesis, controlling the articulator behaviors is difficult. For example, it is hard to collect the data for observing the articulator simulation.

1.5.2 Formant Synthesis

Formant synthesis produces speech using an analog circuit instead of using the human voice, where the controls are generated by a set of rules ([104], [52], and [1]). The rules are generally developed by linguists to replicate the formant structure and other spectral properties of speech. Resonant circuits are used to generate sounds by controlling the parameters like fundamental frequency, voicing, and noise levels. The speech generated is fairly intelligible, which is easily understandable, can be generated using limited computation resources, and is not dependent on collecting large amounts of data. Although the speech generated is intelligible, it still has a lot of unnecessary artifacts and is unnatural. It is also difficult to form the rules.

1.5.3 Concatenative Synthesis

Concatenative synthesis is one such technique that relies on collecting ample amounts of speech data of a person, breaking the speech into small portions, storing them in the database, and concatenating these pieces to form the synthesized speech. The database can usually contain speech utterances from phonemes, syllables, and words to even sentences recorded by a voice actor.

This system is widely accepted and used for commercial purposes like in talking clocks or calculators. Because these systems are dependent on words or phonemes in the database, they cannot be used to generate longer and random utterances. They can only generate a pre-programmed combination of words or phones. Concatenative synthesis can generate audio with high intelligibility and with timbre close to a human voice actor. However, since the synthesis is generated by stitching the small pieces, the synthesis lacks smoothness in stress, prosody, or emotion. Hence, the synthesized audio is less natural and has less prosody.

1.5.4 Parametric Synthesis

To address the problems of concatenative synthesis, parametric synthesis is proposed. Parametric synthesis ([129], [116], [134], and [115]) is a synthesis method that uses Hidden Markov Models (HMM), which is also called statistical parametric speech synthesis. The basic idea is instead of directly generating the waveform by concatenating the small pieces, the model first generates the acoustic features, which are later converted to the audio waveform. This consists of text analysis, an acoustic feature extractor, and a vocoder that converts acoustic features to the audio waveform. Text is converted to linguistic features, and the acoustic model converts these to acoustic features like frequency, spectrum, or cepstrum. These acoustic models are based on HMM, which is trained on paired linguistic features and acoustic features. However, the synthesized audio is less natural, with artifacts like buzzing, muffled or noisy audio. Also, the generated voice is still robotic and can be easily differentiated from the human voice.

1.5.5 Neural Synthesis

With the development of deep learning, neural network TTS systems are proposed. Some works have been proposed to replace the HMM-based acoustic model with a neural network. The first notable work, WaveNet [81] is proposed to generate audio waveform directly from linguistic features. Later various methods like Tacotron [122], Deepvoice [2], FastSpeech [98], FastSpeech2 [97] are proposed, which uses a neural network based acoustic model to convert linguistic features to Mel spectrogram. And these acoustic features are then later converted to audio waveform using vocoders. Vocoders are again developed with neural networks like [92, 60, 55]. Later direct text to waveform end to end systems is also developed like FastSpeech2s [97], ClariNet [84], and EATS [18].

1.6 Emotion Controlled TTS

Emotion is crucial for any conversation to happen. Emotion can totally change the meaning of the dialogue. As shown in Figure 4.1, depending on Romeo's response, Juliet will respond to different things, and it can change the whole trail of the conversation. The existing Neural Text to speech models can achieve human-level performance, but the variations it learns are limited to the data, and generally, it learns to render in the average variations of the training dataset. For example, the models learn only one way of speaking a particular sentence with average variations. Systems like [97] provide controls at the phoneme level to vary a few parameters like pitch, duration, and energy, which will impact the synthesis accordingly. But it is difficult to control these variances per phoneme to render an intended emotion.

Generating any text into the target emotion can be thought of as a supervised task, given the text, audio, and emotion annotation of an utterance. However, collecting such clean data is both time-consuming and costly. Hence it is not reliable to depend on data to train an end to end emotion-controllable TTS systems. However, some unsupervised and semi-supervised systems have been developed to achieve emotion controllability. However, they come with some limitations, where the control of emotion is achieved but at the cost of loss in quality of the rendered speech. Emotion control can also be either discrete emotions (like happy, sad, angry, etc.) or continuous emotions (like arousal, valance, and dominance). Collecting data for discrete emotions is relatively easy than annotating the samples using continuous emotion space because emotion values are highly subjective. But finer control of emotions cannot be achieved with discrete emotions. So, our work focuses on controlling emotions in renderings with continuous emotion space (Valance, Arousal, and Dominance)

This thesis talks about two of our works that can achieve control of emotions in the output renderings for a given text. In chapter 3, we first propose an idea where we try to learn a mapping between emotion value (Arousal and Valance) and the intermediate variances like pitch, duration, and energy of each phoneme. Changing these variances can change the emotion in the output synthesis. The proposed method first trains a multi-speaker TTS system on a clean LibriTTS [82] dataset. Later, it uses annotated emotion corpus MSP Podcast corpus [67] to train a model to learn the mapping between emotion and the intermediate variances. The proposed model also takes speaker embedding from a pre-trained speaker encoder.

In chapter 4, we talk about the shortcomings of the above-proposed method. Since the above method learns the mapping of these variances conditioned on speaker identity, the variances can be highly dependent on speaker embedding, but the speaker embedding used from a pre-trained network is not perfect. We propose a system that can adapt to any existing TTS architecture. The model captures the change in the variances from a neutral emotion instead of capturing the whole variances of any emotion. We show the advantage of this proposed method over the existing methods.

Chapter 2

Related Work

2.1 Neural TTS

Neural network-based TTS have changed the landscape of speech synthesis research and have significantly improved the speech quality over conventional concatenative and statistical parametric approaches [35, 127]. Some of the recent popular neural TTS systems are Tacotron [122], Tacotron 2 [106], Deep Voice 1,2,3 [2, 26] and ClariNet [84]. These approaches first generate Mel-spectrogram autoregressively from text input. The Mel-spectrogram is then synthesized into speech using vocoders like Griffin-Lim [28], WaveNet [81], Parallel WaveNet [80], MelGAN [60], and HiFiGAN [55]. More recently, the FastSpeech [98] and FastSpeech 2 [97] methods approach TTS in a non-autoregressive manner and show extremely high computational gains during training and inference. Despite synthesizing natural-sounding speech, the above-mentioned neural TTS models give little or no control over the emotional expression for a given sentence.

2.2 Multiple Speaker TTS

There has been a major focus on scaling TTS systems to multiple speakers. Early neural multi-speaker TTS models require tens of minutes of training data per speaker. Fan *et al.* [22] proposed a neural network model which uses a shared hidden state representation for multiple speakers and speaker-dependent output layers. Gibiansky *et al.* [26] introduced a multi-speaker variation of Tacotron, which learned low-dimensional speaker embeddings for each training speaker. Their later work [86] scaled up to support over 2,400 speakers. Such systems [26, 86, 22] learn a fixed set of speaker embeddings and therefore only support the synthesis of voices seen during training. More recent approaches decouple speaker modeling from speech synthesis by independently training a speaker-discriminative embedding network [76]. The TTS models are then conditioned on these speaker-discriminative embedding obtainable from a few seconds of speech for the given speaker. Wan *et al.* [119] train speaker verification network, Jia *et al.* [39] condition the Tacotron 2 model on the embeddings of verification network. Our work extends such zero-shot multi-speaker support for a non-autoregressive model, FastSpeech2.

2.3 Prosody and conversational speech

Unlike written text, spoken words contain additional non-verbal information. These cues are collectively termed prosody [62] that include variations in tone, pitch, energy, duration, accents, intonation, stress, etc. [8] showed that prosodic exchange is unavoidable in human dialogue. Various machine learning methods have been proposed to predict emotion in speech from its prosody variations [3, 43]. Variations in pitch accents [78], for example, lead to a significant difference in how the receiver perceives the content. A sentence (like `I said` **`un`**`lock the door, not lock it` from [99]) could be delivered both as a statement and a command by merely changing prosody.

Emotion recognition in conversations has gained increasing attention for developing empathetic machines with emotion-tagged multi-modal data publicly available for modeling like [64, 88, 10]. While most methods like [68, 40] use a combination of text and speech information, some leverage additional side-information from broader context [25] and the topic of conversation [136].

In such labeled data, emotion is often represented as a categorical variable over a discrete space following models like Ekman's basic emotions [20] or the wheel of Plutchik [87]. This choice is largely owing to the ease of annotating data. [101] proposed a continuous two-dimensional space as an alternative called the valence-arousal model for human emotions. Arousal signifies the intensity of the emotion, while valence captures its polarity. It has been extended to add a third dimension of dominance, making it the valence-arousal-dominance (VAD) model. VAD has since been widely used in modeling emotion in music [27, 94], speech [3, 43] and other content [41, 9]. We use the continuous space representation as it is richer and more convenient to handle in our model.

2.4 Expressive TTS and Controllable TTS

Following enormous progress in neural TTS systems, the focus in recent years has shifted to modeling latent aspects of prosody. Humans speak with different styles and tonal variations, but there is an underlying pattern or constraint to these varying styles. The absence of an expected variation or the presence of an unexpected variation is easily detected as uncanny speech by a human listener.

Wang *et al.* [123] proposed a framework to learn a bank of style embeddings called "Global Style Tokens" (GST) that are jointly trained within Tacotron (without any explicit supervision). A weighted combination of these vectors corresponds to a range of acoustic variations. Battenberg *et al.* [5] introduces a hierarchical latent variable model to separate style from prosody. Although such unsupervised methods [123, 5] can achieve prosodic variations, they can be hard to interpret and do not allow a straightforward control for varying the emotional prosody.

Skerry-Ryan *et al.* [108] proposed an end-to-end framework for prosody transfer, where the representation of prosody is learned from reference acoustic signals. The system transfers prosody from one speaker to another in a pitch-absolute manner. Karlapati *et al.* [45] proposed a framework for reference prosody by capturing aspects like rhythm, emphasis, melody, etc., from the source speaker. However, such reference-based methods cannot give the desired level of control as it requires a source reference for each different style of utterance. While such methods can work for scenarios like dubbing, they fall short on audiobook generation and other creative applications.

Habib *et al.* [29] proposed a generative TTS model with a semi-supervised latent variable that can control affect in discrete levels. Data collection involved recording reading text in either a happy, sad or angry voice at two levels of arousal. These six levels of arousal-valence combinations were used for partial supervision of latent variables. The model brings control only over discrete affective states (6 points), only representing a subset of emotions. Our work extends this idea by giving affect control over the continuous space of arousal and valence. Arousal(A) is a measure of intensity, whereas Valance (V) describes emotion's positivity or negativity. Russell *et al.* [101] show that these two parameters can represent various emotions in a 2D plane (Figure 3.2).

By conditioning TTS on AV values, our work allows fine-grained and interpretable control over the synthesized speech. We choose Fastspeech2[97] as the backbone due to its simplicity and ultra-fast inference speed. Fastspeech2 predicts low-level features like pitch, duration, and energy for each phoneme and conditions the decoder on them. Our work facilitates a sentence-level conditioning of these phoneme-level features using scalar values for Arousal-Valence (AV).

Controllable TTS Neural TTS systems are now increasingly popular, improving upon older concatenative statistical systems [73] in synthesized speech naturalness. These are broadly sequence-to-sequence networks with an encoder processing the input text or phoneme sequence followed by a decoder that generates the sequence of Mel frames for output speech. Mel frames are then projected into the time domain by a vocoder [81, 28] to generate the speech. Decoding could be autoregressive with Tacotron-like models [122] or non-autoregressive with Fastspeech-like models [98].

Non-autoregressive models are faster at inference than autoregressive models with about comparable naturalness of speech quality [97]. The trick non-autoregressive models use to generate Mel frames in parallel is to predict the relevant features as an intermediate step and condition the independent decoding of Mels on them. This technique is now increasingly adopted for autoregressive models as well [120] to predict features like phoneme duration that improve decoding stability avoiding alignment issues. Our method is compatible with any architecture that predicts prosodic features of pitch, energy, and duration as an intermediate step before decoding.

Going beyond the naturalness of speech, there has been considerable effort to improve the expressiveness of the renderings. Some focused on learning a linear space of variations in speech expressions for selecting a suitable variation at inference time. [123] learn this space unsupervised by encouraging it to explain all variations in training data not captured in content embedding. A reference encoder maps an input utterance to a style embedding as a linear combination of basis style vectors. Manual analysis is required to understand the prosody feature learned into a basis vector that could include variations like vocal depth or pitch, speaking rate, or even background noise as available in training data. While this offers style control, it does not explicitly learn the prosody variations of interest in the style space. Our work focuses on the same level of control but specifically over the affective state as labeled in some data for supervision.

[107] propose a model similar to [123] with style tokens restricted to valence and arousal. However, the absolute (pitch, energy, duration) feature predictions restrict prosody control, leading to unnatural distortions. Specifically, it skews more towards retaining the speaker's voice identity than the emotion and entangles emotion with other acoustic features. [45] replace the linear style space with a variational reference encoder to generate prosody embedding to condition the decoder. [5] use a similar variational model but instead force its posterior to match that of the reference utterance to copy prosody with a controllable parameter determining the closeness of the match. This trick alleviates certain issues like in pitch-range [130] and transfer to unrelated sentences but exposes a lower degree of control with no explicit levers to operate, as possible in our work.

[29] propose to learn explicit latent representation for various prosodic variables, segregating them into explicitly controllable (like affect, speaking rate, etc.) and implicit (like intonation, rhythm, stress, etc.). While the model offers a higher degree of explicit control, it requires using proprietary studio recorded data with utterances reflecting prompted emotions at specified arousal. Dependence on explicit supervision from studio-recorded data makes it harder to scale this model across languages and other prosodic variations. In contrast, we use publicly available data with emotion labels to train our models.

There are other methods that try to predict suitable prosody features from text content. [95] add a prosody encoder module to the standard TTS network that predicts certain hand-crafted prosody features from text embedding of input. This prosody encoder is used with a small optional bias for affect variations at inference. [33] extend this to replace hand-crafting prosody features with explicit training followed by their prediction from the text. [44] further enriches the textual context using BERT embeddings and parse trees. These methods are limited in expressiveness offering no control over rendering the emotion that our work focuses on.

Chapter 3

Emotion controllable TTS system

Machine-generated speech is characterized by its limited or unnatural emotional variation. Current text-to-speech systems generate speech with either a flat emotion, emotion selected from a predefined set, average variation learned from prosody sequences in training data, or transferred from a source style. We propose a text-to-speech (TTS) system where a user can choose the emotion of generated speech from a continuous and meaningful emotion space (Arousal-Valence space). The proposed TTS system can generate speech from the text in any speaker's style, with fine control of emotion. We show that the system works on emotion unseen during training and can scale to previously unseen speakers given his/her speech sample. Our work expands the horizon of the state-of-the-art FastSpeech2 backbone to a multi-speaker setting and gives it much-coveted continuous (and interpretable) affective control without any observable degradation in the quality of the synthesized speech.

3.1 Introduction

Text-to-speech(TTS) applications strive to synthesize 'human-like speech.' This task not only needs modeling of the human vocal system (to generate the frequencies given a sequence of phonemes) but also captures the prosody and intonation variations present in human speech. Neural network models have made significant improvements in enhancing the quality of generated speech, and most state-of-the-art TTS systems, like Deep Voice[2], Tacotron[122], and Fastspeech2[97] generates natural sounding voice. However, high-level affective controllability still remains a much-coveted property in these speech generation systems and has been a problem of interest in the speech community for well over three decades [11, 103].

Controlling emotional prosody (affective control) is vital for many creative applications (like audiobook generation and virtual assistants) and desirable in almost all speech generation use cases. Affective control is a challenging task, and even with the significant improvements in recent years, TTS systems today do not have high-level interpretable emotion control. The existing systems are restricted to either transfer of prosody from source style[45] or learning prosody globally given a phoneme sequence[97].

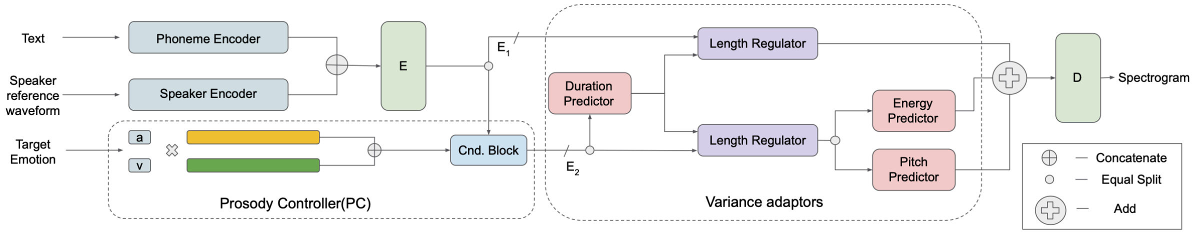


Figure 3.1: Schematic diagram of the proposed model.

Habib *et al.* [29] proposed a system to control affect; however, their method cannot incorporate fine control and is limited to six discrete emotional states.

We propose a TTS system based on FastSpeech2[97] and bring in fine-grain prosody control and multi-speaker control. The improvements are achieved without any observable degradation in the synthesized speech quality and without compromising its ultra-fast inference. Similar to Fastspeech2, our model predicts low-level features from the phoneme sequence (e.g., pitch, energy, and duration). However, the proposed model incorporates high-level and interpretable sentence-level control over the low-level intermediate predictions computed for each phoneme. Our approach has profound implications from a usability perspective because (a) for a human, a phoneme level control is difficult to interact with, and our model allows sentence-level emotional control, and (b) low-level features like pitch, energy, duration, etc. are difficult to interpret and by conditioning them on *arousal valence* values, our model allows an expressible emotional control. We condition the encoder to scale for multiple speakers and transform the encoded vector to incorporate the continuous *arousal-valence* values. Our core contributions are:

- We extend the FastSpeech2 architecture to scale for multiple speakers based on fixed-size speaker embeddings.
- We propose a novel Prosody Control(PC) block into FastSpeech2 architecture to incorporate high-level affective sentence level control using scalar *Arousal-Valence* values on the low-level and phoneme level variance features like pitch, energy, and duration.
- The proposed architecture hence allows to generate speech with fine grain emotional control as they can be chosen from a continuous and interpretable *Arousal-Valence* space.

3.2 Method

Our model uses Fastspeech2 as its backbone[97]. Unlike autoregressive models, Fastspeech2 does not depend on the previous frames to generate the next frames, leading to faster synthesis. The model comprises of mainly three parts, namely: the encoder-decoder block, the prosody control block, and the variance adaptor (Figure 3.1). The encoder block(E) and decoder block(D) are feed-Forward transformers with self-attention and 1-D convolution layers. The model has three inputs:

- Text: The text to be rendered as speech
- Speaker reference waveform: Audio sample of source speaker in whose voice the output will be rendered.
- Target Emotion: Arousal and valence values corresponding to the target emotion

The phoneme encoder gives a vector representation of fixed size for each of the phonemes in the input text. These embeddings are padded along sequence dimensions to match the number of phonemes in all the inputs across a batch. To incorporate speaker information into these embeddings, we condition our encoder(E) on speaker identity by using an embedding trained for speaker verification [119]. These embeddings capture the characteristics of different speakers, invariant to the content and background noise. Given a speaker reference waveform, using the pre-trained model, we generate 256 dimension speaker embedding. We concatenate the phoneme embedding with speaker embedding along the sequence dimension at the zeroth position. i.e., the speaker embedding appears as the first phoneme in the concatenated vector. This technique ensures the constant position of speaker embedding (irrespective of the pad length of phonemes). The encoder(E) learns a representation for each phoneme attending to all other phonemes along with speaker embedding. We call this representation E_1. We observed that conditioning the encoder with speaker embedding gives better results than conditioning the decoder with speaker embedding. In our model, conditioning the decoder with speaker embedding did not capture the speaker's identity. We hypothesize this is because the variance predictions are dependent on speaker embeddings. The encoder's output and predicted variances (pitch, energy, duration) are decoded(at D) to obtain the Mel-spectrogram. The loss is computed between the generated Mel-spectrogram and the spectrogram of target speech(Mel-loss). This end-to-end structure forms the backbone of our system.

The Prosody Control(PC) block generates a latent representation for each phoneme with affective cues from arousal and valence. We use two learnable vectors of length 256 to represent arousal and valance, respectively. The combined emotion is computed as the sum of these two vectors, weighted by arousal and valence inputs. The two vectors are trained with the loss computed at each of the variance predictors along with Mel-loss. The weighted sum is concatenated with E_1 and passed through a linear layer(condition block). The resulting representation is a phoneme embedding incorporating input emotions. We call this representation E_2.

E_2 is passed through the duration predictor, which predicts a duration for each of the phonemes. Based on the duration (d) predicted for each phoneme using E_2, the length regulator expands the hidden states of the phoneme sequence d times for both E_1 and E_2. The total length of the hidden states in the two regulated embeddings now corresponds to the length of the output Mel-spectrograms. Pitch and energy are predicted at corresponding variance predictors using regulated E_2. Each variance predictor is trained with corresponding ground truth extracted from the speech wave. The energy and pitch

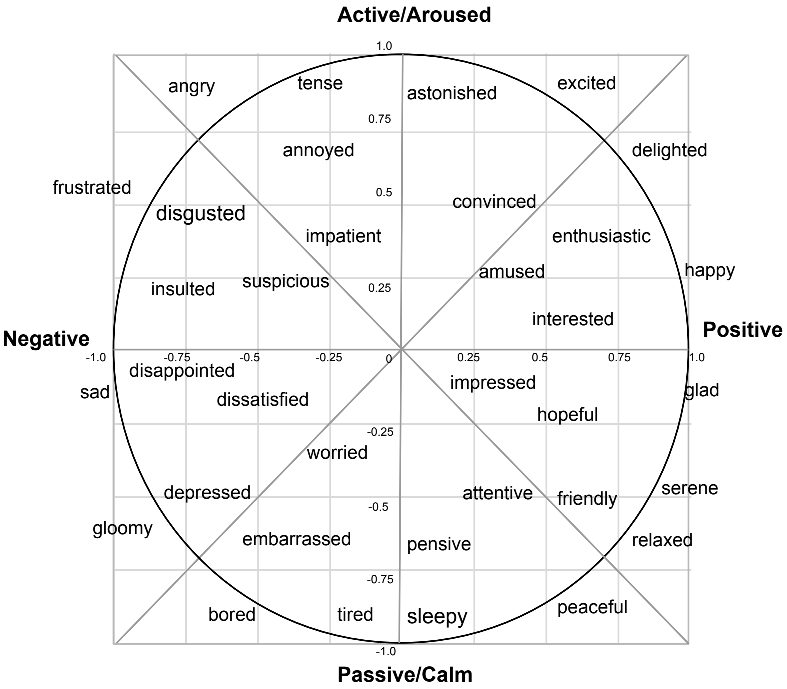


Figure 3.2: The 2-D Emotion Wheel.

computed are added to regulated E_1 and are passed to decoder block(D). The decoder outputs the Mel spectrogram. We use the MelGan vocoder [60] to generate raw speech from the spectrogram.

We use E_2 to predict the variances and use the resultant predictions to modify E_1. Decoder gets E_1 as input, which is not concatenated with affective cues. This strategy ensures that the emotion only modifies the pitch, energy, and duration, and the encoder-decoder module of the TTS can be trained independently of the prosody control block. We propose this strategy to train the backbone and prosody controller block on LibriSpeech and MSP datasets independently. This ensures that errors incurred in transcribing MSP do not affect TTS quality. We train the prosody control block separately after training and freezing the encoder-decoder modules.

3.3 Experiments and Results

3.3.1 Dataset

We use two datasets to train our model. We train our backbone multi-speaker TTS model (leaving out Prosody Controller block) on LibriSpeech [82] dataset. LibriSpeech [82] contains transcripts and corresponding audio samples spoken by multiple speakers. Our model takes phoneme sequences as

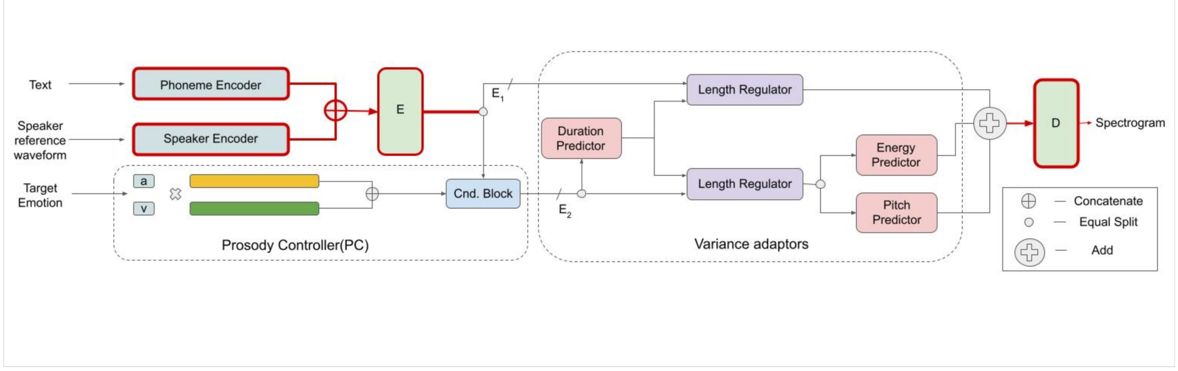


Figure 3.3: Freeze the weights of Encoder-Decoder (in red) and fine-tune the variance adaptors.

input. The text sequence is converted into a phoneme sequence using the method proposed in [114]. We generate Mel spectrogram from the audio file following the work in [106]. This is used to compute loss with a predicted Mel spectrogram. We compute energy, pitch, and duration from the speech to train the corresponding variance predictors. To train the duration predictor, we generate ground truth values of duration per phoneme using Montreal Forced aligner(MFA) [72]. MFA is a speech-text alignment tool used to generate time-aligned versions of audio files from the given transcript. LibriSpeech consists of two clean training sets comprising 436 hours of speech training data. We train on this data and use some of the speaker samples as a validation set. This dataset has no affective annotations.

We train our Prosody Controller(PC) block on MSP Podcast corpus [67] . MSP Podcast is a speech corpus annotated with emotions. It consists of podcast segments annotated with emotion labels and valance arousal values ranging from 1 to 7. The corpus consists of 73K segments comprising 100 hours of speech, split into training and validation data. MSP Podcast corpus does not contain transcripts for the audio segments. To generate transcripts, we use Google speech-to-text API. We use Montreal-Forced-aligner(MFA) [72] to achieve alignment, and if MFA does not find proper alignment for the text and audio pair, the sample is discarded. This accounts for the inaccuracies of the speech-to-text API and background noise in audio samples. After applying MFA and discarding the wrongly transcribed samples, we are left with 55k samples comprising roughly 71 hours of speech. We use this data to train our prosody control module.

3.3.2 Training

We train our model in two stages.

- We first train our multi-speaker model barring Prosody Controller block on LibriSpeech [82] dataset. The encoder-decoder model with variance adaptors is trained together. The total loss consists of Mel loss(computed between the predicted spectrogram and the spectrogram of corresponding ground truth audio), pitch loss, energy loss, and duration loss (each of which is com-

Model	**Mean Opinion Score(MOS)**
Fastspeech2	3.65 ± 0.09
Our Model	3.62 ± 0.13
Speaker Similarity	**Mean Opinion Score(MOS)**
Same speaker set	3.6 ± 0.08
Same gender speaker set	2.55 ± 0.09
Different gender speaker set	1.2 ± 0.04
Affect control	**Avg. rater score in %**
Superlative emotion match	86.0

Table 3.1: The MOS with 95% confidence intervals.

puted directly from the ground truth audio). In this phase, E_1 is directly used as input to variance predictors. The model is trained on 4 GPUs with a batch size of 16. We use Adam optimizer to train the model. The training takes around 200k steps until convergence.

- In the second phase, we train our Prosody Controller block using MSP Podcast[67] corpus. We bring the emotion control by conditioning variance predictors on arousal valance values along with phoneme sequences(E_2). In this phase, we freeze the weights of the encoder-decoder model trained on LibriSpeech and only train the PC and variance adaptors as shown in Figure 3.3. The model is trained on 4 GPUs with a batch size of 16, and it takes 150k steps until convergence.

3.3.3 Model Performance

We measure the naturalness of generated speech, speaker sensitivity, and emotion control of our model through three user studies. We assess the voice's naturalness and speaker similarity using the Mean Opinion Score (MOS) collected from subjective listening tests. We use a Likert scale with a range of 1 to 5 in 1.0 point increments. We evaluate emotion control using the average rater score. The results are reported in Table 3.1.

Naturalness of generated speech: To evaluate the naturalness of the generated speech, we use a set of 30 phrases that do not appear in the training set of either MSP or LibriSpeech and synthesize audio using our model. To compare the MOS of our model, we also synthesize the same phrases using Fastspeech2 [97]. A collection of samples from both these models are provided to users. Twenty proficient English speakers are asked to make quality judgments about the naturalness of the synthesized speech samples and asked to rate on a Likert scale of range 1 to 5, where 1 being 'completely unnatural' and 5 being 'completely natural'. The results in Table 3.1 show that similar scores are obtained for the two models. The results demonstrate that our model does not bring any noticeable distortions in terms of the naturalness of generated speech compared to the Fastspeech2 backbone.

Capturing reference speaker voice : Speaker similarity is evaluated in a similar fashion using MOS. We validate the speaker similarity on three different sets.

- Same speaker set: This set consists of sample pairs synthesized from the same speaker. The pair consists of either a ground truth speech and a synthesized sample or both synthesized samples.
- Same-gender speaker set: Here, we synthesize phrases for a set of speakers of the same gender. We form pairs of samples with the same gender but with different speaker.
- Different gender speaker set: This set is curated by pairing synthesized audio samples generated for speakers from opposite genders.

Given a pair of samples, participants were asked to rate the similarity score of how close the voices sound on a Likert scale of 1 to 5. Where 1 corresponds to 'Not at all similar,' and 5 corresponds to 'extremely similar'. For the same speaker set, we obtained a MOS of 3.6. This shows that our model can synthesize voices that sound close to a given target speaker. The MOS of 2.55 for the same gender set shows that audio generated from different speakers of the same gender has a certain degree of similarity. Furthermore, the low MOS of 1.2 for samples from different genders shows that our model's synthesized speech can be discriminated based on gender.

Affective control: Interpreting affect in rendition is subjective, challenging, and highly correlated with the content. We use user ratings to evaluate affect control. The model being conditioned on the continuous and meaningful space of emotion, the user can change the level of emotion from happy to delighted, sad to depressed, etc., superlatively, during synthesis. We synthesize a set of phrases with different arousal valence(AV) values to evaluate the control obtained by changing AV values.

For our survey, we chose samples consisting of different levels of four emotions: happy, sad, angry, and excited. We provide a pair of samples for each of the above emotions, with one sample corresponding to the lower level of the emotion and the other corresponding to the higher level of the respective emotion (e.g., happy to delighted). We choose appropriate AV values such that particular emotion is expressed in two degrees. Raters are asked not to judge the content and choose the sample expressing the particular emotion strongly (e.g., which one is angrier or happier). Every rater is shown eight different pairs of samples. We choose two pairs from each of the aforementioned emotions. The reported score shows the average percentage score obtained by raters in choosing the stronger emotion.

Compiling results from all the users, we observe that 86% of the raters can correctly choose the sample strongly expressing a particular emotion. The study shows the ability of the model to control prosody using arousal valance values.

3.4 Limitations and Conclusions

Although we were able to achieve reasonable control over the emotions, as in we were able to synthesize the given Arousal and Valance value, the control comes at the cost of quality. And emotion control can also be improved by experimenting with varied levels of emotions in different settings. For example, assessing the emotion level of an utterance is highly subjective and difficult without providing any context. So, the evaluation might not be robust.

Our work addresses the problem of emotional prosody control in machine-generated speech. In contrast to previous prosody control methods, which are either difficult to interpret by humans, require reference audio, or allow selection only among a discrete set of emotions, our method allows a continuous and interpretable variation. We use the FastSpeech2 TTS model as a backbone and add a novel Prosody Control (PC) block. The PC blocks conditions the phoneme level variational parameters on sentence-level Arousal Valance values. We also extend the FastSpeech2 framework to support multiple speakers by conditioning it on a discriminative speaker embedding. Our user study results demonstrate the efficacy of the proposed framework and show that it can synthesize natural-sounding speech, mimic reference speakers, and allow interpretable emotional prosody control.

Audio samples for our experiments are available at: `https://researchweb.iiit.ac.in/~sarath.s/emotts/`

Chapter 4

Enhanced emotion controllable TTS system

We present a method to control the emotional prosody of Text to Speech (TTS) systems by using phoneme-level intermediate features (pitch, energy, and duration) as levers. As a key idea, we propose Differential Scaling (DS) to disentangle features relating to affective prosody from those arising due to acoustics conditions and speaker identity. With thorough experimental studies, we show that the proposed method improves over the prior art in accurately emulating the desired emotions while retaining the naturalness of speech. We extend the traditional evaluation of using individual sentences for a more complete evaluation of HCI systems. We present a novel experimental setup by replacing an actor with a TTS system in offline and live conversations. The emotion to be rendered is either predicted or manually assigned. The results show that the proposed method is strongly preferred over the state-of-the-art TTS system and adds the much-coveted "human touch" in machine dialogue.

4.1 Introduction

> "The text is like a canoe, and the river on which it sits is the emotion. It all depends on the flow of the river, which is your emotion. The text takes on the character of your emotion."
>
> — Sanford Meisner

In natural language processing, vocabulary and grammar tend to take center stage, but those elements of speech only tell half the story. Affective prosody provides context and gives meaning to words, and keeps listeners engaged. Understanding emotional prosody is central to language and social development. Studies suggest that we show remarkable sensitivity to prosody "even as infants" [77, 69]. Recently [59] shows that voice-only communication likely elicits higher empathic accuracy than even multi-sense modes including facial expressions.

[8] shows that any meaningful spoken dialogue cannot happen without some amount of prosodic matching. As humans, we naturally anticipate and adapt with emotional cues in conversing with others, see Figure 4.1 for an example. Celebrated trainer Sanford Meisner employed this to develop *Meisner technique* for theatre actors to react naturally to others in the environment as opposed to *method acting*. The importance of emotional prosody in conversations cannot be overstated and TTS models need to fill this gap to make human-like conversations possible in HCI systems.

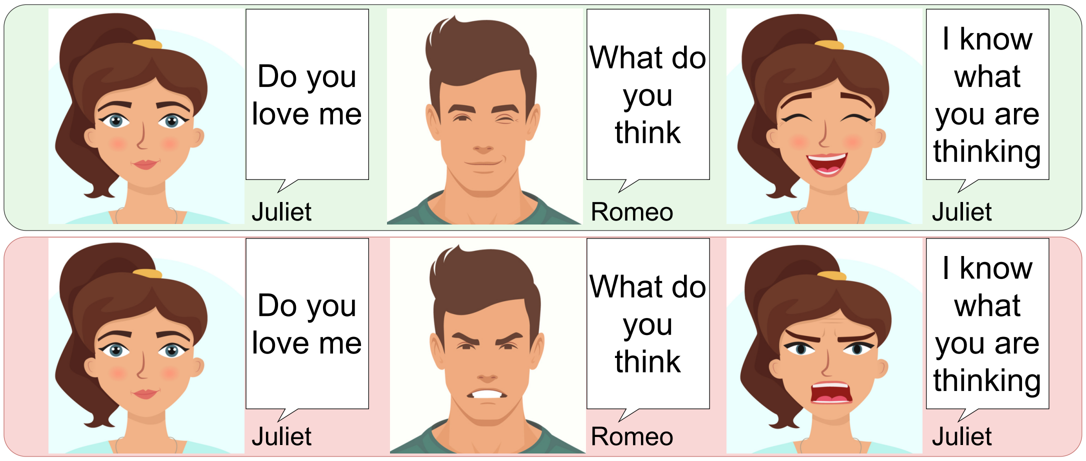


Figure 4.1: Dialogues can have different meanings despite having the same text. Also, starting at the same emotion, Juliet has different emotion post Romeo's response.

[74] study the value of emotional prosody in HCI and emphasize its role in healthcare dialogue systems, improving social interaction skills in people with autism, augmentative and alternative communication devices, and gaming narratives. They explain that successfully incorporating expressive speech into HCI, involves two aspects: (a) prosodic emotion recognition and (b) expression of emotional prosody. Considerable effort has been made towards recognizing and predicting the emotional nuances in human dialogues [48, 90, 136, 64, 89, 118]. However, current TTS systems are yet to improve in rendering emotive or expressive speech for real-world HCI systems.

State-of-the-art TTS systems [97, 122] tend to exhibit average emotions for a given phoneme sequence by taking the mean of utterances from training data. Some efforts towards improving expressiveness [5, 45] provide prosody control using a reference clip. Others like [107] and [29] further focused on controllability, exposing levers that can be manipulated at inference-time to derive the intended expression. However, the quality and stability of synthesized speech heavily depend on various modeling choices. Emotion or prosody modeling, for example, could pick from numerous available discrete or continuous space representations. The encoder network module chosen might vary in its ability to disentangle prosody from other acoustic features like speaker identity and adaptability to content. For example, those relying on reference clips to replicate prosody might perform poorly when input text is unsuitable for rendering with the prosody of reference. Some models feed prosody features with phoneme embeddings directly into the decoder, while others use them to predict intermediate features that are used in conditioning the decoder. It is empirically verified [107] that intermediate features could be suitably manipulated to bring about the desired change in expression.

We take this direction forward to endow the intermediate feature prediction module with affective state control over the final rendering. We propose *Differential Scaling (DS)* of the predicted intermediates to bring about the required change in emotion. The *DS* module is aimed to effect only emotion as intended while remaining agnostic to all other features like speakers' identities or acoustic conditions as seen in train data. We show that this significantly improves the naturalness of the generated speech while allowing finer control over prosody.

In addition to comparing our model's renderings against various others' from literature for naturalness and emotion control on conventional single utterances drawn from disconnected contexts, we also evaluate them in conversations. We curate data with conversational theatre dialogues and replace an actor with a TTS system. We use its response as a proxy to evaluate the empathic accuracy. In another experiment, we had a theatre director control the emotion levers of our TTS model in a live conversation with the actor to evaluate controllability. As demonstrated in the results, our proposed method significantly improves over existing methods in producing suitable prosodic variation lending closer to human-like conversations. The rest of this chapter will elaborate on the following contributions of this work.

- We propose a simple technique of using a *DS* module to better emulate emotions in TTS rendered speech. This works as a plug-and-play with both autoregressive and non-autoregressive TTS models that predict prosodic features as an intermediate step.

- Our work extends the literature of training controllable and expressive TTS models with improved empathic accuracy and without specific studio-recorded data.

- Finally, we present novel methods and data for evaluating TTS models in real conversations with human subjects. The method of evaluation is a useful step towards filling the gap of emulating emotional speech that needs more work.

4.2 Model

Our network uses a backbone TTS that can be borrowed from any model which predicts pitch, energy, and duration as intermediate features from the input phoneme sequence. This network learns to predict the average features for given phonemes. Following the convention in earlier works, we refer to the intermediate features as variances and the module that predicts them as variance adaptor. Prior work improves standard variance adaptors in, say FastSpeech2, by conditioning on emotion variables of valence-arousal in addition to the phoneme sequence to generate expressive speech. We refer to it as Emotional Variance Adaptor (EVA), for which we propose an alternative. Our proposed Differential Scaler (DS) module determines how best to vary the output of the EVA to bring the desired change in emotion. We describe the details of these network choices in this section, specifically, the broader backbone network architecture and the different variance adaptor modules from non-emotive baseline, emotive baseline, and our proposal.

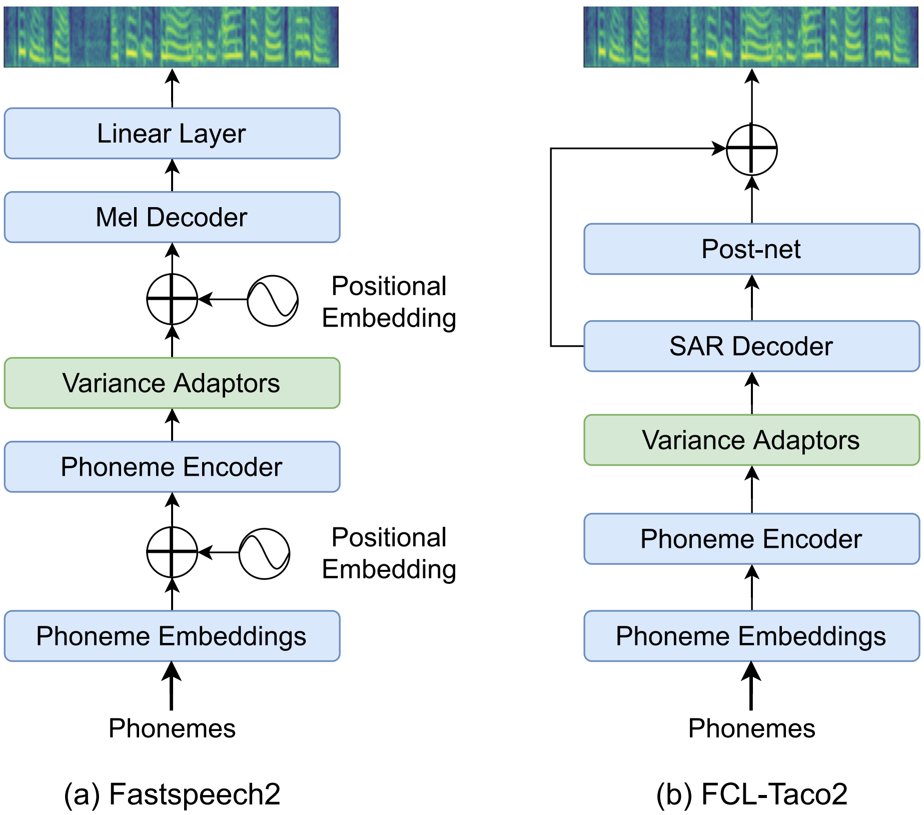


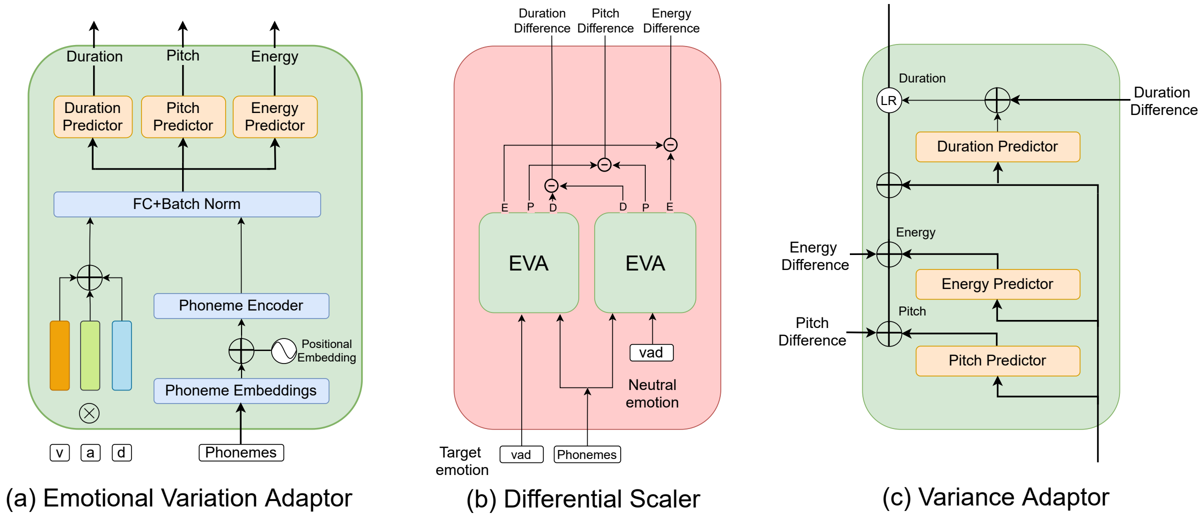
Figure 4.2: Backbone TTS architectures.

4.2.1 Backbone

We present experiments with two suitable choices for our backbone systems, FastSpeech2 and FCL-Taco2. The backbone has three modules; an encoder, variance adaptor, and decoder. The encoder maps an input phoneme sequence to its embedding. Given this representation, the variance adaptor predicts the pitch, energy, and duration for each of the phonemes. These intermediate features are processed by the decoder module downstream to return Mel-spectrogram frames. We reuse the encoder and decoder modules as designed in their original architectures without any changes. We refer readers to the respective papers for details of these networks. Wavenet [81] vocoder is used to map the Mel-spectrogram outputs of the decoder to time-domain raw audio.

4.2.2 Variance adaptor module

Non-emotive baselines. Our baseline models of FastSpeech2 and FCL-Taco2 are trained with the variance adaptors as described by their authors. We also train a derivative of the FastSpeech2 with the variance adaptor modified to make predictions at the phoneme-level and not at frame-level. A duration d_π is predicted for each phoneme π, following which the length regulator repeats the hidden state of



(a) Emotional Variation Adaptor (b) Differential Scaler (c) Variance Adaptor

Figure 4.3: Schematic diagram of the proposed model.

that phoneme π times. Also, unlike Fastspeech2, we use this length regulator after the predicted pitch and energy are added to the encoder output. We refer to this derivative as FastSpeech2π.

Emotive baseline. [107] conditioned the variance adaptor of FastSpeech2 on additional emotion embedding that gives the model control over the prosody of the rendered speech. It generates the emotion embedding as a linear weighted combination of the valence and arousal vectors that are learned from data during training. The weights are valence and arousal values as annotated for training and can be used as control levers to modify emotion during inference. This emotion variance adaptor (EVA) module generates suitable intermediate features of pitch and energy at frame-level and duration at phoneme-level. These features are consumed by the decoder along with the encoder output in generating Mel frames. While this helps control emotional prosody rendered speech, it leads to a significant drop in perceptual quality and naturalness relative to the baselines. Our contribution is an alternative design of the variance adaptor module that improves upon [107]'s FastSpeech2 + EVA model in emotion control and expressiveness and upon the baselines in terms of naturalness.

Differential Scaler. We extend the emotion representation from EVA to include dominance in addition to valence and arousal values. Dominance is the degree of control exerted by an emotion. Including the dominance dimension to the emotion space expands the range of emotions the TTS model can express. For example, by introducing this dimension, we can better distinguish outputs for emotions like 'anger and fear' or 'sad and contempt'.

The *Differential Scaler* module further extends EVA to estimate the change in variances necessary for a pronounced effect of the target emotion relative to its neutral counterpart. As shown in Figure 4.3(b), the variances are estimated using the EVA module for a given phoneme sequence at two different triplets of VAD values. One prediction corresponds to the neutral emotion with VAD values all set to zeros. The other prediction corresponds to the chosen VAD values of the target emotion. We take the difference of these two estimates as the direction along which the variances can be varied for the desired change in emotion without effecting other acoustic features. We are implicitly making two assumptions here. Emotion variations are captured as linear transformations in this space and there is a strong disentangling of emotional prosody with other acoustic features in this space. Results from our empirical evaluation favorably support the above assumptions.

sectionTraining Modeling with intermediate features facilitates training the backbone and the variance adaptors independently on different data. We exploit this to train our variance adaptor on scarcely available VAD annotated data while reusing backbone models trained on abundant transcribed speech data.

Backbone. We train two backbone networks Fastspeech2π (non-autoregressive) and FCL-Taco2 (autoregressive) on Blizzard 2013 dataset [49]. It contains 147 hours of Catherine Bayers's speech, reading books in American English. Due to the style of reading, the dataset is rich in expressiveness and spans different combinations of pitch, energy, and duration. Both models are trained with Mel loss (mean absolute error between predicted and ground truth Mels), pitch loss, energy loss, and duration loss (mean square error between predicted and ground truth features). Both models are trained for 200K iterations using Adam optimizer with warm-up learning rate scheduler and batch size of 16.

EVA. We train EVA on MSP-Podcast corpus [67] annotated with arousal, valance and dominance values. The corpus consists of around 100 hours of speech data, but their transcriptions are not available. We generate transcripts using a speech-to-text model. We use Montreal-Forced-Aligner (MFA) [72] for phoneme alignments. Those transcripts that MFA fails to find a good alignment for are filtered out. The remaining utterances add up to about 71 hours of emotive speech data, which we use to train our EVA. We train pitch, energy, and duration predictors conditioned on VAD values minimizing only the sum of variance losses. For all the experiments, text transcripts are converted to phonemes using [114]. We generate Mel spectrogram from the audio files similar to [122]. Pitch and energy are computed from the Mel spectrogram, and we use MFA for aligning phonemes to train the duration predictor.

4.3 Training

Modeling with intermediate features facilitates training the backbone and the variance adaptors independently on different data. We exploit this to train our variance adaptor on scarcely available VAD annotated data while reusing backbone models trained on abundant transcribed speech data.

Backbone. We train two backbone networks Fastspeech2π (non-autoregressive) and FCL-Taco2 (autoregressive) on Blizzard 2013 dataset [49]. It contains 147 hours of Catherine Bayers's speech, reading books in American English. Due to the style of reading, the dataset is rich in expressiveness and spans different combinations of pitch, energy, and duration. Both models are trained with Mel loss (mean absolute error between predicted and ground truth Mels), pitch loss, energy loss, and duration loss (mean square error between predicted and ground truth features). Both models are trained for 200K iterations using Adam optimizer with warm-up learning rate scheduler and batch size of 16.

EVA. We train EVA on MSP-Podcast corpus [67] annotated with arousal, valance and dominance values. The corpus consists of around 100 hours of speech data, but their transcriptions are not available. We generate transcripts using a speech-to-text model. We use Montreal-Forced-Aligner (MFA) [72] for phoneme alignments. Those transcripts that MFA fails to find a good alignment for are filtered out. The remaining utterances add up to about 71 hours of emotive speech data, which we use to train our EVA. We train pitch, energy, and duration predictors conditioned on VAD values minimizing only the sum of variance losses. For all the experiments, text transcripts are converted to phonemes using [114]. We generate Mel spectrogram from the audio files similar to [122]. Pitch and energy are computed from the Mel spectrogram, and we use MFA for aligning phonemes to train the duration predictor.

4.4 Experiments and user study

We present three experiments; comparison with prior art using conventional evaluation metrics, those for emotional consistency with pre-recorded audio, and finally, live conversations with humans.

4.4.1 Comparisons with prior-art

We compare the proposed approach against four state-of-the-art TTS models. The list includes two non-emotive TTS models (FastSpeech2 and FCL-Taco2), one reference-based method [12], and one AV-conditioned model (FastSpeech2 + EVA). We also compare our method with the modified backbone, Fastspeech2π.

To evaluate the perceptual quality/naturalness, we compare Mean Opinion Score (MOS) [13] averaged across forty subjects proficient in English. We synthesize twenty different sentences from the test set using each of the seven models. We prepare a user study by picking five samples rendered by each model to make a survey. Annotator rates each sample on a Likert scale of one for ‘completely unnatural’ to five for ‘completely natural’.

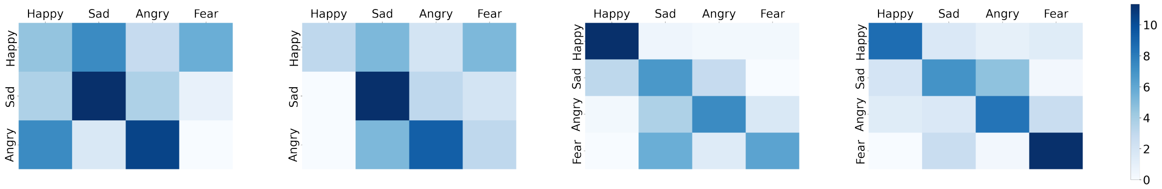
To evaluate the emotional expressiveness of the proposed model, we perform two surveys. In the first survey, given a sample, we ask the user to choose the best perceived emotion from a set of four, namely, ‘Happy’, ‘Sad’, ‘Angry’ and ‘Fear’. We ask the raters not to judge the textual content and annotate the emotion for each sample based on the rendering alone. In the second survey, we evaluate the efficacy of the models to bring about finer control over emotion. We generate two samples with the same broader emotion category but with two levels of intensity. The subject now has to identify the sample with higher intensity. For both surveys, we generate five samples per emotion and twenty samples for each model. We aggregate the rating across forty proficient English language speakers.

4.4.2 Emotional consistency in dialogues

Previous efforts in prosody-controlled TTS have been evaluated on individual sentences without context. We propose a novel evaluation strategy using excerpts from theater recordings. We replace

Model	**MOS**	**Finer Control**	**Coarse Control**				
			Happy	Sad	Angry	Fear	Average
FastSpeech2	3.80±0.13	-	-	-	-	-	-
FCL-taco2	3.39±0.14	-	-	-	-	-	-
FastSpeech2π	3.84±0.13	-	-	-	-	-	-
FastSpeech2π + EVA (Blizzard)	2.95±0.14	-	-	-	-	-	-
Cai et al., [12]	3.08±0.16	80.0	22.7	40.9	52.3	-	38.7
FastSpeech2 + EVA (av)	3.01±0.12	81.2	20.0	68.7	52.9	-	47.2
FastSpeech2 + EVA (avd)	3.05±0.17	80.2	37.5	66.6	50.0	33.3	46.8
FCL-taco2 + DS (our model)	3.30±0.14	83.5	90.1	53.3	56.5	46.8	61.8
FastSpeech2π + DS (our model)	**3.91±0.14**	**85.0**	68.4	50.0	59.5	79.1	**64.2**

Table 4.1: Results for qualitative analysis comparing our model with prior art. The model with (av) only uses arousal and valence for emotion representation while that with (avd) also uses dominance values. See Section 4.5 for details.



(a) [12] (b) [107] (c) Fcl-Taco2 + DS (d) Fastspeech2π + DS

Figure 4.4: Confusion matrices of models performance in the survey to pick the correct emotion. Rows are true emotions and columns are picked emotions. Figure to be viewed in color.

the audio of one of the actors in the conversation with renderings from a TTS model and have a human subject evaluate it for emotional consistency. The emotion for TTS renderings is chosen manually by a theater director. We compare this with TTS rendered with emotion predicted using TodKat [136] from the dialogues spoken so far. This study consolidates the two aspects of HCI we mentioned in the introduction; prosodic emotion recognition and its expression in TTS utterances.

The dataset is curated using segments from four popular plays, namely, ‘Speed-the-Plow’, ‘Night, Mother’, ‘Bobby Gould in Hell’ and ‘Death of a Salesman’. We select 30 dialogue segments collectively from the four plays with an average dialogue length of 90 seconds per segment. Timestamps of segments selected from each play is given in supplementary material. We replace the female voice in the segment with (a) a non-emotive TTS model (Fastspeech2π) (b) our model with emotion predicted for each utterance using TodKat, and (c) our model with a senior theatre director picking the emotion for each utterance. We randomly pick five dialogues from the 30 samples in all three settings for each of our surveys. We ask forty raters to rank the three settings in terms of the emotional consistency of the dialogue *i.e.*, to judge the naturalness and aptness of the emotional prosody in the given context.

4.4.3 Conversation with Meisner trained actor

Consider this hypothetical conversation between two characters called Antony and Cleopatra.

ANTONY: Hi! How are you?
CLEOPATRA: I am fine, thank you.
ANTONY: Good to know.
CLEOPATRA: And how about you?
ANTONY: I am all right, I guess.
CLEOPATRA: Great.

We can imagine the above conversation happening under day-to-day circumstances. However, if Antony just got the news that his paper won the best paper award at ACL, he is likely to start the conversation in a much more lively manner. Even if Cleopatra's paper got rejected, her response, though perhaps a little less lively, is still likely to match Antony's in terms of mood. Alternatively, if Antony just heard that he lost all the money he had invested in an NLP-based startup, he is likely to start the conversation on a sadder note. Even if Cleopatra profited from her latest crypto-currency investment, she could not stress her enthusiasm in response to Antony's sadness. In other words, natural human behavior in conversations is often guided by the behavior of other people involved in the conversation. Humans naturally take into account their own circumstances as well as other people's behavior while reacting in a conversation. Even a commonplace dialogue such as the one quoted above can have different emotional content irrespective of the text because the characters' initial emotional states differ, and their behavior depends on one another.

This insight was observed and successfully used in training actors by the celebrated acting teacher Sanford Meisner. Meisner found that actors were often indulgent in representing the behavior of their own characters rather than listening to and getting affected by other characters in front of them. He proposed that the way in which an actor uttered her lines was not to be predetermined but instead would arise at the moment as a result of the other actor's behavior. For this, he developed exercises that would hone actors' availability to listen and to get affected [70].

Meisner Experiment: A Meisner-trained actor responds to another actor taking into account his/her behavior. In this experiment, we observe how a Meisner-trained actor (Actor M) reacts in a live dialogue initiated by (a) another trained human actor, (b) a non-emotive TTS (Fastspeech2π) and (c) our model (Fastspeech2π with DS). We use the same neutral script with 18 lines in all three cases. We use the behavior of Actor M during interaction with the human as reference. The closeness of Actor M's behavior to this reference while interacting with the two TTS models is used as a measure of the latter's effectiveness in rendering speech expressive enough to evoke an emotive response.

For each of the three scenarios, the conversation is initiated with two different emotional states, viz. (a) highly positive and (b) highly negative. The emotion for our TTS model is chosen live on-the-fly by a theatre director from fourteen bins in the discretized arousal-valence space. The bins are chosen to span the V-shape around high-arousal-high-valence and low-arousal-neutral-valence [17]. We take the majority vote of three listener ratings for each utterance of Actor M on the same discretized arousal-valence space to allow quantitative comparisons.

4.5 Results

4.5.1 Comparing with prior art

Naturalness. Table 4.1 compares the audio quality of the TTS models listed in Section 4.4.1. It can be seen that the proposed model achieves affective control without a drop in perceived audio quality. In contrast, previous SOTA emotive models ([12] and FastSpeech2 + EVA) achieve control over emotion at the cost of naturalness (MOS of 3.08 and 3.01, respectively). This result demonstrates the efficacy of using the DS module over EVA and validates its ability to disentangle affective features from the acoustic ones. The MOS score of FastSpeech2π improves with the addition of DS, as some samples appear more natural when rendered in intended emotions.

Coarse affective control. Results corresponding to emotion detection are presented in Table 4.1. For each sample, the raters were asked to choose one among the four discrete emotions. On average, the FastSpeech2π + DS gives the best results, outperforming the other models by a significant margin. We observe about 17 and 25.5 improvement in percentage points (pp) over FastSpeech2 + EVA and [12] respectively. Figure 4.4 shows the confusion matrix for this survey. Our models are better at differentiating positive valence emotions from negative ones. There is still scope for improvement in distinctly expressing low-valence emotions.

Finer affective control. When asked raters to pick the sample from a pair that expresses a particular emotion better, 85% of the times, they were able to pick the sample that was actually rendered with a higher arousal value (Table 4.1). Our best performing model scores 3.8pp over FastSpeech2 + EVA and 5.0pp over [12].

Efficacy of DS. To further validate the efficacy of DS (over the EVA), we present evaluations to show that the performance gains occur primarily due to the DS module and not the other interventions. We observe that adding 'dominance' to Fastspeech2 + EVA does not improve its MOS and affective controllability, as shown in Table 4.1. Furthermore, we observe a performance drop on Fastspeech2π + EVA when compared against Fastspeech2π + DS when both have their backbones trained on the Blizzard dataset (Table 4.1). The lack of improvement from [107] further highlights that the performance gains by our model does not come from the choice of the dataset on which the backbone is trained. Overall, the two experiments conclusively show that the DS module is the decisive component that brings the improvements in naturalness and controllability to the proposed TTS system.

4.5.2 Emotional consistency in dialogues

As described in Section 4.4.2, we evaluate the emotional consistency of dialogue when a TTS model replaces an actor in excerpts from a play. Figure **??** shows that emotive models bring significant improvement in the emphatic quality of conversations and are picked 80% of the time as the first preference. This result reiterates the hypothesis [123] that prosody averaging, as in non-emotive TTS, is insufficient for emulating emotionally consistent conversations.

Another important observation is how emphatic quality measured as the user's first preference falls from 52% to 27% in moving away from hand-picked to model-predicted emotions. This suggests a scope for improvement in emotion prediction models. Nonetheless, the results present clear evidence that tying together emotion prediction models to expressive TTS is significantly preferable to a non-emotive TTS.

This proposed evaluation methodology is more comprehensive and enables the assessment of a consolidated conversational system as required in expressive HCI that includes various moving parts like causal emotion recognition in conversation and expressive TTS. This is not feasible with the traditional approach of evaluating individual sentences drawn from distinct contexts. We argue that this evaluation with contextual dialogues from a conversation is more coherent to humans as reflected in inter-annotator agreement measured by Fleiss's Kappa Score (FKS). FKS goes up by 34% from 0.43 in traditional coarse affective control (Table 1) to 0.58 for our evaluation strategy (Figure **??**). We hope this will be useful in a more thorough evaluation of expressive HCI systems.

4.5.3 Conversation with Meisner trained actor

As mentioned in Section **??**, we gather the behavioral response of a Meisner-trained human actor to TTS systems (emotive and non-emotive) and compare it against his/her reference response to another human actor. We use Pearson's correlation ρ with reference for valence and compare mean-std (μ, σ) for arousal values.

When the conversation was triggered with a positive initial emotion, we had a high ρ(FastSpeech2π+DS, human) of 0.702 for our model compared to a negative correlation for non-emotive TTS at ρ(FastSpeech2π, human) of -0.282. Similarly, for a negative initial emotion ρ(FastSpeech2π+DS, human) was high 0.838 relative to low ρ(FastSpeech2π, human) of 0.158.

We find that the average arousal for the human response to our TTS (μ=3.5, σ=1.06) is comparable to a human-human conversation (μ=3.94, σ=0.97), as opposed to the response to a non-emotive TTS (μ=2.55, σ=0.49). This indicates that the range of arousal response elicited from a human actor by our TTS is comparable to a human-human conversation as opposed to that of a prosody-unaware TTS.

We also interviewed the human actor about the experience of conversing with the TTS systems. He reported that our TTS gave him "an emotional structure". He felt that the TTS could "dictate the neutral part of the script to change it". He could "remember specific utterances" by our TTS and their emotional content, which "drove him" to respond in an emotional manner. In contrast, he reported that the prosody, unaware TTS gave ”dry answers”, made him feel that it was ”disinterested”, ”auto-generated” and ”did not evoke excitement”. He expressed that he ”could not have a longer conversation with it”.

Audio samples for our experiments and the code are available at: `https://emtts.github.io/tts-demo/`

Chapter 5

Conclusions

This work presents a novel method that leverages prosodic features (pitch, energy and duration) to modify emotions in the output of a TTS system. Our method is model agnostic and can be used with any TTS backbone that predicts prosodic features in an intermediate step. This method outperforms existing approaches by a significant margin in its ability to accurately render desired emotions, while preserving the naturalness of speech. We curated theatre conversation data to evaluate and show that our prosody-aware TTS better maintains the natural flow of emotions in conversations. Our work shows promise in consolidation of prosodic emotional recognition and expression, a coveted pursuit in the field of HCI. We present further qualitative experiments involving professional theatre artists and demonstrate that the proposed TTS method leads to more human-like conversations. While exposing valence, arousal and dominance values as model levers improves control over the final rendering, in reality it is overwhelming for the user to choose them correctly for a desired output. This is further aggravated by the fact that some sentences cannot be suitably spoken with a chosen set of values, degrading output quality. These are limitations that need to be addressed and appropriately deriving these values from semantics of text input or reference clips could be relevant future directions. Affective control is incomplete without explicit levers on the intonations, which is another limitation to be looked upon in the future work.

5.1 Future Work

Manipulating intermediate features as levers has more applications which can be explored. From the experiments it can be observed that, we can adjust pitch, duration and energy for a sequence of phonemes to emulate emotions. We can also manipulate the variances for emulating word emphasis. Word emphasis is a task where we emphasize a particular word in the synthesis. Changing emphasis on words can change the meaning of the whole sentence. For example, in the sentence “He didn’t steal my car”, if we emphasize *He*, it can mean someone else stole the car. If we emphasize *steal*, it can mean he didn’t steal it may be took the car. We experimented with manipulating the variances manually and observe the emphasis in the synthesis. But, there needs to be lot exploration needed to automatically learn the mapping between the words to be emphasised and the corresponding variances.

5.2 Ethical Concerns

This work shares the same concerns as with others in the domain of TTS systems as discussed by [29]. With TTS outputs getting closer to actual human speech, there could be a potential misuse. The threat of abuse of fake voices is particularly high with similar developments in conjugate areas like computer vision. However, the benefits of improvements to emotive TTS technology could significantly benefit HCI and the corresponding applications to problems in healthcare and other domains. Example applications include healthcare dialogue systems, improving social interaction skills in people with autism and augmentative communication devices. TTS systems synthesizing speech with empathy can ease machine interaction in many touchpoint applications. While the benefits seem to outweigh the concerns at this point, we believe the research community should proactively continue to identify methods for detection and prevention of misuse.

Related Publications

- **Empathic Machines: Using Intermediate Features as Levers to Emulate Emotions in Text-To-Speech Systems**. Saiteja Kosgi, Sarath Sivaprasad, Niranjan Pedanekar, Anil Nelakanti, and Vineet Gandhi. **In NAACL**: North American Chapter of the Association for Computational Linguistics 2022.

- **Emotional Prosody Control for Speech Generation** Sarath Sivaprasad, Saiteja Kosgi, and Vineet Gandhi. **In Proc. Interspeech 2021**

OTHER PUBLICATIONS IN MS NOT PART OF THE THESIS:

- **ParrotTTS: Text-to-Speech synthesis by exploiting self-supervised representations** Saiteja Kosgi, Neil kumar shah, Neha Sherin, Anil Nelakanti, and Vineet Gandhi. **Under review at ACL 2022.**

- **Reappraising Domain Generalization in Neural Networks** Sarath Sivaprasad, Akshay Goindani, Vaibhav Garg, Ritam Basu, Saiteja Kosgi, and Vineet Gandhi. Submitted at **ArXiv**.

- **Adversarial Robustness of Mel based speaker recognition systems.** Ritu Srivastava, Saiteja Kosgi, Sarath Sivaprasad, and Vineet Gandhi. **Under review at Signal, Image and Video Processing Journal**

Bibliography

[1] J. Allen, S. Hunnicutt, R. Carlson, and B. Granstrom. Mitalk-79: The 1979 mit text-to-speech system. *The Journal of the Acoustical Society of America*, 65(S1):S130–S130, 1979.

[2] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2017.

[3] M. Asgari, G. Kiss, J. Van Santen, I. Shafran, and X. Song. Automatic measurement of affective valence and arousal in speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 965–969. IEEE, 2014.

[4] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu. Emotion classification based on biophysical signals and machine learning techniques. *Symmetry*, 12(1):21, 2019.

[5] E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby. Effective use of variational embedding capacity in expressive end-to-end speech synthesis. *arXiv preprint arXiv:1906.03402*, 2019.

[6] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[7] J. M. Borst. The use of spectrograms for speech analysis and synthesis. *Journal of the Audio Engineering Society*, 4(1):14–23, 1956.

[8] M. B. Buchholz. Conversational errors and common ground activities in psychotherapy–insights from conversation analysis. *International Journal of Psychological Studies*, 8(3):134–153, 2016.

[9] S. Buechel and U. Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, 2017.

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[11] J. E. Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8(1):1–1, 1990.

[12] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. M. Meng. 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, 2021.

[13] M. Chu and H. Peng. Objective measure for estimating mean opinion score of synthesized speech, Apr. 4 2006. US Patent 7,024,362.

[14] C. H. Coker. Synthesis by rule from articulatory parameters. In *Proceedings of the 1967 conference on speech communication processes*, pages 52–53. IEEE Cambridge, MA, 1967.

[15] F. De Saussure. *Course in general linguistics*. Columbia University Press, 2011.

[16] P. Delattre, F. S. Cooper, A. M. Lieberman, and L. J. Gerstman. 4. speech synthesis as a research technique. In *Eight Decades of General Linguistics*, pages 77–92. Brill, 2013.

[17] R. Dietz and A. Lang. Affective agents: Effects of agent affect on arousal, attention, liking and learning. In *Proceedings of the Third International Cognitive Technology Conference, San Francisco*, 1999.

[18] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*, 2020.

[19] H. Dudley, R. R. Riesz, and S. S. Watkins. A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764, 1939.

[20] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[21] L. Euler. The wonders of the human voice. *Letters of Euler on different subjects in natural philosophy addressed to German Princess, David Brewster, ed., New York: JJ Harper (publisher in 1833)*, pages 76–79, 1761.

[22] Y. Fan, Y. Qian, F. K. Soong, and L. He. Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4475–4479, 2015.

[23] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth annual conference of the international speech communication association*, 2014.

[24] G. Fant and J. Martony. Speech synthesis instrumentation for parametric synthesis (ove ii). *Speech Transmission Laboratory Quarterly Progress and Status Report (KTH)*, 2:18–24, 1962.

[25] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2470–2481, 2020.

[26] A. Gibiansky, S. Ö. Arik, G. F. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *NIPS*, 2017.

[27] J. Grekow. Music emotion maps in arousal-valence space. In *IFIP International Conference on Computer Information Systems and Industrial Management*, pages 697–706. Springer, 2016.

[28] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[29] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby. Semi-supervised generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1910.01709*, 2019.

[30] C. M. Harris. A speech synthesizer. *The Journal of the Acoustical Society of America*, 25(5):970–975, 1953.

[31] C. M. Harris. A study of the building blocks in speech. *The Journal of the Acoustical Society of America*, 25(5):962–969, 1953.

[32] H. L. Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 2009.

[33] Z. Hodari, A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman. Camp: a two-stage approach to modelling prosody in context. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6578–6582. IEEE, 2021.

[34] J. N. Holmes, I. G. Mattingly, and J. N. Shearme. Speech synthesis by rule. *Language and speech*, 7(3):127–143, 1964.

[35] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP*, 1996.

[36] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 93–96. IEEE, 1983.

[37] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (mlsa) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communications)*, 66(2):10–18, 1983.

[38] F. Ingemann. Speech synthesis by rule. *The Journal of the Acoustical Society of America*, 29(11):1255–1255, 1957.

[39] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*, 2018.

[40] W. Jiao, H. Yang, I. King, and M. R. Lyu. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406, 2019.

[41] T. Joshi, S. Sivaprasad, and N. Pedanekar. Partners in crime: Utilizing arousal-valence relationship for continuous prediction of valence in movies. In *AffCon@ AAAI*, 2019.

[42] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[43] N. Kamaruddin and A. W. Abdul Rahman. Valence-arousal approach for speech emotion recognition system. In *2013 International Conference on Electronics, Computer and Computation (ICECCO)*, pages 184–187, 2013.

[44] S. Karlapati, A. Abbas, Z. Hodari, A. Moinet, A. Joly, P. Karanasou, and T. Drugman. Prosodic representation learning and contextual sampling for neural text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6573–6577. IEEE, 2021.

[45] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman. Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech. *arXiv preprint arXiv:2004.14617*, 2020.

[46] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[47] J. Kelly. Speech synthesis. *Proc. 4th Int. Congr. Acoustics, 1962*, pages 1–4, 1962.

[48] T. Kim and P. Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

[49] S. King and V. Karaiskos. The blizzard challenge 2013. 2014.

[50] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000.

[51] D. Klatt. The klattalk text-to-speech conversion system. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1589–1592. IEEE, 1982.

[52] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

[53] W. Koenig, H. Dunn, and L. Lacy. The sound spectrograph. *The Journal of the Acoustical Society of America*, 18(1):19–49, 1946.

[54] W. Koenig and A. Ruppel. Quantitative amplitude representation in sound spectrograms. *The Journal of the Acoustical Society of America*, 20(6):787–795, 1948.

[55] J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[56] R. König. I. on manometric flames. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 45(297):1–18, 1873.

[57] G. Kopp and H. C. Green. Basic phonetic principles of visible speech. *The Journal of the Acoustical Society of America*, 18(1):74–89, 1946.

[58] C. G. Kratzenstein. *Tentamen resolvendi problema ab Academia Scientarum Imperiali Petropolitana ad annum 1780 publice propositum: 1) qualis sit natura et character sonorum litterarum vocalium a, e, i, o, u, tam insigniter se diversorum; 2) Annon construi queant instrumenta ordini tuborum organicorum, sub termino vocis humanae noto, similia, quae litterarum vocalium a, e, i, o, u, sonos exprimant*. 1781.

[59] M. W. Kraus. Voice-only communication enhances empathic accuracy. *American Psychologist*, 72(7):644, 2017.

[60] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[61] W. Lawrence. Parametric artificial talking device. *The Journal of the Acoustical Society of America*, 32(11):1500–1501, 1960.

[62] A. F. Leentjens, S. M. Wielaert, F. van Harskamp, and F. W. Wilmink. Disturbances of affective prosody in patients with schizophrenia; a cross sectional study. *Journal of Neurology, Neurosurgery & Psychiatry*, 64(3):375–378, 1998.

[63] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019.

[64] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, 2017.

[65] A. M. Liberman, F. Ingemann, L. Lisker, P. Delattre, and F. S. Cooper. Minimal rules for synthesizing speech. *The Journal of the Acoustical Society of America*, 31(11):1490–1499, 1959.

[66] B. E. Lindblom and J. E. Sundberg. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4B):1166–1179, 1971.

[67] R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019.

[68] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.

[69] S. Massicotte-Laforge and R. Shi. The role of prosody in infants' early syntactic analysis and grammatical categorization. *The Journal of the Acoustical Society of America*, 138(4):EL441–EL446, 2015.

[70] S. Meisner and D. Longwell. *Sanford Meisner on acting*. Vintage, 2012.

[71] P. Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[72] M. Michael, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: trainable text-speech alignment using kaldi. *In Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[73] C. Michelle and Z. Georgia. Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes. In *2020 ISCA INTERSPEECH*, 2020.

[74] R. L. Mitchell and Y. Xu. What is the value of embedding artificial emotional prosody in human–computer interactions? implications for theory and design in psychological science. *Frontiers in psychology*, 6:1750, 2015.

[75] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.

[76] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[77] T. Nazzi, J. Bertoncini, and J. Mehler. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756, 1998.

[78] E. Nielsen, M. Steedman, and S. Goldwater. The role of context in neural pitch accent detection in english. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7994–8000, 2020.

[79] J. Olive. Rule synthesis of speech from dyadic units. In *ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 568–570. IEEE, 1977.

[80] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*, 2018.

[81] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[82] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, 2015.

[83] G. E. Peterson, W. S.-Y. Wang, and E. Sivertsen. Segmentation techniques in speech synthesis. *The Journal of the Acoustical Society of America*, 30(8):739–742, 1958.

[84] W. Ping, K. Peng, and J. Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*, 2018.

[85] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

[86] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. Deep voice 3: 2000-speaker neural text-to-speech. *ICLR*, 2018.

[87] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[88] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[89] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

[90] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[91] R. K. Potter. Visible patterns of sound. *Science*, 102(2654):463–470, 1945.

[92] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

[93] Y. Qian, Y. Fan, W. Hu, and F. K. Soong. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE, 2014.

[94] F. H. Rachman, R. Samo, and C. Fatichah. Song emotion detection based on arousal-valence from audio and lyrics using rule based method. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–5, 2019.

[95] T. Raitio, R. Rasipuram, and D. Castellani. Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features. In *Proc. Interspeech 2020*, pages 4432–4436, 2020.

[96] H. D. Record. The vocoder. 1940.

[97] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[98] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

[99] A. Rosenberg and J. Hirschberg. Detecting pitch accents at the word, syllable and vowel level. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, page 81–84, USA, 2009. Association for Computational Linguistics.

[100] P. Rubin, T. Baer, and P. Mermelstein. An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2):321–328, 1981.

[101] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[102] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. Atr μ-talk speech synthesis system. In *ICSLP*, volume 92, pages 483–486, 1992.

[103] M. Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[104] P. Seeviour, J. Holmes, and M. Judd. Automatic generation of control signals for a parallel formant speech synthesizer. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 690–693. IEEE, 1976.

[105] C. H. Shadle and R. I. Damper. Prospects for articulatory synthesis: A position paper. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[106] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE*

international conference on acoustics, speech and signal processing (ICASSP), pages 4779–4783. IEEE, 2018.

[107] S. Sivaprasad, S. Kosgi, and V. Gandhi. Emotional Prosody Control for Speech Generation. In *Proc. Interspeech 2021*, pages 4653–4657, 2021.

[108] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *ICML*, 2018.

[109] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio. Char2wav: End-to-end speech synthesis. 2017.

[110] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. D. Richards. Normalization of non-standard words. *Computer speech & language*, 15(3):287–333, 2001.

[111] R. Sproat and N. Jaitly. Rnn approaches to text normalization: A challenge. *arXiv preprint arXiv:1611.00068*, 2016.

[112] J. Steinberg and N. French. The portrayal of visible speech. *The Journal of the Acoustical Society of America*, 18(1):4–18, 1946.

[113] J. Q. Stewart. An electrical analogue of the vocal organs. *Nature*, 110(2757):311–312, 1922.

[114] H. Sun, X. Tan, J.-W. Gan, H. Liu, S. Zhao, T. Qin, and T.-Y. Liu. Token-level ensemble distillation for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1904.03446*, 2019.

[115] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[116] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

[117] J. Trouvain and F. Brackhane. Wolfgang von kempelen's speaking machine as an instrument for demonstration and research. In *Proceedings of the 17th International. Congress of Phonetic Sciences. 17-21 August 2011, Hong Kong*, pages 164–167, 2011.

[118] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[119] L. Wan, Q. Wang, A. Papir, and I. L. Moreno. Generalized end-to-end loss for speaker verification. In *ICASSP*, 2018.

[120] D. Wang, L. Deng, Y. Zhang, N. Zheng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng. Fcl-taco2: Towards fast, controllable and lightweight text-to-speech synthesis. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5714–5718, 2021.

[121] W. Wang, S. Xu, B. Xu, et al. First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention. In *Interspeech*, pages 2243–2247, 2016.

[122] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[123] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*, 2018.

[124] Wikipedia contributors. Speech synthesis — Wikipedia, the free encyclopedia, 2021. [Online; accessed 28-October-2022].

[125] Wikipedia contributors. Brazen head — Wikipedia, the free encyclopedia, 2022. [Online; accessed 28-October-2022].

[126] R. D. William D Stanley, Gary R Dougherty and H. Saunders. *Digital signal processing.* Journal of Vibration and Acoustics-transactions of The Asme, 1988.

[127] Z. Wu, O. Watts, and S. King. Merlin: An open source neural network speech synthesis system. In *SSW*, pages 202–207, 2016.

[128] T. Yoshimura. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems. *PhD diss, Nagoya Institute of Technology*, 2002.

[129] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

[130] L. Younggun and K. Taesu. Robust and fine-grained prosody control of end-to-end speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[131] H. Ze, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.

[132] H. Zen. Acoustic modeling in statistical parametric speech synthesis-from hmm to lstm-rnn. 2015.

[133] H. Zen and H. Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474. IEEE, 2015.

[134] H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *speech communication*, 51(11):1039–1064, 2009.

[135] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337, 2019.

[136] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online, Aug. 2021. Association for Computational Linguistics.