MANIPULATING EMOTIONS: GENERATIVE MODELING OF PROSODY FOR EMOTIONAL SPEECH SYNTHESIS

by

Ravi Shankar

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

October, 2023

Abstract

This thesis focuses on data-driven emotional speech generation using prosodic elements. Speech, a universal mode of communication, carries vital information beyond semantics, such as speaker identity and emotion. This work emphasizes intonation, intensity variation, and rhythm modulation as key prosodic elements for emotional speech understanding and generation.

Throughout this thesis, we combine probabilistic modeling with deep neural networks to transform neutral speech into emotional speech. We explore supervised, unsupervised, and reinforcement learning paradigms and rigorously evaluate these techniques against state-of-the-art models. The VESUS corpus, a reference dataset collected by us, ensures our methods generalize across multiple speakers and unseen vocabulary.

Chapter 1 introduces speech production, emotion models, and the importance of prosody in emotional speech perception. It sets the stage for emotional speech generation using prosodic transformation.

Chapter 2 provides essential technical background on diffeomorphic mapping, variational inference, and graphical models. These concepts are crucial for understanding the subsequent chapters.

Chapter 3 and Chapter 4 focus on supervised models for modifying

prosody (F0 and energy) using the VESUS corpus. The former presents two models: one based on a highway network with gender embeddings and another employing diffeomorphic regularization. Chapter 4 extends the model to predict F0 and energy contour at the utterance level, leveraging segmental and supra-segmental properties.

Chapter 5 introduces the Variational CycleGAN framework for unsupervised prosody modification, addressing the limitations of vanilla CycleGAN.

Chapter 6 presents a supervised rhythm modulation algorithm combining generative modeling and dynamic time warping (DTW) to align input speech with a hypothetical desired output. It uses latent variable modeling for attention maps and DTW similarity matrices.

Finally, Chapter 7 discusses an unsupervised duration modification method employing reinforcement learning. This approach identifies important segments within an utterance using a masking strategy with a first-order Markov property, with the agent learning a distribution over transformation options.

In summary, this thesis employs diverse techniques, from supervised to unsupervised learning, to enhance emotional speech using prosodic elements, culminating in comprehensive evaluation and applicability to various speakers and vocabulary.

Thesis Committee

Primary Readers

Archana Venkataraman (Primary Advisor)
Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Mounya Elhilali (Second Reader)
Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Rama Chellappa
Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Nicolas Charon
Assistant Professor
Department of Applied Math and Statistics
Johns Hopkins Whiting School of Engineering

Anurag Kumar
Research Scientist
Meta Reality Labs, Redmond, WA

Acknowledgments

I would like to express my deepest gratitude to the members of my dissertation committee for their invaluable guidance, support, and expertise throughout this journey.

First and foremost, I extend my heartfelt appreciation to Prof. Rama Chellappa, Prof. Mounya Elhilali, Prof. Archana Venkataraman, Dr. Nicolas Charon, and Dr. Anurag Kumar. Your unwavering commitment to my research and your insightful feedback have been instrumental in shaping this dissertation. I am truly fortunate to have had the opportunity to learn from each of you.

I want to acknowledge my family for their unconditional love and encouragement. To my mom and dad, your constant belief in my abilities has been a driving force behind my success. I also remember my beloved brother, who tragically succumbed to COVID-19 in 2021. His memory continues to inspire me, and I dedicate this work to his loving memory.

To my friends, both in the lab and beyond, your support has been a source of strength and motivation. To my lab members – your camaraderie and collaboration have made this journey enjoyable and rewarding.

Special thanks go to my girlfriend, Divya, for her support, understanding,

and patience throughout this challenging endeavor. Your love has been my anchor.

Lastly, I want to acknowledge my faithful companion, Buddy Barnes, whose wagging tail and boundless enthusiasm provided much-needed breaks and moments of joy during long hours of research.

To all those mentioned and the countless others who have contributed to my personal and academic growth, I extend my deepest gratitude. Your support and belief in me have been indispensable, and I am truly humbled by your presence in my life.

Thank you all for being a part of this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Effective and natural communication has played a crucial role in the evolution of mankind. It is by the virtue of communication via languages that we have been successful at organizing in groups, and thwarted off enemies for survival. At the heart of natural communication is our ability to speak. Speaking is the easiest way to manifest our thoughts, purpose and desires. We as human beings are exposed to a plethora of sounds long before we can read and see, which makes it fundamental to the understanding of world around us.

Decades of research have led to a very deep understanding of the way speech is generated by humans. In fact, we can easily mimic the speech production process by a computer using very simple techniques rooted in basics of signal processing. These primitive models generate intelligible speech, but usually have a buzziness (called vocoder characteristics) which is often considered unpleasing. Lately, the quality of such machine generated speech has seen tremendous improvements, credit to the deep neural networks and ability to train them, which allow us to approximate complex distributions in very high dimensions. Broadly, there are two main components of human

speech: first is the content which provides semantic meaning to the words spoken, and second is the manner or style of speaking. The manner or style of speech signal carries para-linguistic information about speaker's identity, mood and intent. Therefore, speech represents a more rich and varied form of information than text. It is therefore natural to demand the same amount of richness from a machine synthesised speech for natural conversations. Unfortunately, it is a difficult problem which has eluded many speech and audio researchers. The difficulty arises from the subjectivity of speaker and emotion specific characteristics in speech. While research in the domain of speaker identification has been largely successful, emotions are hard to predict from short speech utterances due to factors such as context, and simply due to variations in manifesting them. Additionally, emotion perception from speech is closely tied to speaker's knowledge. Therefore, disentangling an emotion representation from speech is a challenging task worth exploring.

Having said that, prior research in the domain of speech perception have identified key para-linguistic aspects of speech that are crucial for underlying emotional intent. These features formally belong to the group of prosodic features. Intonation, voice quality, intensity, timbre and rhythm are some of the important prosodic features that provides uniqueness to a speech signal and characterize speaking style or manner of speech.

1.1 Speech Production

The speech production process starts with the formulation of thoughts into words with the intended meaning. The brain, then sends out signals to the

articulators in order to produce each sound of the utterance. The physical activity begins with the lungs pushing out air through the oral and nasal cavity. The volume of air pushed out encounters the vocal folds which is either tensed or relaxed. When the vocal folds are tensed, it vibrates at a regular rhythm to produce voiced sounds, such as, /a/, /i/, etc. The frequency of vibration is determined by the amount of tension in vocal folds. When the folds are relaxed, there is no vibration and the air passes straight through it.

The vocal tract includes larynx, pharynx and oral cavities. The shape of vocal tract is responsible for certain resonances and anti-resonances in the acoustic produced. These resonance frequencies determine the identity of phonemes or sounds. Voiced sounds have prominent resonances in the 0-3000Hz range, whereas unvoiced sounds have peaks in the tail of the frequency spectrum. The vocal tract is modelled as a combination of tube of changing widths and lengths. A cylinder is the simplest approximation of the vocal tract for which as linear system can be designed [1].

1.2 Models of Emotion

There are many models of emotional categories that also depict the relationship between then. Some of the popular ones are shown in Fig. 1.1.

1.2.1 Plutchik's Model

The Plutchik Wheel of Emotions, also known as the Plutchik Emotion Circumplex, is a model that visually represents various human emotions and their relationships [2]. It was developed by the American psychologist Robert

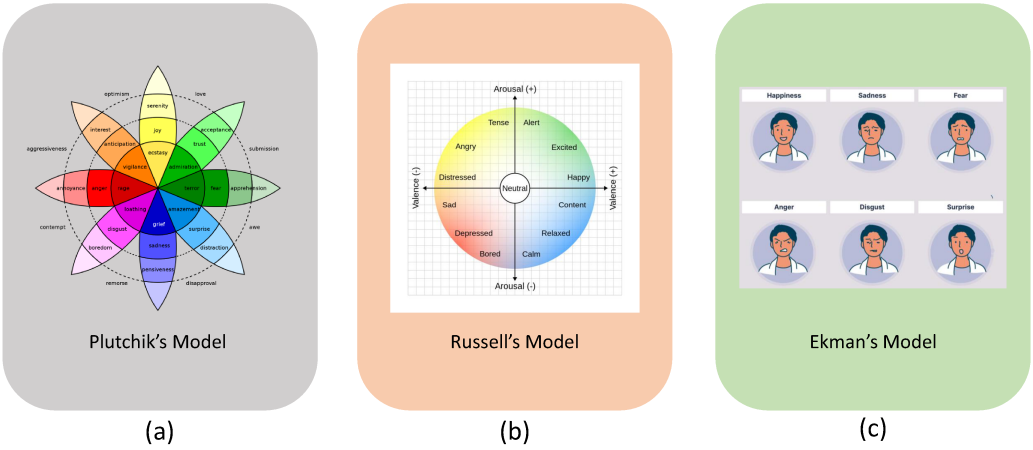

Figure 1.1: Models of emotion proposed by psychologists.

Plutchik in the 1980s as a way to conceptualize and organize the complex array of human emotions. The model is often depicted as a circle with eight primary emotions placed equidistant from each other, forming a color wheel-like diagram. The eight primary emotions in the Plutchik Wheel are:

- Joy
- Trust
- Fear
- Surprise
- Sadness
- Disgust
- Anger
- Anticipation

Plutchik argued that these eight primary emotions are fundamental and can combine in various ways to produce secondary and tertiary emotions. He also proposed that these emotions can be arranged in pairs of opposites, such as joy and sadness or trust and disgust, to create additional emotional states. Furthermore, Plutchik's model suggests that emotions can intensify or mellow through varying degrees, leading to a more nuanced understanding of emotional experiences. For example, the emotion of anger can range from mild irritation to intense rage.

1.2.2 Russell's Model

Russell's model of emotions [2], also known as the circumplex model of affect, is a psychological framework that aims to understand and represent emotions based on two primary dimensions: valence and arousal. The two main dimensions in Russell's model are:

- **Valence**: This dimension reflects whether an emotion is experienced as positive (pleasant) or negative (unpleasant). Emotions with a positive valence are typically associated with feelings of happiness, joy, and contentment, while those with a negative valence are linked to emotions like sadness, anger, and fear.

- **Arousal**: Arousal refers to the level of physiological activation or intensity associated with an emotion. Emotions can vary in terms of their arousal, ranging from low arousal (calm, relaxed) to high arousal (excited, agitated). For example, calmness and boredom are low-arousal states, while excitement and anxiety are high-arousal states.

Russell's model places emotions within this two-dimensional space, allowing for a more nuanced understanding of emotional experiences. Emotions are positioned in relation to their valence (positive to negative) and arousal (low to high) levels. This results in a circular diagram where emotions can be located based on their characteristic valence and arousal levels.

1.2.3 Ekman's Model

Ekman's model of emotion is based on the idea that there are several basic or primary emotions that are universally recognized across cultures [2]. These basic emotions are considered to be biologically hardwired and share common facial expressions and physiological responses. Ekman initially identified six basic emotions:

- **Happiness:** Associated with a smiling facial expression.
- **Sadness:** Characterized by a frowning facial expression.
- **Anger:** Recognized by a furrowed brow and clenched jaw.
- **Fear:** Marked by wide eyes and a tense facial expression.
- **Disgust:** Typically shown with a wrinkled nose and raised upper lip.
- **Surprise:** Displayed with raised eyebrows and widened eyes.

These basic emotions, according to Ekman, are considered to be universal because they are recognizable across different cultures and are associated with distinct and specific facial expressions. Ekman's work also highlighted the importance of micro-expressions, which are very brief and often involuntary

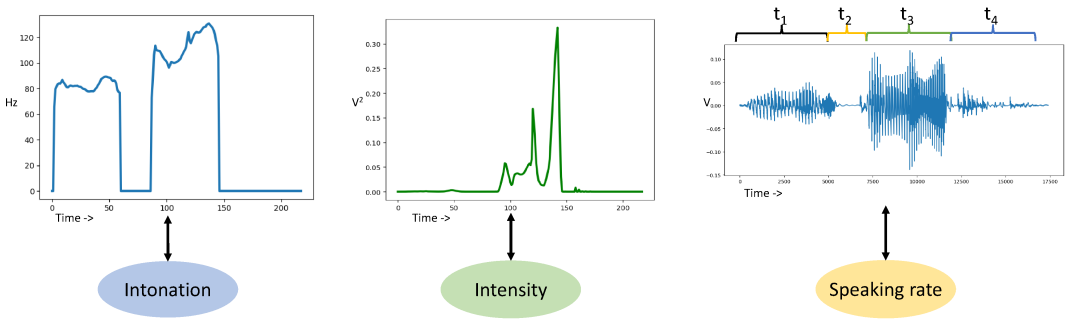

Figure 1.2: Prosodic features in speech that characterize emotion perception.

facial expressions that can reveal concealed emotions. He developed the Facial Action Coding System (FACS) as a comprehensive tool for analyzing and categorizing facial muscle movements associated with emotions.

1.3 Prosodic Features for Emotion

In this work, we will address the problem of emotional speech generation via emotion conversion of neutral sounding speech. Note that, this is a passive model of emotional speech synthesis which relies on the availability of a vocoder that can generate neutral speech of sufficiently high quality. We argue that such vocoders are easily available and can be run cheaply in real-time [3, 4, 5]. Therefore, our task boils down to learning an appropriate transformation of prosodic features (i.e., intonation, speaking rate and intensity) from neutral emotion to angry, happy or sad.

1.3.1 Intonation

Intonation is defined as the rise and fall of the note while singing. Therefore, it is characterized by the variation of pitch of frequency of vocal cords. This

is also true while speaking. The fundamental frequency of vibration (F0) of vocal cords determines the pitch at any instantaneous moment and its fluctuation over time is called the F0/pitch contour. F0/pitch contour is a very strong correlate of intonation so modifying pitch contour significantly affects underlying intent in speech signal.

1.3.2 Intensity

Intensity variations is determined by the variation of energy of the signal in small windows of 10-30 ms. Speech is a short-term stationary signal, meaning the properties of signal change every few milliseconds. Therefore, analysis window of size 10-30 ms with some overlap typically works well in practice for short-term feature extraction such as energy. The change in energy signature over time is called the energy/intensity contour which is yet another prosodic feature affecting speech emotion perception.

1.3.3 Speaking Rate

Finally, speaking rate or the variation of speaking rate can denote the urgency of the situation while speaking therefore, it can affect the underlying intent. Speaking rate modulation is primarily used to emphasize certain portion of the speech and de-emphasize others. Hence, learning a speaking rate modification function from neutral to emotional speech is of importance. However, unlike intonation and intensity, speaking rate modulation does not have any parameterization which makes it challenging to develop a transformation function for it. Next, we will see the standard prosody feature extraction

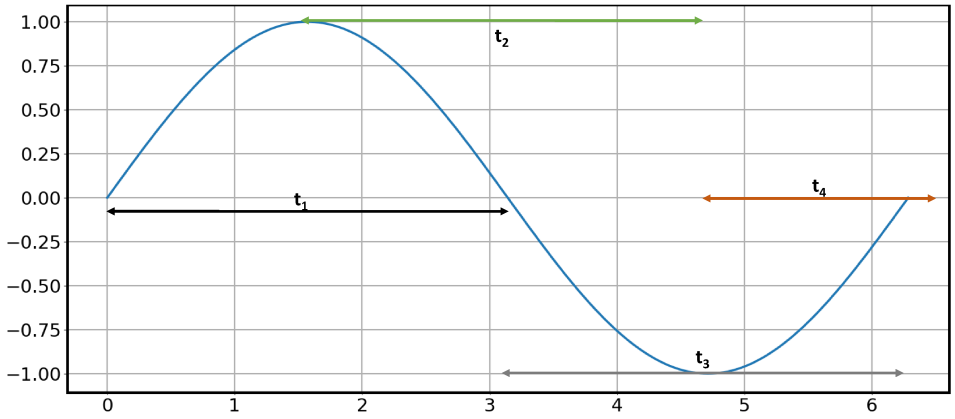

Figure 1.3: F0 candidates elimination in DIO algorithm.

pipeline and methods for pitch and intensity modification.

1.4 Feature Extraction

To extract the pitch and intensity information, we use the WORLD vocoder [6] based decomposition. The decomposition is based on source-filter modeling of speech. The first step in WORLD vocoder decomposition is the estimation of F0 values in windows of specified length. DIO [7] algorithm. DIO has three steps: first is low-pass filtering with multiple cutoff frequencies followed by calculation of F0 candidates and their reliability. Finally, selection of the candidate with highest reliability. The main idea is: if any filtered speech signal consist purely of fundamental component, the corresponding reliability measure based on sine wave will have low variation over multiple measures of the time period. Fig. 1.3 depicts the candidate filtering process in DIO.

After extracting F0 values for each short-time window, the next step is the estimation of spectral envelope of the signal using CheapTrick algorithm. It uses pitch-synchronous analysis of the utterance with a Hanning window of

length 3 x time-period. This facilitates power stabilization of each window. Denoting the speech signal by $y(t)$ and the Hanning window with $w(t)$, the windowing results in:

$$\int_0^{3T_0} [y(t)w(t)]^2 dt = 1.125 \int_0^{T_0} [y(t)]^2 dt \tag{1.1}$$

Suppose the pitch of a window is denoted by ω_0, the power spectrum of each window is smoothed by a rectangular window of length $2\omega_0/3$ which gives a smoothed power spectrum:

$$P_s(\omega) = \frac{3}{2\omega_0} \int_{-\omega_0/3}^{\omega_0/3} P(\omega + \lambda)\, d\lambda \tag{1.2}$$

Finally, a set of filtering operation (liftering) is carried out to compensate for the effect of zeroing out multiples of ω_0 frequency. Therefore, the overall process can be written as:

$$p_s(\tau) = \mathscr{F}^{-1}(\log(P_s(\omega))) \text{ then, } P_l(\omega) = \exp \mathscr{F}\left[ł_s(\tau) ł_q(\tau) p_s(\tau)\right]$$

$$\text{here, } ł_s(\tau) = \frac{sin(\pi f_0 \tau)}{\pi f_0 \tau} \text{ and, } ł_q(\tau) = q_0 + 2q_1 cos\left(\frac{2\pi\tau}{T_0}\right)$$

$\mathscr{F}$ represents the Fourier transform operation and $ł_q(\tau)$ is the recovery filter which removes effect of rectangular smoothing of power spectrum in the previous stage.

Finally, the aperiodicity is obtained via application of PLATINUM [8] algorithm. The aperiodicity components captures the randomness in the quasi-periodic nature of excitation signal. It is useful for reconstructing speech of good quality, but does not affect the underlying mood/intent of speech.

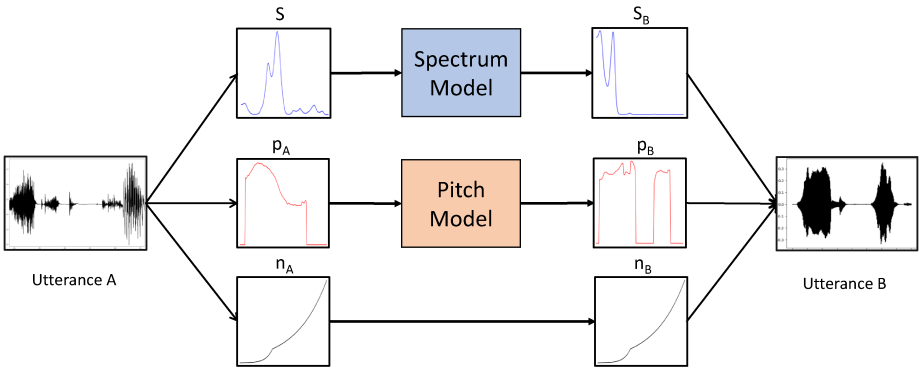

Figure 1.4: A typical prosody conversion pipeline which has been widely used in the past to inject emotion into neutral speech. The first step is to decompose speech into spectrum, pitch and aperiodic components (using WORLD vocoder). A separate model for spectrum and pitch converts is learned using available data followed by an overlap-add based synthesis.

We defer the discussion on aperiodicity to the corresponding paper. For re-synthesizing speech, given the pitch contour: $\mathbf{p}_t$, spectral envelope: $\mathbf{S}_{t,f}$ and aperiodicity: $\mathbf{A}_{t,f}$, we use the overlap add method to stitch the individual segments of speech generated at each time index t. The spectral envelope component is used to generate a minimum-phase system which is convolved with the excitation signal to produce speech. We represent the synthesis operation by following: $\tilde{y}(t) = WORLD(\mathbf{S}, \mathbf{p}, \mathbf{A})$ where the subscript t has been removed to denote usage of each frame for overlap-add.

1.5 Prosody Modification for Emotional Speech

In this section, we discuss the general idea of modifying speech prosody for emotional speech generation. A very general conversion process is shown in Fig. 1.4. The following subsections describe each component in detail.

1.5.1 Pitch and Intensity Modification

Owing to the similarity between intonation and intensity parameterization (1-D curves), we will discuss their transformation strategy together. Once F0 and spectral envelope is extracted, the energy contour can be determined using $\mathbf{e}_t = \sum_f \mathbf{S}_{t,f}^2$. Therefore, the dimensionality of pitch contour $\mathbf{p}_t$ and energy contour $\mathbf{e}_t$ is same, given by the number of analysis window on the utterance. For a pair of domains A and B, the prosody transformation function is denoted by $F : (\mathbf{P}_A \times \mathbf{E}_A) \rightarrow (\mathbf{P}_B \times \mathbf{E}_B)$. Using this function, we can generate the appropriate pitch and energy contour in target domain (such as neutral to angry). Then, we can generate the corresponding speech by first creating the energy adjusted spectral envelope, i.e., $\mathbf{S}_B = \mathbf{S}_A \times \sqrt{\frac{\mathbf{E}_B}{\mathbf{E}_A}}$. The generated speech in domain B is then $y_B(t) = WORLD(\mathbf{S}_B, \mathbf{p}_B, \mathbf{A})$.

1.5.2 Speaking Rate Modulation

As mentioned in the last section, speaking rate is determined by the # of words/syllables spoken per minute. Therefore, there is no direct way to compute the changes in speaking rate in an utterance without speech-to-text decoding [9]. Blind word/syllable segmentation procedures can be employed, but their segmental property will make the learning function discrete. In this thesis, we will handle this issue in an indirect manner. We defer the discussion to chapters 6 and 7 where we will learn about a supervised and an unsupervised method for speaking rate modulation.

References

[1] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, and Paavo Alku. *Introduction to Speech Processing*. 2nd ed. 2022. DOI: 10.5281/zenodo.6821775. URL: https://speechprocessingbook.aalto.fi.

[2] Wikipedia contributors. *Emotion classification — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=Emotion_classification&oldid=1168249679.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis". In: *CoRR* abs/2010.05646 (2020). arXiv: 2010.05646. URL: https://arxiv.org/abs/2010.05646.

[4] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499 (2016). arXiv: 1609.03499. URL: http://arxiv.org/abs/1609.03499.

[5] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *CoRR* abs/1712.05884 (2017). arXiv: 1712.05884. URL: http://arxiv.org/abs/1712.05884.

[6] Masanori MORISE, Fumiya YOKOMORI, and Kenji OZAWA. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". In: *IEICE Transactions on Information and Systems* E99.D.7 (2016), pp. 1877–1884. DOI: 10.1587/transinf.2015EDP7457.

[7] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. “Fast and Reliable F0 Estimation Method Based on the Period Extraction of Vocal Fold Vibration of Singing Voice and Speech”. In: 2009. URL: https://api.semanticscholar.org/CorpusID:60600726.

[8] Masanori Morise. “PLATINUM: A method to extract excitation signals for voice synthesis system”. In: *Acoustical Science and Technology* 33.2 (2012), pp. 123–125. DOI: 10.1250/ast.33.123.

[9] F. Grosjean and Harlan Lane. “How the listener integrates the components of speaking rate.” In: *Journal of experimental psychology. Human perception and performance* 2 4 (1976), pp. 538–43. URL: https://api.semanticscholar.org/CorpusID:44576527.

Chapter 2

Background

Speech is perhaps our primary mode of communication as humans. It is a rich medium, in the sense that both semantic information and speaker intent are intertwined together in a complex manner. The ability to convey emotion is an important yet poorly understood attribute of speech. Common work in speech analysis focuses on decomposing the signal into compact representations and probing their relative importance in imparting one emotion versus another. These representations can be broadly categorized into two groups: acoustic features and prosodic features. Acoustic features (e.g., spectrum) control resonance and speaker identity. Prosodic features (e.g., F0, energy contour) are linked to vocal inflections that include the relative pitch, duration, and intensity of each phoneme. Together, the prosodic features encode stress, intonation, and rhythm, all of which impact emotion perception. For example, expressions of anger often exhibit large variations in pitch, coupled with increases in both articulation rate and signal energy. In this paper, we develop an automated framework to transform an utterance from one emotional class to another. The problem, known as *emotion conversion*, is an important stepping

stone to affective speech synthesis.

Broadly, the goal of emotion conversion is to modify the perceived affect of a speech utterance without changing its linguistic content or speaker identity. This setting allows the user to control the speaking style, while allowing the model to be trained on limited data. Emotion conversion is a particularly challenging problem due to the inherent ambiguity of emotions themselves [1, 2]. The boundaries between emotion classes are also blurry, and prior knowledge about the speaker can sometimes play a major role in the emotion perception. That being said, one of the main application of emotion conversion is to evaluate the quality of human-machine dialog systems [3]. Here, intonation changes can indicate the level of naturalness of a conversation between a machine and a person. Emotion conversion can also be helpful in studying neurodevelopmental disorders such as autism, which is characterized by poor emotion perception capability. On the technical front, being able to control the granularity of the emotion expression in synthesized speech is an important step towards developing an intelligent conversational system. Finally, emotion conversion can be used for data augmentation when training emotion classification or speaker recognition systems [4, 5].

Early work in emotion conversion traces its roots to neuroscientific experiments, which were designed to study the influence of emotions in the brain. Interestingly, many of the implicated features tend to generalize across languages. For example, the work of [6] determined the F0 (i.e., pitch) contour and the energy (loudness) profile as the main factors responsible for primary emotions. Additionally, voice quality and utterance duration have

also been identified as features affecting emotion perception [7]. Voice quality is a function of the spectrum representation and duration can be called as a proxy for the speaking rate. A comprehensive study was conducted by [8] to understand the impact of systematically changing acoustic and prosodic features on emotional perception. These experiments were performed on a Japanese language database with some consistency shown for English.

2.1 Prior Works

Algorithms for emotion conversion fall into three general categories. The first approach relies on constructing a statistical model of the source and target prosodic features to allow inference from one domain to another. One example of this approach is the work of [9], which uses classification and regression trees (CART) to modify the F0 contour in Mandarin. An alternate strategy uses a Gaussian Mixture Model (GMM) for voice and emotion conversion. The central idea is to learn a GMM that captures the joint distribution of the source and target emotional speech features during training. Inference of a new conversion is done via the conditional mean of the target features given the test source features. Mathematically, let $\mathbf{z}_i = [\mathbf{x}_i \quad \mathbf{y}_i]^{\mathrm{T}}$ denote the concatenated source and target features and c_i denote the latent cluster assignment for utterance i. From here, we have:

$$P(\mathbf{z}_i|c_i) = \sum_{k=1}^{K} P(\mathbf{z}_i|c_i = k)P(c_i = k) \tag{2.1}$$

where, $P(\mathbf{z}_i|c_i = k) \sim N(\mathbf{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and its parameters are estimated via the Expectation-Maximization (EM) algorithm, along with the latent prior

$P(c_i = k)$. Using properties of the Gaussian distribution, it can be shown that the conditional mean of the target features $\mathbf{y}_i$ given the source features $\mathbf{x}_i$ is given by the expression

$$E[\mathbf{y}_i|\mathbf{x}_i] = \sum_{k=1}^{K} P(c_i = k|\mathbf{x}_i)\left[\boldsymbol{\mu}_k^y + \boldsymbol{\Sigma}_k^{xy}(\boldsymbol{\Sigma}_k^{xx})^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k^x)\right] \tag{2.2}$$

where, $P(c_k|x)$ can be computed via Bayes' Rule. One of the main drawback of this approach is the over-smoothing of spectral parameters in inference stage due to averaging effect. To counter this, a global variance constraint based inference proposed by [10] was adopted for emotion conversion by [11].

The second approach for emotion conversion is based on sparse recovery [12]. This technique entails learning an over-complete dictionary of both acoustic and prosodic features for each emotion class. During conversion, the input utterance is first decomposed using the source emotion dictionary by estimating a coefficient matrix with sparsity prior. These coefficients are then used for reconstruction using the target emotion dictionary elements/atoms. The authors of [12] used active Newton-set [13] based non-negative matrix factorization [14] to estimate the sparse coding. Mathematically, given a non-negative matrix input $\mathbf{X}$ (e.g., spectrogram magnitude), we seek non-negative matrices $\mathbf{U}$ and $\mathbf{V}$ to minimize:

$$\mathcal{J} = \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \lambda \sum_j \|\mathbf{V}(:, j)\|_1 \tag{2.3}$$

The first term in Eq. (2.3) enforces the data fidelity, whereas the second term encourages sparsity of the learned encoding $\mathbf{V}$. The variable $\mathbf{U}$ denotes the overcomplete dictionary.

The third approach for emotion conversion relies on deep neural networks to automatically learn complex and nonlinear speech modifications. For example, a bidirectional LSTM approach has been suggested by [15, 16] for modifying the prosodic features. The authors further proposed using a continuous wavelet transform based parameterization for the F0 and energy contour to decompose into segmental and supra-segmental components. Our prior work proposed an alternative method for prosodic modification based on highway neural networks [17, 18], which maximize the representation log likelihood in an EM algorithm setting. We further proposed an F0 modification scheme using the principle of diffeomorphic curve warping as a smoothness prior for the transformed F0 contour [19]. This diffeomorphic parameterization was extended to spectrum modification in [20]. Specifically, we used a latent variable regularization technique to sequentially modify the F0 contour and the spectrum.

The methods discussed so far belong to the domain of supervised learning. Namely, they rely on labeled parallel speech data to learn the requisite emotion conversion. Curating parallel corpora is expensive, which explains why there are only a handful of such databases [21] available online. Beyond data scarcity, most supervised emotion conversion methods require the parallel utterances to be time-aligned using dynamic time warping (DTW) [22] prior to analysis. This alignment procedure allows us to learn a frame-wise mapping between the source and target utterances. While simple and apt for smaller corpora, DTW is prone to errors, particularly during periods of silence or unvoiced sounds.

The current iteration of methods focus on unsupervised emotion conversion and do not require parallel data. These models rely on expressiveness of neural networks to learn a parametric distribution for each pair of emotions. One of the most prominent model in this space is Generative Adversarial Network (GAN). Mathematically, let G and D denote the generator and discriminator, respectively. The objective of the GAN is a minimax loss given by the following:

$$\mathcal{L}_{adv} = \min_{G} \max_{D} \; E_{x \sim P(X)}[\log(D(x)] + E_{z \sim P(Z)}[\log(1 - D(G(z))] \tag{2.4}$$

where $P(X)$ denotes the data distribution and $P(Z)$ denotes a noise density which is usually Normal i.e, $N(0, I)$.

The Cycle-GAN architecture goes one step beyond Eq. 2.4 by tying two separate GANs together via a cycle consistency objective. Formally, let A and B denote the domains of the source and target data distributions. The two generators in Cycle-GAN are tasked with learning transformation from $A \rightarrow B$ and $B \rightarrow A$, respectively. The cycle consistency loss connects the generators by enforcing that the sequence of transformations, i.e. $A \rightarrow B \rightarrow A$ should look similar to the original input. For clarity, we will refer to these generators as the “forward" and “backward" transformations of the Cycle-GAN and use the notation G_γ (forward) and G_θ (backward). Mathematically, the cyclic objective is written as:

$$\mathscr{L}_{cycle} = E_{x \sim P(X)} \left[\| x - G_\theta(G_\gamma(x)) \|_1 \right] \tag{2.5}$$

The algorithm of [23] uses a Cycle-GAN to disentangle the content and

style of a speech utterance into two separate variables based on *a priori* information embedded into the network architecture. Another approach proposed by [24] uses a Cycle-GAN to transform the F0 contour and spectrum, as parameterized by a discrete wavelet transform, for emotion conversion. A Star-GAN [4] model proposed by [5] relies on a multi-task discriminator and a single generator for conversion between multiple emotional classes. Due to the poor quality of generated samples, the authors used this method for data augmentation to improve emotion classification accuracy, rather than for speech synthesis. While all these methods show tremendous promise, one common drawback is that they have been trained and evaluated on single speaker datasets. Thus, it is unclear how they will perform in either a multi-speaker or an out-of-sample generalization setting.

2.2 Dataset: VESUS

Varied Emotion in Syntactically Uniform Speech (VESUS) [21] repository as a new resource for the speech community. VESUS is a lexically controlled database, in which a semantically neutral script is portrayed with different emotional inflections. In total, VESUS contains over 250 distinct phrases, each read by ten actors in five emotional states. The authors use crowd sourcing to obtain ten human ratings for the perceived emotional content of each utterance. Its unique database construction enables a multitude of scientific and technical explorations.

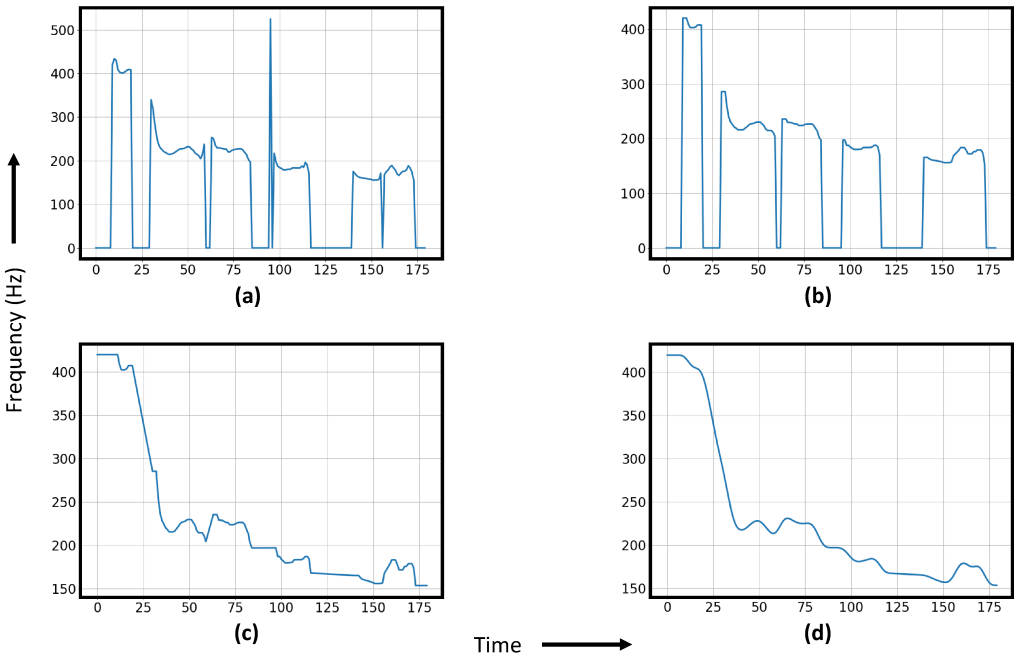

Figure 2.1: Pre-processing of prosodic features before learning any transformation function. (a) F0 extracted using DIO algorithm, (b) removing erroneous spikes using median filtering of kernel size 3, (c) removing the zeros using linear interpolation, and (d) smoothing the interpolated pitch using mean filtering.

2.3 Pre-processing of Prosodic Features

Prosodic features such as F0 contour and energy contour have discontinuities at word boundaries. This discontinuous nature is tedious to deal with in machine learning models, so we develop a separate pre-processing pipelines for F0 and energy contour to make them amenable for data-driven learning (see Fig. 2.1). After extracting the pitch contour, the first step in pre-processing is to remove sharp transitions that appear due to imperfect extraction algorithm. We use a median filtering with a window of size 5 to remove false voiced/unvoiced detection. A linear interpolation of the F0 contour in regions of unvoiced signal is applied followed by a mean smoothing operation using a kernel of size 13. Note that, this pre-processing step removes the information about unvoiced portion of speech which can be handled by storing the unvoiced frames indices in a database.

For energy contour, we carry out the same pre-processing but remove the zero interpolation operation. This is done because zeros in energy contour specifies regions of silence instead of unvoiced frames.

2.4 Diffeomorphic Transformation and LDDMM

A diffeomorphic mapping is a smooth and invertible mapping between two manifolds. It is a shape preserving transformation. Two manifolds M and N are said to be diffeomorphic is there exist a one-to-one continuously differentiable function f such that $f(M) = N$. For the purpose of function identifiability, we are typically interested in specification of manifolds up to a

diffeomorphism. This search can be made coarser by using the homeomorphic criterion instead of diffeomorphism. These mapping functions can be used to transition from one manifold to another, for example, the manifold of pitch contour in neutral emotion to angry emotion.

The idea of diffeomorphism stems from the concepts in abstract algebra where a vector representation in $\mathbb{R}^n$ represents a collection of points, a curve or a surface. For shape analysis and matching, high dimensional diffeomorphism is generated via smooth flows $\phi_t \forall t \in [0,1]$ satisfying the following ordinary differential equation:

$$\frac{d}{dt}\phi_t = v_t \circ \phi_t, \text{ where } \phi_0 = id \tag{2.6}$$

Here, v_t is the vector field (Eulerian) that determines the flow. The vector fields are continuously differentiable, at least once. They are modelled as belonging to a Hilbert space $(V, \|\cdot\|)$ using Sobolev embedding theorems. Therefore, diffeomorphic maps are defined via composition of smooth vector fields. The diffeomorphism group are all those flows which have absolutely integrable vector fields in Sobolev norm, i.e.,

$$Diff_V = \{\varphi = \phi_1 : \dot{\phi}_t = v_t \circ \phi_t,\ \phi_0 = id,\ \int_0^1 \|v_t\|_V\, dt < \infty\} \tag{2.7}$$

Large deformation diffeomorphic mapping is a specific type of algorithm for diffeomorphic mapping where the objective is to learn a map when the differences between landmark points (on curves/surfaces) are large. For landmark matching between a paired collected of points $\{x_i, y_i\}_{i=1}^N$, [25] proposed the

Figure 2.2: Graphical model: observed variable X is generated from latent variable Z.

following formulation:

$$\min_{v:\dot{\phi}_t = v_t \circ \phi_t} J(v) = \frac{1}{2} \int_0^1 \int_{\mathbb{R}^3} Av_t \cdot v_t \, dx \, dt + \frac{1}{2} \sum_i (\phi_1(x_i) - y_i)^T (\phi_1(x_i) - y_i) \tag{2.8}$$

Here, A is the differential operator $A : V \to V^*$ which determines the norm via $\|v\|_V^2 = \int_{\mathbb{R}^3} Av \cdot v \, dx, \; v \in V$. V^* is the dual of V and Av is the generalized function in the dual space. This formulation will been used later in this thesis to estimate the deformation field for F0/energy contour estimation.

2.5 Variational Inference

Very often our data is generated by conditioning on some additional factors. For example, in a Gaussian mixture model, the underlying assumption is that there exist a random variable sampling the cluster index to generate the data from. In a more general sense, the data generation process can have certain hidden variable that we might never see in practice but need to infer about. Fig. 2.2 shows the a very simple generative process where the observed data is represented by X and the latent variable (e.g. cluster index in GMM) is denoted by Z. Our goal is to infer the posterior distribution of Z conditioned

on X, i.e. $P(Z|X)$ which using the Bayes' rule can be written as:

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}, \text{ where } P(X) = \int P(X|Z=z)P(z)\, dz \tag{2.9}$$

The function $P(X)$ is called the evidence as it is the likelihood of observed data which we would like to maximize. Computing the evidence is intractable in general except for cases where the integral can be estimated in a closed form. Therefore, we need a better way to get the posterior distribution of Z. One way to solve this problem is to approximate the posterior by a simpler distribution which is easier to handle, such as exponential family [26, 27]. However, to estimate the parameter of this approximate distribution, we still need a metric to quantify the notion of closeness in distribution sense.

This problem is solved by variational inference using KL-divergence as the distance metric between two sets of distribution. KL divergence between two distributions $P(X)$ and $P(Z)$ is defined as:

$$D_{KL}(P(X)\|P(Z)) = \int P(X) \log \frac{P(X)}{P(Z)}\, dX \tag{2.10}$$

One way to think of KL divergence is that it is the remaining uncertainty in Z after observing X. Variational inference, therefore, converts the problem of estimating a distribution to an optimization problem given by:

$$q^*(Z) = \arg\min_{q \in \mathscr{Q}} D_{KL}\big[P(Z|X)\|q(Z)\big] \tag{2.11}$$

where, $\mathscr{Q}$ denoted the family of tractable distribution. The expectation in this case is taken w.r.t the density $P(Z|X)$ which we do not have access to in general. A solution to this problem is to consider $D_{KL}\big[q(Z)\|P(Z|X)\big]$ for

optimization which is easier to solve. Note that, KL divergence metric is not symmetric, therefore, we incur a loss in the approximation which is reasonable for practical applications.

2.5.1 Evidence Lower Bound

The KL divergence term in Equation 2.11 can be written as:

$$D_{KL} = E\big[\log q(Z)\big] - E\big[\log P(Z|X)\big] \qquad \text{or}$$

$$-D_{KL} + \log P(X) = E\big[P(X,Z)\big] - E\big[q(Z)\big]$$

which allows us to define the evidence lower bound (ELBO) via:

$$ELBO(q) = E\big[P(X,Z)\big] - E\big[q(Z)\big] = E\big[P(X|Z)\big] - D_{KL}\big[q(Z)\|P(Z)\big] \quad (2.12)$$

Thus, the ELBO lower bounds the likelihood of data and can be seen as a regularization for the data fit [28]. A popular application of ELBO is in the variational auto-encoder where the prior $P(Z)$ is assumed to be a normal Gaussian distribution.

2.5.2 Mean Field Variational Approximation

To simplify the optimization problem, the variational distribution is sometimes chosen to be of the form:

$$q(Z|X) = \prod_{i=1}^{D} q(z_i) \quad (2.13)$$

i.e., as a collection of independent random variable. This is called a mean field approximation and the resulting distribution does not depend on X. However,

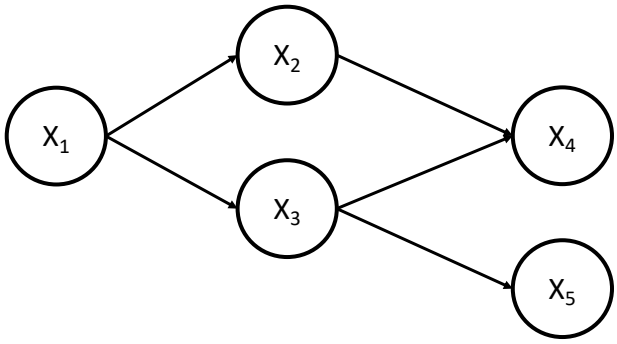

Figure 2.3: A simple graphical model representing local relationships.

it is still a lower bound to the evidence function.

2.6 Directed Acyclic Graphical Models

Graphical models are a succinct way to represent joint probability distribution over a collection of random variables using local relationships between them. Consider a collection of binary random variables $\{X_1, X_2, X_3, X_4, ...X_N\}$, in the absence of any local relationships, the joint distribution can be written as:

$$P(X_1, X_2, X_3....., X_N) = P(X_1) \times \prod_{i=2}^{N} P(X_i|X_{1:i-1}) \tag{2.14}$$

To completely specify this joint distribution, we need to specify about 2^N parameters. This will be infeasible if the variable X_i are defined over k classes.

Now, let's suppose that these random variables have some local relationships among each other such that, for each random variable X_i in the set, we have an additional information about its parents π_i. This relationship can be represented in a graphical format having N nodes and the edges corresponding to each element in π_i. A simple example is shown in Fig. 2.3 with 5

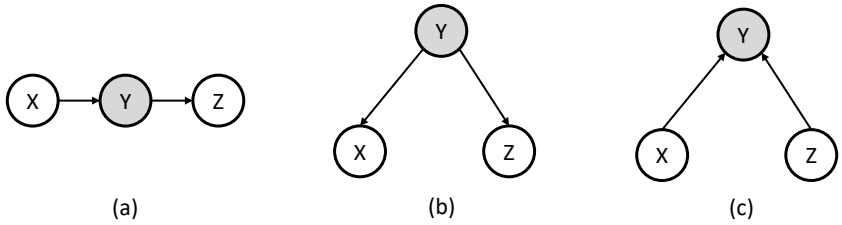

Figure 2.4: Conditional independence relationships induced by graphical structure.

nodes and 5 edges. Note that, each edge ending at a node represents a parent assignment for that specific node. For example, node X_2 and X_3 have X_1 as their parent, node X_4 has X_2 and X_3 as its parents, and so on. Using this local relationship, we can now define the joint density over the set of random variables $\{X_1, X_2, X_3, X_4, X_5\}$ as:

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3)P(X_5|X_3) \tag{2.15}$$

Notice that, now we only have to specify $\sum_i 2^{|\pi_i|}$ parameters which is typically much less than before.

2.6.1 Conditional Independence in Directed Graphs

Another important advantage of directed graphical models is the conditional property embedded in it. We are often interested in knowing whether a pair of pair of random variable are conditionally independent or not given some other random variables. This can be easily determined using Bayes' Ball algorithm. Fig. 2.4 specifies the three most common types of situations that arise while determining the conditional independence relationship. These three cases cover all necessary information required to answer questions pertaining to

conditional independences. Fig. 2.4(a) implies $P(X, Z|Y) = P(X|Y)P(Z|Y)$, i.e., X and Z are conditionally independent given Y. Fig. 2.4(b) implies the same, i.e., X and Z are conditionally independent given Y. Note that, in the absence of any knowledge about Y, X and Z are not marginally independent. Finally, Fig. 2.4(c) implies $X \not\perp\!\!\!\perp Z$ given Y but X and Z are marginally independent. Knowing conditional independences does not allow us to infer about the graphical structure. For example, both Fig. 2.4(a) and Fig. 2.4(b) imply the same conditional independence, but the graphical relationships are different.

2.7 Gumbel Softmax

For unsupervised learning, the ability to sample from a categorical distribution is an important constraint that one needs to worry about. However, it is not easy to backpropagate through such samplers due to the discrete nature of the distribution. One can define a proxy for the gradient function in such cases but it is often incorrect. A simple example of categorical sampling in neural network can be modeling a Gaussian mixture model as the latent space representation. The uni-modal Gaussian distribution has an easy re-parameterization trick which is widely employed, it is not the case for the mixture model. Authors in [29] proposed a Gumbel distribution based re-parameterization for categorical distribution. Defining z to be a categorical random variable having $\pi_1, \pi_2, .., \pi_k$ as the class probabilities, the Gumbel-max trick [30] allows sampling z using:

$$z = \arg\max_i [g_i + \log \pi_i] \tag{2.16}$$

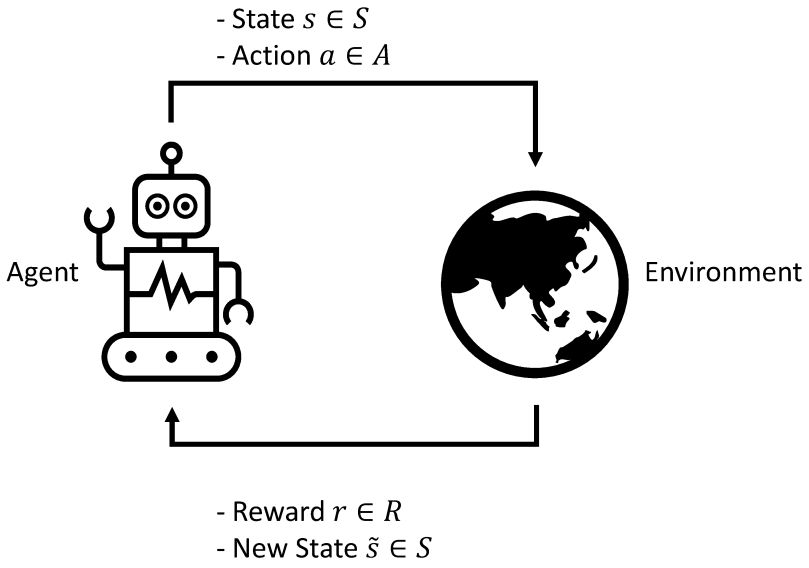

Figure 2.5: Basic reinforcement learning framework.

Here, $g_i's$ are drawn i.i.d from Gumbel(0,1) distribution. The softmax version of this max-trick can be expressed as:

$$y_i = \frac{\exp\left(\log \pi_i + g_i\right)/\tau}{\sum_j \exp\left(\log \pi_j + g_j\right)/\tau} \qquad \forall i \in 1, 2, ...k \tag{2.17}$$

2.8 Reinforcement Learning

Reinforcement learning is employed in situations where we want an agent (machine learning model) to learn from experience. These experiences are derived from the agent's interaction with the environment which provides a feedback via reward. Therefore, the objective is to learn a strategy of interaction with the environment in order to maximize the reward. Since the agent is acting in an environment, the environment specification may or may not be available to us. The agent is summarized by the current state $s \in S$, in which

the agent can take one of many actions $a \in A$. Upon taking an action, the agent receives a reward $r \in R$ and registers a change in its state $s' \in S$. Fig. 2.5 summarizes this mechanism in a loop structure.

2.8.1 Value Functions

The agent's policy, $\pi(s)$ determines the rule of interaction, and can be probabilistic or deterministic. Each state has a value function $V_\pi(s)$ which is the expected reward an agent will achieve in future starting from s following the policy π. It measures the goodness of state w.r.t the reward obtained. Our goal is to learn the optimal policy and the value function, simultaneously. A single interaction of the agent with an environment at time t can be characterized by the tuple (S_t, A_t, R_{t+1}). The agent receives the reward R_{t+1} after taking an action A_t in state S_t. The environment is described by a model which specifies the reward function and state transition. Specifically, the transition function F denotes probability of transitioning from state s to s' after taking action a while receiving reward r, i.e.:

$$F(s', r|s, a) = P(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a) \tag{2.18}$$

The reward function R predicts the expected reward at next time-step if the action a is taken:

$$R(s, a) = E\big[R_{t+1}|S_t = s, A_t = a\big] = \sum_{r \in R} r \sum_{s' \in S} P(s', r|s, a) \tag{2.19}$$

Next, the policy function can be deterministic $\pi(s)$ or stochastic $\pi(a|s) = P(A = a|S = s)$. The return from time step t is computed via:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2.20}$$

The variable $\gamma \in [0, 1]$ is called the discount factor which penalizes the future rewards. Finally, the state value of a state s under policy π is defined as:

$$V_\pi(s) = E_\pi\big[G_t|S_t = s\big] \tag{2.21}$$

and the state-action value is defined as:

$$Q_\pi(s, a) = E_\pi\big[G_t|S_t = s, A_t = a\big] \tag{2.22}$$

This leads to another definition of state value as:

$$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s, a) \tag{2.23}$$

2.8.2 Optimal Policy

The optimal value function yields the maximum return, i.e.:

$$V_*(s) = \max_\pi V_\pi(s) \textbf{ and } Q_*(s, a) = \max_\pi Q_\pi(s, a) \tag{2.24}$$

Thee optimal policy is one which achieves these optimal value functions:

$$\pi_* = \arg\max_\pi V_\pi(s) \textbf{ or } \pi_* = \arg\max_\pi Q_\pi(s, a) \tag{2.25}$$

Defining $P^a_{ss'} = \sum_{r \in R} P(s', r|s, a)$, the Bellman's Optimality equations are :

$$V_*(s) = \max_{a \in A} Q_*(s, a)$$

$$Q_*(s, a) = R(s, a) + \gamma \sum_{s' \in S} P^a_{ss'} V_*(s')$$

$$V_*(s) = \max_{a \in A} \Big(R(s, a) + \gamma \sum_{s' \in S} P^a_{ss'} V_*(s') \Big)$$

$$Q_*(s, a) = R(s, a) + \gamma \sum_{s' \in S} P^a_{ss'} \max_{a' \in A} Q_*(s', a')$$

2.8.3 Policy Gradient and REINFORCE Algorithm

Policy gradient methods learn the policy $\pi_\theta(a|s)$ using a parametric form of the policy function. In the discrete action case, the reward function can be defined as:

$$\mathcal{J}(\theta) = V_{\pi_\theta}(S_1) = E_{\pi_\theta}\Big[V_1\Big]$$

$$\mathcal{J}(\theta) = \sum_{s \in S} d_{\pi_\theta}(s) V_{\pi_\theta}(s) = \sum_{s \in S} \Big(d_{\pi_\theta}(s) \sum_{a \in A} \pi(a|s, \theta) Q_\pi(s, a) \Big)$$

We can compute the gradient of the reward function w.r.t θ and get the following result:

$$\begin{aligned}
\mathcal{J}(\theta) &= \sum_{s \in S} d(s) \sum_{a \in A} \pi(a|s;\theta) Q_\pi(s,a) \\
\nabla \mathcal{J}(\theta) &= \sum_{s \in S} d(s) \sum_{a \in A} \nabla \pi(a|s;\theta) Q_\pi(s,a) \\
&= \sum_{s \in S} d(s) \sum_{a \in A} \pi(a|s;\theta) \frac{\nabla \pi(a|s;\theta)}{\pi(a|s;\theta)} Q_\pi(s,a) \\
&= \sum_{s \in S} d(s) \sum_{a \in A} \pi(a|s;\theta) \nabla \ln \pi(a|s;\theta) Q_\pi(s,a) \\
&= E_{\pi_\theta}\left[\nabla \ln \pi(a|s;\theta) Q_\pi(s,a)\right]
\end{aligned}$$

Here, we replace $d_{\pi_\theta}(s)$ with $d(s)$ which has a theoretical support provided in [31].

The REINFORCE algorithm is a Monte-Carlo approach to estimate the parameters θ. The steps involved are:

- Initialize θ randomly.
- Generate an episode $S_1, A_1, R_2, S_2, A_2, R_3, ...S_T$.
- Estimate return G_t for all time steps and update $\theta \leftarrow \theta + \alpha \gamma^t G_t \nabla \ln \pi(A_t|S_t,\theta)$.

References

[1] James A. Russell, Jo-Anne Bachorowski, and José-Miguel Fernandez-Dols. "Facial and Vocal Expressions of Emotion". In: *Annual Review of Psychology* 54 (2003), pp. 329–349. DOI: 10.1146/annurev.psych.54.101601.145102.

[2] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition)*. Worth Publications, 2011.

[3] Marc Swerts and Emiel Krahmer. "On the Use of Prosody for On-line Evaluation Spoken Dialogue Systems". In: (2000).

[4] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation". In: *CoRR* abs/1711.09020 (2017). arXiv: 1711.09020. URL: http://arxiv.org/abs/1711.09020.

[5] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 3502–3506. DOI: 10.1109/ICASSP40776.2020.9054579.

[6] Juliane Schmidt, Esther Janse, and Odette Scharenborg. "Perception of Emotion in Conversational Speech by Younger and Older Listeners". In: *Frontiers in Psychology* 7 (2016), p. 781. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2016.00781.

[7] Zeynep Inanoglu and Steve Young. "A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 1. 2007, pp. 490–493.

[8] Yasuki Hashizawa, Shoichi Takeda, M. D. Hamzah, and G. Ohyama. "On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion". In: *Proc. Interspeech*. 2004.

[9] Jianhua Tao, Yongguo Kang, and Aijun Li. "Prosody conversion from neutral speech to emotional speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1145–1154. DOI: 10.1109/TASL.2006.876113.

[10] T. Toda, A. W. Black, and K. Tokuda. "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), pp. 2222–2235. ISSN: 1558-7916. DOI: 10.1109/TASL.2007.907344.

[11] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". In: *American Journal of Signal Processing* 2 (2012), pp. 134–138. DOI: 10.5923/j.ajsp.20120205.06.

[12] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki. "Exemplar-based emotional voice conversion using non-negative matrix factorization". In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. 2014, pp. 1–7. DOI: 10.1109/APSIPA.2014.7041640.

[13] Tuomas Virtanen, Bhiksha Raj, Jort Gemmeke, and Hugo Van hamme. "Active-set newton algorithm for non-negative sparse coding of audio". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2014, pp. 3092–3096. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854169.

[14] Patrik O. Hoyer. "Non-negative matrix factorization with sparseness constraints". In: *CoRR* cs.LG/0408058 (2004). URL: http://arxiv.org/abs/cs.LG/0408058.

[15] M. Schuster and K.K. Paliwal. "Bidirectional Recurrent Neural Networks". In: *Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. ISSN: 1053-587X. DOI: 10.1109/78.650093.

[16] Huaiping Ming, Dong-Yan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion". In: *Proc. Interspeech 2016*. 2016, pp. 2453–2457. DOI: 10.21437/Interspeech.2016-1053.

[17] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway Networks". In: *CoRR* abs/1505.00387 (2015). arXiv: 1505.00387. URL: http://arxiv.org/abs/1505.00387.

[18] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective". In: *Proc. Interspeech 2019*. 2019, pp. 2848–2852. DOI: 10.21437/Interspeech.2019-2512.

[19] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Automated Emotion Morphing in Speech Based on Diffeomorphic Curve Registration and Highway Networks". In: *Proc. Interspeech 2019*. 2019, pp. 4499–4503. DOI: 10.21437/Interspeech.2019-2386.

[20] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Multi-Speaker Emotion Conversion via Latent Variable Regularization and a Chained Encoder-Decoder-Predictor Network". In: *Proc. Interspeech 2020*. 2020, pp. 3391–3395. DOI: 10.21437/Interspeech.2020-1323.

[21] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proc. Interspeech 2019*. 2019, pp. 316–320. DOI: 10.21437/Interspeech.2019-1413.

[22] "Dynamic Time Warping (DTW)". In: *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*. Dordrecht: Springer Netherlands, 2008, pp. 570–570. ISBN: 978-1-4020-6754-9. DOI: 10.1007/978-1-4020-6754-9_4969.

[23] Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye. "Nonparallel Emotional Speech Conversion". In: *Proc. Interspeech 2019*. 2019, pp. 2858–2862. DOI: 10.21437/Interspeech.2019-2878.

[24] Kun Zhou, Berrak Sisman, and Haizhou Li. "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data". In: *CoRR* abs/2002.00198 (2020). arXiv: 2002.00198.

[25] Sarang C. Joshi and Michael I. Miller. "Landmark matching via large deformation diffeomorphisms." In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 9 8 (2000), pp. 1357–70. URL: https://api.semanticscholar.org/CorpusID:6659707.

[26] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080%2F01621459.2017.1285773.

[27] Kevin P. Murphy. "Machine learning - a probabilistic perspective". In: *Adaptive computation and machine learning series*. 2012. URL: https://api.semanticscholar.org/CorpusID:17793133.

[28] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. "An Introduction to Variational Methods for Graphical Models". In: *Mach. Learn.* 37.2 (1999), 183–233. ISSN: 0885-6125. DOI: 10.1023/A:1007665907178. URL: https://doi.org/10.1023/A:1007665907178.

[29] Eric Jang, Shixiang Gu, and Ben Poole. *Categorical Reparameterization with Gumbel-Softmax*. 2017. arXiv: 1611.01144 [stat.ML].

[30] K. D. Tocher. "Statistical Theory of Extreme Values and Some Practical Applications; Probability Tables for the Analysis of Extreme Value Data". In: *Journal of the Royal Statistical Society: Series A (General)* 118.1 (1955), pp. 106–106. DOI: https://doi.org/10.2307/2342529. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2342529. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2342529.

[31] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

Chapter 3

Frame-wise models for Prosody

In this chapter, we introduce a new model for emotion conversion in speech based on highway neural networks. Our model uses the contextual pitch, energy and spectral information of a source emotional utterance to predict the frame-wise fundamental frequency and signal intensity under a target emotion. We also incorporate a latent gender representation to promote cross-speaker generalizability. Our neural network is trained to maximize the error log-likelihood under an assumed Laplacian distribution. We validate our model on the VESUS repository collected at Johns Hopkins University, which contains parallel emotional utterances from 10 actors across 5 emotional classes. The proposed algorithm outperforms three state-of-the-art baselines in terms of the mean absolute error and correlation between the predicted and target values. We evaluate the quality of our emotion manipulations via crowd-sourcing. Finally, we apply our emotion morphing model to utterances generated by Wavenet to demonstrate our unique ability to inject emotion into synthetic speech.

We circumvent the data limitations by learning a multi-speaker model that

transforms a neutral utterance to one of the three target emotions. We focus on modifying two prosodic features namely, pitch and signal energy [1]. Pitch, being a correlate of the fundamental frequency, controls the intonation. In general, pitch tends to rise for anger and happiness, and it tends to fall for sadness and fear. Energy, on the other hand is a correlate of the intensity and controls the fluctuations in loudness profile. Typically, the loudness is higher when speaker is excited and is lower when in sad emotional state. However, beyond these general trends, the actual relationship between pitch/energy and emotion is highly complex and is governed by both local and global speech properties. Our strategy is to learn a mapping function for these two prosody features from a neutral state to an emotional state by factoring in both the segmental and supra-segmental nature of speech.

Several previous works have explored the problem of emotion morphing. For example, the work of [2] explicitly models the fundamental frequency (F0) contour using a linear model, a Gaussian mixture model (GMM), and a classification-regression tree (CART). In contrast, the work of [3] develops an independent transformation model for pitch, duration and spectrum. A GMM model constrained by global variance [4] was introduced by [5]. This framework estimates the joint distribution of the source and target spectral and prosody features. Another strategy relies on dictionary learning and sparse recovery to estimate the emotional transfer function. For example, the work of [6] uses parallel exemplars aligned using dynamic time warping [7], a greedy optimization based sequence alignment procedure, to create a source and a target dictionary. Going one step further, the work of [6] estimates

a sparse encoding using an active set Newton method based non-negative matrix factorization (NMF) [8]. The sparse encoding is used only to estimate the contextual spectrum envelope, whereas the fundamental frequency is directly copied from the corresponding frame of target dictionary. A more recent approach in emotion conversion is the application of bi-directional long-short term memory networks (Bi-LSTM) [9]. LSTMs are particularly suited for time series data, such as speech. Simultaneous conversion of both spectral and prosody features is carried out in [10]. In this method, the F0 and energy contour are parameterized using 10 scales of continuous wavelet transform [11]. An approximate reconstruction of converted F0 and energy values synthesizes the final speech signal using the STRAIGHT [12] module.

Unlike prior work, our approach converts the prosodic features without any explicit parameterization. We rely on a highway network architecture which is faster to train than the Bi-LSTM and more robust on small datasets. Our highway network input consist of the smoothed spectrogram averaged within the standard Mel-frequency bands, along with the F0 values in a 360 ms context window, and a novel gender embedding. The highway network uses a likelihood based loss function to predict the frame-wise pitch and energy for the target emotion. We do not change the spectrum of the signal itself to maintain speaker identity. Our model is trained from scratch using the VESUS emotion dataset collected at Johns Hopkins [13]. We perform both objective and subjective evaluation to compare the results of our proposed model with three state-of-the-art baseline methods. Finally, we apply the emotion morphing model to synthetic utterances generated by Google Wavenet [14].

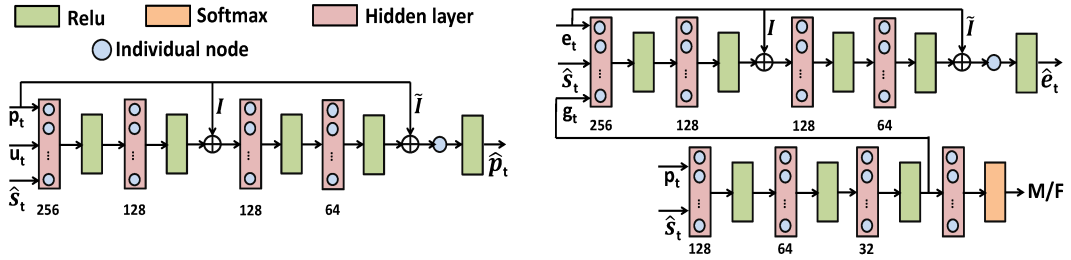

Figure 3.1: (Left) shows the highway network architecture for pitch prediction and (right) shows the model used for prediction of energy. The gender embedding $\mathbf{g_t}$ is obtained from the smaller network trained for gender classification on the same dataset.

3.1 Highway Network for Emotion Warping

We use the STRAIGHT vocoder [12] to extract the F0 and energy contours. During training, we align the source and target emotional utterance using dynamic time warping [7]. This process allows us to learn a frame-wise transformation for pitch and energy values. However, it is not applied to any of the test utterances during conversion. We incorporate a novel latent representation for gender to improve the generalizability of our model across multiple speakers. Finally, we again use STRAIGHT to re-synthesize the modified utterances.

3.1.1 Feature Extraction

As described above, our model predicts the frame-wise pitch and energy variations from source to target emotion. The features used for pitch prediction include a compressed form of smoothed spectral envelope, the utterance-level normalized fundamental frequency, and un-normalized pitch values with a context of 180 ms on both sides of the frame. We use only the voiced

portion of each utterance to extract the pitch normalization parameters. The reason behind using a long contextual window for fundamental frequency is to account for both local and global properties. In other words, prosody is affected by both segmental (phonetic level) and supra-segmental (syllable or word level) characteristics of an utterance. A context of 360 ms ensures that the pitch information for mapping function is provided over on average two syllables. All features are extracted using a window of size 10 ms and a 10 ms stride.

To reduce the dimensionality of the input space, we compress the spectral envelope using the standard Mel frequency filterbanks. Namely, we first compute a 1,024 point FFT for each frame, resulting in a 513 dimensional magnitude spectrum $S \in \mathbb{R}^{513\times 1}$ (frequency range 0 to π). We then use the normalized Mel filterbank matrix to obtain a 128-dimensional input representation. The filterbank matrix preserves the shape of the spectrum while accelerating the training of our deep highway network. Reducing the size to below 128 dimensions leads to noticeable loss in the shape of spectral envelope.

The utterance-normalized pitch (zero mean, unit variance), $\mathbf{u_t}$ allows us to capture the extreme values of target distribution. Conceptually, this feature acts as a flag forcing the neural network to sample from the tails of output distribution.

The computation of energy for each frame t is done by squaring and summing the short-time spectrum $S \in \mathbb{R}^{513\times 1}$:

$$\mathbf{e_t} = \sqrt{\sum_{k=1}^{513} S_{k,t}^2} \quad t \in 1, 2, ...T \tag{3.1}$$

where T is the total number of frames extracted from an utterance. Similar to the pitch contour, a context of 360 ms is used for energy as well. The VESUS repository contains parallel emotional utterances across ten different speakers. We obtain a frame-wise correspondence between source and target prosody features using DTW for training the neural network.

3.1.2 Highway Network Architecture

We employ a highway neural network with one input layer, four hidden layers and one output layer along with multiple skip connections [15]. Fig. 3.1 shows the schematic diagram of the highway network architectures used for predicting pitch and energy. The input spectral features $\hat{\mathbf{s}}_\mathbf{t}$ are normalized to mean zero and unit variance while the pitch contours $\mathbf{p_t}$ are fed in without any normalization. The output of highway network is given by:

$$\hat{\mathbf{p}}_\mathbf{t} = \phi[W_{45} \times (\phi[W_{34} \times (\phi[W_{23} \times \phi[W_{12} \times \phi[W_{01} \times \{\hat{\mathbf{s}}_\mathbf{t}, \mathbf{u_t}, \mathbf{p_t}\} + b_1] + b_2] \oplus \mathbf{I}\mathbf{p_t}) + b_3] + b_4] \oplus \tilde{\mathbf{I}}\mathbf{p_t}) + b_5] \tag{3.2}$$

The variables W_{ij} denote the weights going from layer i to layer j, and ϕ is the Relu non-linearity [16] applied at the output of each hidden node. The terms $\mathbf{I}\mathbf{p_t}$ and $\tilde{\mathbf{I}}\mathbf{p_t}$ represent the skip connections to the output of the second and fourth hidden layer, respectively. While $\mathbf{I}$ is the identity matrix, $\tilde{\mathbf{I}}$ denotes just the three central rows of the identity matrix $\mathbf{I}$ which provides a short pitch

context of 30 ms to the neural network before the final output. As designed, the highway network learns a perturbation on top of the input pitch values conditioned on the source spectrum and pitch contour. The skip connections add the correct bias, i.e, source pitch, back into the signal to better match the ground truth target. Closer to the input layer, the full 360 ms contextual pitch information is provided to extract important features from the contour, but as we go deeper, a shorter context proves to be sufficient. We use $\hat{\mathbf{p}}_\mathbf{t}$ to denote the pitch predicted for the input source frame at time t. A log transformation of the pitch tends to collapse its dynamic range and makes the predictions to saturate at the mean of the training samples.

A similar architecture is used for the energy prediction. We replace the contextual pitch contour $\mathbf{p_t}$ by contextual energy contour $\mathbf{e_t}$. Unlike pitch, which inherently carries gender information, predicting energy requires an auxiliary gender input, as illustrated in Fig. 3.1. Here, we train a relatively shallow neural network having three hidden layers with the smoothed spectrum and pitch contour as input. The output of the final hidden layer is used as a latent embedding for the gender $\mathbf{g_t}$ to predict energy at time index t. Denoting the input by $\{\hat{\mathbf{s}}_\mathbf{t}, \mathbf{e_t}\} \oplus \mathbf{g_t}$, the predicted energy is:

$$\hat{\mathbf{e}}_\mathbf{t} = \phi[W_{45} \times (\phi[W_{34} \times (\phi[W_{23} \times \phi[W_{12} \times \phi[W_{01} \times \{\hat{\mathbf{s}}_\mathbf{t}, \mathbf{e_t}\} \oplus \mathbf{g_t} + b_1] + b_2] \oplus \mathbf{I}\mathbf{e_t}) + b_3] + b_4] \oplus \tilde{\mathbf{I}}\mathbf{e_t}) + b_5] \quad (3.3)$$

During training we use a dropout [17] rate of 0.3 and batch normalization [18] after every hidden layer and before the skip connections with identity

map are concatenated. These implementation details help us to improve the generalization capability of our highway network. We use the Adam optimizer [19] with a fixed learning rate of 0.01 and mini-batches of size 500.

3.1.3 Maximum Likelihood Objective

Since the dynamic range of pitch is very high, the standard l_2 loss is not appropriate because of its sensitivity towards penalizing extreme values in the difference. In contrast, mean absolute error (i.e., l_1 penalty) allows the highway network to evenly focus on the less extreme values of pitch (such as around 200 Hz which occur more frequently in the data). We train the highway networks by maximizing the likelihood of the error for each training sample in a mini-batch [20]. In particular, we assume that the error function defined by $\mathscr{E}_n = y_n - \hat{y}_n$, where y_n is the true value and $\hat{y}_n$ is the model estimate, is drawn from a Laplacian distribution with mean 0 and variance b:

$$\mathscr{E}_n \sim \frac{1}{2b} \exp\left\{ -\frac{\|y_n - \hat{y}_n\|_1}{b} \right\} \tag{3.4}$$

The parameters of highway network, denoted by θ get updated via standard backpropagation algorithm. From here, the variance of the error distribution b is updated after every epoch of the highway network update in a maximum likelihood framework similar to the expectation maximization (EM) algorithm:

$$\hat{b} = \frac{1}{N} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|_1 \tag{3.5}$$

The algorithm for training the model parameters and estimating the Laplacian variance alternates between the following steps:

- Update θ to minimize $\sum_{n=1}^{N} \|y_n - \hat{y}_n\|_1$ while b fixed.
- Update b using Eq. (3.5) while θ is fixed.

At a high level, our maximum likelihood strategy acts as a learning rate scheduler by re-scaling the step size by variance in each epoch. In practice, this approach improves the correlation observed between the ground truth and predicted pitch/energy beyond the standard minimum absolute error objective.

3.1.4 Reconstruction

In the reconstruction stage, the predicted pitch and energy values over the input frames are smoothed using a mean filter to ensure the continuity in pitch and energy contour. While the pitch is directly used for synthesis, the energy values are implicitly used by re-scaling the spectrum using the equation:

$$\hat{S}_t = S_t \times \frac{\hat{\mathbf{e}}_\mathbf{t}}{\mathbf{e}_\mathbf{t}} \quad for \quad t = 1, 2, ...T \tag{3.6}$$

Here, $\mathbf{e_t}$ is the original energy value of frame t while $\hat{\mathbf{e}}_\mathbf{t}$ is the predicted energy value. The aperiodicity component of the STRAIGHT vocoder is copied directly from the source speech.

3.2 Experiments and Results

We carry out both the quantitative and qualitative evaluations to compare our performance with the current state-of-the-art techniques for emotion and prosody conversion in speech.

3.2.1 Dataset and Experimental Setup

Our training and evaluation relies on the VESUS emotional dataset collected at Johns Hopkins University [13]. VESUS contains a set of parallel emotional utterances spoken by a mix of amateur and professional actors. The original database has 2500 utterances for each of the five emotional classes: happiness, anger, sadness, fear and neutral. The repository also contains an emotion perception rating for each utterance provided by ten Amazon Mechanical Turk (AMT) raters.

For the proposed model, we use only those utterances from VESUS repository which are agreed upon by more than 50% of the AMT raters. We also omit the fear category from our experiments because of its high confusion with sad and neutral emotions. The total numbers in our experiment are:

- For **Neutral to Angry**: 1534 utterances for training, 72 for validation and, 61 for testing.
- For **Neutral to Happy**: 790 utterances for training, 43 for validation and, 43 for testing.
- For **Neutral to Sad**: 1449 for training, 75 for validation and, 63 for testing.

Objective evaluation includes the mean absolute error and the Pearson's correlation coefficient measure between the predicted value of pitch and energy and their ground truth counterparts. For subjective evaluation, we ask raters on AMT to classify each of the converted test samples for perceived emotion. Our designed survey asks AMT workers to listen to two speech files.

One of them is the baseline neutral speech and the other one is the speech converted into some target emotion. The order of neutral and emotional speech is randomized to weed out any non-diligent raters or bots. After they finish listening, we ask them to classify the emotion in both audio files. We find this type of bias correction using source (neutral) speech to be important because emotion perception is highly dependent on the knowledge about speaker articulation and speaking style.

3.2.2 Baseline methods

We compare our proposed model with three state-of-the-art baseline methods. The first baseline fits a Gaussian mixture model (GMM) [5] to the joint distribution of the source and target STRAIGHT cepstral features and fundamental frequency. We further incorporate the Global variance constraint proposed by [4] to improve the GMM based conversion model.

The second baseline uses the sparse Non-Negative Matrix Factorization (NMF) method developed in [6]. Here, two parallel dictionaries of STRAIGHT spectrum are constructed from the training dataset. An active Newton set based NMF estimates the sparse coding of input spectral features over the source dictionary. This encoding is then used to construct the converted spectrum and fundamental frequency from the target dictionary.

The third baseline is the Bi-LSTM model [10] which is pre trained for voice conversion using the CMU-ARCTIC corpus [21] and then fine-tuned for emotion conversion on the VESUS database. This method simultaneously converts both spectral and prosodic (pitch, energy) features. The prosodic

Table 3.1: MAE and Pearson's Correlation measures for pitch and energy across target emotions using universal model.

Alg.	MAE($\hat{p}_t$)	Cor($\hat{p}_t$)	MAE($\hat{e}_t$)	Cor($\hat{e}_t$)
Neutral-to-Angry				
GMM	44.3	0.54	4.24	0.57
NMF	94.2	0.22	4.2	0.22
Bi-LSTM	57.4	0.34	5.77	0.56
Proposed	**39.6**	**0.64**	**1.9**	**0.6**
Neutral-to-Sad				
GMM	29.1	0.8	5.87	0.53
NMF	65.3	0.4	7.9	0.32
Bi-LSTM	29.6	0.78	5.23	0.5
Proposed	**22.2**	**0.83**	**3.4**	**0.67**
Neutral-to-Happy				
GMM	53.8	0.51	4.24	0.53
NMF	106.7	0.25	6.5	0.23
Bi-LSTM	67.6	0.48	4.8	0.52
Proposed	**49.8**	**0.54**	**2.5**	**0.68**

features are parameterized by continuous wavelet transform [11] coefficients. The intention behind such parameterization is to consider both short-term and long-term pitch and energy trajectories by using multiple scales for the wavelet transform.

3.2.3 Results

Table 3.1 reports the quantitative performance of all four methods. Like the baseline algorithms, we train separate models for each target emotion category. Note that the proposed model outperforms all the baselines by a significant

margin. The global variance based GMM model is the second best algorithm for emotion conversion on the VESUS dataset. The high performance of GMM compared to the Bi-LSTM can be attributed to its simplicity, which makes it less prone to overfitting, and the large number of speakers in the VESUS repository. Our results also suggest that the procedure used in [10] of fine-tuning an Bi-LSTM model can not achieve the good performance for emotion conversion. Further, the assumption that local optima for emotion conversion should be close to the voice conversion solution on the error surface may not necessarily be true. NMF based sparse recovery and reconstruction performs the worst among all four models. This result is expected because there is no explicit constraint on the estimation of sparse coding. Specifically, there are multiple acoustic units that have very similar spectral envelopes and hence the algorithm also does not guarantee a smooth transition going from one frame to another.

In contrast to the baselines, our proposed pitch and energy prediction model is more robust because it focuses on learning a single, highly relevant transformation, rather than attempting to modify the entire spectrum. In addition, the highway network architecture allows us to learn a perturbation model that translates easily across multiple speakers. From a technical standpoint, it also facilitates for a smooth flow of gradients during the back-propagation [22]. Further, the EM type update of variance and weights of highway network in each iteration has a scaling effect on the mini-batch loss. This indirectly adjusts the learning rate during training, thereby helping the network converge to a better local optima than the Bi-LSTM model.

We evaluate the subjective quality of our emotion conversion using AMT. Empirically, we found the reconstructed speech from the GMM and NMF models to be highly distorted and unintelligible. Therefore, we only obtain crowd-sourced ratings for our highway network and the Bi-LSTM model. We crowd-source the same utterances spoken by same speakers for the highway network and Bi-LSTM model to get a uniform comparison between the two. Fig. 3.2 (top) shows the emotion classification accuracy on the testing utterances. Compared to the baseline model, our proposed model has higher classification accuracy across all three emotions. Further, the classification for neutral-to-sad is best followed by neutral-to-angry and then neutral-to-happy. This result is in line with the objective measures for pitch prediction (see Table 3.1). The Bi-LSTM model performs poorly because it fails to capture important prosody variations that contribute to emotion perception.

The final experiment examines our ability to inject emotional cues into synthetic speech generated by Google Wavenet [14]. We use the text-to-speech API provided by Google to generate same utterances as spoken in the VESUS repository. The utterances are generated for a female American English speaker. For this experiment, we fine tune the highway network by picking a speaker from the VESUS dataset who has the most expressive emotional utterances. We use 220/120/220 samples for fine tuning the neutral to angry/happy/sad model, respectively. The fine tuning procedure runs for only 50 epochs starting from the mixed speaker model weights. The crowd sourcing setup is unchanged from the previous case. Fig. 3.2 (bottom) shows the emotion classification result on speech generated by Wavenet. We see that speech modified by the

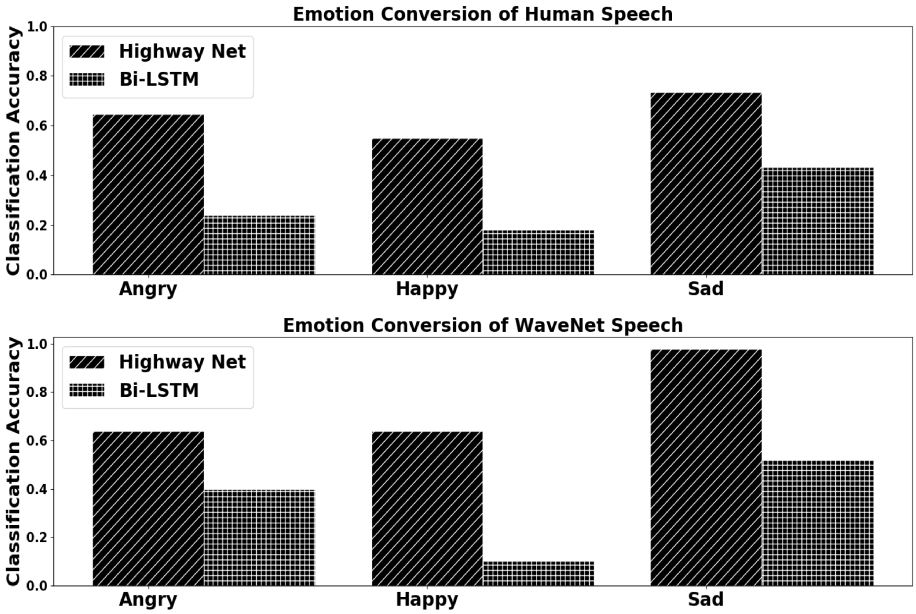

Figure 3.2: Emotion Classification accuracy for human (top) and Wavenet's speech (bottom) obtained via crowd-sourcing.

highway network is clearly perceived as emotional, in contrast to the Bi-LSTM. This first-of-its-kind demonstration shows that our model is highly adaptable to new (and even synthetic) speakers with minimal training data for fine tuning. In contrast, the Bi-LSTM does worse for the same setup due to its complex architecture and attempt to modify the entire spectral range.

In summary, our quantitative and qualitative results together show the markedly improved performance for our proposed model over three competing baselines. Our results also suggest that modifying just the pitch and energy contours is sufficient for emotion conversion. Finally, our experiment on Wavenet demonstrates that we can infuse emotions into a synthetic speech by fine tuning our cross-speaker model.

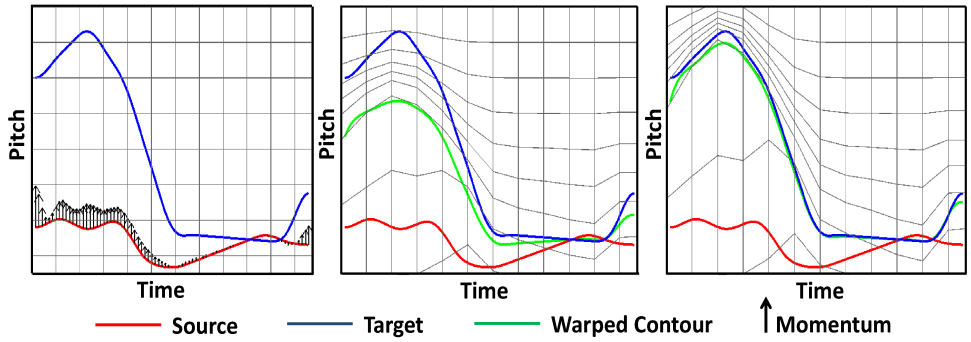

Figure 3.3: Illustration of 2-D diffeomorphic registration for emotion conversion. **Left:** source (neutral) and target (emotional) pitch contours from parallel utterances. **Middle:** intermediate output as source moves towards target. **Right:** final curve alignment.

3.3 Momentum-Based Emotion Conversion

VESUS contains parallel emotional utterances, which allows us to draw frame-wise correspondences. We use the STRAIGHT vocoder [12] to extract pitch contours from the utterances. During training, we align the source and target emotional utterances using dynamic time warping [7]. From here, we use the formulation in the next section to estimate the frame-wise momenta for each utterance pair. We then train an H-Net to predict these momenta based on the pitch and spectral information in the original utterance. During testing, we estimate the frame-wise pitch momentum using our trained H-Net and apply the diffeomorphic transformation to obtain the new pitch contour. We re-synthesize the modified utterance again using STRAIGHT.

3.3.1 Diffeomorphic Registration for 2-D Curves

Our goal in this work is to learn a *transformation* on pitch contours that alters the perceived emotional content of the reconstructed utterance. We adopt the

Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [23, 24], which provides global convergence and optimality guarantees. At a high level, LDDMM is based on an underlying vector field that acts on the source contour. This vector field is parameterized by an exponential map, which provides a smooth transition. For simplicity, we assume that the signals have been aligned using dynamic time warping (DTW). In this case, the vector field acts only in the vertical direction to locally change the pitch values. Fig. 3.3 illustrates this warping process on two pitch contours.

Mathematically, let $\mathbf{p_t}$ and $\hat{\mathbf{p}}_\mathbf{t}$ be the source and target pitch contours, respectively. The time index $\mathbf{t}$ corresponds to the discrete sampling of the contours from $\mathbf{t} = 0, \ldots, T$. Our approach is related to the landmark LDDMM setting of [25, 26] and [27] with a vertical constraint on the vector field. In particular, let $\mathbf{v_t}(\mathbf{x}; s)$ be a non-stationary and finite norm vector field across time $\mathbf{t}$ and pitch values $\mathbf{x}$. These vector fields generate the dynamical deformations with respect to the second evolution argument s. Namely, for a fixed point in time $\mathbf{t}$, we can consider the continuous flow $\mathbf{x} \mapsto \varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s)$ of the vector field for $s \in [0, 1]$ defined by $\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; 0) = \mathbf{p_t}$ and the ordinary differential equation (ODE) $\partial_s \varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s) = \mathbf{v_t}(\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s); s)$. Here, the initial condition specifies that we begin the evolution process from the source pitch contour. The ODE specifies that the displacement at every new pitch value is given by the vector field $\mathbf{v_t}(\mathbf{x}; s)$. The evolution process terminates at $s = 1$.

We now formulate the registration problem between the source pitch contour $\mathbf{p_t}$ and the target pitch contour $\hat{\mathbf{p}}_\mathbf{t}$ through the following optimal

control problem:

$$\min_{\mathbf{v}\in\mathcal{V}} \frac{1}{2}\int_0^1 \|\mathbf{v_t}(\cdot;s)\|_V^2 \, ds + \lambda \sum_{t=1}^{T} \left(\varphi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p_t};1) - \hat{\mathbf{p}}_{\mathbf{t}}\right)^2 \tag{3.7}$$

The first term of Eq. (3.7) is a smoothness constraint on the underlying vector field. The Hilbert norm $||\cdot||_V$ is implicitly defined through a 2-D exponential kernel that operates across time and pitch. The second term of Eq. (3.7) is the data matching term, which enforces that the warped source contour should be close to the target contour. Notice that the parameter λ controls the trade-off between smoothness and registration fidelity.

The Pontryagin maximum principle of optimal control [27] allows us to derive necessary conditions for the solution to Eq. (3.7). In this case, the theory shows that there exist variables $\mathbf{m}_{\mathbf{t}}^{s}$ for $s \in [0,1]$ that we call *momenta*. These momenta behave like hidden state variables in the continuous-time Kalman filter framework. The "observed" variables in this analogy are the pitch values of the warped contour. The Hamiltonian dynamics associated with the state/observer model allow us to reformulate Eq. (3.7) as a minimization over initial momenta $\mathbf{m}_{\mathbf{t}}^{0}$.

Formally, let $\mathbf{z}_t(s) = [\mathbf{t} \quad \varphi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p_t};s)]^{\mathrm{T}}$ be a two-dimensional vector of the time and deformed pitch value, and let $\gamma_{ij}(s)$ be the kernel evaluated at the pair of vectors $\mathbf{z}_i(s)$ and $\mathbf{z}_j(s)$. The quadratic objective for the collection of initial momenta can be written as follows:

$$\mathcal{J}(\mathbf{m}^0) = \frac{1}{2}\sum_{i,j=1}^{T} \gamma_{ij}(0)\mathbf{m}_{\mathbf{i}}^{0}\mathbf{m}_{\mathbf{j}}^{0} + \lambda \sum_{t=1}^{T} \left(\varphi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p_t};1) - \hat{\mathbf{p}}_{\mathbf{t}}\right)^2 \tag{3.8}$$

subject to Hamiltonian equations. A standard approach to solve such a problem numerically is given by *shooting algorithms* [28]. We apply a quasi-Newton descent method on $\mathcal{J}$, where the gradient w.r.t $\mathbf{m}^0$ of the second term in Eq. (3.8) is computed via the adjoint Hamiltonian equations.

Our strategy is to use Eq. (3.8) to solve directly for the initial momenta in the training dataset, where we have access to parallel emotional utterances. We will then train a neural network to predict these momenta directly from the signal characteristics. This neural network will be applied to the testing utterances to predict the (unknown) initial momenta. The contour registration process is completely specified once we have these values.

3.3.2 Input Features for Momentum Prediction

As described above, our model predicts the initial displacement (i.e., momenta) to transform a source utterance to the target emotion. We use two classes of features to predict the frame-wise momentum: a compressed form of the raw spectrum and the original pitch contour with a 200 ms context on both sides of the frame. Our rationale for using a long contextual window for pitch is to account for both local and global properties. Since pitch is affected by both segmental (phonetic level) and supra-segmental (syllable or word level) characteristics, a context of 360 ms ensures that the pitch information is provided over on average two syllables. All input features are extracted using a frame period of 5 ms and a 5 ms window stride.

To reduce the input dimensionality, we compress the raw spectral envelope using the normalized Mel frequency. Specifically, we first compute a 1,024

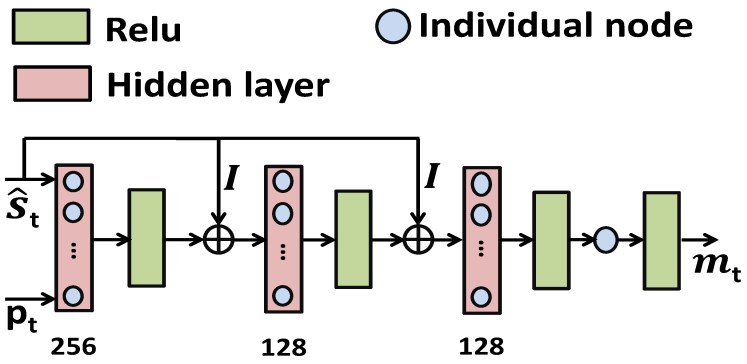

Figure 3.4: H-Net architecture for initial momentum prediction.

point FFT for each time frame, which results in a 513 dimensional magnitude spectrum $F_t \in \mathbb{R}^{513\times 1}$ (frequency range 0 to π). We use the normalized Mel filterbank matrix to obtain a 128-dimensional input representation $\hat{S}_t \in \mathbb{R}^{128\times 1}$. The filterbank matrix preserves the shape of the spectrum while preserving the acoustic information present in the frame. Our compression scheme is highly effective in accelerating the training times for our deep neural networks. Empirically, we find that further compression beyond 128 dimensions leads to undesirable distortions in the spectral envelope.

3.3.3 Highway Neural Network Architecture

We employ an artificial neural net with skip connections between the input and hidden layers. This architecture is known as a highway network (H-Net). Our model contains one input layer, three hidden layers, and one output layer, as illustrated in Fig. 3.4. The input spectral features $\hat{\mathbf{s}}_\mathbf{t}$ are normalized to mean 0 and unit variance while the pitch contours $\mathbf{p_t}$ are fed in without any normalization. The output of neural network, i.e., the initial momentum $\mathbf{m_t}$,

is given by the following expression:

$$\mathbf{m_t} = \phi[W_{34} \times \phi[W_{23} \times (\phi[W_{12} \times (\phi[W_{01} \times \{\hat{\mathbf{s}}_\mathbf{t}, \mathbf{p_t}\} + b_1] \oplus \mathbf{I}\hat{\mathbf{s}}_\mathbf{t}) + b_2] \oplus \mathbf{I}\hat{\mathbf{s}}_\mathbf{t}) + b_3] + b_4] \quad (3.9)$$

The variables W_{ij} in Eq. (3.9) denote the weights going from layer i to layer j, and ϕ is the ReLU non-linearity ([16]) applied at each hidden layer and the output. The variable b_i is the bias related to the layer i. The term $\mathbf{I}\hat{\mathbf{s}}_\mathbf{t}$ denotes the skip connections concatenated to the second and third hidden layer output, respectively. The variable $\mathbf{I}$ is the identity matrix showing there is no transformation of the features being carried out in skip connections. Variable $\mathbf{m_t}$ is the momentum predicted for the input source frame t. We use a dropout [17] rate of 0.3 and batch normalization [18] after every hidden layer and before the skip connections with identity. We use the Adam optimizer [19] with a fixed learning rate of 0.01 and mini-batch sizes of 500.

3.3.4 Reconstruction

The predicted momenta are used to transform the entire source pitch contour. The aperiodicity and spectrogram components are copied directly from the source speech. We reconstruct the modified utterance using STRAIGHT by replacing the source pitch contour with the transformed version.

3.4 Experimental Setup

We performed both an objective and subjective evaluation of our momentum prediction framework. The results are compared to three state-of-the-art emotion conversion baseline algorithms.

3.4.1 Emotional Speech Dataset and Evaluation

Our training and evaluation again relies on the VESUS emotional dataset. We consider three emotion conversion models: neutral to angry, neutral to sad, and neutral to happy. These conversions span both high- and low-arousal emotions to test the limits of our diffeomorphic registration approach. We sub-select the VESUS utterances based on $\geq 50\%$ agreement between raters. The total numbers in our experiment are:

- **Neutral to Angry:** 1534 utterances for training, 72 for validation, and 100 for testing.
- **Neutral to Happy:** 790 utterances for training, 43 for validation, and 43 for testing.
- **Neutral to Sad:** 1449 utterances for training, 63 for validation, and 70 for testing.

We follow the same objective and subjective evaluation protocol as the previous approach. Our baseline methods are once again: global variance constrained GMM model [5], NMF technique [6] and the LSTM model for conversion [10].

Table 3.2: MAE and Pearson's Correlation measures for pitch across target emotions using multi-speaker model.

Algorithm	MAE(F0)	Corr(F0)
Neutral-to-Angry		
GMM	44.3	0.54
NMF	94.2	0.22
Bi-LSTM	57.4	0.34
Proposed	**40.5**	**0.61**
Neutral-to-Happy		
GMM	53.8	0.51
NMF	106.7	0.25
Bi-LSTM	67.6	0.48
Proposed	**49.8**	**0.54**
Neutral-to-Sad		
GMM	29.1	**0.8**
NMF	65.3	0.4
Bi-LSTM	29.6	0.78
Proposed	**27.7**	0.74

3.5 Experimental Results

Table 3.2 summarizes the objective results obtained for baseline and proposed methods. Our algorithm is uniformly better at approximating the target pitch contour in absolute error sense. The results demonstrate that our parameterization of pitch deformation by initial momentum does work effectively.

The GMM based prosody and spectrum conversion comes a close second, beating both NMF and Bi-LSTM based models. The reason for this can be attributed to the simplicity of GMM which allows it to learn the parameters

i.e., mean and covariances in high dimensional space. However, the speech reconstructed by GMM is poor because of the averaging effect that mixture models have. It fails to conditionally sample from the tails of joint distribution and hence the predicted pitch wiggles about the mean of the training data. NMF does a poor job in prediction of prosody because of the lack of any global constraint while estimating sparse coding. The cepstral features are not a unique representation of an acoustic unit and there exist a many-to-one mapping. This further results in discontinuities in the converted spectrum going from one frame to the next. In the end, the reconstructed speech is very distorted and sometimes completely unintelligible. Bi-LSTM does worse compared to our method of pitch approximation because of its over parameterization. The multi-scale wavelet transform used for encoding the prosodic features leads to a very rough estimate of the predicted pitch and energy contour. Furthermore, the underlying assumption about the existence of local minima for emotion conversion being close to the voice conversion optima is not always true.

In contrast, our proposed model predicts only one value which is the initial momentum parameter. Besides, we design our H-Net to appropriately learn this regression function by minimizing the l_1 penalty which, unlike l_2 loss allows the model to evenly focus on the less extreme parts of the target distribution.

Our subjective evaluations are based on five crowd-sourced ratings for each converted speech via AMT. A majority voting decides the final emotion label of the converted utterances. We found the reconstructed speech from the

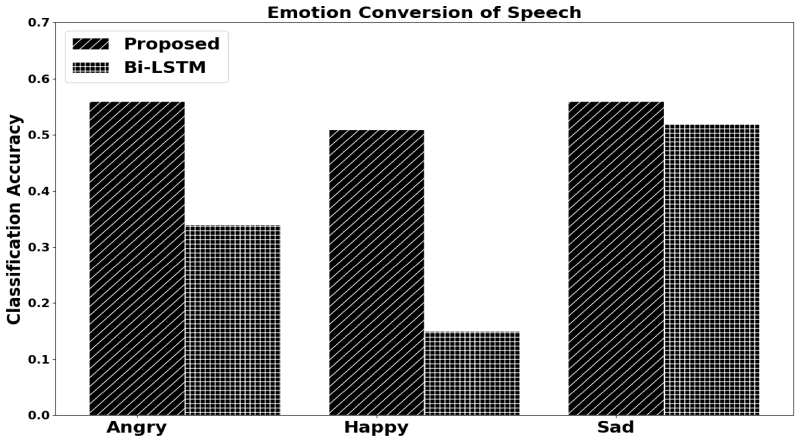

Figure 3.5: Momenta: comparison of emotion classification accuracy.

GMM and NMF models are be highly distorted and unintelligible. Therefore, we only obtain crowd-sourced ratings for our HNet and the Bi-LSTM model. To get a uniform comparison between the proposed method and Bi-LSTM based conversion, we crowd-source the ratings for exact same utterances spoken by same speakers. Fig.3.5 shows the emotion classification accuracy on the testing utterances. Compared to the baseline model, our proposed model has higher classification accuracy across all three emotions. Further, the classification for neutral-to-angry is the best followed by neutral-to-happy and then neutral-to-sad. Comparatively, the high arousal emotions like angry or happy are easier to discern than low arousal emotions like sad. This effect is evident in the Fig.3.5 as the difference in classification accuracy is lowest for neutral-to-sad conversion. Our method, unlike the Bi-LSTM model, only modifies the pitch and still does remarkably better on the listening tasks. This proves that the proposed method is very robust for carrying out emotion morphing. Another point to be noted is that, since we only modify the pitch and not the spectral envelope, the speaker information is retained and the

converted speech is distortion-less.

3.6 Conclusion

In this chapter, We have demonstrated the first multi-speaker emotion conversion model based on modifying pitch and energy. Our novel highway network based prosody prediction model has the lowest mean absolute error and highest correlation with the ground truth values when trained and tested on the VESUS emotional dataset. We trained our highway network in an alternating fashion by maximizing the error likelihood. A Laplacian assumption on the residual distribution in each mini-batch was made and was motivated by the data itself. Our algorithm outperformed the state-of-the-art methods for emotion conversion on subjective listening tasks by significant margins thereby proving the effectiveness of our procedure. Finally, we showed that our model is capable of injecting emotion into vocoder output which has not been done before in the literature.

We further proposed a method for emotion conversion based on estimating a curve warping function for pitch contours. The warping was based on a diffeomorphic registration technique that generates a sequence of smooth and invertible time-varying vector fields in an iterative fashion. We trained a highway network to predict the deformation parameter, also called as the initial momentum, for every point on a given pitch contour. The warped curve was used to reconstruct speech for three target emotions. Our experiments showed that the speech generated by modified pitch contours were perceived more emotional than speech generated by the baseline algorithm. Our proposed

model retained the speaker characteristics and the quality of speech by not changing the spectral envelope of the source audio.

References

[1] Robert W. Frick. "Communicating Emotion. The Role of Prosodic Features". In: *Psychological Bulletin* 97 (1985), pp. 412–429. DOI: 10.1037/0033-2909.97.3.412.

[2] Yongguo Kang, Jianhua Tao, and Bo Xu. "Applying Pitch Target Model to Convert F0 Contour for Expressive Mandarin Speech Synthesis". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 1. 2006, pp. I –I. DOI: 10.1109/ICASSP.2006.1660125.

[3] Zeynep Inanoglu and Steve Young. "A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 1. 2007, pp. 490–493.

[4] T. Toda, A. W. Black, and K. Tokuda. "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), pp. 2222–2235. ISSN: 1558-7916. DOI: 10.1109/TASL.2007.907344.

[5] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". In: *American Journal of Signal Processing* 2 (2012), pp. 134–138. DOI: 10.5923/j.ajsp.20120205.06.

[6] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki. "Exemplar-based emotional voice conversion using non-negative matrix factorization". In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. 2014, pp. 1–7. DOI: 10.1109/APSIPA.2014.7041640.

[7] "Dynamic Time Warping (DTW)". In: *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*. Dordrecht: Springer Netherlands, 2008, pp. 570–570. ISBN: 978-1-4020-6754-9. DOI: 10.1007/978-1-4020-6754-9_4969.

[8] Tuomas Virtanen, Bhiksha Raj, Jort Gemmeke, and Hugo Van hamme. "Active-set newton algorithm for non-negative sparse coding of audio". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2014, pp. 3092–3096. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854169.

[9] M. Schuster and K.K. Paliwal. "Bidirectional Recurrent Neural Networks". In: *Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. ISSN: 1053-587X. DOI: 10.1109/78.650093.

[10] Huaiping Ming, Dong-Yan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2016, pp. 2453–2457. DOI: 10.21437/Interspeech.2016-1053.

[11] Manuel Sam Ribeiro and Robert A. J. Clark. "A multi-level representation of f0 using the continuous wavelet transform and the Discrete Cosine Transform". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2015, pp. 4909–4913. DOI: 10.1109/ICASSP.2015.7178904.

[12] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech Communication* 27 (1999), pp. 187–207. DOI: 10.1016/S0167-6393(98)00085-5.

[13] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkatarman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2019.

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499 (2016). arXiv: 1609.03499. URL: http://arxiv.org/abs/1609.03499.

[15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway Networks". In: *CoRR* abs/1505.00387 (2015). arXiv: 1505.00387. URL: http://arxiv.org/abs/1505.00387.

[16] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 978-1-60558-907-7.

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[18] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[19] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2015).

[20] Yannan Wang, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "A Maximum Likelihood Approach to Deep Neural Network Based Nonlinear Spectral Mapping for Single-Channel Speech Separation". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2017, pp. 1178–1182. DOI: 10.21437/Interspeech.2017-830.

[21] John Kominek and Alan W Black. "The CMU Arctic speech databases". In: *SSW5-2004* (2004).

[22] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient Descent Finds Global Minima of Deep Neural Networks". In: *CoRR* abs/1811.03804 (2018). arXiv: 1811.03804.

[23] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. "Computing large deformation metric mappings via geodesic flows of diffeomorphisms". In: *International journal of computer vision* 61.139-157 (2005).

[24] L. Younes. *Shapes and diffeomorphisms*. Springer, 2010.

[25] Sarang C Joshi and Michael I Miller. "Landmark matching via large deformation diffeomorphisms". In: *IEEE transactions on image processing* 9.8 (2000), pp. 1357–1370.

[26] M. Miller, A. Trouvé, and L. Younes. "Hamiltonian Systems and Optimal Control in Computational Anatomy: 100 Years Since D'Arcy Thompson." In: *Annual Review Biomedical Engineering* 7.17 (2015), pp. 447–509.

[27] S. Arguillère, E. Trélat, A. Trouvé, and L. Younes. "Registration of Multiple Shapes using Constrained Optimal Control". In: *SIAM Journal on Imaging Sciences* 9.1 (2016), pp. 344–385.

[28] François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. "Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation". In: *International Journal of Computer Vision* 97.2 (2012), pp. 229–241.

Chapter 4

Supervised Encoder-Decoder-Predictor for F0 and Spectrum

In this chapter, we propose a new method for emotion conversion in speech based on a chained encoder-decoder-predictor neural network architecture. Unlike Chapter 3, we will not treat the F0 and energy value for each frame as an i.i.d sample. Instead, we learn a model to convert the F0 contour and the spectral envelope completely. The encoder constructs a latent embedding of the fundamental frequency (F0) contour and the spectrum, which we regularize using the Large Diffeomorphic Metric Mapping (LDDMM) registration framework. The decoder uses this embedding to predict the modified F0 contour in a target emotional class. Finally, the predictor uses the original spectrum and the modified F0 contour to generate a corresponding target spectrum. Our joint objective function simultaneously optimizes the parameters of three model blocks. We show that our method outperforms the existing state-of-the-art approaches on both, the saliency of emotion conversion and

the quality of resynthesized speech. In addition, the LDDMM regularization allows our model to convert phrases that were not present in training, thus providing evidence for out-of-sample generalization.

4.1 Background and Prior Works

The quality of machine-generated speech has improved phenomenally in the last decade, largely due to the representational power of deep neural networks [1, 2, 3], which are trained on hundreds of hours of transcribed human speech. However, controlling the expressiveness of synthetic speech remains an open challenge. Recent works in emotional speech synthesis include [4], which generates singing voice conditioned on the input rhythm, pitch and linguistic features. A disentangled model for style and content is proposed by [5, 6] to infer the latent representations responsible for expressiveness. While these models represent seminal contributions to emotional speech synthesis, the latent representations are learned in an unsupervised manner, which makes it difficult for the user to control the output emotion. Another problem is the poor rate of speech generation due to the auto-regressive nature of these models [7]. These challenges motivate the study of emotion conversion as an alternative to end-to-end synthesis approaches. Notably, emotion conversion methods provide controllability over the generated affect, they require much less data to train, and the processing speed is high enough for real-time applications.

Several interesting approaches for emotion conversion have been proposed in the recent past. For example, the work of [8] uses a Gaussian Mixture

Model with global variance constraint (GMM-GV) to modify the fundamental frequency (F0) contour and the spectrum. A bidirectional long-short term memory (Bi-LSTM) based architecture has been proposed by [9] to estimate the F0 contour and the spectral features of the target emotion utterance. Another approach by [10] converts the pitch contour and energy contour of the source utterance using a highway neural network which maximizes the error log likelihood in an expectation-maximization scheme. The same authors further proposed a curve registration based method [11] to modify only the F0 contour. Finally, a cycle-consistent generative adversarial network (cycle-GAN) proposed by [12] learns to sample the pitch contour and the spectrum from the target emotional class in an unsupervised manner. While these methods have been successful in single-speaker settings, many of them fail on multi-speaker dataset due to the larger overlap of F0 and spectral features between emotional classes. In this chapter, we propose a novel approach to model the relationship between the F0 contour and the spectral features, deriving it from the basic knowledge of these two representations. Furthermore, unlike other existing methods, our chained estimation also minimizes the mismatch between F0 and the corresponding spectral harmonics. Our second contribution in this chapter is to implicitly model the target pitch contour as a smooth and invertible warping of source F0 contour. This is done by learning a latent embedding based on the Large Diffeomorphic Metric Mapping (LDDMM) [13, 14] framework. In essence the embedding serves as an intermediary between the source and target emotions. We demonstrate that imposing this constraint improves the prediction of the pitch contour significantly.

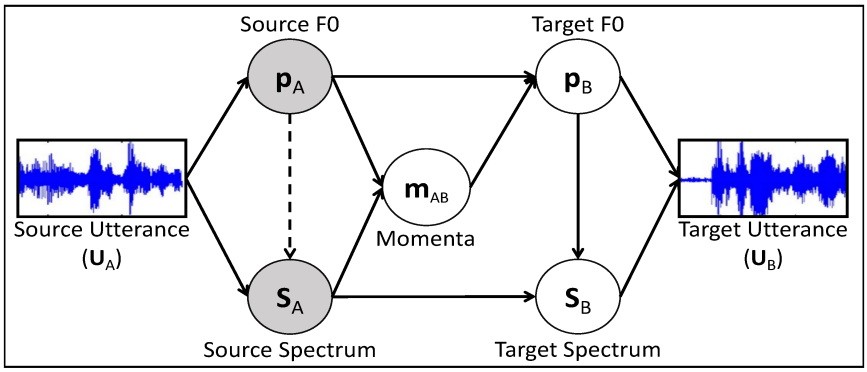

Figure 4.1: Graphical model of our emotion conversion strategy. $\mathbf{m}_{AB}$ is the intermediary between emotion classes.

Our architecture consists of three separate convolutional neural networks for predicting the embedding, the pitch contour, and the spectrum, respectively. These networks are trained in an end-to-end fashion from a unified objective function. We compare our model against three state-of-the-art baseline methods using the multispeaker VESUS dataset [15]. We further demonstrate that our model does well on sentences, which are not part of the training set, establishing its generalization capability. Finally, in addition to emotion conversion, we show that the proposed model generates better quality of speech than the baselines from both supervised and unsupervised domain.

4.2 Method

Our novel method uses a chained encoder-decoder-predictor network architecture to modify both the spectrum and the F0 contour of an utterance. The three components of the architecture are jointly optimized through a unified loss function.

Fig. 4.1 describes the relationship between the random variables in our

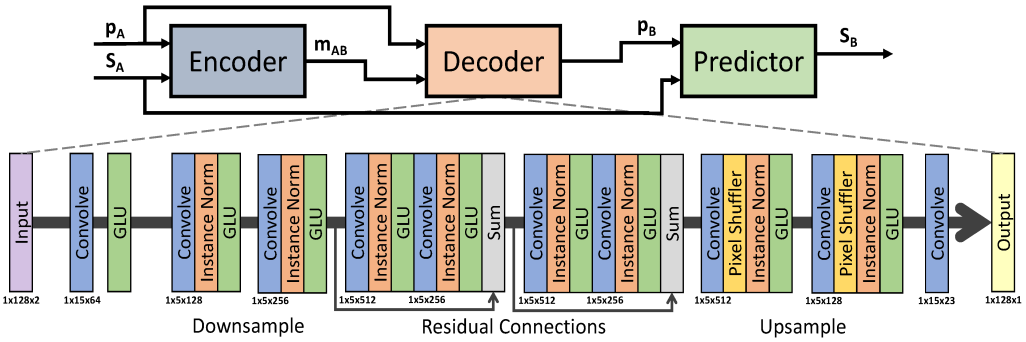

Figure 4.2: Block model representation of the encoder-decoder-predictor. Encoder and decoder use the same architecture whereas predictor has an extra residual block. GLU in the model stands for the gated linear unit. We use instance normalization due to small mini-batch size and pixel shuffling for up-sampling. The size and number of kernels are indicated below each convolution block.

model. We use WORLD vocoder [16, 17] for the analysis and synthesis of speech. Given a source-target pair of emotional utterances denoted by $\mathbf{U}_A$ and $\mathbf{U}_B$, respectively, the source utterance is decomposed into its components: the spectrum ($\mathbf{S}_A$) and the F0 contour ($\mathbf{p}_A$). These components allow us to estimate an intermediate parameter, known as the momenta ($\mathbf{m}_{AB}$). From here, the target F0 contour ($\mathbf{p}_B$) is modeled as a function of the source F0 contour ($\mathbf{p}_A$) and the momenta ($\mathbf{m}_{AB}$). Next, we estimate the target spectrum ($\mathbf{S}_B$) given the target F0 contour ($\mathbf{p}_B$) and the source spectrum ($\mathbf{S}_A$). Finally, the estimated variables are used to synthesize the target emotion utterance. The joint distribution shown in Fig. 4.1 factorizes as:

$$P(\mathbf{p}_A, \mathbf{S}_A, \mathbf{m}_{AB}, \mathbf{p}_B, \mathbf{S}_B) = P(\mathbf{p}_A)P(\mathbf{S}_A|\mathbf{p}_A)P(\mathbf{m}_{AB}|\mathbf{p}_A, \mathbf{S}_A)P(\mathbf{p}_B|\mathbf{p}_A, \mathbf{m}_{AB})P(\mathbf{S}_B|\mathbf{S}_A, \mathbf{p}_B) \tag{4.1}$$

4.2.1 Regularization via latent representation

We use an explicit prior on the latent variable to improve the prediction of F0 and spectrum. Specifically, we model the target F0 contour as a smooth and invertible deformation of the source F0 contour. The idea of smooth deformations has been used extensively for images [18], but here we use it for 2-D curves. Mathematically, let $\mathbf{p}_A^{\mathbf{t}}$ and $\mathbf{p}_B^{\mathbf{t}}$ denote a pair of source and target F0 contours, respectively. The variable $\mathbf{t}$ corresponds to the location of the analysis window as it moves across a given speech utterance. The objective of this deformation process is to estimate a series of small vertical displacements $\mathbf{v_t}(\mathbf{x};\mathbf{s})$ [13] over frequency and time. The variable $\mathbf{s}\ \epsilon\ [0,1]$ controls the evolution of these small displacements in the discrete setting. The registration problem can thus be formulated as:

$$\min_{\mathbf{v}\epsilon V} \frac{1}{2}\int_0^1 \|\mathbf{v_t}(\cdot;\mathbf{s})\|_V^2 ds + \lambda \sum_{t=1}^{T} \|\phi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p}_A^{\mathbf{t}};1) - \mathbf{p}_B^{\mathbf{t}}\|_2^2 \tag{4.2}$$

Here, $\|\cdot\|_V$ denotes the Hilbert norm which is implicitly defined in our case by a Gaussian kernel. The variable $\phi_{\mathbf{t}}^{\mathbf{v}}$ denotes the net displacement field i.e, $\phi_{\mathbf{t}}^{\mathbf{v}} = \int_0^1 \mathbf{v_t}(\cdot;s)ds$.

Further, it has been theoretically shown in [19, 20] that the objective in Eq. (4.2) can be reformulated in terms of variables $\mathbf{m}_{\mathbf{t}}^0$, known as the initial momenta, according to:

$$\Gamma(\mathbf{m}^0) = \frac{1}{2}\sum_{i,j=1}^{T} \gamma_{ij}\mathbf{m}_{\mathbf{i}}^0\mathbf{m}_{\mathbf{j}}^0 + \lambda \sum_{t=1}^{T} \|\phi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p}_A^{\mathbf{t}};1) - \mathbf{p}_B^{\mathbf{t}}\|_2^2 \tag{4.3}$$

The variable γ_{ij} is an exponential smoothing kernel evaluated on pairs of time points of the source contour $\mathbf{p}_A^{\mathbf{t}}$.

During training, we solve Eq. (4.3) for every pair of source and target F0

contours to generate the ground truth momenta. This variable summarizes the transformation between emotion pairs. Since the momenta and source F0 contour uniquely specify the transformation, we use it as an intermediary between any given pair of utterances. In comparison, [11] predicts a momentum for every frame of the pitch contour and then warps it over several iterations specified by variable **s**. It is a sub-optimal strategy, as there is no temporal coherence constraint in predicting the momenta. Note that we do not have access to the ground truth momenta during testing and run the network in an open loop fashion without intermediate regularization.

4.2.2 Encoder-Decoder-Predictor Network

Current methods in emotion conversion modify the F0 and spectrum without imposing any explicit relationship between the features. As a result, there are significant residual harmonics present in the spectrum, which results in the poor quality of resynthesised speech. Our approach overcomes this limitation via the conditional relationships modeled in Fig. 4.1. Here, the conditional spectrum estimate is given by:

$$\hat{\mathbf{S}}_B = \arg\max_{\mathbf{S}_B} P(\mathbf{S}_B|\mathbf{S}_A, \mathbf{p}_A) \tag{4.4}$$

Using rules of probability, we can rewrite Eq. (4.4) as:

$$
\begin{aligned}
\hat{\mathbf{S}}_{\mathbf{B}} &= \arg\max_{\mathbf{S}_B} \int_{\mathbf{p}_B} P(\mathbf{S}_B, \mathbf{p}_B | \mathbf{S}_A, \mathbf{p}_A)\, d\mathbf{p}_B \\
&= \arg\max_{\mathbf{S}_B} \int_{\mathbf{p}_B} P(\mathbf{S}_B | \mathbf{S}_A, \mathbf{p}_B) P(\mathbf{p}_B | \mathbf{S}_A, \mathbf{p}_A)\, d\mathbf{p}_B \\
&= \arg\max_{\mathbf{S}_B} \int_{\mathbf{p}_B} P(\mathbf{S}_B | \mathbf{S}_A, \mathbf{p}_B) \times \int_{\mathbf{m}_{AB}} P(\mathbf{p}_B | \mathbf{m}_{AB}, \mathbf{p}_A) \\
&\qquad \times P(\mathbf{m}_{AB} | \mathbf{S}_A, \mathbf{p}_A)\, d\mathbf{m}_{AB}\, d\mathbf{p}_B \\
&= \arg\max_{\mathbf{S}_B} \int_{\mathbf{m}_{AB}} P(\mathbf{m}_{AB} | \mathbf{S}_A, \mathbf{p}_A) \times \int_{\mathbf{p}_B} P(\mathbf{p}_B | \mathbf{m}_{AB}, \mathbf{p}_A) \\
&\qquad \times P(\mathbf{S}_B | \mathbf{S}_A, \mathbf{p}_B)\, d\mathbf{p}_B\, d\mathbf{m}_{AB}
\end{aligned}
$$

where we have used Eq. (4.1) to derive the above expression. The first term term we encounter is $P(\mathbf{m}_{AB}|\mathbf{S}_A, \mathbf{p}_A)$ which is the probability density of the intermediate latent representation i.e., momenta. It is conditioned on both, the source F0 contour and the spectrum. The second term, $P(\mathbf{p}_B|\mathbf{m}_{AB}, \mathbf{p}_A)$ is the density over the target F0 contour given the momenta and the source F0 contour. Finally, $P(\mathbf{S}_B|\mathbf{S}_A, \mathbf{p}_B)$ is the target spectrum conditioned on the target pitch contour and the source spectrum. Note that the expression requires multiple integrations, and is hence, intractable. However, we can make point estimates for each density function using a deep convolutional neural network [21] (CNN) thereby, allowing us to write:

$$\begin{aligned}
\hat{\mathbf{m}}_{AB} &= \arg\max_{\mathbf{m}_{AB}} P(\mathbf{m}_{AB}|\mathbf{S}_A, \mathbf{p}_A; \theta_e) \\
\hat{\mathbf{p}}_B &= \arg\max_{\mathbf{p}_B} P(\mathbf{p}_B|\hat{\mathbf{m}}_{AB}, \mathbf{p}_A; \theta_d) \\
\hat{\mathbf{S}}_B &= \arg\max_{\mathbf{S}_B} P(\mathbf{S}_B|\mathbf{S}_A, \hat{\mathbf{p}}_B; \theta_p)
\end{aligned} \tag{4.5}$$

The CNN approximating $P(\mathbf{m}_{AB}|\mathbf{S}_A, \mathbf{p}_A; \theta_e)$ is called an encoder because it distills information about the input data. The CNN modeling $P(\mathbf{p}_B|\mathbf{m}_{AB}, \mathbf{p}_A; \theta_d)$ is called the decoder because it estimates the output pitch from the latent embedding and source pitch contour. The encoder-decoder portion is a basic sequence-to-sequence model for pitch contours. Finally, the CNN modeling $P(\mathbf{S}_B|\mathbf{S}_A, \mathbf{p}_B; \theta_p)$ is called a predictor as it generates the spectrum for the converted speech.

The architecture of these CNNs is shown in Fig. 4.2. We adapt the architecture from [22] by reducing the number of residual layers in each block. The entire sequence of three neural networks is trained together from a unified objective. The loss function for optimizing the parameters is given by:

$$\begin{aligned}
\mathscr{L} &= -\log\left(P\Big(\mathbf{m}_{AB}, \mathbf{p}_B, \mathbf{S}_B|\mathbf{S}_A, \mathbf{p}_A; \theta_e, \theta_d, \theta_p\Big)\right) \\
&= \lambda_e\|\hat{\mathbf{m}}_{AB} - \bar{\mathbf{m}}_{AB}\|_1 + \lambda_d\|\hat{\mathbf{p}}_B - \bar{\mathbf{p}}_B\|_1 + \lambda_p\|\hat{\mathbf{S}}_B - \bar{\mathbf{S}}_B\|_1
\end{aligned} \tag{4.6}$$

During training, we minimize the negative log likelihood of momenta and the target features with respect to θ. We model the conditional distribution of

each variable by Laplace density function. The corresponding ground truths ($\bar{\mathbf{m}}_{AB}, \bar{\mathbf{p}}_B, \bar{\mathbf{S}}_B$) are used as the mean while the variances are assumed to be constant. This in turn is equivalent to minimizing the mean absolute error of each target variable with an appropriate scaling, defined by λ_e, λ_d and λ_p, which are the hyperparameters in our model.

One benefit of coupling the neural networks is that the encoder and the decoder become aware of the downstream task of spectrum prediction. We train the neural network [23] using Adam optimizer [24] with a learning rate of 1e-5 and a mini-batch of size one. 23 dimensional MFCC features are used as spectrum representation extracted by an analysis window of length 5ms. During training, the context size is fixed at 640ms which results in dimensionality of 128×1 for F0 contour and 128×23 for spectrum. The dimensions of momenta are same as the F0 contour. The hyperparameters, λ_e, λ_d and λ_p are set to 0.01, 1e-4 and 1e-4, respectively. We do not normalize the input and output features during training to preserve their scale. Code can be downloaded from: https://engineering.jhu.edu/nsa/links/.

4.3 Experiments and Results

We carry out an ablation study for the momenta $\mathbf{m}_{AB}$ and a qualitative evaluation of emotional salience and quality.

4.3.1 Emotional Speech Dataset

We evaluate our algorithm on the VESUS dataset [15] which contains 250 parallel utterances spoken by 10 actors (gender balanced) in neutral, sad, angry

and happy emotional classes. Each spoken utterance has a crowd-sourced emotional saliency rating provided by 10 workers on Amazon Mechanical Turk (AMT). These ratings represent the ratio of workers who correctly identify the intended emotion in a recorded utterance. For robustness, we restrict our experiments to utterances that were correctly and consistently rated as emotional by at least 5 of the 10 AMT workers. As a result, the total number of utterances used are as follows:

- **Neutral to Angry conversion**: 1534 utterances for training, 72 for validation and, 61 for testing.
- **Neutral to Happy conversion**: 790 utterances for training, 43 for validation and, 43 for testing.
- **Neutral to Sad conversion**: 1449 utterances for training, 75 for validation and, 63 for testing.

Our subjective evaluation includes both an emotion perception test and, a quality assessment test. These experiments are carried out on Amazon Mechanical Turk (AMT); each pair of speech utterances is rated by 5 workers. The perception test asks the raters to identify the emotion in the converted speech sample, and the quality assessment test asks them to rate the quality of speech sample on a scale of 1 to 5. We include both the neutral and converted utterances to account for the speaker bias. Further, the samples were randomized to mitigate the effects of non-diligent raters and to identify bots.

4.3.2 Baselines

We compare our encoder-decoder-predictor model to three state-of-the-art baseline methods. The first approach learns a Gaussian mixture model using concatenated source and target features [8]. During inference, a maximum likelihood estimate of target features is made given the source features. A global variance constraint ensures that the estimate is not over-smooth, which is a common problem in joint modeling techniques.

The second baseline is a Bi-LSTM supervised learning approach [9]. Since Bi-LSTMs generally require considerable data to train, we adopt the strategy in [9] of training the model on a voice conversion task [25] and then fine-tuning it for emotion conversion. This method encodes the prosody features via a Wavelet transform to represent both short-term and long-term trajectory information of F0 and energy contours.

The third baseline is a recently proposed unsupervised method for emotion conversion [12]. This algorithm uses cycle-GANs to inject emotion into neutral utterances. A set of cycle-GAN transforms the spectrum while the other set transforms the prosody features. Once again, prosodic features are parameterized using Wavelet basis similar to the Bi-LSTM.

4.3.3 Experimental Results

As a sanity check, we carry out an ablation study to understand the effect of latent variable regularization via the LDDMM momenta. Fig. 4.3 shows the resulting mean absolute error in pitch prediction for each emotion pair. As seen, the F0 prediction is statistically significantly better in two emotional

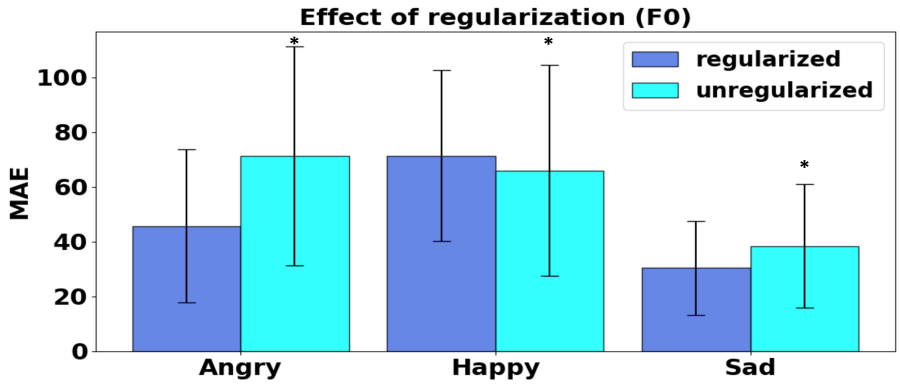

Figure 4.3: Effect of latent variable regularization on the prediction of fundamental frequency (F0) for each emotion pair. Marker $*$ indicates $p < 10^{-2}$ for paired t-test scores.

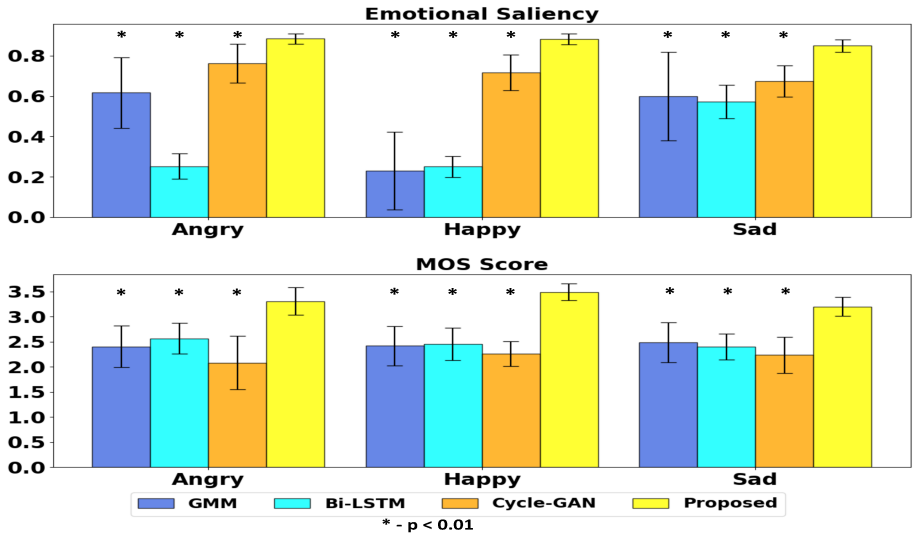

Figure 4.4: Confidence of emotion conversion (top) and the quality of reconstruction (bottom) for VESUS test samples.

pairs. Neutral to happy conversion is an exception to this general trend, but we conjecture that this is due to the smaller training dataset (∼800 samples compared to >1400 for angry and sad). The error bars in all three emotion pairs are however, tighter than the un-regularized model, indicating robustness.

4.3.3.1 Mixed Speaker Evaluation

Fig. 4.4 illustrates crowd-sourcing results on the VESUS test dataset. Our proposed method has the highest emotional saliency rating in comparison to the baselines. The GMM did not produce intelligible speech when trained in a multi-speaker setting, as the F0 and spectral features do not exhibit distinct clusters when aggregated across speakers. Hence, the results in Fig. 4.4 correspond to single-speaker training/testing. We note that our GMM evaluation is unfairly optimistic, and yet, the performance is worse than our method and the cycle-GAN. The Bi-LSTM model which simultaneously predicts the wavelet coefficients for F0 and energy, along with the spectrum has very poor conversion results for angry and happy. It is likely that the Bi-LSTM focuses on a subset of the features to minimize the overall loss. The cycle-GAN, on the other hand does produce reasonable results even though it is unsupervised. This is likely due to the implicit regularization produced by cyclic consistency and identity loss [26]. Lastly, our proposed model has the best conversion score for all three emotion pairs and the tightest error bars in comparison to the baselines. Thus, our approach of combining the local and global task in a chained model works extremely well by allowing the individual pieces to train efficiently without losing oversight of the end goal.

The bottom plot in Figure 4.4 shows the subjective quality of speech reconstruction after emotion conversion measured using mean opinion score (MOS). The chained neural network is uniformly better than the baseline algorithms on the VESUS dataset. It means that the proposed approach not only converts the emotion with a high degree of confidence but also manages to keep the

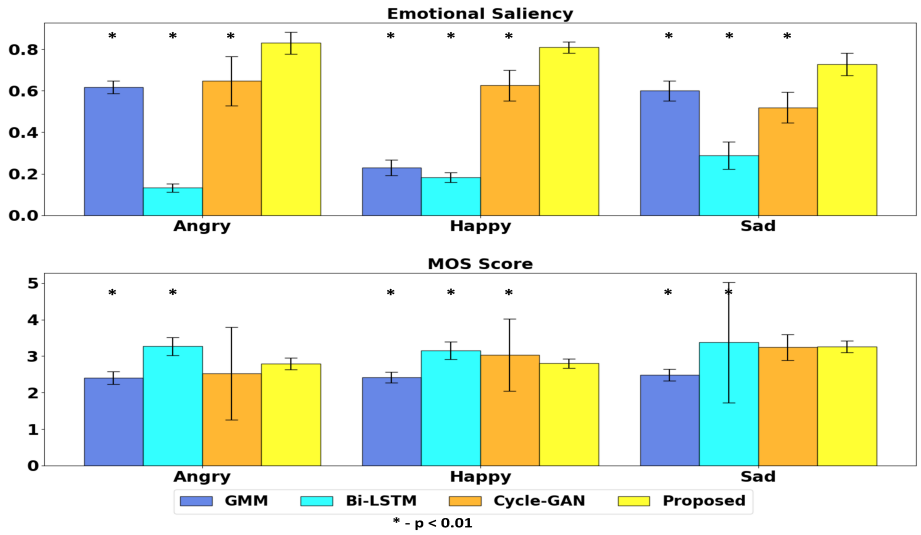

Figure 4.5: Confidence of emotion conversion (top) and the quality of reconstruction (bottom) on unseen samples.

quality of speech intact after conversion.

4.3.3.2 Out-of-Sample Generalization

We further conduct an out-of-vocabulary emotion conversion experiment. Here, we set aside 7 randomly selected phrases per speaker from each emotion category. These phrases are not part of the training set to simulate unseen utterances during testing. Fig. 4.5 shows the results of this experiment. The GMM results are based on single-speaker evaluation. Once again, the proposed model has the best conversion performance with narrow error bounds. The Bi-LSTM does worse on unseen utterances demonstrating a lack of generalization capability. On the other hand, the cycle-GAN degrades a little but the saliency stays above 0.5 for all three emotion pairs. This is mainly due to the non-parallel nature of the Cycle-GAN model which makes no assumption on the speakers or the utterances. Our approach achieves this by not normalizing the input features using cohort statistics. Taken together, conditioning the

spectrum estimation on the pitch can learn a complex relationship between the two which can be efficiently exploited as in our case.

The MOS in Fig. 4.5 show that Bi-LSTM has the best quality of reconstruction among the three. Empirically, it does not modify the speech at all, thereby, making it sound more natural by default. There is a tie for the second place between Cycle-GAN and the proposed model. Our proposed approach has much smaller error bars than Cycle-GAN due to training with un-normalized features and momenta regularization.

4.4 Conclusions

We have proposed a novel method for emotion conversion that modifies pitch and spectrum using a chained neural network. Our proposed approach used a latent variable to regularize the F0 estimation, which in turn affects the spectrum prediction. We showed that using a diffeomorphic prior on the F0 contour and conditioning of spectrum on it leads to better generalization on unseen utterances. The experiments were carried out on the VESUS dataset and results on converted test samples were statistically significant. We concluded that our proposed algorithm did not degrade the quality of speech during conversion.

References

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499 (2016). arXiv: 1609.03499.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *CoRR* abs/1712.05884 (2017). arXiv: 1712.05884.

[3] Jean-Marc Valin and Jan Skoglund. "LPCNET: Improving Neural Speech Synthesis through Linear Prediction". In: *ICASSP* (2019), pp. 5891–5895. DOI: 10.1109/ICASSP.2019.8682804.

[4] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens". In: *CoRR* abs/1910.11997 (2019). arXiv: 1910.11997.

[5] Yuxuan Wang, R. J. Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous. "Uncovering Latent Style Factors for Expressive Speech Synthesis". In: *CoRR* abs/1711.00520 (2017). arXiv: 1711.00520.

[6] Eric Battenberg, Soroosh Mariooryad, Daisy Stanton, RJ Skerry-Ryan, Matt Shannon, David Kao, and Tom Bagby. "Effective Use of Variational Embedding Capacity in Expressive End-to-End Speech Synthesis". In: *CoRR* abs/1906.03402 (2019). arXiv: 1906.03402.

[7] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. "A Review of Deep Learning Based Speech Synthesis". In: *Applied Sciences* 9 (2019), p. 4050. DOI: 10.3390/app9194050.

[8] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". In: *American Journal of Signal Processing* 2 (2012), pp. 134–138. DOI: 10.5923/j.ajsp.20120205.06.

[9] Huaiping Ming, Dong-Yan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion". In: *Proc. Interspeech 2016*. 2016, pp. 2453–2457. DOI: 10.21437/Interspeech.2016-1053.

[10] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective". In: *Proc. Interspeech 2019*. 2019, pp. 2848–2852. DOI: 10.21437/Interspeech.2019-2512.

[11] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Automated Emotion Morphing in Speech Based on Diffeomorphic Curve Registration and Highway Networks". In: *Proc. Interspeech 2019*. 2019, pp. 4499–4503. DOI: 10.21437/Interspeech.2019-2386.

[12] Kun Zhou, Berrak Sisman, and Haizhou Li. "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data". In: *CoRR* abs/2002.00198 (2020). arXiv: 2002.00198.

[13] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. "Computing large deformation metric mappings via geodesic flows of diffeomorphisms". In: *International journal of computer vision* 61.139-157 (2005).

[14] Sarang C Joshi and Michael I Miller. "Landmark matching via large deformation diffeomorphisms". In: *IEEE transactions on image processing* 9.8 (2000), pp. 1357–1370.

[15] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proc. Interspeech 2019*. 2019, pp. 316–320. DOI: 10.21437/Interspeech.2019-1413.

[16] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech Communication* 27 (1999), pp. 187–207. DOI: 10.1016/S0167-6393(98)00085-5.

[17] Masanori Morise, Fumiya YOKOMORI, and Kenji Ozawa. “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications”. In: *IEICE Transactions on Information and Systems* E99.D (2016), pp. 1877–1884. DOI: 10.1587/transinf.2015EDP7457.

[18] A. Sotiras, C. Davatzikos, and N. Paragios. “Deformable Medical Image Registration: A Survey”. In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1153–1190.

[19] Laurent Younes. *Shapes and Diffeomorphisms*. Springer-Verlag Berlin Heidelberg, 2010.

[20] Hsi-Wei Hsieh and Nicolas Charon. “Diffeomorphic registration of discrete geometric distributions”. In: *CoRR* abs/1801.09778 (2018). arXiv: 1801.09778.

[21] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *Proc. ICLR 2016*. 2016.

[22] Takuhiro Kaneko and Hirokazu Kameoka. “Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks”. In: *CoRR* abs/1711.11293 (2017). arXiv: 1711.11293.

[23] Martín Abadi and Ashish Agarwal et.al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: https://www.tensorflow.org/.

[24] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2015).

[25] John Kominek and Alan W Black. “The CMU Arctic speech databases”. In: *SSW5-2004* (2004).

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *Proc. ICCV 2017*. IEEE Computer Society, 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

Chapter 5

Unsupervised Variational CycleGAN for F0 and Energy

This chapter introduces a new framework for non-parallel emotion conversion in speech. Our framework is based on two key contributions. First, we propose a stochastic version of the popular Cycle-GAN model. Our modified loss function introduces a Kullback–Leibler (KL) divergence term that aligns the source and target *data distributions* learned by the generators, thus overcoming the limitations of sample-wise generation. By using a variational approximation to this stochastic loss function, we show that our KL divergence term can be implemented via a paired density discriminator. We term this new architecture a variational Cycle-GAN (VCGAN). Second, we model the prosodic features of target emotion as a smooth and learnable deformation of the source prosodic features. This approach provides implicit regularization that offers key advantages in terms of better range alignment to unseen and out-of-distribution speakers. We conduct rigorous experiments and comparative studies to demonstrate that our proposed framework is fairly robust with high performance against several state-of-the-art baselines.

5.1 Background

Emotional cues in speech are conveyed through vocal inflections known as prosody. Key attributes of prosody include the fundamental frequency (F0) contour, the relative energy of the signal, and the spectrum [1]. Many supervised and unsupervised algorithms have been proposed for emotion conversion. For example, the work of [2] proposed a Gaussian mixture model (GMM) to jointly model the source and target prosodic features. During inference, the target features are estimated from the source via a maximum likelihood optimization. A recent approach by [3] uses a Bidirectional LSTM (Bi-LSTM) to predict the spectrum and F0 contour. To overcome the data limitation, the authors pre-train their model on a voice conversion dataset and then fine-tune it for emotion conversion. The prosodic manipulation proposed by [4, 5] uses a highway neural network to predict the F0 and intensity for each frame of the input utterance. While these models have made significant contributions to the field, they require parallel emotional speech data for training, which limits their generalizability.

An unsupervised technique to disentangle style and content from speech has been proposed by [6]. This algorithm uses architecture based priors to separate style and content from spectrum while modifying the F0 using a linear Gaussian model. The authors of [7] offer a simpler cycle-GAN model for non-parallel emotion conversion, which independently modifies the spectrum and F0 contour. The latter is parameterized via a wavelet transform, which expands the input feature dimensionality. These approaches, however, are trained and evaluated on single speakers, with no validation on multispeaker

conversion.

In this chapter we propose a novel technique for emotion conversion using a variational formulation of the Cycle-GAN. Our novel loss formulation leads to a joint density discriminator which minimizes the upper bound on KL-divergence between the target data density and its parameterized counterpart. Our method further learns the target emotion F0 and energy contour by modeling them as a smooth deformation of the source emotion features. A preliminary version of this work appeared in Interspeech 2020 [8]. This chapter provides the following novel contributions. First, we model the transformation of F0 and energy contours of an utterance jointly using intermediate hidden variables. This is in contrast with the previous approach where we modify the F0 contour and spectrum, independently. Second, our graphical model for the conversion strategy allows us to disentangle the discriminator's objective for energy and F0 contour using conditional independencies directly inferred from the graph. Finally, we evaluate our proposed framework in both, a multi-speaker setting as well as on out-of-distribution speakers which the model does not see during training. We further provide comparative studies about the distribution and stability properties of our technique with a state-of-the-art baseline.

5.2 Method

Our strategy is to manipulate two key prosodic features: the F0 (pitch) contour, and the energy (loudness) contour. Fig. 5.1 shows the relationship between the features during the inference step of the process. We begin with taking

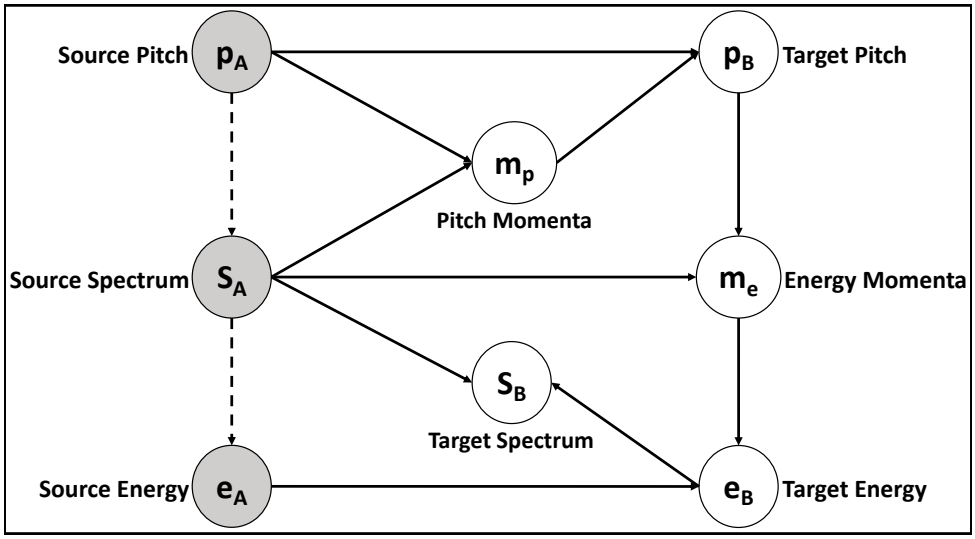

Figure 5.1: Graphical representation of our emotion conversion strategy. $\mathbf{m_p}$ and $\mathbf{m_e}$ serve as an intermediaries for pitch and energy contours, respectively.

an utterance in source emotion A from which we extract the F0 contour $(\mathbf{p}_A)$ and the mel-cepstral features $(\mathbf{S}_A)$ using the WORLD vocoder [9]. The energy contour $(\mathbf{e}_A)$ is extracted directly from the spectral features. We define latent variables called momenta $(\mathbf{m}_p, \mathbf{m}_e)$, which serve as intermediaries between the two emotion classes under consideration. The F0 contour in target emotion $(\mathbf{p}_B)$ is a deterministic function of the momenta $(\mathbf{m}_p)$ and source F0 contour, through a diffeomorphic warping process that we describe in Section 5.2.2. The estimated F0 contour and the source spectrum together generate the momenta $(\mathbf{m}_e)$ for the energy contour which is then further used to generate the cepstral features $(\mathbf{S}_B)$. The estimated F0 contour and cepstral features combine together to give the converted utterance in the target emotion B.

We take an unsupervised approach to model training and evaluation using a Cycle-GAN framework. This strategy allows us to handle non-parallel and multi-speaker datasets. For robustness, we introduce a novel KL-divergence

loss to align the *distribution* of the source and target emotional classes, as described in Section 5.2.1. The KL-divergence gives rise to a new class of discriminators that operate on pairs of samples.

To summarize, our technical innovations are as follows:

- We propose a joint model for F0 and energy modification which uses latent variables called momenta as an intermediary between source and target emotion features.
- We highlight several shortcomings of cyclic consistency loss which is the backbone of our baseline reference model and analyze them theoretically.
- We propose a new KL-divergence penalty and minimize its upper bound to address the limitations of cyclic loss. We verify its advantages through multiple experiments.
- We evaluate our model on multiple experiment paradigms i.e, single speaker, mixed speaker, leave-one-fold and Wavenet to paint a complete picture of our model.

5.2.1 Variational Cycle-GAN

The cycle consistency loss of a traditional Cycle-GAN is given by Eq. (2.5) and repeated below for convenience:

$$\mathscr{L}_{cycle} = E_{x\sim P(X)}\left[\|x - G_\theta(G_\gamma(x))\|_1\right] \tag{5.1}$$

This formulation imposes just a point-wise regularization on the input X and the cyclic converted sample $G_\theta(G_\gamma(x))$.

It is easy to show that Eq. (5.1) is not a well-behaved loss function (Propositions 1 and 2 in [10]). Specifically,

- It only enforces a first-order moment matching between the generated and target data distributions.
- The expectation in Eq. (5.1) depends on the sampling variance, which leads to a noisy gradient estimate when optimizing the parameters of the generator.

The first point establishes a weak coupling between the two generators. In addition, the discriminators D_θ and D_γ do not have information about the complementary generators when training a traditional Cycle-GAN. At a high level, the min-max game played by the generators and discriminators is operating on incomplete information about the underlying data.

The second point often results in poor calibration of the gradients under scenarios where the target distribution is perfectly learnable. Practically speaking, this sampling variance is unknown, which can lead to instability during the optimization. For example, it may prompt the generator to take a step that does not reduce the cycle consistency loss (e.g., overshooting the local optimum). Further, because this variance is inherently tied to the parameters of the neural network, the generators can potentially end up learning a null or an identity function in order to minimize the expected cycle consistency loss (e.g., mode collapse). Finally, due to the expected loss being a function of the dimensionality of the data, it scales the gradients computed during backpropagation making the impact of sampling variance more pronounce.

We approach these problems by considering KL-divergence based penalty on the input data distribution and the cyclic transformation. Formally, let $(\mathbf{S}_A, \mathbf{p}_A)$ and $(\mathbf{S}_B, \mathbf{p}_B)$ be the source and target cepstrum and F0 contours of two *non-parallel* utterances in emotion A and B, respectively. The generators are denoted by $G_\gamma : (\mathbf{S}_A, \mathbf{p}_A) \rightarrow (\mathbf{S}_B, \mathbf{p}_B)$ and $G_\theta : (\mathbf{S}_B, \mathbf{p}_B) \rightarrow (\mathbf{S}_A, \mathbf{p}_A)$. The corresponding distributions learned by the generator functions are given by $P_\gamma(\mathbf{S}_B, \mathbf{p}_B)$ and $P_\theta(\mathbf{S}_A, \mathbf{p}_A)$. Our new penalty for the generator G_γ is:

$$\mathscr{L}_{G_\gamma} = KL\Big(P(\mathbf{S}_A, \mathbf{p}_A) \| P_\theta(\mathbf{S}_A, \mathbf{p}_A)\Big) \tag{5.2}$$

Using the law of total probability, we can write:

$$P_\theta(\mathbf{S}_A, \mathbf{p}_A) = \int\int P_\theta(\mathbf{S}_A, \mathbf{p}_A | \mathbf{S}_B, \mathbf{p}_B) \times P(\mathbf{S}_B, \mathbf{p}_B)\, d\mathbf{S}_B\, d\mathbf{p}_B \tag{5.3}$$

Eq. (5.3) is generally intractable, but we can derive an upper bound on the loss in Eq. (5.2) that can be optimized easily [10]. Effectively, we can minimize:

$$\tilde{\mathscr{L}}_{G_\gamma} = E_{(\mathbf{S}_A, \mathbf{p}_A)}\Big[E_{(\mathbf{S}_B, \mathbf{p}_B) \sim P_\gamma}\Big[\log\Big(P_\gamma(\mathbf{S}_B, \mathbf{p}_B | \mathbf{S}_A, \mathbf{p}_A) \times P(\mathbf{S}_A, \mathbf{p}_A)\Big)\Big]\Big] \tag{5.4}$$

Eq. (5.4) highlights an important difference between traditional Cycle-GAN and our variational approach. Namely, our min-max objective leverages higher-order relationships by comparing the joint density of source and target data factorized by the two generators. This transparency is noticeably absent in the traditional Cycle-GAN, in which the discriminator operates

on the marginal densities $P(\mathbf{S}_A, \mathbf{p}_A)$ and $P_\theta(\mathbf{S}_A, \mathbf{p}_A)$ to determine whether the sample is “real" or “fake". Finally, we implement the spectrum modification module solely by changing the energy contour; this strategy avoids degradation in speech quality due to errors in spectrum prediction. We have conducted an experiment [10], which demonstrates no difference in user preference for speech generated with the original (mismatched) spectrum and speech generated with a modified spectrum based on the new F0 contour.

5.2.2 Prosodic Regularization via Momenta

As shown in the Fig. 5.1, we use two intermediate representations (denoted by $\mathbf{m_p}$ and $\mathbf{m_e}$) to model the transition of prosodic features from the source to target emotion. This technique can be viewed as an implicit regularization on the conversion procedure. Practically, we model the target prosodic contours as a smooth deformation of the source F0/energy contours. This idea stems from the domain of image registration where a moving image is iteratively deformed to align or match with a fixed image [11]. We adapt this registration framework from 2-dimensional image surfaces to 1-dimensional curves in the Euclidean space.

While there are multiple ways to represent the deformation process, one popular technique is known as the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [12, 13]. These functions are defined as a smooth and invertible mapping between two topological manifolds. An important feature of this LDDMM model is the ability to parameterize diffeomorphic transformations by low-dimensional embeddings known as momenta [14]. Effectively,

the source prosodic contour specifies the initial state, while the momenta ($\mathbf{m_p}$) specifies the initial trajectory of the dynamical system. Thus, specifying the input curve and momenta are sufficient to generate the final state of a target curve.

Mathematically, let $\mathbf{p}_A^{\mathbf{t}}$ and $\mathbf{p}_B^{\mathbf{t}}$ denote a pair of source and target F0 contours, respectively. The variable $\mathbf{t}$ corresponds to the location of the analysis window as it moves across a given speech utterance. The goal of the deformation process is to estimate a series of small vertical displacements $\mathbf{v_t}(\mathbf{x}; \mathbf{s})$ over frequency and time. The integral of these small displacements produces a final large vector field denoted by $\phi_{\mathbf{t}}^{\mathbf{v}} = \int_0^1 \mathbf{v_t}(\cdot; s)ds$ [12]. Representing the momenta variable by $\mathbf{m}_p$, the LDDMM objective function can be written as:

$$\Gamma(\mathbf{m}_p) = \frac{1}{2} \sum_{i,j=1}^{T} \gamma_{ij}[\mathbf{m}_p]_i[\mathbf{m}_p]_j + \lambda \sum_{t=1}^{T} \|\phi_{\mathbf{t}}^{\mathbf{v}}(\mathbf{p}_A^{\mathbf{t}}; 1) - \mathbf{p}_B^{\mathbf{t}}\|_2^2 \tag{5.5}$$

The variable γ_{ij} is an exponential smoothing kernel evaluated on pairs of time points of the source contour $\mathbf{p}_A^{\mathbf{t}}$ whereas, λ is the trade-off between smoothness of momenta and the difference between the source and target F0 contours.

Rather than solving Eq. (5.5) explicitly to obtain the momenta, we estimate it blindly via sampling from the generators. From a practical standpoint, the continuous time process specified by LDDMM can be easily discretized to run for a fixed number of iterations. The main advantage of using a latent regularizer is that it allows the F0 and energy contours to be generated in a dynamically controlled fashion. Adversarial training can be susceptible to mode collapse due to imbalance between generator-discriminator losses,

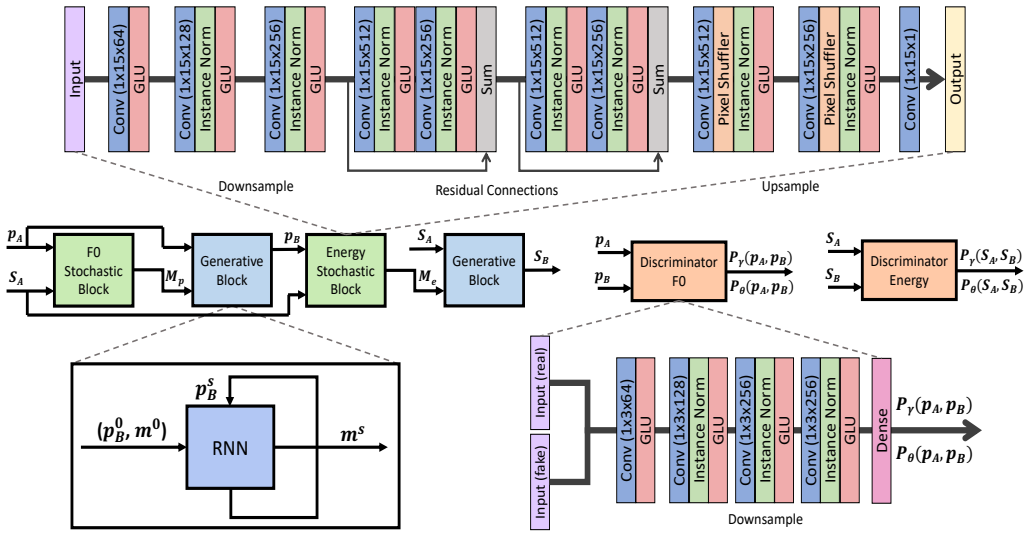

Figure 5.2: Architecture of the neural network for F0 and energy prediction. The output of F0 prediction is fed as input for energy estimation. Each generator has two blocks: a stochastic block for sampling momenta and a generative/deterministic block for curve warping (represented as an RNN).

learning rates, and the architecture of the neural networks. Deformation based F0 estimation stabilizes the generative process and prevents it from swinging wildly and leading to mode collapse. We will also demonstrate that this latent regularization improves the generalization capabilities of our framework to unseen speakers. Algorithm 1 outlines the warping process given a momenta, an F0 contour and an exponential smoothing kernel having a scale σ. This scale parameter controls the smoothness of the velocity vector fields and is fixed for all our experiments.

5.2.3 Hybrid Generative Architecture

Our F0/energy conversion is a two-step process: first, we estimate the momenta, then, we modify the source prosodic contours via a deterministic warping using momenta. Our generators mimic this process by integrating a

stochastic component with trainable parameters and a deterministic component with fixed/static parameters. The stochastic component for F0 momenta prediction takes the spectrum and source F0 as its inputs. For energy momenta prediction, the stochastic component relies on the source spectrum (which implicitly contains the energy information) and converted F0. The dimensions of the momenta are the same as F0 and energy contour. We empirically fix the smoothness parameter, σ at 50 for F0 and at 2 for energy contour to span the appropriate ranges. We adapt the 1-D convolutional architecture from [15] for the stochastic block of the generators as shown in Fig. 5.2. It has been experimentally verified that fully convolutional networks are more stable in a GAN framework than including fully-connected layers [16]. The deterministic LDDMM warping function can be represented as a recurrent neural network (RNN) with a fixed set of parameters due to its iterative nature.

We constrain the generators to sample smoothly varying momenta by adding a Laplacian penalty $\mathscr{L}_m = E[\|\nabla \mathbf{m}_p\|^2] + E[\|\nabla \mathbf{m}_e\|^2]$ to the overall generator loss. The gradient of this term is approximated by the first-order difference of the momenta along the time axis. The final objective to minimize for the loss of generator G_γ is as follows:

$$\begin{aligned}\mathscr{L}_{G_\gamma} &= \lambda_{c_1} E\Big[\|\mathbf{p}_A - \mathbf{p}_A^c\|\Big] + \lambda_m E\Big[\|\nabla \mathbf{m}_p\|_2^2 + \|\nabla \mathbf{m}_e\|_2^2\Big] \\ &+ \lambda_i E\Big[\|\mathbf{e}_A - \mathbf{e}_A^I\|\Big] + \lambda_{c_2} E\Big[\|\mathbf{e}_A - \mathbf{e}_A^c\|\Big] \\ &+ \lambda_d E_{(\mathbf{S}_A, \mathbf{p}_A)}\Big[E_{(\mathbf{S}_B, \mathbf{p}_B)\sim P_\gamma}\Big[\log\Big(D_\gamma(\mathbf{S}_A, \mathbf{p}_A, \mathbf{S}_B, \mathbf{p}_B)\Big)\Big]\Big] \end{aligned} \tag{5.6}$$

In the case of energy contour modification, we add an identity loss to

Algorithm 1: Warping to generate the target F0 contour given the momenta and source F0 contour

function GenerateF0 $(\mathbf{m}_p, \mathbf{p}_A)$;
Input : momenta $(\mathbf{m}_p)$ and source F0 $(\mathbf{p}_A)$
Output: target F0 $(\mathbf{p}_B)$
Set $s = 0$, $[\mathbf{p}_B]^0 = \mathbf{p}_A$ and $[\mathbf{m}_p]^0 = \mathbf{m}_p$;
if $s < 5$ **then**
    $d_{i,j} \leftarrow [\mathbf{p}_A]_i^s - [\mathbf{p}_A]_j^s$;
    $K_{i,j} \leftarrow \exp -\frac{(d_{i,j})^2}{\sigma^2}$;
    $[\mathbf{p}_B]_i^{s+1} \leftarrow [\mathbf{p}_B]_i^s + \sum_l K_{i,l} \cdot [\mathbf{m}_p]_l^s$;
    $[\mathbf{m}_p]_i^{s+1} \leftarrow [\mathbf{m}_p]_i^s + 2 \cdot \sum_j \frac{-K}{\sigma^2}\, d_{i,j} \cdot [\mathbf{m}_p]_i^s\, [\mathbf{m}_p]_j^s$;
    $s \leftarrow s + 1$;
else
    return $[\mathbf{p}_B]^s$;
end

the generator, which keeps the modified contour “close" to the original. The superscripts I and c denote the identity and cyclic components, respectively. Identity loss has been proposed by [17] in Cycle-GANs to make the generators more robust and allow them to reduce distortion when presented with a sample from target density itself. We omit the identity loss for the F0 conversion, as this contour tends to vary widely across utterances and emotional classes.

Finally, we update the parameters of the stochastic block of the generators by back-propagating through the deterministic LDDMM transformation, as implemented by an RNN.

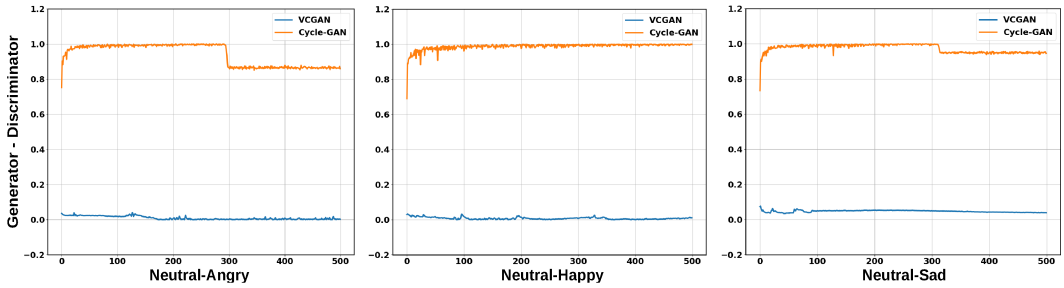

Figure 5.3: Comparing Cycle-GAN with its variational counterpart. On Y-axis, we denote the difference between the generator and discriminator loss. On X-axis, we denote the number of epochs. The plots represent the mismatch between the adversarial losses which is an indicator of instability in training [18].

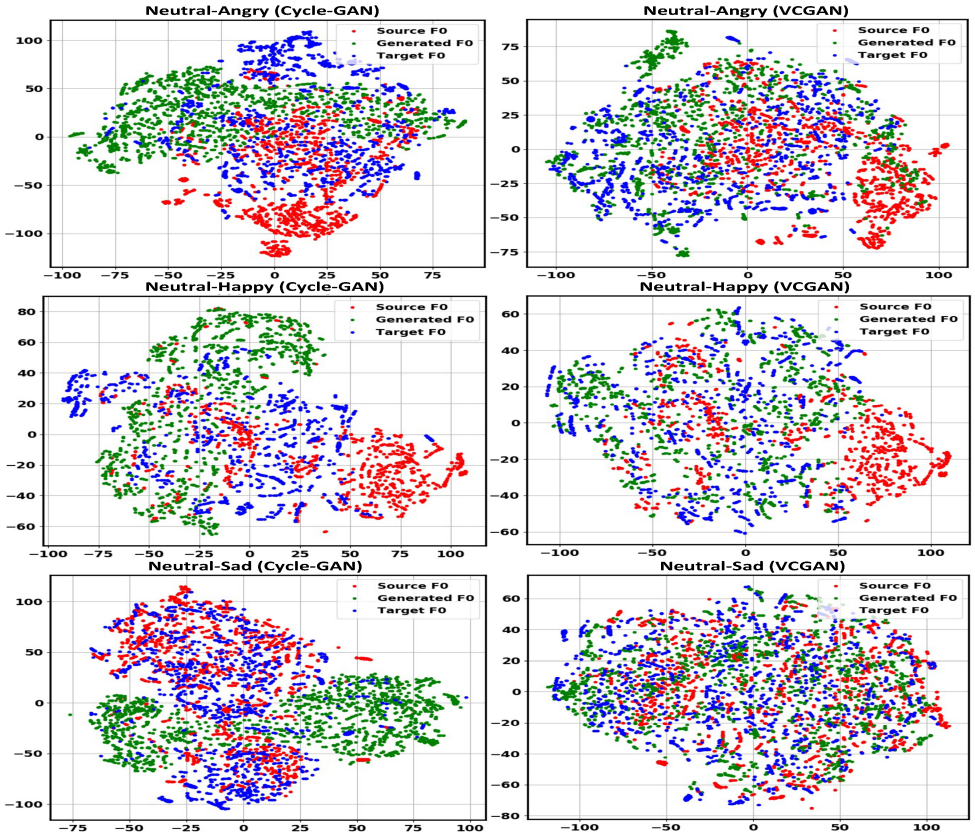

Figure 5.4: Visualizing t-SNE embeddings of source, converted and target F0 contours. The left column shows the embeddings generated using Cycle-GAN and the right column shows the same for variational model.

5.2.4 Discriminator Loss and Architecture

We model the probability ratio term in Eq. (5.4) by a discriminator denoted by D_γ. Conceptually, this discriminator distinguishes between the joint distributions of $(\mathbf{S}_A, \mathbf{p}_A)$ and $(\mathbf{S}_B, \mathbf{p}_B)$ learned by generators G_γ and G_θ, respectively.

During training of the discriminator D_γ, we minimize:

$$\mathscr{L}_{D_\gamma} = -E_{(\mathbf{S}_A, \mathbf{p}_A)}\Big[E_{(\mathbf{S}_B, \mathbf{p}_B)\sim P_\gamma}\Big[\log\Big(D_\gamma(\mathbf{S}_A, \mathbf{p}_A, \mathbf{S}_B, \mathbf{p}_B)\Big)\Big]\Big]$$

$$- E_{(\mathbf{S}_B, \mathbf{p}_B)}\Big[E_{(\mathbf{S}_A, \mathbf{p}_A)\sim P_\theta}\Big[\log\Big(1 - D_\gamma(\mathbf{S}_A, \mathbf{p}_A, \mathbf{S}_B, \mathbf{p}_B)\Big)\Big]\Big] \tag{5.7}$$

A similar discriminator has been proposed to train autoencoders in an adversarial setting [19, 20]. We use this discriminator to establish a macro connection between the two generators by providing them complete information about the generators. Another way to interpret Eq. (5.7) is that it classifies between different factorizations of the complete data distribution by the two generators. In fact, the optimal discriminators train the corresponding generators to minimize the Jensen-Shannon divergence between $P_\gamma(\mathbf{S}_A, \mathbf{p}_A, \mathbf{S}_B, \mathbf{p}_B)$ and $P_\theta(\mathbf{S}_A, \mathbf{p}_A, \mathbf{S}_B, \mathbf{p}_B)$ (derived in [10]).

We split each discriminator in two partial discriminators which separately provide the feedback for F0 and energy conversion task [10]. There are three advantages to splitting up the discriminator's loss into an F0 and an energy contribution. First, this strategy provides greater flexibility, as the user can decide whether or not to perform energy conversion without altering the F0 transformation. This scenario may be useful in cases of limited training data, as we empirically observe greater variability in energy across utterances, thus making it harder to learn a conversion model. Second, as noted in this section,

decoupling the F0 and energy backpropagation procedures prevents either variable from dominating the joint distribution during training. Third, we found empirically that training a unified discriminator results in an unstable model [10]. This is because the gradient information must backpropagate through the energy generator to update the F0 model parameters. Thus, the error signal suffers from a vanishing gradient problem, which makes it challenging to properly train the generative models. In contrast, splitting the discriminator allows F0 model to get a direct feedback from its corresponding discriminator for improved learning. We refer to this combined framework for F0 and energy conversion as a Variational Cycle-GAN (VCGAN).

5.2.5 Modifying the Spectrum via Energy

The spectral envelope is highly sensitive to changes in the location and filter response of the resonance frequencies. In fact, even minor changes can substantially degrade the quality and intelligibility of resynthesized speech. Our VCGAN framework circumvents this problem by modifying just the energy profile of the spectral envelope, i.e., the energy contour.

First, we extract the energy contour of the given speech signal from its spectral representation using:

$$\mathbf{e}_A = \sum_{f=0}^{\frac{F_s}{2}} [\mathbf{S}_A]_f^t \tag{5.8}$$

where, f corresponds to the frequency and t is the time. Once the energy contour has been modified through the VCGAN, denoted as $\mathbf{e}_B$, then the

converted spectrum $\mathbf{S}_B$ is given by:

$$\mathbf{S}_B = \mathbf{S}_A \times \sqrt{\frac{\mathbf{e}_B}{\mathbf{e}_A}} \tag{5.9}$$

This operation scales the frequency bins uniformly and simply modifies the overall intensity profile of the speech utterance.

During training, we use 23-dimensional MFCC features for spectrum representation over a context of 128 frames extracted using a 5ms windows. The dimensionality of F0/energy contour is 128x1 while that of spectrum is 128x23. The smoothing kernel for registration is chosen to be $[6, 50]$ and $[6, 2]$ for the F0 and energy contour, respectively. The generator and discriminator networks are optimized alternately in every mini-batch update. We fix the mini-batch size to 2 and the learning rates are fixed at 1e-5 and 1e-7 for the generators and discriminators, respectively. We use Adam optimizer [21] with an exponential decay of 0.5 for the first moment. Sampling process in the generators is implemented via dropout [22] rate of 0.3 during both training and testing.

5.3 Experimental Results: Demonstrating Model Stability

In this section, we demonstrate the desirable properties of our variational formulation, as compared to the traditional Cycle-GAN proposed in [6]. We also demonstrate the effectiveness of momenta regularization over the standard discrete wavelet transform representation. These experiments highlight the benefits of our VCGAN for emotion conversion.

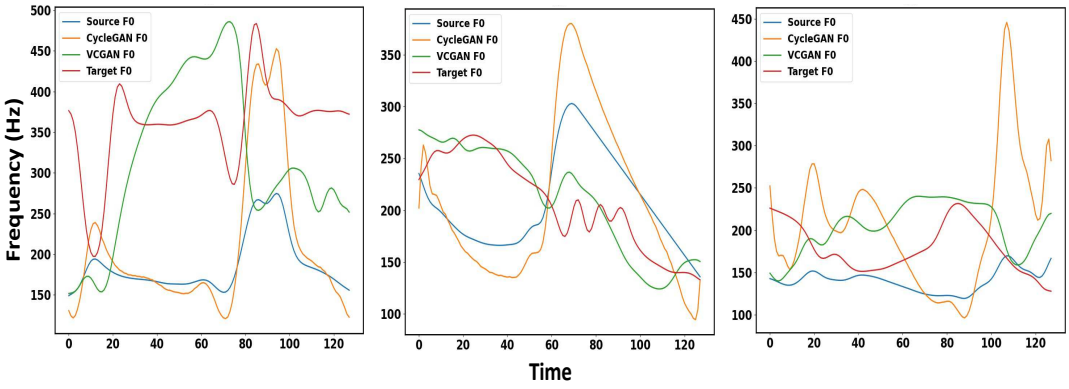

Figure 5.5: Comparing the F0 contours generated by Cycle-GAN and our momenta regularized variational model. Using diffeomorphic warping as a regularizer leads to more stable F0 contour generation in comparison to wavelet based regularization.

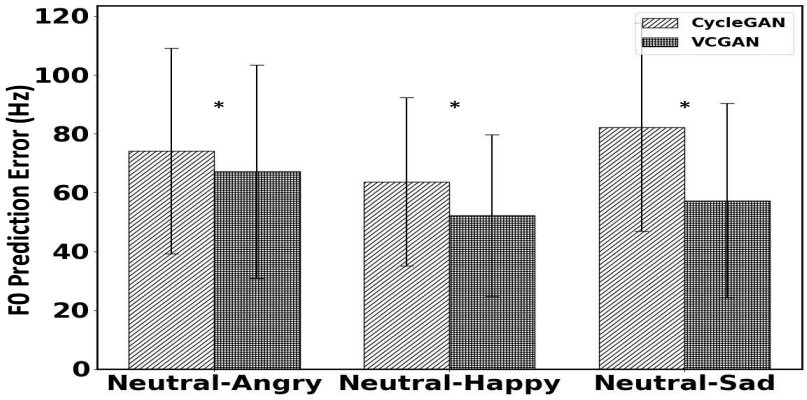

Figure 5.6: F0 RMSE comparison between Cycle-GAN and VCGAN. The results are statistically significant at level 0.05 (* denote p-value $\leq 1e-10$).

5.3.1 VESUS Dataset

We evaluate our algorithms on the VESUS dataset [23] collected at Johns Hopkins University. VESUS contains 250 utterances/phrases spoken by 10 different actors (gender balanced) in neutral, sad, angry and happy emotional classes. Each spoken utterance has a crowd-sourced emotional saliency rating collected from 10 workers on Amazon Mechanical Turk (AMT) [24]. These ratings represent the ratio of workers who correctly identify the intended

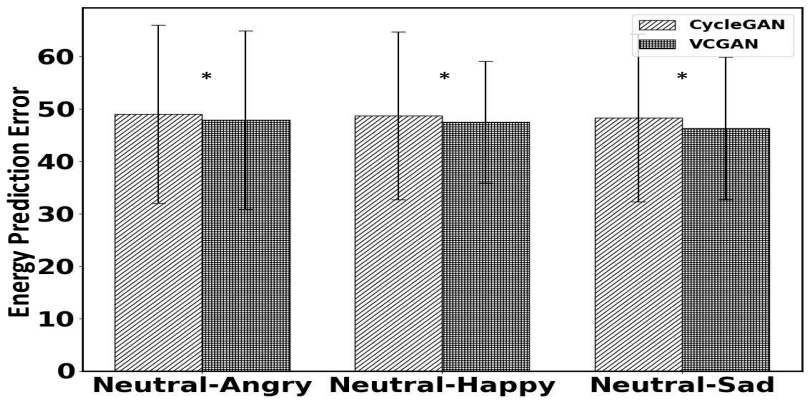

Figure 5.7: Energy RMSE comparison between Cycle-GAN and VCGAN. Results are statistically significant at level 0.05 (* denote p-value $\leq 1e-10$).

emotion in a recorded utterance. For robustness, we restrict our experiments in this section and the next to utterances that were correctly and consistently rated as emotional by at least 5 out of the 10 AMT workers. The total number of utterances for each emotion class are:

- **Neutral to Angry conversion**: 1667 utterances.
- **Neutral to Happy conversion**: 876 utterances.
- **Neutral to Sad conversion**: 1587 utterances.

5.3.2 Stability of Training

We first evaluate model stability during training. Here, we borrow from game theory to quantify performance. Namely, the optimal outcome of an adversarial game occurs when both participants achieve the Nash equilibrium [25]. Translating this idea into generative adversarial training implies equality of generator and discriminator losses. While a strict equality is difficult to achieve in practice, similar losses typically indicate better quality of the

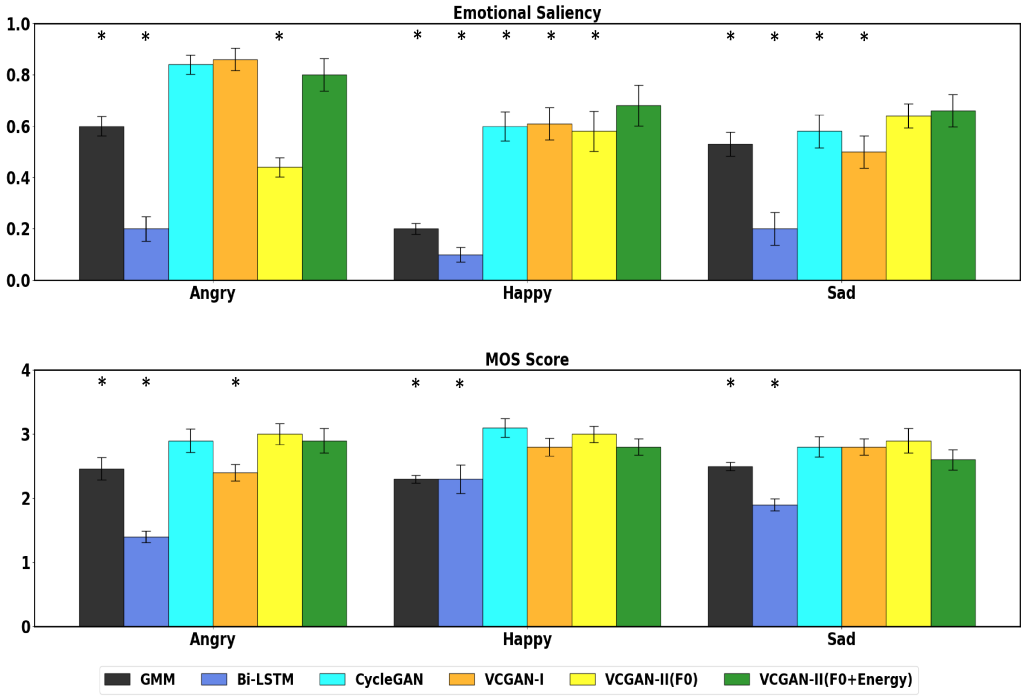

Figure 5.8: Single-Speaker Evaluation: we create the training, validation and testing sets for each emotion pair based on the speaker from VESUS with the highest number of emotionally salient utterances. The asterisk (*) denotes statistical significance for the test (VCGAN-II (F0+Energy) $>$ Method) at $p < 0.05$.

generated samples. Fig. 5.3 shows the difference between the generator and discriminator losses for the Cycle-GAN (orange) and our proposed VCGAN (blue). We note that the VCGAN achieves better calibration of the generator and discriminator objectives (i.e., near equality), whereas traditional Cycle-GAN fails to do so. Thus, we conclude that our training algorithm exhibits better stability in practice. Another important aspect of our proposed strategy is that computed training loss wiggles around the optimal point. It is crucial for adversarial training as the absence of this variance can sometimes signal mode collapse [26].

To illustrate the improved generator calibration, Fig. 5.4 shows the tSNE

plots [27] of the source, generated, and target emotion F0 values extracted over 640ms long windows. This duration typically encompasses multiple syllables in conversational English, often corresponding to words, and is therefore supra-segmental in nature. Notice that the point cloud of generated F0 values by the Cycle-GAN shows poor overlap with the target F0 distribution. We hypothesize that, as the Cycle-GAN focuses on just the first-order moments, the generators ultimately learn a mapping function whose output lies on a completely different manifold than the actual data distribution. This further indicates that the Cycle-GAN acts as a poor estimator of the target data density due to the weak constraint imposed by cycle-consistency loss. The VCGAN, on the other hand, does a much better job of approximating the target data density. This is because the KL-divergence penalty between the given data distribution and its cyclic counterpart enforces a stronger global dependency between the two generators. This macro connection in the form of feedback from the joint-density discriminator facilitates learning a better mapping function, especially given the limited data.

5.3.3 Effect of Momenta Regularization

The second critical component of our proposed VCGAN framework is the momenta based regularization for modeling the target prosodic contours. As discussed, the momenta specify an iterative warping process. In contrast, the works of [3, 7] use a continuous wavelet transform to parameterize the F0 contour to stabilize its generative process. Empirically, we observe

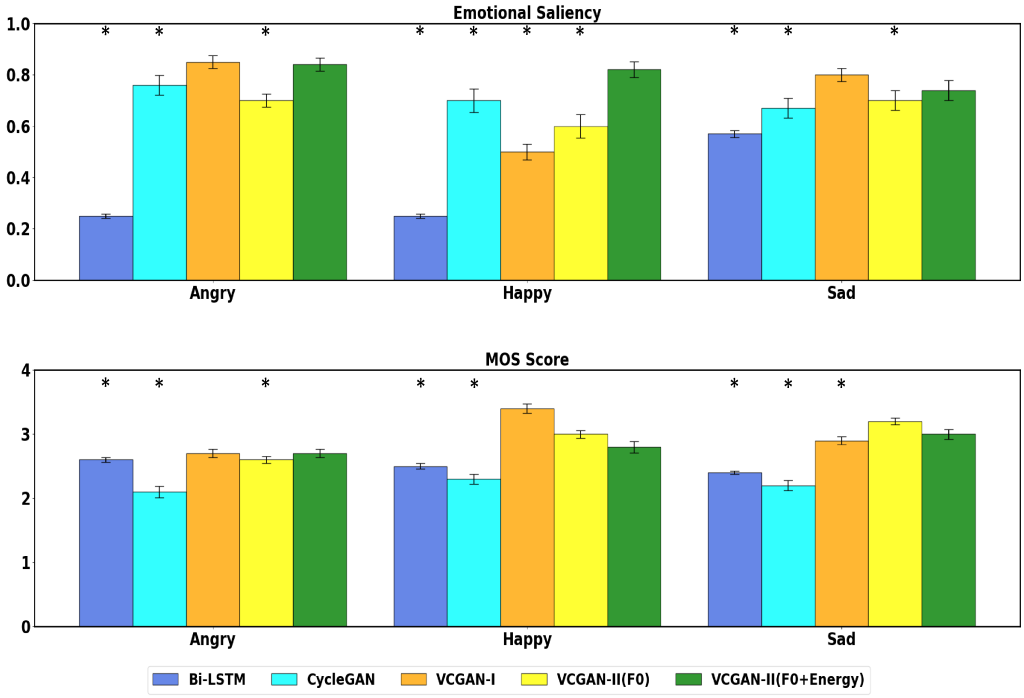

Figure 5.9: Mixed Speaker Evaluation: we create the training, validation and testing sets by randomly sampling utterances from VESUS across all speakers. The asterisk (*) denotes statistical significance for the one-tailed t-test (VCGAN (F0+Energy) > Method) at $p < 0.05$.

that our proposed momenta-based warping allows more flexible transformations and better scale matching between the generated and target contours. Fig. 5.5 shows example pitch contours *generated during testing*. As seen, the momenta-based VCGAN is less sensitive to extreme local fluctuations in the generated contours due to the iterative warping process. Moreover, our warping approach takes the source F0 contour as a baseline curve and estimates a perturbation on top of it. This results in a better alignment of the scale of F0 values in going from one emotion to the other (see Fig. 5.5).

To establish our claim objectively, we use paired samples from the VESUS dataset to compute root mean square error (RMSE) between the generated

and target frames of the F0 (Fig. 5.6) and energy (Fig. 5.7) contours. As seen, our momenta-based warping is significantly better than wavelet based regularization used in [3, 6] for all three emotion conversion tasks. The overall F0 loss is slightly higher for neutral-angry and neutral-sad conversion in comparison to the neutral-happy conversion. This is because the sad and angry emotions are portrayed in a more diverse manner in the VESUS dataset.

5.4 Experimental Results: Emotion Conversion

In this section we evaluate the emotion conversion performance against several supervised and unsupervised baseline algorithms. We train a separate model for each pair of emotions. However, the model architecture remains fixed in each case. Our subjective evaluation includes both an emotion perception query and a quality assessment test carried out on Amazon Mechanical Turk (AMT). Specifically, each pair of speech utterances (neutral and converted) is rated by 5 workers on AMT. The perception test asks the raters to identify the emotion in the converted speech sample after listening the corresponding neutral utterance. The quality assessment test asks them to rate the quality of the speech sample on a 1-5 scale, also called as mean opinion score or MOS. The reason we include both the neutral and converted utterances is to account for the speaker bias. Given the known variability in emotional perception across people, we collect 5 ratings for each converted sample and report the average. Finally, some samples were randomly and intentionally corrupted to mitigate the effects of non-diligent raters and to identify/flag bots.

We conduct four evaluations of increasing level of difficulty. The simplest scenario is single-speaker emotion conversion, in which we train and evaluate the model on utterances from the same speaker. Next is a mixed-speaker evaluation, in which we pool the utterances across speakers for each emotion class and randomly divide them into training, validation, and testing. The third assessment is out-of-speaker evaluation; here the models are trained and tested on different speakers. Finally, our Wavenet evaluation is the most difficult and queries how well the models generalize to synthetic speech.

5.4.1 Baseline Models

We compare our proposed VCGAN with several state-of-the-art algorithms from supervised and unsupervised learning domains. The first baseline is the global variance constrained GMM used for voice conversion, which learns the join density of source and target emotion features [2]. The second baseline uses a Bi-LSTM model [3] to learn the conditional density of the target emotion features namely, the F0 and energy contours. This method uses the wavelet decomposition of the prosodic features to control the segmental and supra-segmental nature of prosody. The third technique is a recently proposed Cycle-GAN framework [7] to modify the F0 contour using its wavelet parameterization. Further, the authors learn a secondary set of Cycle-GANs to modify the mel-cepstral features for every pair of source-target emotions. Our fourth baseline is a simplified version of the proposed VCGAN model [8] (referred in experiments as VCGAN-I). It is a mixed approach in the sense that it learns a variational Cycle-GAN for the F0 conversion and a traditional Cycle-GAN for

converting the Mel-cepstral features. In essence, all of the baseline techniques in this work modify the F0 and energy contour (as extracted from the spectrum or MFCC features). Finally, we compare our complete F0+energy modification framework with just F0 modification to understand the role of energy contour.

5.4.2 Single Speaker Evaluation

We first evaluate how well our VCGAN framework can convert emotions for a single speaker. Note that, this is the simplest setting in which our goal is to show generalization on a single speaker. To maximize the amount of data, we select the VESUS speaker with the highest number of consistently rated utterances (see Section III-A) for each emotion pair. This yields the following sample sizes:

- **Neutral to Angry Conversion**: 200 utterances for training, 25 for validation and, 10 for testing.
- **Neutral to Happy Conversion**: 100 utterances for training, 5 for validation and, 10 for testing.
- **Neutral to Sad Conversion**: 200 utterances for training, 25 for validation and, 10 for testing.

Fig. 5.8 illustrates the performance across all models in this single speaker setting. We notice that the Bi-LSTM suffers due to the limited training utterances, which suggests that the model cannot learn an appropriate mapping with this amount of data. GMM model fares better because it has the least amount of parameters among all the competing methods. It is capable of

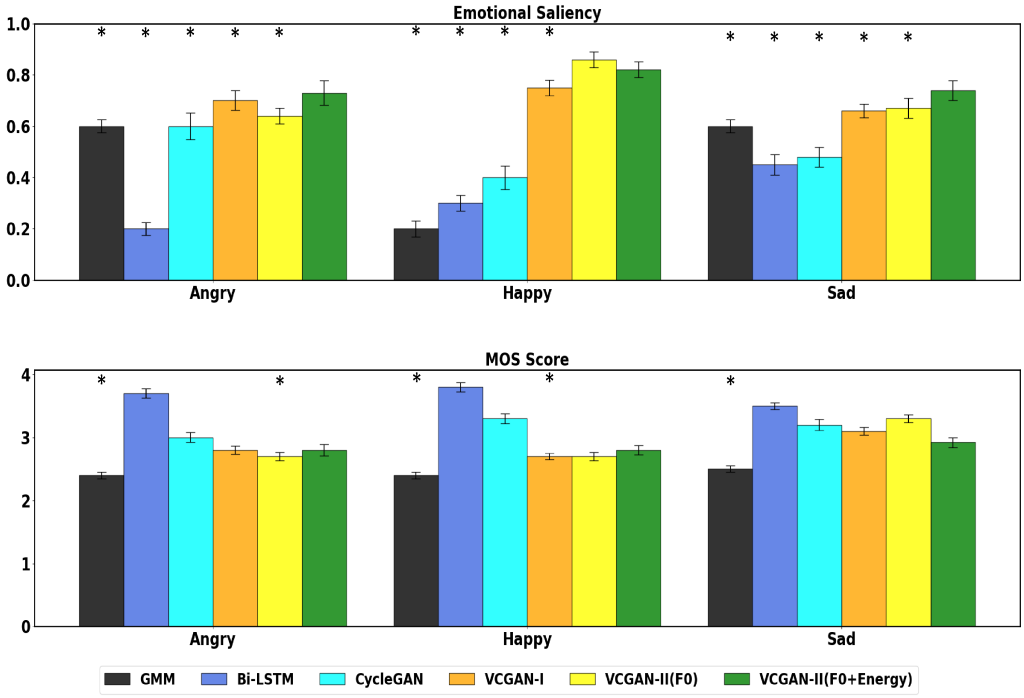

Figure 5.10: Out-of-Speaker Evaluation: we create 5 folds from VESUS, each comprising of a male and a female speaker. We train the VCGAN model on four folds and evaluate its performance on the fifth. The asterisk (*) denotes statistical significance for the test (VCGAN-II (F0+Energy) $>$ Method) at $p < 0.05$.

learning some aspects of the transfer function in a data-starved scenario. The Cycle-GAN achieves comparable performance to VCGAN-II (F0+Energy) on emotional saliency and outperforms our method on the MOS score. This behavior is unsurprising, as the Cycle-GAN architecture was designed for and evaluated on single speaker conversion tasks. VCGAN-II(F0+Energy) achieves the most robust performance across the three VCGAN models. We posit that this may be due to its reduced parameterization and focus on both F0 and energy.

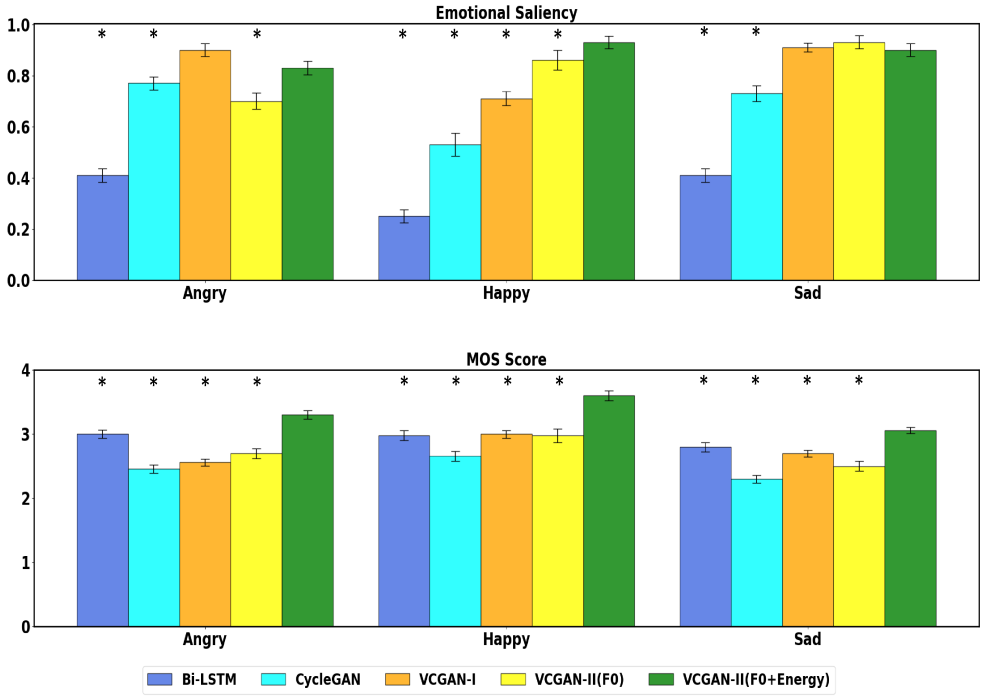

Figure 5.11: Wavenet Evaluation: We apply our mixed speaker models (without fine-tuning) to modify speech generated by the Wavenet model. The asterisk (*) denotes statistical significance for the one-tailed t-test (VCGAN (F0+Energy) $>$ Method) at $p < 0.05$.

5.4.3 Mixed Speaker Evaluation

To evaluate the performance of our model in a mixed speaker setting, we split the VESUS corpus as follows:

- **Neutral to Angry Conversion**: 1534 utterances for training, 72 for validation and, 61 for testing.
- **Neutral to Happy Conversion**: 790 utterances for training, 43 for validation and, 43 for testing.

- **Neutral to Sad Conversion**: 1449 utterances for training, 75 for validation and, 63 for testing.

We use the training and validation set to learn the parameters of the models and then evaluate using the test set. Empirically, we observe that the GMM does not produce intelligible speech in this setting due to the wide variation of speakers. Therefore, we have removed from the analysis. Fig. 5.9 shows the results of the crowd-sourcing experiment on test dataset. The mixed speaker setting is more challenging than the single speaker case because of variability across speakers in terms of estimating the dynamic range of the prosodic features. There is a high chance of learning an average mapping by the model.

We note that the Bi-LSTM has the worst emotion conversion accuracy (below 50%) for all three pairs of emotions. However, it also generates reasonably good audio quality. Empirically, we observe that the Bi-LSTM learns an near-identity mapping, meaning it does not perform any emotion conversion, but simply reconstructs the (already high quality) input utterance. The CycleGAN [6] model fairs reasonably well in terms of emotional saliency; however, the speech reconstruction quality is significantly lower than all three of our proposed models. VCGAN-II (F0 and F0+energy), in comparison, shows a uniformly consistent performance across the three emotion classes and does an extremely good job of retaining the speech naturalness post-conversion. We attribute this all-round performance to the momenta based regularization and to the variational formulation. The VCGAN-I model comes close to our proposed F0+energy framework for angry and sad emotions, but its conversion accuracy falls below 50% for the happy emotion, making it the

Table 5.1: Data splits used for the Out-of-Speaker Evaluation

Emotion Pair	Fold	Train	Validation	Test
Neutral-Angry	1	1347	320	123
	2	1212	455	125
	3	1610	57	122
	4	1310	357	125
	5	1189	478	107
Neutral-Happy	1	710	166	285
	2	581	295	289
	3	779	97	278
	4	833	43	284
	5	601	275	220
Neutral-Sad	1	1357	230	169
	2	1340	247	174
	3	1154	433	172
	4	1329	258	173
	5	1167	420	153

least consistent.

5.4.4 Out-of-Speaker Evaluation

We now tackle the more challenging task of out-of-speaker generalization. Here, we create five folds from VESUS, each one consisting of a single male and a single female speaker. We then train five separate models for each neutral $\Longleftrightarrow$ emotional pair corresponding using four of these folds and then test on the fifth remaining fold. Note that, this task tests the model's ability to generalize and learn transformation for speakers which are not part of the training set. Since each speaker has a different number of consistently rated emotional utterances, the data splits are fold-dependent, as shown in Table 5.1.

We sample 10 utterances for each fold and each emotion pair to collect the final ratings. Fig. 5.10 shows the average performance across the folds for all

methods. Once again, we evaluate two variants of our proposed framework: VCGAN-II(F0) and VCGAN-II(F0+Energy). Once again, the GMM fails to produce intelligible speech for the out-of-speaker experiment. Therefore, we have trained it for each speaker individually rather than fold-wise. Ultimately, the GMM model is not suitable for a real-world application, where the speakers may be unknown or vary between training and deployment.

At a first glance, we can see that the unsupervised models (GANs) generally outperforms the supervised method (Bi-LSTM). In fact, the Bi-LSTM model has the worst emotion conversion accuracy (below 50%) for all three pairs of emotions. However, it also generates the best audio quality among all the competing models, likely due to the minimal conversion. This result suggests a trade-off, which requires balancing the “strength" of the emotion conversion but not distorting the spectrum and F0 contour by too much modification.

Among the unsupervised models, Cycle-GAN has uniformly poor conversion accuracy across all emotion pairs. It fails to generalize to unseen speakers due to its weak generator-discriminator coupling and the variation between training and testing speakers. The generated speech quality is however higher, which is consistent with the behavior of the Bi-LSTM model. VCGAN models (VCGAN-I and VCGAN-II) outperform the remaining models in the fold-wise evaluation in emotion conversion task. They also achieve a good trade-off between emotion conversion and maintaining the naturalness of speech. The VCGAN-II(F0+Energy) model has the best balance among the three and is consistent on 2 out of 3 emotion conversion tasks. The difference between

Table 5.2: Performance across the four evaluation paradigms: Single-speaker, Mixed-speaker, Out-of-speaker, and Wavenet for Neutral to Angry conversion.

Evaluation	Algorithm	Neutral-angry	
		Acc.	MOS
Single Speaker	GMM [2]	0.6±0.1	2.5±0.6
	Bi-LSTM [3]	0.2±0.2	1.4±0.3
	Cycle-GAN [6]	0.84±0.1	2.9±0.6
	VCGAN-I [8]	**0.86±0.1**	2.4±0.4
	VCGAN-II(F0)	0.44±0.1	**3.0±0.5**
	VCGAN-II(F0+Energy)	0.8±0.2	2.9±0.6
Mixed Speaker	Bi-LSTM [3]	0.25±0.1	2.6±0.3
	Cycle-GAN [6]	0.76±0.3	2.1±0.7
	VCGAN-I [8]	**0.85±0.2**	**2.7±0.5**
	VCGAN-II(F0)	0.7±0.2	2.6±0.4
	VCGAN-II(F0+Energy)	0.84±0.2	**2.7±0.5**
Out-of-Speaker	GMM [2]	0.6±0.2	2.4±0.4
	Bi-LSTM [3]	0.2±0.2	**3.7±0.6**
	Cycle-GAN [6]	0.6±0.4	3.0±0.6
	VCGAN-I [8]	0.7±0.3	2.8±0.5
	VCGAN-II(F0)	0.64±0.2	2.7±0.5
	VCGAN-II(F0+Energy)	**0.73±0.3**	2.8±0.7
Wavenet	Bi-LSTM [3]	0.41±0.2	3.0±0.5
	Cycle-GAN [6]	0.77±0.2	2.46±0.5
	VCGAN-I [8]	**0.9±0.2**	2.56±0.4
	VCGAN-II(F0)	0.7±0.24	2.7±0.6
	VCGAN-II(F0+Energy)	0.83±0.2	**3.3±0.5**

VCGAN-II(F0) and VCGAN-II(F0+Energy) demonstrates how variation in energy plays an important role in the perception of emotion. Angry and sad emotions seem to be affected the most by this variation. Angry emotion is often characterized by a significant rise in the loudness, whereas sad emotion is exactly the opposite. Our VCGAN-II(F0+Energy) model is able to capture and encapsulate this information to some extent.

Table 5.3: Performance across the four evaluation paradigms: Single-speaker, Mixed-speaker, Out-of-speaker, and Wavenet for Neutral to Happy conversion.

Evaluation	Algorithm	Neutral-happy	
		Acc.	MOS
Single Speaker	GMM [2]	0.2±0.1	2.3±0.2
	Bi-LSTM [3]	0.1±0.1	2.3±0.7
	Cycle-GAN [6]	0.6±0.2	3.1±0.5
	VCGAN-I [8]	0.6±0.2	2.8±0.4
	VCGAN-II(F0)	0.58±0.2	**3.0±0.4**
	VCGAN-II(F0+Energy)	**0.68±0.2**	2.8±0.4
Mixed Speaker	Bi-LSTM [3]	0.25±0.1	2.5±0.3
	Cycle-GAN [6]	0.7±0.3	2.3±0.5
	VCGAN-I [8]	0.5±0.2	**3.4±0.5**
	VCGAN-II(F0)	0.6±0.3	3.0±0.4
	VCGAN-II(F0+Energy)	**0.82±0.2**	2.8±0.6
Out-of-Speaker	GMM [2]	0.2±0.2	2.4±0.4
	Bi-LSTM [3]	0.3±0.2	**3.8±0.6**
	Cycle-GAN [6]	0.4±0.3	3.3±0.6
	VCGAN-I [8]	0.75±0.2	2.7±0.4
	VCGAN-II(F0)	**0.86±0.2**	2.7±0.5
	VCGAN-II(F0+Energy)	0.82±0.2	2.8±0.6
Wavenet	Bi-LSTM [3]	0.25±0.2	2.98±0.5
	Cycle-GAN [6]	0.53±0.3	2.66±0.5
	VCGAN-I [8]	0.71±0.2	3.0±0.4
	VCGAN-II(F0)	0.86±0.3	2.98±0.7
	VCGAN-II(F0+Energy)	**0.93±0.2**	**3.6±0.5**

5.4.5 Wavenet Evaluation

Our final evaluation is on synthetic speech. In this case, we use the models trained in the mixed speaker evaluation (Section IV-C) without any fine tuning. This paradigm is more challenging because the test speaker characteristics are completely different from the training set. We generate “neutral" utterances using the Wavenet API provided by Google [43]. The utterances are based on randomly sampled phrases from the VESUS dataset to preserve syntactic

Table 5.4: Performance across the four evaluation paradigms: Single-speaker, Mixed-speaker, Out-of-speaker, and Wavenet for Neutral to Sad conversion.

Evaluation	Algorithm	Neutral-sad	
		Acc.	MOS
Single Speaker	GMM [2]	0.53±0.2	2.5±0.4
	Bi-LSTM [3]	0.2±0.2	1.9±0.3
	Cycle-GAN [6]	0.58±0.2	2.8±0.5
	VCGAN-I [8]	0.5±0.2	2.8±0.4
	VCGAN-II(F0)	0.64±0.2	**2.9±0.6**
	VCGAN-II(F0+Energy)	**0.66±0.2**	2.6±0.5
Mixed Speaker	Bi-LSTM [3]	0.57±0.1	2.4±0.2
	Cycle-GAN [6]	0.67±0.3	2.2±0.6
	VCGAN-I [8]	0.8±0.2	2.9±0.5
	VCGAN-II(F0)	0.7±0.3	**3.2±0.4**
	VCGAN-II(F0+Energy)	**0.74±0.3**	3.0±0.6
Out-of-Speaker	GMM [2]	0.6±0.2	2.5±0.4
	Bi-LSTM [3]	0.45±0.3	**3.5±0.4**
	Cycle-GAN [6]	0.48±0.3	3.2±0.7
	VCGAN-I [8]	0.66±0.2	3.1±0.5
	VCGAN-II(F0)	0.67±0.3	3.3±0.5
	VCGAN-II(F0+Energy)	**0.74±0.3**	2.9±0.6
Wavenet	Bi-LSTM [3]	0.41±0.2	2.8±0.6
	Cycle-GAN [6]	0.73±0.2	2.3±0.5
	VCGAN-I [8]	0.9±0.1	2.7±0.4
	VCGAN-II(F0)	**0.93±0.2**	2.5±0.6
	VCGAN-II(F0+Energy)	0.9±0.2	**3.1±0.4**

similarity between training and testing. The number of testing utterances is the same as in Section IV-C: 61 for neutral $\rightarrow$ angry, 43 for neutral $\rightarrow$ happy, and 63 for neutral $\rightarrow$ sad. Since, the Wavenet model generates audio in time domain directly, we use the WORLD vocoder to extract acoustic and prosodic features.

Fig. 5.11 shows that our VCGAN-II models are extremely good at converting the emotions in synthesized speech. While VCGAN-I matches the

proposed model in terms of emotional saliency, the quality of generated audio trends significantly lower in comparison. This is likely due to the secondary spectrum modification, which is not harmonically matched with the modified F0 contours. This experiment further demonstrates our VCGAN-II framework is robust even when the unseen speaker has completely different characteristics than the dataset on which the model has been trained. This is a first model in our knowledge that generalizes so well to a synthetic speaker (simulated by Wavenet).

5.4.6 Summary of Results

Table 5.2, Table 5.3 and Table 5.4 summarize the crowd sourcing results across the different evaluation paradigms, i.e., single speaker, mixed speaker, out-of-speaker, and Wavenet. Right away, we observe an inconsistency in performance as we progress from one experiment setting to another. This variation is expected, due to the increasing levels of difficulty of each evaluation. Specifically, our single speaker evaluation queries the performance of each model on utterances from the same speaker. In the mixed-speaker case, we train and test the models on the same collection of speakers, but randomly split the utterances between the two sets. This evaluation is more challenging because the models must learn characteristics of multiple speakers. In the out-of-speaker evaluation, we train the models on a set of four male/female speaker pairs and test on the remaining pair. Thus, the models never see utterances from the test speakers during training, which is a more difficult task. Finally, the Wavenet evaluation queries how well the models generalize to

synthetic speech, which by default is produced under different environmental (and physiological) conditions.

The asterisks (*) in Figs. 5.8-5.11 denote significantly improved performance between the VCGAN-II (F0+Energy) and the alternate methods. This analysis was conducted via a one-sided t-test for each emotion pair at significance level $p < 0.05$. We observe that while the three VCGAN models perform similarly, VCGAN-II tends to have more robust performance across evaluation settings. The traditional Cycle-GAN does well on the single-speaker evaluation, likely because this architecture was developed for voice conversion and can capitalize on individual speaker characteristics. However, it achieves significantly lower emotional saliency as the evaluation becomes more difficult (i.e., multi-speaker, out-of-speaker, Wavenet). The GMM has variable emotion conversion performance in the single-speaker setting, but fails to generate intelligible speech in the multi-speaker paradigms, and performs poorly in the other two evaluations. Finally, the Bi-LSTM achieve low emotional saliency but consistently high MOS score. This is due to the fact that it collapses into an identity transformation and fails to modify the utterance at all. From Table 5.2, 5.3 and 5.4, we conclude that our VCGAN-II models achieve the best trade off between emotional saliency and speech reconstruction quality. Thus, combining F0 contour and spectrum modification (via energy) into a single unified framework can achieve much better performance on emotion conversion and reconstruction quality assessment tasks than modeling them separately.

By using diffeomorphic registration for the F0 and energy contour, our

novel framework offers some key advantages over the standard wavelet parameterization. Furthermore, our momenta-based approach does not require any speaker/cohort specific normalization to match the range of loudness and fundamental frequency. The deformation process takes care of the individual ranges, thereby, allowing the VCGAN to automatically adapt to the test speaker. Additionally, the KL divergence penalty between the target data density and the generator estimated density constrains the model to behave in a predictable manner. The conditional independence of the target spectrum and target F0 contour (Fig. 5.1) is another notable aspect of our approach; empirically, it helps preserve the naturalness of the modified speech.

5.5 Conclusion

In this chapter, we have proposed a novel method for robust emotion conversion. Our technique uses a modified version of Cycle-GAN called variational Cycle-GAN (VCGAN). VCGAN was derived as an upper bound on the KL-divergence penalty between the target data distribution and the generator estimated distribution. We showed that this led to a new joint density discriminator which constrained the forward-backward generators at the distribution level. Empirically, we demonstrated that this distributional matching was better at learning the target densities for emotion conversion. In addition, we modeled the features in the target utterance as a smooth warped version of the source. This allowed the algorithm to adaptively adjust the F0 and loudness range of a test speaker without any feature normalization. We showed that our approach led to a consistent performance across four emotion conversion

tasks. Further, our framework achieved a good balance between the emotion conversion accuracy and the naturalness of synthesized speech, as demonstrated by real-world crowd sourcing experiments. We also compared our proposed framework against state-of-the-art emotion conversion baselines from supervised and unsupervised learning domain. Our method universally outperformed these techniques.

References

[1] Klaus R Scherer. "Vocal communication of emotion: A review of research paradigms". In: *Speech Communication* 40.1 (2003), pp. 227 –256. ISSN: 0167-6393. DOI: https://doi.org/10.1016/S0167-6393(02)00084-5.

[2] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". In: *American Journal of Signal Processing* 2 (2012), pp. 134–138. DOI: 10.5923/j.ajsp.20120205.06.

[3] Huaiping Ming, Dong-Yan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion". In: *Proc. Interspeech 2016*. 2016, pp. 2453–2457. DOI: 10.21437/Interspeech.2016-1053.

[4] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective". In: *Proc. Interspeech 2019*. 2019, pp. 2848–2852. DOI: 10.21437/Interspeech.2019-2512.

[5] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Automated Emotion Morphing in Speech Based on Diffeomorphic Curve Registration and Highway Networks". In: *Proc. Interspeech 2019*. 2019, pp. 4499–4503. DOI: 10.21437/Interspeech.2019-2386.

[6] Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye. "Nonparallel Emotional Speech Conversion". In: *Proc. Interspeech 2019*. 2019, pp. 2858–2862. DOI: 10.21437/Interspeech.2019-2878.

[7] Kun Zhou, Berrak Sisman, and Haizhou Li. "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data". In: *CoRR* abs/2002.00198 (2020). arXiv: 2002.00198.

[8] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "Non-Parallel Emotion Conversion Using a Deep-Generative Hybrid Network and an Adversarial Pair Discriminator". In: *Proc. Interspeech 2020*. 2020, pp. 3396–3400. DOI: 10.21437/Interspeech.2020-1325.

[9] Masanori Morise, Fumiya YOKOMORI, and Kenji Ozawa. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". In: *IEICE Transactions on Information and Systems* E99.D (2016), pp. 1877–1884. DOI: 10.1587/transinf.2015EDP7457.

[10] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. *A Diffeomorphic Flow-based Variational Framework for Multi-speaker Emotion Conversion*. 2022. arXiv: 2211.05071 [eess.AS].

[11] A. Sotiras, C. Davatzikos, and N. Paragios. "Deformable Medical Image Registration: A Survey". In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1153–1190.

[12] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. "Computing large deformation metric mappings via geodesic flows of diffeomorphisms". In: *International journal of computer vision* 61.139-157 (2005).

[13] Sarang C Joshi and Michael I Miller. "Landmark matching via large deformation diffeomorphisms". In: *IEEE transactions on image processing* 9.8 (2000), pp. 1357–1370.

[14] Hsi-Wei Hsieh and Nicolas Charon. "Diffeomorphic registration of discrete geometric distributions". In: *CoRR* abs/1801.09778 (2018). arXiv: 1801.09778.

[15] Takuhiro Kaneko and Hirokazu Kameoka. "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1711.11293 (2017). arXiv: 1711.11293.

[16] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2015. arXiv: 1511.06434 [cs.LG].

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proc. ICCV 2017*. IEEE Computer Society, 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved Techniques for Training GANs". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, 2234–2242. ISBN: 9781510838819.

[19] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles A. Sutton. "VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning". In: *Proc. NIPS, 2017*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 3308–3318.

[20] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. "Adversarially Learned Inference". In: *CoRR* abs/1606.00704 (2016). arXiv: 1606.00704.

[21] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2015).

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[23] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proc. Interspeech 2019*. 2019, pp. 316–320. DOI: 10.21437/Interspeech.2019-1413.

[24] Amazon. "Amazon Mechanical Turk". In: (). URL: https://www.mturk.com.

[25] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M. Dai, Shakir Mohamed, and Ian Goodfellow. *Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step*. 2018. arXiv: 1710.08446 [stat.ML].

[26] Ricard Durall, Avraam Chatzimichailidis, Peter Labus, and Janis Keuper. "Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues". In: *CoRR* abs/2012.09673 (2020). arXiv: 2012.09673. URL: https://arxiv.org/abs/2012.09673.

[27] L. V. D. Maaten and Geoffrey E. Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

Chapter 6

Supervised Open-Loop Framework for Duration

In the previous chapters, we saw how pitch and energy modifications can be used to inject emotional cues into the neutral speech or to change the overall speaking style [1, 2, 3, 4, 5]. The speaking rate modulation also known as rhythm variation, plays a crucial role in conveying emotions [6] and in diagnosing human speech pathologies [7]. While there are many approaches for automated pitch and energy modification [8, 9, 10, 11, 12], comparatively little progress has been made in changing the rhythm of a speech utterance. In fact, rhythm is difficult to manipulate because, unlike pitch or energy, there is no explicit coding for the relative duration of phonemes across the utterance. Rather, this information is implicitly defined and varies dramatically across speakers and utterances. As a result, rhythm modification methods either require considerable user supervision or they are geared towards aligning to known speech signals. Even prior work on quantifying the transitory behavior of rhythm [13] is limited and requires *a priori* alignment of the audio files.

Perhaps the earliest duration modification method is the time-domain

pitch synchronous overlap and add (TD-PSOLA) algorithm [14]. TD-PSOLA modifies the pitch and duration of a speech signal by replicating and interpolating between individual frames centered at the peaks of auto-correlation signal. However, the user must manually specify both the portion of speech to modify and the exact manner in which it should be altered. Methods such as [15, 16] take a more user-friendly and performative approach to modify the pitch and rhythm, but they still require manual input to guide the process. An alternate approach to changing rhythm is a frame-wise alignment between a source utterance and a given target. Here, the most common approach is Dynamic Time Warping (DTW) [17]. It is a dynamic programming approach to align two sequences of different lengths. DTW requires both, the source and target speech which renders it unusable for generative modeling.

Finally, recent advancements in deep learning have led to a new generation of neural vocoders that disentangle the semantic content from the speaking style [18, 19, 20]. These vocoders can alter the speaking rate via the learned style embeddings. While these models represent seminal contributions to speech synthesis, the latent representations are learned in an unsupervised manner, which makes it difficult for the user to control the output speaking style. Another drawback is that these methods require large amounts of data and computational resources for adequate model training and speech generation [21, 22].

In this chapter, we introduce an automated and adaptive speech duration modification scheme. Our approach combines the structured simplicity of dynamic decoding with the representation capabilities of deep neural networks.

Namely, we model the alignment between a source and target utterance via a latent attention map; these maps are used as replacement of the similarity matrix for backtracking. We train a masked convolutional encoder-decoder network to estimate these attention maps using a stochastic mean absolute error (MAE) formulation. Unlike the conventional DTW [17] algorithm, once trained our framework operates in an open-loop fashion on the source utterance without needing access to the target. We demonstrate our framework on a voice conversion task using the CMU-Arctic dataset [23] and on three multi-speaker emotion conversion tasks using the VESUS dataset [24]. Our experiments confirm that the proposed model can perform adaptive duration modification with limited training data and minimal distortion.

6.1 Method

Our technique uses an attention based encoder-decoder framework to process an input sequence and produce another sequence as output. Specifically, the input sequences used in our model are the Mel-frequency representation of a speech signal. We further inject domain knowledge or prior into the neural network model by restricting the scope of the attention map between the encoded and decoded representations and strategically leverage DTW to generate intelligible speech. We provide a brief description of the training/testing strategy followed by a discussion of the baseline methods at the end of this section.

Fig. 6.1 illustrates our underlying generative process. Given an utterance X, we first estimate the length T of the (unknown) target utterance Y and

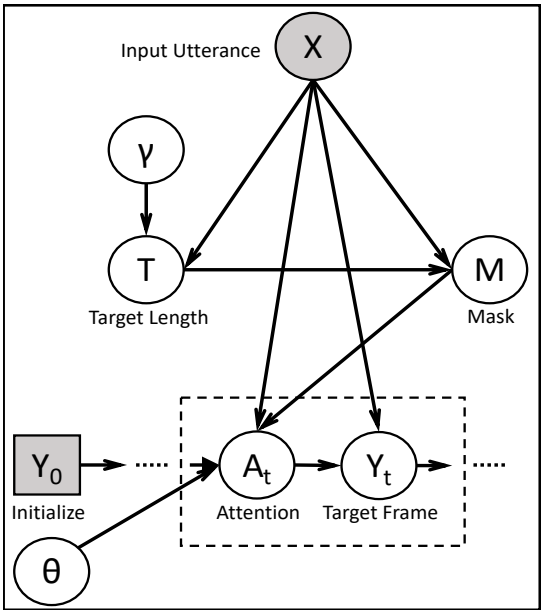

Figure 6.1: Graphical model for rhythm modification. γ and θ are the model parameters inferred during training. Attention A_t is conditionally independent of target length T given X and M

subsequently use it to estimate a mask M for the attention map. The mask restricts the domain of the attention vectors A_t at each frame t during the inference stage to mitigate distortion. We use paired data (X_{tr}, Y_{tr}) to train a convolutional encoder-decoder network to generate the attention vectors. During testing, we first generate the attention map from the input X and use it to produce the target speech Y.

6.1.1 Loss Function

Formally, let $X \in \mathbb{R}^{D \times T_s}$ denote the frame-wise Mel filter-bank energies extracted from the input speech. Here, D is the number of filter banks, and T_s is the number of temporal frames in the utterance. Similarly, we denote the target speech as $Y \in \mathbb{R}^{D \times T}$, where the target length T is usually different from

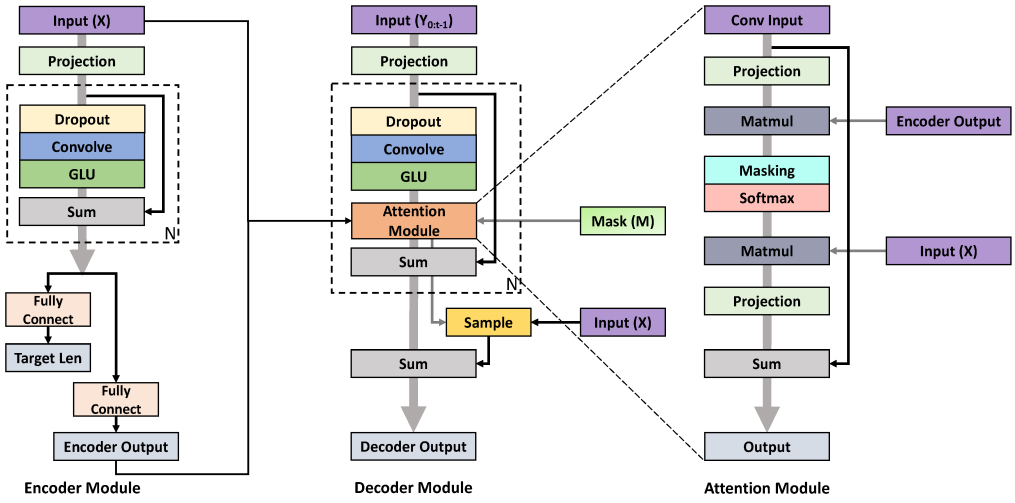

Figure 6.2: Model architecture used for the sequence-to-sequence speech generation. The encoder and decoder modules consist of 10 identical blocks. Projection layers are simple feed-forward layers without any non-linearity to project input features in high dimension.

T_s.

Our generative process for the target speech is as follows:

$$T \sim \text{Laplace}(T^0, b_T) \quad \textit{and} \quad Y_t \sim \text{Laplace}(Y_t^0, b_y) \;\forall t, \tag{6.1}$$

where Y_t is the target Mel filter-bank energy features at time t. We use Laplace distributions to leverage sparse nature of filter-bank energies. The parameters $\{T^0, b_T, Y_t^0, b_y\}$ of the distributions are unknown and implicitly estimated via a deep neural network, which is parameterized by γ and θ (see Fig. 6.1).

By treating the unknown parameters as functions of the input X, we obtain the following estimating equations for the target sequence length and frame-wise Mel filter-bank energies:

$$\hat{T} = f_\gamma(X) \quad \textit{and} \quad \hat{Y}_t = X \cdot A_t + f_\theta(X, \hat{Y}_{0:t-1}). \tag{6.2}$$

The functions $f_\gamma(\cdot)$ and $f_\theta(\cdot,\cdot)$ correspond to the length prediction and energy estimation component of the neural network. The variable $A_t \in \mathbb{R}^{T_s}$ is an attention vector that combines frame-wise features of the source utterance X to generate the target frame $\hat{Y}_t$. Our model differs from standard sequence-to-sequence model by treating the neural network predictions as residuals added to input sequence itself, where these residuals depend on input and the history of predictions $\hat{Y}_{0:t-1}$. This autoregressive property allows the neural network to learn both segmental and supra-segmental variations that can potentially distinguish between different speakers or emotions.

During training, we use paired data (X, Y) and maximize the likelihood of the target speech signal with respect to the neural network weights $\{\theta, \gamma\}$. This likelihood can be written

$$P(\hat{Y}, \hat{T}|X) = P(\hat{T}|X) \prod_{t=1}^{\hat{T}} P(\hat{Y}_t|X, \hat{T}, \hat{Y}_{0:t-1}), \tag{6.3}$$

where, the second term in Eq. (6.3) can be obtained by introducing a deterministic attention mask M and marginalizing A_t:

$$\begin{aligned} P(\hat{Y}_t|X, \hat{T}, \hat{Y}_{0:t-1}) &= \sum_{A_t} P(\hat{Y}_t, A_t|X, \hat{T}, \hat{Y}_{0:t-1}, M) \\ &= \sum_{A_t} P(\hat{Y}_t|X, \hat{T}, A_t, \hat{Y}_{0:t-1}) P(A_t|X, \hat{Y}_{0:t-1}, M) \end{aligned} \tag{6.4}$$

The variable M here denotes the attention mask. We introduce M for mathematical convenience, as it is a deterministic function of the source length T_s and the estimated length $\hat{T}$. We encode the attention A_t as a one-hot vector across the T_s frames of the source speech. Thus, it follows a categorical distribution. For simplicity, we model A_t as conditionally independent of the target

length $\hat{T}$ given the mask M and the input X. Taking the $\log(\cdot)$ of likelihood term and combining with Eq. (6.4) yields:

$$\begin{aligned}
\mathcal{L}(\theta, \gamma) &= -\log\big(\sum_{A_t} P(\hat{Y}_t, A_t | X, \hat{T}, \hat{Y}_{0:t-1}, M)\big) - \log\big(P(\hat{T}|X)\big) \\
&= -\log\Big(\sum_{A_t} \frac{q_\theta(A_t|X, \hat{Y}_{0:t-1}, M)}{q_\theta(A_t|X, \hat{Y}_{0:t-1}, M)} P(\hat{Y}_t, A_t | X, \hat{T}, \hat{Y}_{0:t-1}, M)\Big) \\
&\qquad\qquad - \log\big(P(\hat{T}|X)\big) \\
&\leq -\sum_{A_t} q_\theta(A_t|X, \hat{Y}_{0:t-1}, M) \log\big(P(\hat{Y}_t|X, A_t, \hat{Y}_{0:t-1})\big) \\
&\qquad\qquad - \log\big(P(\hat{T}|X)\big) + KL(q_\theta(A_t)||P(A_t)) \\
&= -\sum_{A_t} q_\theta(A_t|X, \hat{Y}_{0:t-1}, M) \log\big(P(\hat{Y}_t|X, A_t, \hat{Y}_{0:t-1})\big) \\
&\qquad\qquad - \log\big(P(\hat{T}|X) - H(q_\theta) + const. \\
&\leq -\sum_{A_t} q_\theta(A_t|X, \hat{Y}_{0:t-1}, M) \log\big(P(\hat{Y}_t|X, A_t, \hat{Y}_{0:t-1})\big) \\
&\qquad\qquad - \log\big(P(\hat{T}|X) + const. \qquad (6.5)
\end{aligned}$$

The distribution $q_\theta(\cdot)$ above is an approximating distribution for the attention vectors implemented by a convolutional network. The first inequality uses the convexity of the $-\log$ function, and the second inequality comes from the fact that entropy $H(q_\theta) \geq 0$. Notice that we have implicitly assumed $P(A_t|X, \hat{Y}_{0:t-1}, M)$ has a uniform distribution over the masked region as a non-informative prior. This is a reasonable assumption given that the masking process reduces the attention domain to a small region. However, q_θ is **not**

penalized for deviating from this uniform distribution prior during training. This flexibility allows the network to learn realistic attention vectors during autoregressive decoding. Eq. (6.5) can be easily translated into a neural network loss function which we minimize for $\{\theta, \gamma\}$:

$$\begin{aligned}\mathcal{L}(\theta, \gamma) &= \lambda_1 \times E_{A_t \sim q_\theta}\big[\log\big(P(\hat{Y}_t|X, A_t, \hat{Y}_{0:t-1})\big)\big] + \lambda_2 \times \log\big(P(\hat{T}|X)\big) \\ &= \lambda_1 \times E_{A_t}\big[\|\hat{Y}_t - Y_t^0\|_1\big] \; + \; \lambda_2 \times \|\hat{T} - T^0\|_1 \end{aligned} \tag{6.6}$$

where λ_1 and λ_2 are the model hyperparameters that adjusts the trade-off between the two objectives and contains the variances of the Laplace distributions. Notice that the loss in Eq. (6.6) computes an expectation over the attention maps. We use the Monte-Carlo estimate by sampling from the attention map at each time-step. The training procedure is therefore stochastic in nature due to this random sampling from the attention map.

6.1.2 Convolutional sequence-to-sequence model

We use a masked convolutional sequence-to-sequence model to learn the duration transformation from one domain to another. Fig. 6.2 shows the interplay between the encoder, decoder and modified attention modules of our deep neural network. The architecture is adapted from [25] by adding residual connections to the final layer and reconfiguring the attention module. The encoder in Fig. 6.2 is a stack of gated convolutions which performs two tasks: (i) approximating the length of the target sequence and (ii) learning appropriate representation for the decoding process. We insert an attention module between the encoder and decoder layers to leverage locality constraint during

Figure 6.3: Binary attention masks with 3 different slopes.

the generation of the output sequence. Notice that, a general attention map uses the entire input sequence to decode a single frame of the output sequence. Therefore, we apply a masking strategy inspired by Itakura parallelogram of DTW framework which acts as a prior knowledge over the difference in speaking rates between the source and target domains. Masked convolutions were initially proposed by [25] for language modeling. This architecture allows the network to exploit local continuity of speech and can be trained faster than a conventional RNN or LSTM while also requiring fewer learnable parameters. These advantages are important in cases of limited training data.

6.1.3 Masking

We use the mask M to constrain the scope of the attention mechanism to be similar in time-scale to the input. This procedure is important for two reasons. From a speech quality perspective, large local swings in speaking rate may generate unintelligible speech. From an estimation perspective, the speech utterances contains hundreds (sometimes thousands) of frames. It is difficult to robustly train a deep network to generate such long attention vectors based on smaller datasets.

These masks are derived from Itakura parallelogram [26], as illustrated in

Fig. 6.3 and is different from [27] due to hard cut-off in scope. The slope of the Itakura parallelogram specifies the minimum and maximum speaking rates that the reconstructed utterances are allowed to possess in comparison to the input speech. In this chapter, we fixed the minimum and maximum variation in speaking rate to 0.8 and 1.25, respectively, based on empirical observations of the training data.

6.1.4 DTW Back-Tracking

Our final step uses the learned attention map as a proxy for the DTW similarity matrix. This strategy allows us to train the model on a relatively small dataset (e.g., 2-3 hours) and still generate intelligible speech during open-loop modification of new utterances. Formally, we apply a dynamic programming operation to the attention maps produced by the neural network to get a path of alignment from source to target. To avoid skipping phonemes, we constrain the dynamic programming path to take at most one horizontal or vertical step at a time while backtracking. Once estimated, the path informs a reorganization of the source utterance frames via localized contraction and dilation operations. Following this reorganization, the target speech is synthesized via the WORLD vocoder [28].

We train our model using mini-batch gradient descent and the Adam optimizer [29] with a fixed learning rate of 10^{-4} and a batch size of 16. The input X are 80-dimensional Mel-filterbank energies spanning 0-8 kHz. The projection layer expands this input to 256 dimensions. Both the encoder and decoder consist of 10 convolutional layers, each followed by a gated linear

unit. Given the small dataset size, we use data augmentation to mitigate overfitting. Specifically, we reverse the input-output sequences and randomly extract intervals of variable size (with probability 0.5) from the full speech utterance.

Algorithm 2: Strategy for model training

function trainModelParameters (X, Y);
Input : filterbank energies ($X \in \mathbb{R}^{D \times T_s}$, $Y \in \mathbb{R}^{D \times T_t}$)
Output: model parameters (θ, γ)
if *epoch* $<$ *MaxEpochs* **then**
  for *minibatch* **do**
    Predict target length $\hat{T} = f_\gamma(X)$ and create the mask $M \in \mathbb{R}^{T_s \times T_t}$;
    Estimate $A \in \mathbb{R}^{T_s \times T_t}$ using masked convolution and sample $u \sim U(0, 1)$;
    if $u < 0.2$ **then**
      Sample $a \in \mathbb{R}^{T_s}$ from A_{T_s};
      Reconstruct using: $\hat{Y}_t = X \cdot a + f_\theta(X, Y_{0:t-1})$;
    else
      Reconstruct using: $\hat{Y}_t = X \cdot A_{T_s} + f_\theta(X, Y_{0:t-1})$;
    end
    Compute prediction errors and update parameters;
  end
  epoch $\leftarrow$ epoch + 1;
end
return trainedModel;

6.1.5 Training and Testing Strategy

During training, we optimize Eq. 6.6 based on the Mel filterbank energies Y and utterance durations T from paired input-output utterances. The forward pass through the network (Fig. 6.2) processes the input frames and generates

Algorithm 3: Strategy for model testing (i.e., open-loop duration modification)

function modifyDuration (X);
Input : filter-bank energy ($X \in \mathbb{R}^{D \times T_s}$ and Y_0)
Output: alignments $((x_1, y_1), (x_2, y_2), ...)$
Predict length of target sequence $\hat{T}_t = f_\gamma(X)$;
Create attention mask $M \in \mathbb{R}^{T_s \times \hat{T}_t}$ and Set $t = 0$;
if $t < \hat{T}_t$ **then**
    Using mask M_t, X, and $Y_{0:t-1}$ estimate A_t;
    Using X, $Y_{0:t-1}$, and A_t, predict Y_t;
    $t \leftarrow t + 1$;
end
Run DTW backtracking on the attention matrix A;
return (alignments $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$);

an embedding to predict the target sequence length $\hat{T}$. The embedding is also used to generate an attention vector as a categorical distribution at each decoder step inside the specified masked region. We use a stochastic sampling procedure for the attention vector, in which we randomly mix between a single sample from the distribution q_θ and the MAP estimate. Empirically, this strategy provides robustness to sub-optimal local minima (see Alg. 2).

During testing, we rely on the predicted length to generate the attention map and the target frames. We also use a MAP strategy, rather than the stochastic mixing procedure. Once generated, we use the attention map as a proxy for the DTW similarity matrix; using a Viterbi alignment procedure, we rearrange the input frames to produce the modified speech (Alg.3).

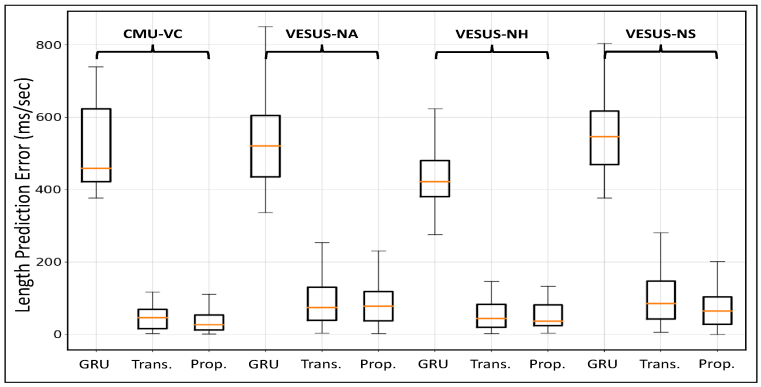

Figure 6.4: Length prediction errors ($\downarrow$) across different models.

6.1.6 Baseline Comparison Methods

We compare our convolutional encoder with two commonly used sequence-to-sequence frameworks: (i) Gated Recurrent Unit or GRU model [30], and (ii) Transformer model [31]. Due to space limitations, further details of the baseline architectures and training strategy have been omitted from the chapter.

6.2 Experimental Results

We evaluate our rhythm modification framework on two publicly available multi-speaker datasets: CMU-ARCTIC [23] for voice morphing and VESUS [24] for emotion conversion.

6.2.1 Data and Conversion Tasks

The CMU-ARCTIC database has 4 American English speakers (two male, two female), who we paired by gender for voice conversion. Of the resulting 2264 sentence pairs, we train our model and the baselines using 2164 utterances and reserve the remaining 100 utterances (random 50-50 split) for validation

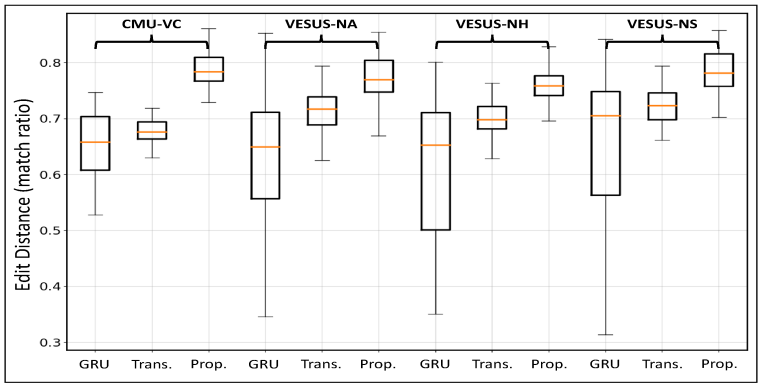

Figure 6.5: Alignment similarity (↑) between attention and DTW.

and testing of the open-loop modification properties.

VESUS is an emotional speech corpus containing utterances in 4 emotion classes: neutral, angry, happy, and sad. Each utterance contains 10 crowd-sourced ratings of emotional saliency. For robustness, we only use utterances that are correctly annotated by at least half of the listeners. We consider three neutral-emotional conversion tasks as follows:

- **Neutral to Angry**: 2385 utterances for training, 72 for validation and, 61 for testing.
- **Neutral to Happy**: 2431 utterances for training, 43 for validation and, 43 for testing.
- **Neutral to Sad**: 2371 utterances for training, 75 for validation and, 63 for testing.

Due to the smaller sample size, we pretrain the models on CMU-ARCTIC and fine-tune it for emotion conversion.

6.2.2 Length Prediction

As a sanity check, we compare the predicted utterance length by our framework with that of the ground truth parallel utterance. Fig. 6.4 shows the error in predicting the length ratio in a ms/sec format. Notice that, our framework mispredicts the utterance lengths by only 40ms/sec and 65ms/sec (on average) on CMU-ARCTIC and VESUS, respectively. Duration prediction is particularly challenging on VESUS due to marked differences between neutral and emotional utterances. The median prediction error for GRU model is in the range of $400 - 600$ms per second of the input utterance. The Transformer fares relatively well in comparison to GRU because of its ability to establish long-range dependency. However, our framework performs slightly better, perhaps due to the multi-task setup and the fusion of deep representation with Bayesian regularization.

6.2.3 Attention Alignment

Next, we compare the open-loop alignment estimated via the attention map with the supervised DTW algorithm where both utterances are known. To compare the warping paths, we code the horizontal, diagonal, and vertical moves of the backtracking procedure into three classes. We then compute the edit distance between the attention map and DTW-based alignment schemes. Fig. 6.5 illustrates the match ratio normalized by the average length of sequences. As seen, the match ratio varies between 0.70 and 0.85, which suggests that our convolutional model can readily learn the general characteristics of duration modification. The GRU model performs poorly in this task due

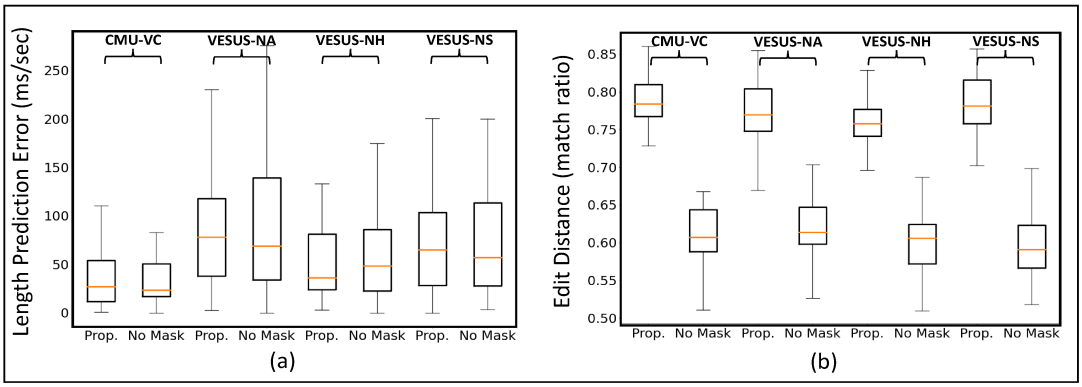

Figure 6.6: (a) Length prediction of target utterances (↓) and (b) measuring similarity of attention map (↑) to DTW cost matrix. Model is trained without mask constraint on attention map.

to its inability to learn sequence transformations across 100s of frames. The Transformer model does a little better than the GRU on this task, but still underperforms our method, likely due to the small training dataset. Our proposed model performs best because of the Itakura masking constraint and its reduced parameterization, which permits learning in small-data regimes. Thus, our method can be used as a tool for manipulation of speaking rate at both, local and global scale.

6.2.4 Ablation Analysis: Removing Itakura masking

There are multiple components in the proposed model which work in synchronised manner to produce naturally sounding speech. In addition to the generative modeling, the two most important augmentations we have made to the masked convolutional network pipeline are: (i) using Itakura masking for attention map and (ii) using an attention weighted residual connection in the final layer. Therefore, we perform ablation experiments to understand the relative significance of each of these augmentations. Our first experiment

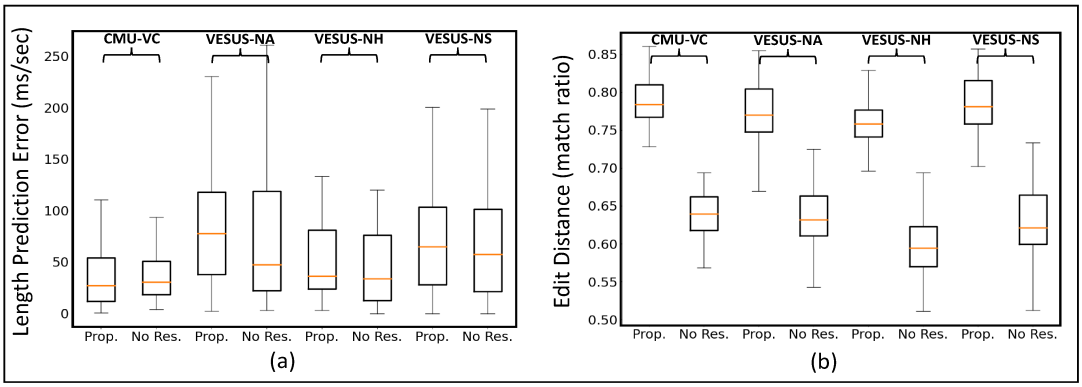

Figure 6.7: (a) Length prediction of target utterances (↓) and (b) measuring similarity of attention map (↑) to DTW cost matrix. Model is trained without residual connection in decoder layer.

removes masking from the attention layers. Fig. 6.6 shows the model's performance on target length prediction and approximating the DTW similarity matrix. The results in Fig. 6.6(a) indicate that the length prediction performance is roughly similar to the proposed model. This is expected because, the encoder part of the network is exactly same. The attention map is constrained only in the encoder-to-decoder transition. Hence, it estimates length with a relatively small error (in ms/sec). The match ratio metric however, (shown in Fig. 6.6(b)) is considerably worse. Itakura masking procedure acts as a good inductive bias/prior on the attention map because the speech rate do not fluctuate drastically in human conversations. Therefore, our localization scheme for the attention map is crucial to improve the edit distance in our model.

6.2.5 Ablation Analysis: Removing Residual connection

Our second ablation experiment involves removing the attention weighted residual connection from the final layer of decoder. Fig. 6.7(a) shows that the

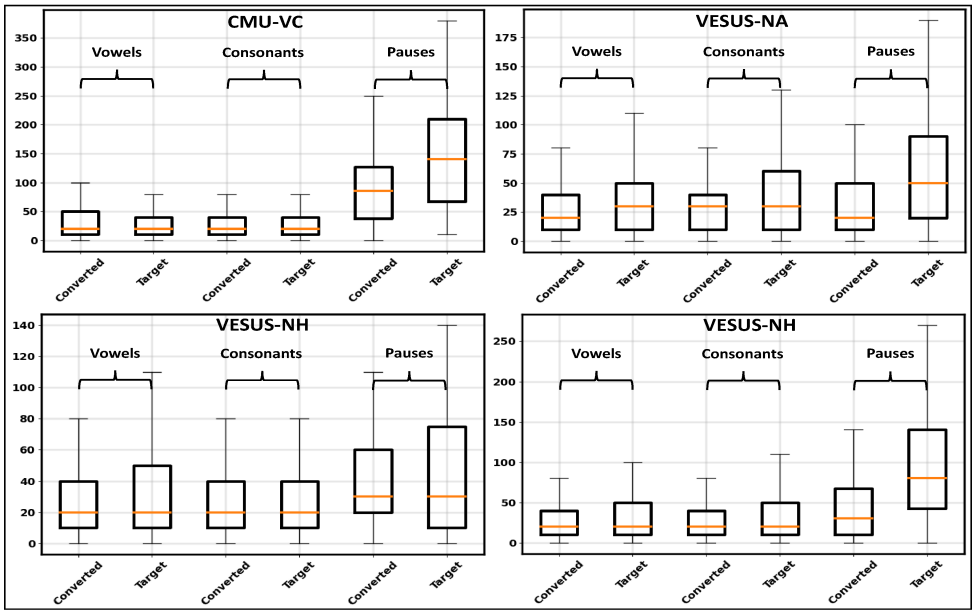

Figure 6.8: Duration differences between source/target and source/converted pairs for vowels, consonants, and pauses.

model is able to estimate the target sequence lengths with a relatively low error rate. We attribute this to the fact that the encoder portion is same as proposed model. The match ratio (Fig. 6.7(b)) in this experiment is slightly better than the no-masking results but, worse than the proposed model. Therefore, we can confidently say both Itakura masking scheme and residual connection helps in approximating DTW similarity matrix. Further, the presence of residual connection is extremely useful in providing a good gradient signal for the convolutional network to learn prediction of target frames. Since the linguistic content of input and target utterances are same, it further allows the neural network to inherit input speech properties which is helpful in auto-regressive generation mode.

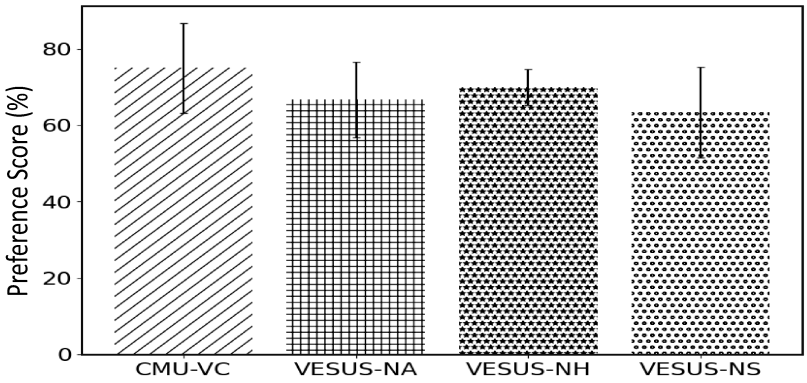

Figure 6.9: Preference score (in %) of proposed method (↑) relative to the input with ground-truth as reference (crowd-sourced).

6.2.6 Component-Wise Duration Analysis

Fig. 6.8 compares the differences in duration between the converted utterances and the ground truth targets for vowels, consonants and short pauses. We use the Penn Phonetic Forced Alignment tool [32] to get the text and speech alignment. As seen in Fig. 6.8, our method faithfully modifies the duration of vowels and consonants, but it is less effective with short pauses. This trend is intuitive, as our model relies on replication of the frames determined by the backtracking on similarity map. Therefore, it cannot create pauses if these frames do not exist in the source utterances. Nonetheless, our model consistently estimates the difference between vowels and consonants duration across multiple tasks, which corroborates our claim of developing a general purpose speech rate manipulation framework.

6.2.7 Rhythm Similarity Assessment

To evaluate the rhythm of modified speech, we design a crowd-sourcing based preference test scheme. In this experiment, the evaluators are asked to listen

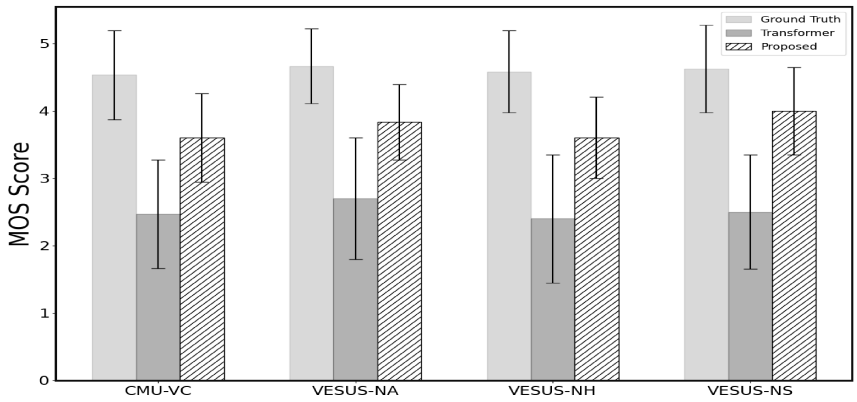

Figure 6.10: Crowd-sourced MOS (↑) of generated speech (hatched bars) vs the ground-truth samples from each task (shaded left) and baseline transformer model (shaded middle).

the ground-truth speech as a reference. It is then followed by a selection task between the unmodified (source) and the modified utterance, whichever has the highest perceptual similarity to the reference in terms of speaking rate modulation. The results of this experiment are demonstrated in Fig. 6.9. We can note that, the similarity scores are in the range of 60-80% which is relatively high, considering that source and ground-truth utterances have duration difference in the order of 100-200ms only. CMU-VC task has the highest similarity score mainly because of the long utterances that allow the listeners to discern the differences in an effective manner.

6.2.8 Speech Reconstruction Quality

Finally, we use crowd sourcing to obtain a mean opinion score (MOS) for the re-synthesized speech quality of the testing utterances. The crowd sourcing was performed using Amazon mechanical turk (AMT). We collect 5 listener ratings for each converted utterance in the test set, and we also add clean (ground-truth) along with some noisy/distorted utterances to the converted

samples set to get the baseline scores and flag non-invested listeners and bots on AMT. As seen in Fig. 6.10, our method achieves an average MOS between $3.7 - 4.0$ across the four tasks (rightmost bars). Further, the ground-truth baseline score of each task (leftmost bars) are in the range of $4.5 - 5$, whereas the MOS score of speech generated by transformer model (middle bars) are in the range of $2 - 3$. It shows the superiority of proposed model over transformer baseline. We note that CMU-ARCTIC task has the lowest MOS, possibly due to longer and more complex utterances. Interestingly, the MOS is unaffected by errors in length prediction, as evidenced by the VESUS neutral-angry emotion conversion task. Thus, our model provides a robust way to alter speech characteristics.

6.3 Conclusions

We have introduced a new framework for adaptive rhythm modification. Our model used an attention based convolutional encoder-decoder architecture to estimate attention maps which associate frames of the input speech with frames of the target speech. The attention maps are modeled as latent variables in a graphical framework, which lead to a stochastic formulation of the mean absolute error (MAE) loss for model training. During testing, the attention map is directly used as an approximation of the similarity matrix for a DTW-style backtracking procedure. We evaluated our framework on a voice conversion and three separate emotion conversion tasks using CMU-ARCTIC and VESUS corpora. Our evaluation metrics are the L1 distance for target length prediction, and an edit distance based matching ratio for path similarity.

Our proposed model outperformed existing seq-2-seq models designed solely on transformer and LSTM architectures in both metrics. Further, we ablate our proposed model's performance against simpler versions of it, i.e., no residual connection and no Itakura masking scheme. These ablations showed that removing either of these components leads to poor match ratio performance. Overall, our framework produced similar duration modification as the vanilla DTW, but *without requiring access to the target utterance*. Finally, we showed that the re-synthesized speech had similar naturalness to most state-of-the-art neural vocoders.

References

[1] James A. Russell, Jo-Anne Bachorowski, and José-Miguel Fernandez-Dols. "Facial and Vocal Expressions of Emotion". In: *Annual Review of Psychology* 54 (2003), pp. 329–349. DOI: 10.1146/annurev.psych.54.101601.145102.

[2] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition)*. Worth Publications, 2011.

[3] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Automated Emotion Morphing in Speech Based on Diffeomorphic Curve Registration and Highway Networks". In: *Proc. Interspeech 2019*. 2019, pp. 4499–4503. DOI: 10.21437/Interspeech.2019-2386.

[4] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective". In: *Proc. Interspeech 2019*. 2019, pp. 2848–2852. DOI: 10.21437/Interspeech.2019-2512.

[5] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. *Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens*. 2019. arXiv: 1910.11997 [cs.SD].

[6] Juliane Schmidt, Esther Janse, and Odette Scharenborg. "Perception of Emotion in Conversational Speech by Younger and Older Listeners". In: *Frontiers in Psychology* 7 (2016), p. 781. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2016.00781.

[7] Sebastian P. Bayerl, Florian Hönig, Joelle Reister, and Korbinian Riedhammer. *Towards Automated Assessment of Stuttering and Stuttering Therapy*. 2020. arXiv: 2006.09222 [q-bio.QM].

[8] T. Toda, A. W. Black, and K. Tokuda. "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory". In:

IEEE Transactions on Audio, Speech, and Language Processing 15.8 (2007), pp. 2222–2235. ISSN: 1558-7916. DOI: 10.1109/TASL.2007.907344.

[9] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". In: *American Journal of Signal Processing* 2 (2012), pp. 134–138. DOI: 10.5923/j.ajsp.20120205.06.

[10] Takuhiro Kaneko and Hirokazu Kameoka. "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1711.11293 (2017). arXiv: 1711.11293.

[11] Ravi Shankar, Jacob Sager, and Archana Venkataraman. "Non-Parallel Emotion Conversion Using a Deep-Generative Hybrid Network and an Adversarial Pair Discriminator". In: *Proc. Interspeech 2020*. 2020, pp. 3396–3400. DOI: 10.21437/Interspeech.2020-1325.

[12] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. "Multi-Speaker Emotion Conversion via Latent Variable Regularization and a Chained Encoder-Decoder-Predictor Network". In: *Proc. Interspeech 2020*. 2020, pp. 3391–3395. DOI: 10.21437/Interspeech.2020-1323.

[13] Amalia Arvaniti. "Measuring Speech Rhythm". In: *The Cambridge Handbook of Phonetics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2021, 312–335. DOI: 10.1017/9781108644198.013.

[14] F. Charpentier and M. Stella. "Diphone synthesis using an overlap-add technique for speech waveforms concatenation". In: *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing* 11 (1986), pp. 2015–2018.

[15] Sylvain Le Beux, Christophe d'Alessandro, Albert Rilliard, and Boris Doval. "Calliphony: a system for real-time gestural modification of intonation and rhythm". In: *Proc. Speech Prosody 2010*. 2010, paper 101.

[16] Samuel Delalez and Christophe d'Alessandro. "Adjusting the Frame: Biphasic Performative Control of Speech Rhythm". In: *Proc. Interspeech 2017*. 2017, pp. 864–868. DOI: 10.21437/Interspeech.2017-396.

[17] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49. DOI: 10.1109/TASSP.1978.1163055.

[18] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499 (2016). arXiv: 1609.03499.

[19] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *CoRR* abs/1712.05884 (2017). arXiv: 1712.05884.

[20] Yuxuan Wang, R. J. Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous. "Uncovering Latent Style Factors for Expressive Speech Synthesis". In: *CoRR* abs/1711.00520 (2017). arXiv: 1711.00520.

[21] Yusuke Yasuda, Xin Wang, and Junichi Yamagishi. "Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis". In: *Computer Speech and Language* 67 (2021), p. 101183. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2020.101183.

[22] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. "Efficient Neural Audio Synthesis". In: *Proc. ICML 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. PMLR, 2018, pp. 2415–2424.

[23] John Kominek and Alan W Black. "The CMU Arctic speech databases". In: *SSW5-2004* (2004).

[24] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proc. Interspeech 2019*. 2019, pp. 316–320. DOI: 10.21437/Interspeech.2019-1413.

[25] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. "Convolutional Sequence to Sequence Learning". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, 1243–1252.

[26] F. Itakura. "Minimum prediction residual principle applied to speech recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23.1 (1975), pp. 67–72. DOI: 10.1109/TASSP.1975.1162641.

[27] Rohan Badlani, Adrian Lancucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. "One TTS Alignment To Rule Them All". In: *CoRR* abs/2108.10447 (2021). arXiv: 2108.10447. URL: https://arxiv.org/abs/2108.10447.

[28] Masanori Morise, Fumiya YOKOMORI, and Kenji Ozawa. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". In: *IEICE Transactions on Information and Systems* E99.D (2016), pp. 1877–1884. DOI: 10.1587/transinf.2015EDP7457.

[29] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2015).

[30] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *CoRR* abs/1406.1078 (2014). arXiv: 1406.1078. URL: http://arxiv.org/abs/1406.1078.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[32] Jiahong Yuan and Mark Liberman. "Speaker identification on the SCOTUS corpus". In: *Proceedings of Acoustics '08* (2008).

Chapter 7

Unsupervised Markov Model for Duration

7.1 Introduction

In the previous chapter, we discussed a supervised method for learning speaking rate modulation in speech. The discussed approach relied on having parallel data (same phrase spoken by same speaker across multiple categories), which can be expensive to obtain or sometimes infeasible. An ideal algorithm should be able to make acoustic correspondence between collection of utterances in different categories. This was the premise of Cycle-GAN [1] framework which we studied in Chapter 5. Unsupervised methods are typically employed to extract meaningful representation from data for some down-stream task. This usually requires a huge collection of unlabelled data but the refined features can be used to train a downstream task-specific model in a data-efficient manner. This technique is employed in recent strategies for speech recognition where, a large self-supervised neural network (SSL) is

trained to extract features followed by a small speech recognition models using only a fraction of the total data. Wav2Vec [2], HuBert [3, 4], and WavLM [5] are some of the popular examples that currently employ this strategy.

The explicit parameterization of intonation and intensity variations via F0 contour and short-time energy allowed us to learn a density mapping function via Cycle-GAN. This approach is not possible for speaking rate modulation as we cannot summarize it efficiently at segmental and supra-segmental level. The supervised model discussed in Chapter 6 adopted a dynamic time warping procedure which carries out local and global rhythm modification [6]. To solve this problem in an unsupervised manner, we will make some simplifications to our problem statement and adopt a reinforcement learning strategy. To be more specific, the rhythm modification for emotion/voice conversion has three sub-problems:

- Identifying segments of speech that are informative of emotions.
- Predicting a constant factor of modification for the segments discovered.
- Modifying the length of important segments without hurting quality.

Here, we have made an assumption that, we do not need to modify the rhythm of every single phoneme/syllable in an utterance. We can modify only a subset of these components (most important ones) for the purpose of emotion conversion. Further, these components need not be a complete syllable or word, they can be a collection or mix of any of these components including short pause and silences. After identifying such segments (countable in practice), we can process them sequentially to predict a length modification

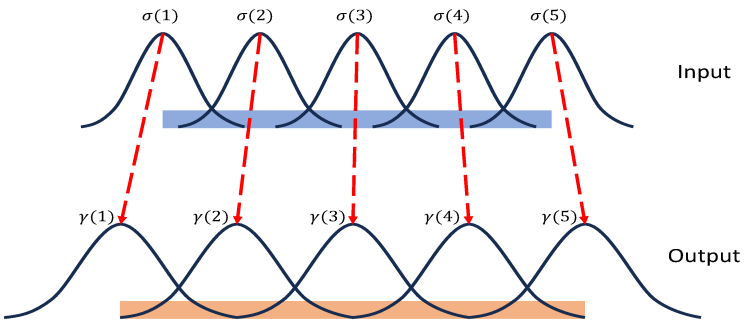

Figure 7.1: Overlap add operation to stretch the input signal by a factor > 1.

factor. These factors are uniform for the entire length of the segment. Finally, the segment can be appropriately modified by its corresponding factor.

In the rest of this chapter, we will go in a reverse manner solving each of the three sub-problems mentioned above. First, we will discuss the mechanism of length modification in speech, followed by our reinforcement learning strategy for predicting factor of modification. Finally, we will explain the technique for discovering important segments in a weakly supervised manner. The segment discovery approach will rely on prediction of human perception of emotional saliency using a masking framework based on Markov process.

7.2 Mechanism of Modification

There are many algorithms which allows speech length modification such as overlap-add (OLA) [7, 8], wave similarity overlap-add (WSOLA), phase vocoder, to name a few. The underlying principle is to split the input audio into chunks of a fixed length and use overlap add to expand or shorten them. In previous chapter, we used local replication and deletion of frames in the input signal to modify its duration. While this approach works in practice, it

introduces discontinuities in phase resulting in a choppy effect for listeners. WSOLA reduces this choppiness via a correlation based local search to find the best segment for reconstruction at any given instant.

To modify duration of a signal using overlap-add, the first step is to lay-down the type of window function $w(n)$, its width l and overlap factor η. Hanning window of width more than the lowest fundamental frequency is a popular choice for this operation. Then, the length of output signal $z(n)$ and the overlap factor decides the time-stamps where the window's center would appear in the input signal $y(n)$. Specifically, let $\tau(n)$ be the time-stretching function, the position of window on output signal can be derived via:

$$\gamma(1) = 1 \text{ and } \gamma(k) = \gamma(k-1) + \eta \tag{7.1}$$

Here the total number of γ is $\lceil |z|/\eta \rceil$. Knowing the $\gamma(k)$, we can figure out the window position on input signal y by $\sigma(k) = \tau^{-1}(\gamma(k))$. Finally, the reconstruction by overlap-add is given by:

$$z(n) = \frac{\sum_{k=1}^{len(\sigma)} w(n - \gamma(k)) \cdot y(n - \gamma(k) + \sigma(k))}{\sum_{k=1}^{len(\sigma)} w(n - \gamma(k))} \tag{7.2}$$

The choice of Hanning window with an overlap factor of 0.5 ensures that the denominator in Equation 7.2 sums to 1. Fig. 7.1 represents the schematic diagram of this operation.

However, incorporating WSOLA in a loss function for optimization is infeasible due to its non-differentiable nature. Therefore, we use WSOLA based modification as a part of the environment description when deriving

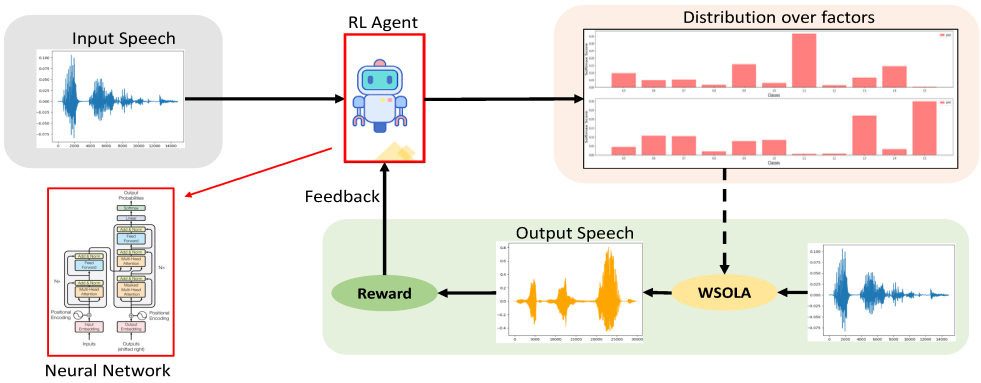

Figure 7.2: RL Strategy: Reinforcement learning framework for predicting factor of modification. The grey panel summarizes the state of the observer, the red panel is the action space, and the green panel represents the environment model.

up with the reinforcement learning strategy for factor learning.

7.3 Factor of Modification: Policy Gradient

In the previous section, our assumption was that we had the segment and factor of modification that affects emotion perception in a given audio. Knowing these details, we used the WSOLA algorithm to modify the length. Now, we will discuss our approach to get a distribution over the factors of modification using reinforcement learning strategy. We assume that we have access to the input speech and the salient regions of that input signal. The first step is to discretize the space of possible duration factors. We choose a range of 0.25-1.9 in steps of 0.15. Note that, this covers a very wide range for WSOLA operation and the extreme values can potentially create some distortions in generated audio. By bucketing all the possible factors of modification, we create a finite number of classes over which we can learn a categorical distribution.

We employ an offline policy gradient to estimate the factor of modification

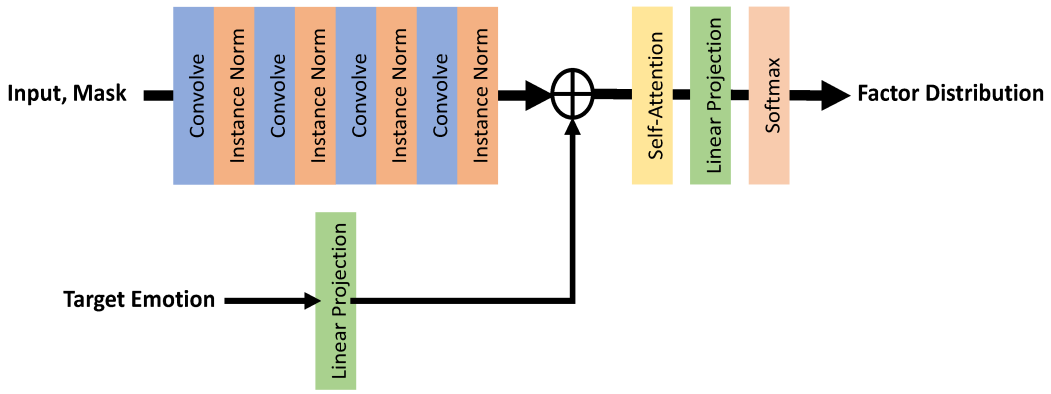

Figure 7.3: Neural network architecture of the RL agent used for factor prediction.

given the speech signal and information about the salient segments [9, 10]. Denoting the speech utterance by $y(t)$ and the segment information variable by $M_t \in \{0, 1\}$, the state of the system is characterized by the tuple $S = (y_t, M_t)$. Mask variable M_t is an indicator (same length as y) denoting which segments of the signal are important (value 1) and which ones are not (value 0). Our reinforcement learning agent takes this state tuple and the target emotion description in a one-hot vector representation. It predicts a distribution over the discrete set of factor which describes the action space A. After sampling from the categorical distribution over the action space, we modify the length of corresponding segment using WSOLA and a reward signal r is generated. This reward signal measures the goodness of learned distribution over the actions for emotion modification, i.e., the increase in score of the target emotion category. An episode in this case is only one time-step long. Fig. 7.2 shows the complete RL framework at a higher level. We will now dive into the details of the reinforcement learning agent used here.

7.3.1 RL Agent

The reinforcement learning (RL) agent is a deep neural network [11] consisting a stack of convolution and transformer layers to learn appropriate distribution over the actions A. It is conditioned on three quantities: (a) the input speech signal (in time domain), (b) segment mask through indicator variables, and (c) the target emotion code corresponding to which a prediction has to be made. Fig. 7.3 shows the neural network architecture used for estimating a probability distribution over the allowable set of factors. Since, the distribution is over an entire utterance, we use max-pooling on the output of transformer layer to feed into the final softmax layer. Therefore, the policy function is a neural network parameterized by θ. The objective of this neural network is to maximize the expected reward which can be written as:

$$\mathcal{L}(\theta) = E_{\pi}\Big[r(s)\Big] \Rightarrow \nabla\mathcal{L}(\theta) = \nabla \sum_{a \in A} \pi(a|s) r(s)$$

$$\begin{aligned}
\nabla\mathcal{L}(\theta) &= \sum_{a \in A} \nabla\pi(a|s) r(s) \\
&= \sum_{a \in A} \pi(a|s) \nabla \log \pi(a|s) r(s) \\
&= E_{\pi}\Big[r(s) \nabla \log \pi(a|s)\Big]
\end{aligned}$$

This is known as the policy gradient theorem [12]. It suggests that to train the neural network, we do not need estimation of gradients through the reward function. It is specially helpful in our case because the reward framework uses WSOLA operation which is not differentiable.

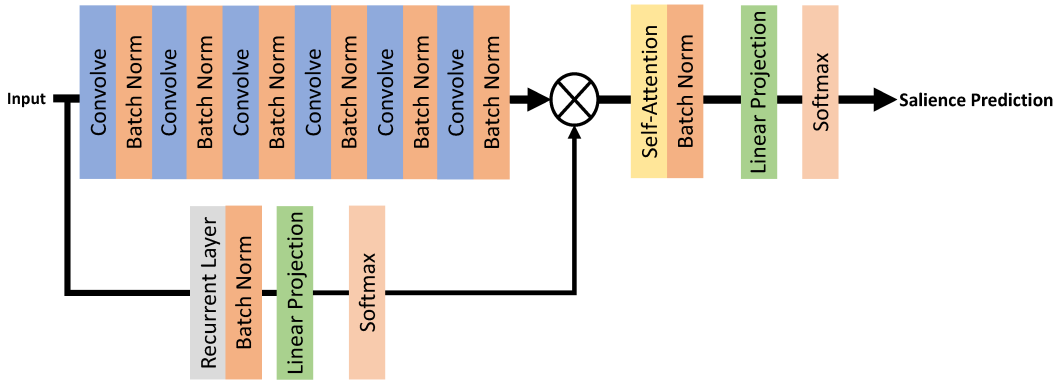

Figure 7.4: Neural network model used for prediction of human perception of emotional saliency. The architecture has three components: (a) feature extraction from raw waveform using stack of convolutions, (b) posterior prediction of Bernoulli masking random variable and (c) salience prediction using masked features.

7.4 Salience Prediction

Finally, getting the salient regions for emotion perception is the last piece of the puzzle that we have not solved yet. We use a simple masking strategy (similar to attention maps) in order to get a continuous segment of speech responsible for human perception of emotion [13]. The VESUS corpus consists of utterances in 5 emotion categories, namely: neutral, angry, happy, sad and fearful. In addition to the audio file, each utterance has an annotation obtained from 10 listeners on Amazon mechanical turk asking them to classify emotion in the corresponding utterance. The ratings provided by these listeners provide a categorical distribution over the emotion classes. Our task is to predict the perception score for each emotion using only the masked portion of input speech.

Attention mechanism is a straightforward approach to solve this problem. However, without additional constraints, it will discover non-contiguous

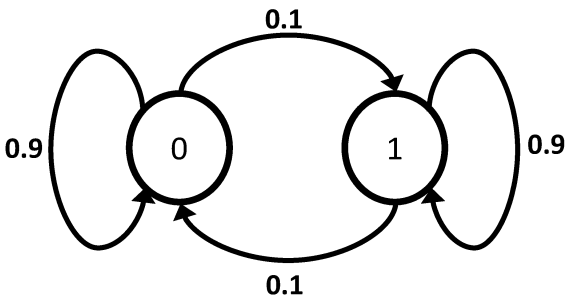

Figure 7.5: Markov states for the prior on the Bernoulli Mask random variables.

segments of speech (sometimes single frames) which are informative. Our objective is to find continuous segments (syllable/word level) that we can be manipulated by WSOLA. We devise a clever masking strategy that allows us to discover such segments. Specifically, we design a neural network with three components: (a) feature extractor module made entirely of downsampling convolution layers, (b) a mask generator module that generates Bernoulli mask over the features and (c) a salience predictor module to predict emotional saliency using information contained in the masked region. Fig. 7.4 shows the neural network architecture in detail.

7.4.1 Masking Variable

We generate mask via sampling from its approximate posterior learned by a combination of recurrent and linear projection module (Fig. 7.4). Given a sequence of frames $X_t \ \forall \ t \in [1, ...T]$, the binary mask $[M_1, M_2, M_3....M_T]$ is a collection of T Bernoulli random variable with the following property imposed as a prior:

$$P(M_t|M_{t-1}) = \begin{cases} Ber(0.9), & \text{if } M_t = M_{t-1} \\ Ber(0.1), & \text{otherwise} \end{cases} \qquad \forall t = 2, 3, ...T$$

Further, $P(M_1) = Ber(0.01)$, this constraint specifies that the masking follows a first-order Markov property, i.e. the future is independent of past given present. The mask distribution as time t is dependent on the mask at time $t-1$. It is largely similar (0 or 1) to its previous time-step which ensure a continuity in the segments of importance. Fig. 7.5 shows the state transition diagram of the corresponding Markov chain. While this prior constraint on the mask helps identify continuous segments, it can happen that the mask takes the value 1 for the entire duration of the speech utterance. First, we define the distribution over the mask variables learned by the neural network as:

$$q_\theta(M_1, M_2, M_3...., M_T|\mathbf{X}) = q_\theta(M_1|\ \mathbf{X})q_\theta(M_2|\ \mathbf{X})...q_\theta(M_T|\ \mathbf{X}) \tag{7.3}$$

where, we have used the mean-field approximation [14, 15] for the variational posterior learned by the neural network (parameterized by θ). We add a sparsity penalty at each time step via KL divergence loss with a Bernoulli distribution of very small success. Specifically, the sparsity penalty can be written as:

$$\mathcal{L}_{sparse} = \sum_{t=1}^{T} D_{KL}\left[q_\theta(M_t\|\mathbf{X})|Ber(0.01)\right] \tag{7.4}$$

Adding the sparsity penalty to approximate posterior resolves the problem of the mask being triggered for the entire speech duration. Finally, the Markov prior is imposed on the posterior using KL divergence penalty which can be

written as:

$$
\begin{aligned}
\mathcal{L}_{prior} &= D_{KL}\big[q_\theta(\mathbf{M}|\mathbf{X})\|P(\mathbf{M})\big] \\
&= D_{KL}\big[q_\theta(M_1, M_2 ... M_T|\mathbf{X})\|P(M_1, M_2, ... M_T)\big] \\
&= \sum_{M_T}\sum_{M_{T-1}} ... \sum_{M_1} q_\theta(M_1|\mathbf{X}) q_\theta(M_2|\mathbf{X}) ... q_\theta(M_T|\mathbf{X}) \\
&\qquad \times \log \frac{q_\theta(M_1|\mathbf{X}) q_\theta(M_2|\mathbf{X}) ... q_\theta(M_T|\mathbf{X})}{P(M_1)P(M_2|M_1)..P(M_T|M_{T-1})} \\
&= \sum_{M_1} q_\theta(M_1|\mathbf{X}) \log \frac{q_\theta(M_1|\mathbf{X})}{P(M_1)} \sum_{M_2} q_\theta(M_2|\mathbf{X}) \log \frac{q_\theta(M_2|\mathbf{X})}{P(M_2|M_1)} \\
&\qquad ... \sum_{M_T} q_\theta(M_T|\mathbf{X}) \log \frac{q_\theta(M_T|\mathbf{X})}{P(M_T|M_{T-1})}
\end{aligned}
$$

Therefore, the KL-divergence penalty [16] is decomposed into T terms where each term can be computed conditioned on the past which has only two values, i.e., 0 or 1. Therefore, the computation is tractable and the operation can be vectorized easily for efficiency. Finally, $q_\theta(M_t|\mathbf{X})$ is by parameterized by a Bernoulli parameter between 0 and 1. We use sigmoid activation in the mask generator module to get the posterior distribution. Then, a sampling process generates the mask which is element-wise multiplied to the extracted features (from convolutional stack) and fed into the salience prediction module. The saliency loss is an L-1 penalty over the predicted softmax (5 classes: neutral, angry, happy, sad and fear) and the ground truth.

Finally, since we sample from the variational posterior to generate the

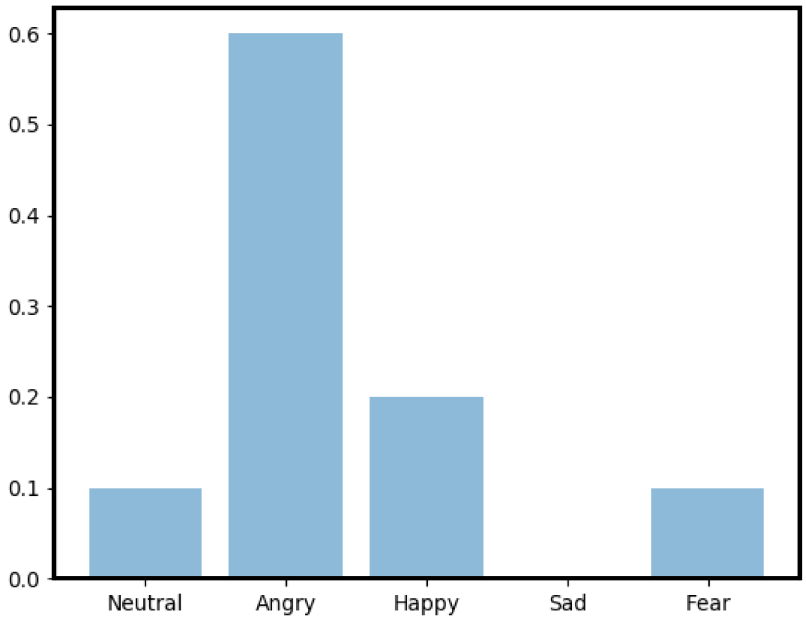

Figure 7.6: An example showing how salience score obtained from AMT looks like for an utterance.

mask for salience prediction, we use Gumbel softmax (Chapter 2) [17, 18] for backpropagation through the sampler module.

7.5 Experiments and Results

In this section, we will discuss the results of saliency prediction using Markov masking and analyze the phonetic/syllabic structure in masked region for each emotion class. We will begin with description of the dataset.

7.5.1 Dataset

Once again, we use the VESUS [19] corpus to carry out the experiments. VESUS comes with a crowd-sourced annotation obtained from 10 listeners on

Amazon Mechanical Turk for each utterance. This allows us to create a soft assignment over mixture of emotion (neutral/angry/happy/sad/fearful) rather than one single emotion category for prediction (See Fig. 7.6). Therefore, our salience predictor predicts the human perception of emotion during training and inference stage. We split the VESUS according to the following scheme:

- 11000 samples are randomly selected for training (mixed across speakers)
- 250 samples are randomly selected for validation
- 750 samples are randomly chosen for evaluation/testing

7.5.2 Emotion Recognition Accuracy

We evaluate the human perception prediction on VESUS testset. The results are summarized in Table 7.1. We can see that the top-1 results (weighted F1 and accuracy) are above 75% which shows that the designed feature extractor and salience predictor modules (Fig. 7.4) are good at predicting soft score over the emotion classes. We further evaluated top-2 accuracy of the proposed model by checking the presence of target distribution mode in the top-2 scores of prediction. The accuracy score corresponding to this evaluation is $> 90\%$. This is particularly important because, in many cases the ground truth saliency score is a tie between two or more emotions. Top-1 prediction ignores this issue whereas top-2 captures it to some extent.

Fig. 7.7 shows the confusion matrix on the test set obtained from proposed model. We can see that the diagonal elements are relatively higher for most emotions. Fear category has some confusion with sad emotion which is

Table 7.1: Emotion recognition performance on VESUS test set.

Mode	F1 (macro)	F1 (weighted)	Accuracy
Top-1	0.72	0.76	0.78
Top-2	-	-	0.93

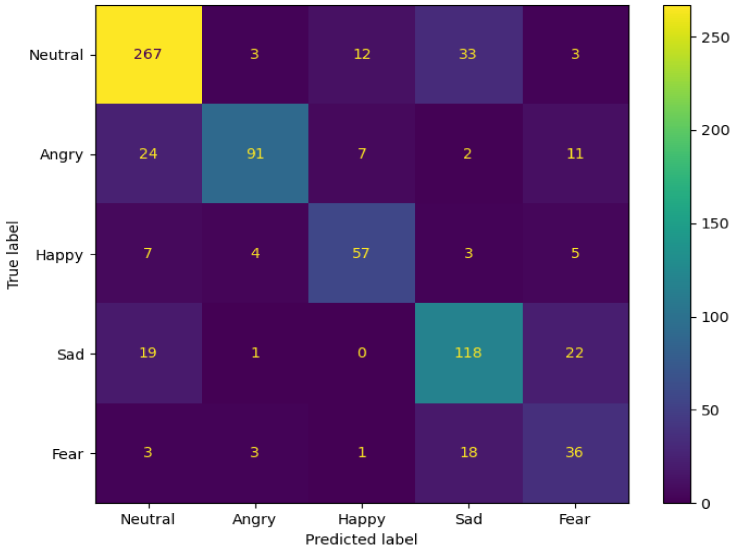

Figure 7.7: Confusion matrix corresponding to top-1 accuracy on VESUS testset.

expected because fear and sadness are usually expressed with shivering nature of the voice.

Finally, Fig. 7.8 shows the distribution/spread of mutual information between the predicted distribution over emotional categories "vs" the ground truth. To compute the joint density, we use the empirical estimate obtained from the pair of ground-truth and predicted softmax scores. Fig. 7.8(a) shows the empirical joint density over the test set while Fig. 7.8(b) shows the mutual information (MI) between the two distribution. Note that, we cannot directly compare this or make any inference based on the MI score as it does not have any baseline.

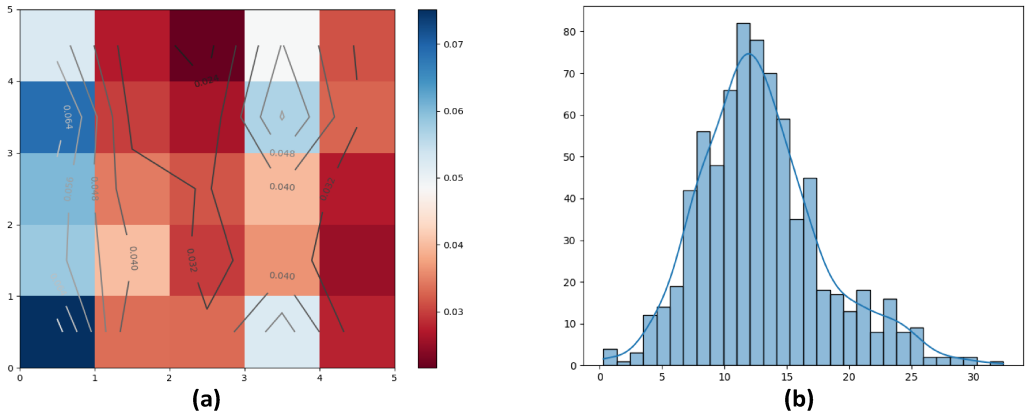

Figure 7.8: (a) Empirical joint density estimated from test set and (b) mutual information estimated from the predicted softmax scores 'vs' ground-truth annotation obtained from Mechanical Turk.

Figure 7.9 shows some examples of the discovered segments and the corresponding predictions of six different utterances from VESUS test set. The top plot is the audio signal (in red) while the second plot is the extracted features from the convolutional encoder part of salience predictor as shown in Fig. 7.4. The third plot is the variational posterior predicted (in blue) for the masking random variable and the corresponding mask sample (in blue) obtained via sampling. The sampled mask is also overlaid on extracted features (in second) as shown in white. Finally, the last bar plot shows the ground-truth and predicted salience score over emotion categories.

7.5.3 Emotion Conversion

In order to change the duration of segments estimated by salience predictor, we train the RL agent by sampling one of the contiguous chunks during training. During inference, the individual chunks are separately processed

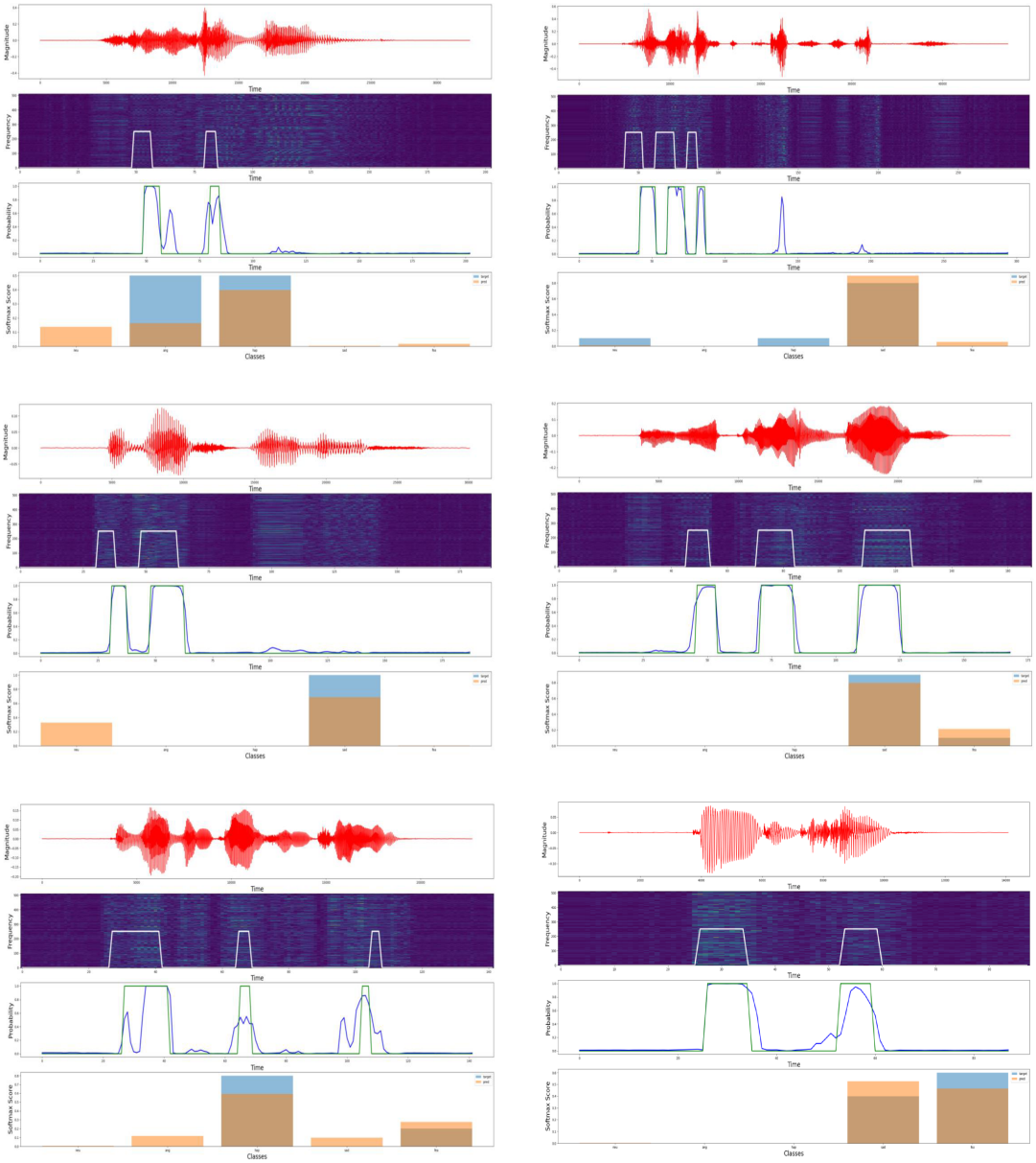

Figure 7.9: Some examples of discovered segments important for prediction of the corresponding emotion classes.

which provides more flexibility in terms of rhythm manipulation as different segments can undergo varying degrees of modification. Fig. 7.10 depicts the fraction of samples that register increase in the target emotion class in the

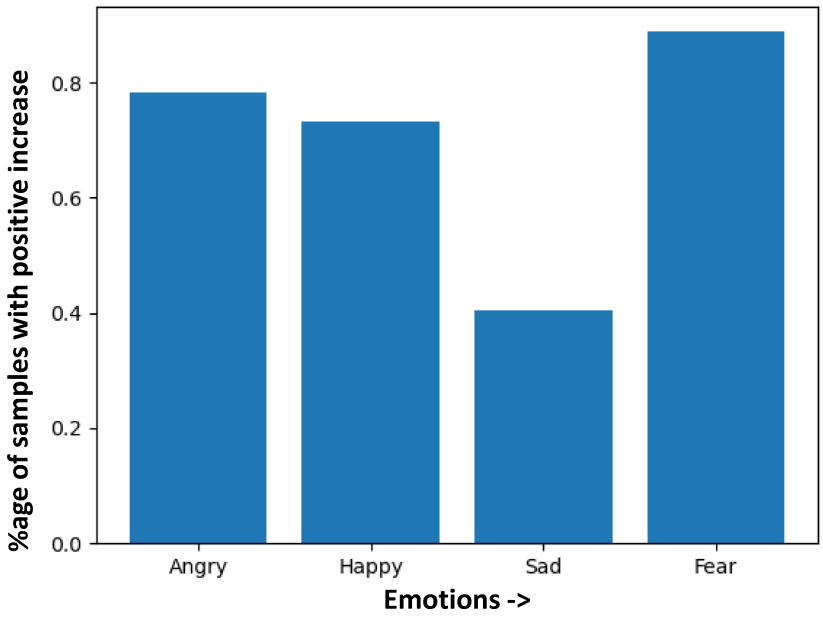

Figure 7.10: Percentage of test samples with positive increase in the target emotion score post-modification.

test set. It can be observed that more than 75% of the samples see an increase in the target emotion saliency score when the targets are: angry, happy or fear. When the target is sad, the RL agent fails to increase the corresponding salience score in the majority of cases. This is because the underlying salience predictor has poor robustness for sad emotion and small variations in the signal results in prediction changing to fear category. Due to this instability, the loss feedback is less reliable to train the RL model.

Finally, Fig. 7.11 provides an insight into how the individual emotion class scores are affected when the target emotions are: (a) angry, (b) happy, (c) sad and (d) fearful. Notice that, the model does extremely well on reducing the neutral score demonstrating it is easier to convert neutral samples for at least 3 different emotion categories. Therefore, this model can be used to

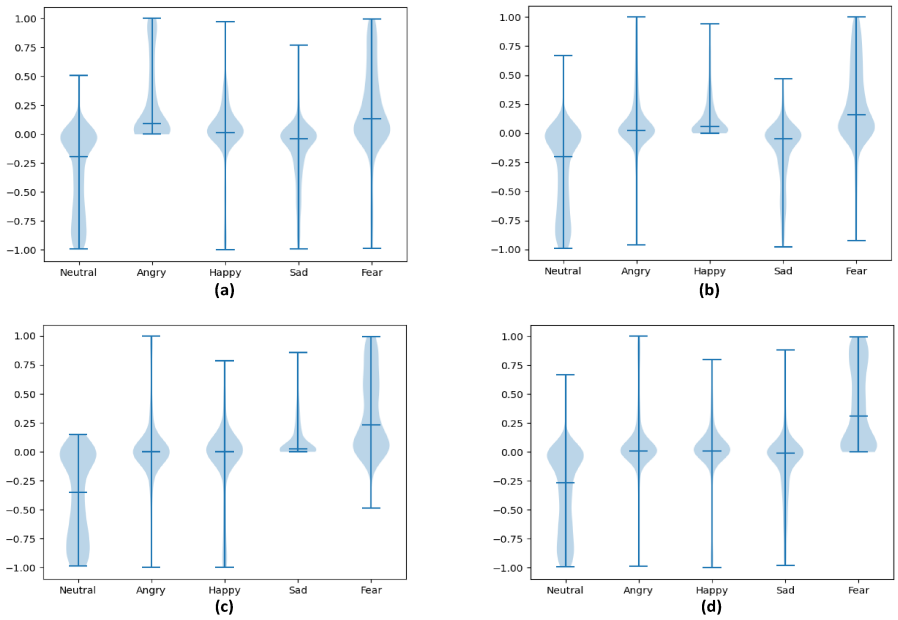

Figure 7.11: Relative saliency score changes when the target emotion is: (a) angry, (b) happy, (c) sad and (d) fearful.

manipulate/inject emotion into neural utterances which is produced by most state-of-the-art neural vocoders. Further, sad emotion category witnesses a small change in the score due to unreliable salience prediction. Overall, our proposed model can be a good starting point for duration modification and can be combined with powerful pitch/energy modification models to achieve the desired effect.

7.6 Conclusion

In conclusion, the reinforcement learning model developed for rhythm modification in speech for emotional speech synthesis represents a significant advancement in the field of emotional speech generation. By effectively identifying contiguous segments crucial for emotion perception through the

innovative Markov masking strategy and implementing KL divergence-based sparsity loss, the model not only excels in emotion recognition on the VESUS corpus but also provides valuable insights into the identification of speech segments for duration modification. As we look to the future, the potential to train the RL agent to estimate a distribution over discrete modification factors and utilize WSOLA opens up exciting avenues for further refinement and enhancement in emotional speech synthesis research.

References

[1] Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. *A Diffeomorphic Flow-based Variational Framework for Multi-speaker Emotion Conversion*. 2022. arXiv: 2211.05071 [eess.AS].

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12449–12460.

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460.

[4] William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. "Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute". In: *arXiv preprint arXiv:2306.06672* (2023).

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *arXiv preprint arXiv:2110.13900* (2021).

[6] Ravi Shankar and Archana Venkataraman. "Adaptive Duration Modification of Speech using Masked Convolutional Networks and Open-Loop Time Warping". In: *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*. 2023, pp. 177–183. DOI: 10.21437/SSW.2023-28.

[7] Jonathan Driedger. *Time-Scale Modification Algorithms For Music Audio Signals*. Saarland University, Faculty of Natural Sciences and Technology I, Department of Computer Science, 2011.

[8] W. Verhelst and M. Roelands. "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech". In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. 1993, 554–557 vol.2. DOI: 10.1109/ICASSP.1993.319366.

[9] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. MIT Press, 1999. URL: https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

[10] Lilian Weng. "Policy Gradient Algorithms". In: *lilianweng.github.io* (2018). URL: https://lilianweng.github.io/posts/2018-04-08-policy-gradient/.

[11] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. *AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders*. 2022. arXiv: 2211.09120 [cs.CV].

[12] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018. ISBN: 0262039249.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[14] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080%2F01621459.2017.1285773.

[15] Kevin P. Murphy. "Machine learning - a probabilistic perspective". In: *Adaptive computation and machine learning series*. 2012. URL: https://api.semanticscholar.org/CorpusID:17793133.

[16] Wikipedia contributors. *Kullback–Leibler divergence — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=Kullback%E2%80%93Leibler_divergence&oldid=1173344518.

[17] Eric Jang, Shixiang Gu, and Ben Poole. *Categorical Reparameterization with Gumbel-Softmax*. 2017. arXiv: 1611.01144 [stat.ML].

[18] K. D. Tocher. “Statistical Theory of Extreme Values and Some Practical Applications; Probability Tables for the Analysis of Extreme Value Data”. In: *Journal of the Royal Statistical Society: Series A (General)* 118.1 (1955), pp. 106–106. DOI: https://doi.org/10.2307/2342529. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2342529. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2342529.

[19] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. “VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English”. In: *Proc. Interspeech 2019*. 2019, pp. 316–320. DOI: 10.21437/Interspeech.2019-1413.

Chapter 8

Conclusion

In conclusion, this thesis embarked on the ambitious journey of injecting emotion into neutral and monotone speech, guided by the principles of prosody modeling. Our overarching goal was to bridge the emotional gap in speech by focusing on the fundamental prosodic features that shape our perception of emotion: pitch contour, intensity contour, and rhythm variation.

While pitch and intensity could be parameterized by F0 contour and energy contour, respectively, rhythm remained a challenge without explicit parameterization. Through meticulous research and innovative approaches, we set out to learn a mapping function that could transform neutral speech into the emotional states of anger, happiness, and sadness.

Our methodology leveraged the World vocoder to decompose speech into its essential components: the spectral envelope, F0 contour, and aperiodicity. By focusing on the spectral envelope, we successfully extracted the energy contour, leaving aperiodicity untouched as it was deemed less significant in the context of emotion perception.

In Chapter 3, we delved into the development of a framewise model for

predicting F0 and energy values using a highway neural network. This was achieved through a maximum likelihood estimation framework, employing a Laplace distribution loss with a zero mean and specified variance. Our iterative strategy, reminiscent of an E-M algorithm, enabled us to estimate the parameters of the highway network effectively. We also introduced a novel variation of this approach, involving intermediate latent variables known as Momenta.

The concept of Momenta variables, elucidated in the background section, served as a unique parameterization of diffeomorphic mappings between two manifolds. This intricate process involved a flow composed of multiple velocity vector fields, facilitating a smooth transition between the manifolds. The low-dimensional embeddings encapsulated by Momenta fully characterized the high-dimensional flow, paving the way for their application in our study.

Our use of Momenta as intermediaries in learning a smooth flow from a collection of F0 and energy contours in neutral emotion to the states of anger, happiness, and sadness introduced a valuable form of regularization. This regularization not only enabled us to capture variations in pitch across different speakers and emotional states but also yielded lower mean absolute errors and closer alignment with ground-truth emotional states. Our comprehensive evaluation, which included comparisons with several state-of-the-art baselines, convincingly demonstrated the robustness and versatility of our model, especially in multi-speaker settings.

In addressing the limitations of framewise models, which tend to overlook the segmental and supra-segmental properties of speech, Chapter 4 of this

thesis introduces a novel approach - the Encoder-Decoder-Predictor model for prosody modification. This innovative model takes a holistic view of speech, converting both pitch and spectrum information for the entire utterance simultaneously. It serves as an end-to-end framework, constructed with gated convolution layers featuring residual connections to facilitate efficient backpropagation.

One of the key advancements in our approach was the utilization of momenta variables as an intermediate representation bridging the source and target emotional states. This intermediate step plays a crucial role in regularizing the prediction of the F0 contour, enhancing the overall accuracy of our model. To train the system effectively in a multi-task setting, we employed the mean absolute error as our loss function. Furthermore, we incorporated constraints from the WORLD vocoder into the generation process, ensuring that the generated speech aligned more closely with natural speech patterns. Specifically, the generation of the spectral envelope was conditioned on the predicted pitch contour, utilizing the pitch-synchronous analysis technique to improve the quality of synthesized speech.

Our experiments yielded compelling results, demonstrating that the regularization introduced via momenta variables significantly improved the accuracy of pitch estimation. Furthermore, the versatility and robustness of our proposed technique were evident as it successfully generalized to multiple speakers and performed well even in unseen phrase settings. These findings underscore the potential of the Encoder-Decoder-Predictor model as a valuable tool for enhancing the emotional expressiveness of speech and its

applicability in real-world scenarios involving various speakers and speech contexts.

In response to the requirement for parallel data time-aligned using Dynamic Time Warping (DTW) for supervised learning, Chapter 5 introduces a groundbreaking approach known as the Variational CycleGAN (VCGAN). This novel formulation addresses the limitations of traditional methods by offering a more comprehensive and efficient solution for prosody modification.

VCGAN represents a significant advancement in the field, as it allows for the simultaneous modification of both the F0 contour and energy contour in a single pass. This simultaneous modification capability enhances the model's ability to capture and recreate the complex prosodic variations necessary for expressing emotions in speech. Integral to the VCGAN framework is the utilization of Diffeomorphic flow, particularly the momenta parameterization, as a critical component of the generator. This choice of representation aids in regularizing the modification process, resulting in more accurate and expressive prosody transformations. To further enhance the performance of the VCGAN model, we addressed several limitations present in the vanilla CycleGAN model. These enhancements included tackling higher-order moment matching and addressing gradient calibration issues, which ultimately led to more stable and precise prosody modifications.

A pivotal development in VCGAN was the incorporation of a joint density discriminator, which effectively reduced the mismatch between generator and discriminator losses in the adversarial setting. This strategic addition played a crucial role in improving the overall quality and fidelity of the modified

prosody. Extensive experimentation was conducted to evaluate the performance of VCGAN, encompassing both seen and unseen speakers. These experiments included comprehensive comparisons against various baselines, showcasing the model's robustness and adaptability. Particularly noteworthy was the model's ability to handle unseen speakers, which was convincingly demonstrated through testing on speech generated by the Wavenet model. These findings underscored the VCGAN's potential as a versatile tool for prosody modification, especially in scenarios involving a wide range of speakers and speech contexts.

In Chapter 6, our focus shifted towards the realm of rhythm modulation, where we introduced a supervised technique that showcased our ability to effectively modify the rhythm of speech. This innovative approach aimed to tackle the nuanced aspects of prosody that contribute to the rhythmic variations observed in speech. Central to our approach was the incorporation of Dynamic Time Warping (DTW) alignment into an encoder-decoder model, with an attention map serving as an intermediary. This sophisticated architecture allowed us to capture the temporal relationships between speech segments accurately, a fundamental requirement for rhythm modulation. A key breakthrough in our methodology was the modeling of the attention map as a latent variable with a non-informative prior. This novel approach enabled us to approximate the cost matrix required for the DTW procedure, effectively enhancing the efficiency and effectiveness of the rhythm modulation process.

Our masked convolutional encoder-decoder framework played a pivotal role in auto-regressively decoding the frames of the target sequence, even

when they did not resemble intelligible speech. Remarkably, we demonstrated that this approximate target sequence was sufficiently capable of modifying the duration and rhythm of the input utterance, showcasing the model's adaptability and versatility. To further refine the rhythmic qualities of the modified speech, we introduced the Itakura parallelogram-based masking technique. This method effectively constrained the speaking rate in the target sequence, contributing to a more natural and expressive rhythmic transformation.

To validate the capabilities of our model, we conducted comprehensive experiments on a range of tasks, including speaker conversion using the CMU-ARCTIC dataset and emotion conversion using the VESUS dataset. These experiments not only demonstrated our model's ability to learn a robust rhythm mapping function but also highlighted its efficiency in doing so with a minimal dataset and a low computational footprint. Overall, Chapter 6 represents a significant step forward in the realm of rhythm modulation, showcasing the potential for highly effective and resource-efficient prosody modification techniques.

In Chapter 7, we introduced a novel and sophisticated framework for rhythm modification within the realm of prosody. This innovative approach leveraged reinforcement learning to tackle the challenge of modifying the duration of specific segments within given speech while preserving the overall emotional expression. At the core of our methodology was the utilization of the WSOLA (Waveform Similarity Overlap and Add) technique, which is highly effective for time-stretching or compressing audio segments. However,

WSOLA poses a challenge in that it is non-differentiable, making it unsuitable for direct integration into traditional differentiable neural networks. To address this limitation, we turned to policy gradient methods, a powerful technique for estimating optimal policy functions conditioned on segments of importance within the speech.

A critical aspect of our approach was the need to identify salience regions within the speech that are essential for conveying emotion. To achieve this, we devised a Markov masking framework, allowing us to pinpoint continuous regions of speech that carry crucial emotional information. These salience regions served as the foundation for our rhythm modification policy. To guide the reinforcement learning process effectively, we established a reward function for policy gradient. This reward function was derived from a saliency prediction module that we trained a priori on the VESUS corpus, a rich dataset for emotion-related speech. Importantly, the training data for the saliency predictor was augmented with soft emotional ratings obtained through Amazon Mechanical Turk, providing a valuable source of human-annotated emotional perception data.

In conclusion, this comprehensive thesis journeyed through the intricate landscape of prosody modification, presenting a multifaceted approach to inject emotion into neutral and monotone speech. Chapters 4 and 5 unveiled the Encoder-Decoder-Predictor model and Variational CycleGAN (VCGAN), offering novel solutions for pitch, energy contour, and rhythm modulation. Chapter 6 introduced a supervised technique with Dynamic Time Warping

(DTW) alignment for rhythm modification, showcasing its efficacy in a variety of tasks. In Chapter 7, a reinforcement learning framework, guided by salience regions identified through a Markov masking framework, tackled rhythm modification, demonstrating remarkable adaptability and emotion preservation. These advances collectively represent a significant stride toward enhancing speech expressiveness and emotion perception, fostering the potential for more engaging human-computer interactions while showcasing the versatility and robustness of these innovative prosody modification techniques.

Ravi Shankar, Ph.D. Candidate
Department of Electrical and Computer Engineering, Johns Hopkins University
rshanka3@jhu.edu | (+1) 443-842-2025 | github.io/ravi-shankar

---

Research Summary

Research associate with 8+ years of research experience in **machine learning**, **speech/audio processing** and **deep learning**. I have co-authored **11 research articles** with $\sim$ **100 research citations** in the domain of speech and audio processing. In the past, I have worked as an **iOS** and **Rails developer** at housing.com and CaRPM, respectively.

Education

Johns Hopkins University, Baltimore *(2017 - 2023)*
Ph.D. candidate, Department of Electrical and Computer Engineering
Advisor: Dr. Archana Venkataraman

Johns Hopkins University, Baltimore *(2022 - 2023)*
M.S. in Applied Math and Statistics (Statistics concentration)
Advisor: Prof. Amitabh Basu

Indian Institute of Technology (IIT), Guwahati *(2011 - 2015)*
Bachelors in Electrical Engineering
Advisors: Prof. S.R.M Prasanna (Dean, RnD) and Prof. S. Sundaram

Skills

Statistical Modeling, Python, Matlab, C++, Shell Scripting, TensorFlow, PyTorch, Generative Modeling, GANs, VAEs, Transformers, CNNs, LSTMs, Spoken Keyword Detection, TTS, Reinforcement Learning, ASR, Multimodal learning, Self/Un-supervised learning

Work Experience

Research Scientist Intern, Meta Reality Labs, Redmond *(May'22 - Aug'22)*
- Knowledge distillation of Wav2Vec embeddings for speech enhancement.
- Implementation of Causal GCRN model for complex domain enhancement.

Research Staff, IDIAP Institute, Martigny *(Jan'17 - Jun'17)*
- Analyzed effect of temporal continuity of acoustic features in ASR models.

Research Assistant, IIT Guwahati *(Sep'16 - Dec'16)*
- Proposed joint DTW-CNN framework for keyword spotting in speech.

Ruby-on-Rails Developer, CaRPM, Gurgaon *(Jan'16 - Aug'16)*
- Developed back-end module for analysis of used cars to estimate resale value.
- Developed accelerometer data based trip quality evaluation for maintenance.

Research Assistant, AICML, UofA *(Sep'15 - Jan'16)*
- Developed patient specific survival prediction using ordinal classifier.

iOS Developer, Housing.com, Mumbai *(Jun'15 - Sep'15)*
- Implemented new features for real estate rental in the native iOS app.

Selected Research articles

A Closer Look at Wav2Vec2 Embeddings for Single-channel Speech Enhancement
Ravi Shankar, Ke Tan, Buye Xu, Anurag Kumar (Meta Reality Labs)
Under Submission.

Adaptive Duration Modification of Speech using Masked Convolution and Time Warping
Ravi Shankar, Archana Venkataraman
ISCA Speech Synthesis Workshop, 2023

A Diffeomorphic Flow-based Variational Model for Emotion Conversion
Ravi Shankar, Hsi-Wei Hsieh, Nicholas Charon, Archana Venkataraman
IEEE/ACM Transactions for Audio, Speech and Language Processing (2022)

Non-parallel Emotion Conversion using a Deep-Generative Hybrid Network and an Adversarial Pair Discriminator
Ravi Shankar, Jacob Sager, Archana Venkataraman
Proceedings of Interspeech 2020

More papers: https://scholar.google.com/citations?user=uGtWx6EAAAAJ&hl=en

Honours & Awards

MINDS Data Science Research Fellowship (2019-20 and 2020-21)
Received for **'Diffeomorphic Time Warping for Duration Modification'**. Awarded annually for mathematical contribution in advancing machine learning and data science.

ISCA 2020 Travel Award
Recognized for our technical contributions in the chained Encoder-Decoder-Predictor model based on reviewer's comments.

NVIDIA Research Fellowship
Our proposal titled **'AI for Mental Health and Speech Disorder'** featured among the top 5% PhD proposals in 2021.

Graduate Research Fellowship, JHU
JHU award for ECE PhD students to recognize their outstanding research contribution in the domain of Computer Engineering.

UofA Computing Science Research Fellowship
Awarded by University of Alberta to fund research in Computer Science (**Declined**).

Institute Merit Scholarship, 2012-13
Annual award for best academic performance in Electrical Engineering.

DAAD-WISE Fellowship, 2014
Fellowship for funding a summer research internship in TU Darmstadt, Germany.

Merit-Cum-Means Scholarship, 2012-13, 2013-14 and, 2014-15
Award to recognize students from lower income group and unprivileged households having good academic credentials.

Leadership Roles

NSA Lab Seminar Organizer, JHU (2022-23)
Technical Hiring Lead, CaRPM (2016)
Student Representative, EEE, IIT Guwahati (2013-15)
Electronics Club Secretary, IIT Guwahati (2014-15)

Mentoring

Abdouh Harouna Kenfack, MSE in Dept. of Applied Math, JHU
Yi-Te Hsu, MSE in Dept. of Computer Science, JHU
Arjun Somayazulu, BS in Dept. of Computer Science, JHU
Jacob Sager, BS in Dept. of Electrical and Computer Engineering, JHU

Services

TA, **Probabilistic Machine Learning** (EN.520.651) (Fall'21, Fall'22)
Reviewer, **NeuRips** 2022, 2023
Reviewer, **UAI** 2023
Reviewer, **ICLR** 2022
Reviewer, **Interspeech** 2021, 2022, 2023
Reviewer, **CISS** 2021

References

Dr. Archana Venkataraman, **Associate Professor, ECE**, BU
Dr. Amitabh Basu, **Associate Professor, AMS**, JHU
Dr. Anurag Kumar, **Research Scientist Lead**, Facebook/Meta
Abhishek Maitreyi, **CEO**, CaRPM