A Diffeomorphic Flow-Based Variational Framework for Multi-Speaker Emotion Conversion

Ravi Shankar, *Student Member, IEEE*, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman, *Member, IEEE*

***Abstract*—This paper 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.**

***Index Terms*—Cycle-GAN, diffeomorphic registration, nonparallel emotion conversion, variational approximation.**

I. Introduction

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

Manuscript received 4 October 2021; revised 11 May 2022 and 29 July 2022; accepted 4 September 2022. Date of publication 26 September 2022; date of current version 2 December 2022. This work was supported in part by the NSF CAREER Award 1845430 (PI Venkataraman) and in part by the NSF under Grants DMS-1945224 and DMS-1953267 (PI Charon). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Corresponding author: Ravi Shankar.)*

Ravi Shankar and Archana Venkataraman are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: rshanka3@jhu.edu; archana.venkataraman@jhu.edu).

Hsi-Wei Hsieh and Nicolas Charon are with the Department of Applied Math and Statistics, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: hhsieh@cis.jhu.edu; charos@cis.jhu.edu).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Johns Hopkins Homewood Institutional Review Board under Application No. FWA 00005834.

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2022.3209948, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2022.3209948

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.

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 $\mathbf{z}_i = [\mathbf{x}_i \quad \mathbf{y}_i]^T$ denote the concatenated source and target features and c_i denote the latent cluster assignment for utterance i. From here, we have:

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

where, $P(\mathbf{z}_i / c_i = k) \sim N(\mathbf{z}_i; \mu_k, \Sigma_k)$ 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

$$E[\mathbf{y}_i / \mathbf{x}_i] = \sum_{k=1}^{K} P(c_i = k / \mathbf{x}_i) \left[\mu^y{}_k + \Sigma^{xy}{}_k (\Sigma^{xx}{}_k)^{-1} (\mathbf{x}_i - \mu_x)_k \right] \tag{2}$$

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 overcomplete 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:

$$J = \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \lambda \sum_j \|\mathbf{V}(:, j)\|_1 \tag{3}$$

The first term in (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 suprasegmental 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:

$$\begin{aligned} L_{adv} = \min_G \max_D \; & E_{x \sim P(X)}[\log(D(x))] \\ & + E_{z \sim P(Z)}[\log(1 - D(G(z)))] \end{aligned} \tag{4}$$

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 (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:

$$L_{cycle} = E_{x \sim P(X)} [\|x - G_\vartheta(G_\gamma(x))\|_1] \tag{5}$$

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.

In this paper 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 [25]. This paper provides the following novel contributions above the conference paper. 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.

II. Method

Our strategy is to manipulate two key prosodic features: the F0 (pitch) contour, and the energy (loudness) contour. Fig. 1 shows the relationship between the features during the inference step of the process. We begin with taking an utterance in source emotion A from which we extract the F0 contour (p_A) and the mel-cepstral features (S_A) using the WORLD vocoder [26]. The energy contour (e_A) is extracted directly from the spectral features. We define latent variables called momenta (m_p, m_e), which serve as intermediaries between the two emotion classes under consideration. The F0 contour in target emotion (p_B) is a deterministic function of the momenta (m_p) and source F0 contour, through a diffeomorphic warping process that we describe in Section II-B. The estimated F0 contour and the source spectrum together generate the momenta (m_e) for the energy contour which is then further used to generate the cepstral features (S_B). The estimated F0 contour and cepstral features

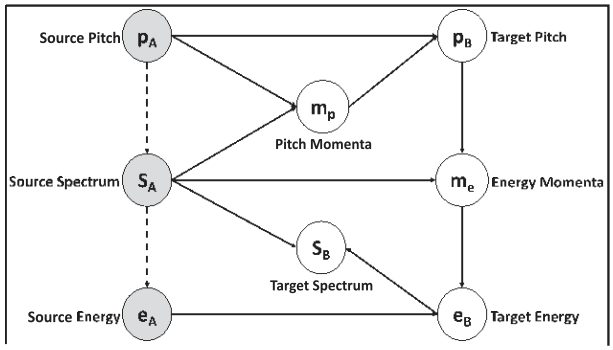

Fig. 1. Graphical representation of our emotion conversion strategy. m_p and m_e serve as an intermediaries for pitch and energy contours, respectively.

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 II-A. 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.

A. Variational Cycle-GAN

The cycle consistency loss of a traditional Cycle-GAN is given by (5) and repeated below for convenience:

$$L_{cycle} = E_{x \sim P(X)}\left[kx - G_\vartheta(G_\gamma(x))k_1\right] \tag{6}$$

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

It is easy to show that (6) is not a well-behaved loss function (Propositions 1 and 2 in Appendix A). Specifically,

1) It only enforces a first-order moment matching between the generated and target data distributions.
2) The expectation in (6) 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 (S_A, p_A) and (S_B, 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 $G_\gamma : (S_A, p_A) \rightarrow (S_B, p_B)$ and $G_\vartheta : (S_B, p_B) \rightarrow (S_A, p_A)$. The corresponding distributions learned by the generator functions are given by $P_\gamma(S_B, p_B)$ and $P_\vartheta(S_A, p_A)$. Our new penalty for the generator G_γ is:

$$L_{G_\gamma} = KL(P(S_A, p_A) \| P_\vartheta(S_A, p_A)) \tag{7}$$

Using the law of total probability, we can write:

$$P_\vartheta(S_A, p_A) = \int\int P_\vartheta(S_A, p_A | S_B, p_B) \times P(S_B, p_B)\, dS_B\, dp_B \tag{8}$$

Equation (8) is generally intractable, but we can derive an upper bound on the loss in (7) that can be optimized easily (see Appendix B). Effectively, we can minimize:

$$\tilde{L}_{G_\gamma} = E_{(S_A, p_A)} E_{(S_B, p_B)|P_\gamma} [\log (P_\gamma(S_B, p_B | S_A, p_A) \times P(S_A, p_A))] \tag{9}$$

Equation (9) 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 $P(S_A, p_A)$ and $P_\vartheta(S_A, p_A)$ 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 (see Appendix G), 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.

B. Prosodic Regularization via Momenta

As shown in the Fig. 1, we use two intermediate representations (denoted by m_p and 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 [27]. 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) [28], [29]. 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 [30]. Effectively, the source prosodic contour specifies the initial state, while the momenta (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. Fig. 2 shows an example of momenta acting on a source F0 contour to match it with a target F0 contour via the LDDMM registration process.

Mathematically, let p_A^t and p_B^t denote a pair of source and target F0 contours, respectively. The variable 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 $v_t(x; s)$ over frequency and time. The integral of these small displacements produces a final large vector field denoted by $\varphi_t^v = \int_0^{R_1} v_t(\cdot; s) ds$ [28]. Representing the momenta variable by m_p, the LDDMM objective function can be written as:

$$\Gamma(m_p) = \frac{1}{2} \sum_{i,j=1}^{T} \gamma_{ij} [m_p]_i [m_p]_j + \lambda \sum_{t=1}^{T} \| \varphi_t^v(p_A^t; 1) - p_B^t \|_2^2 \tag{10}$$

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

Rather than solving (10) 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, 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

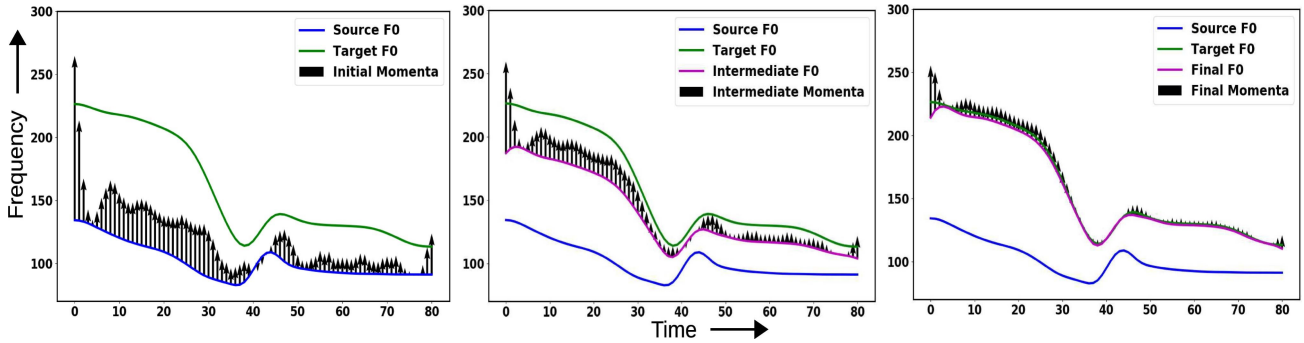

Fig. 2. Demonstration of the warping procedure. The leftmost figure shows the source and target F0 contours and the initial momenta. The middle figure shows the F0 contours at an intermediate time step. The rightmost figure is the final result of warping where modified F0 contour matches with the target F0.

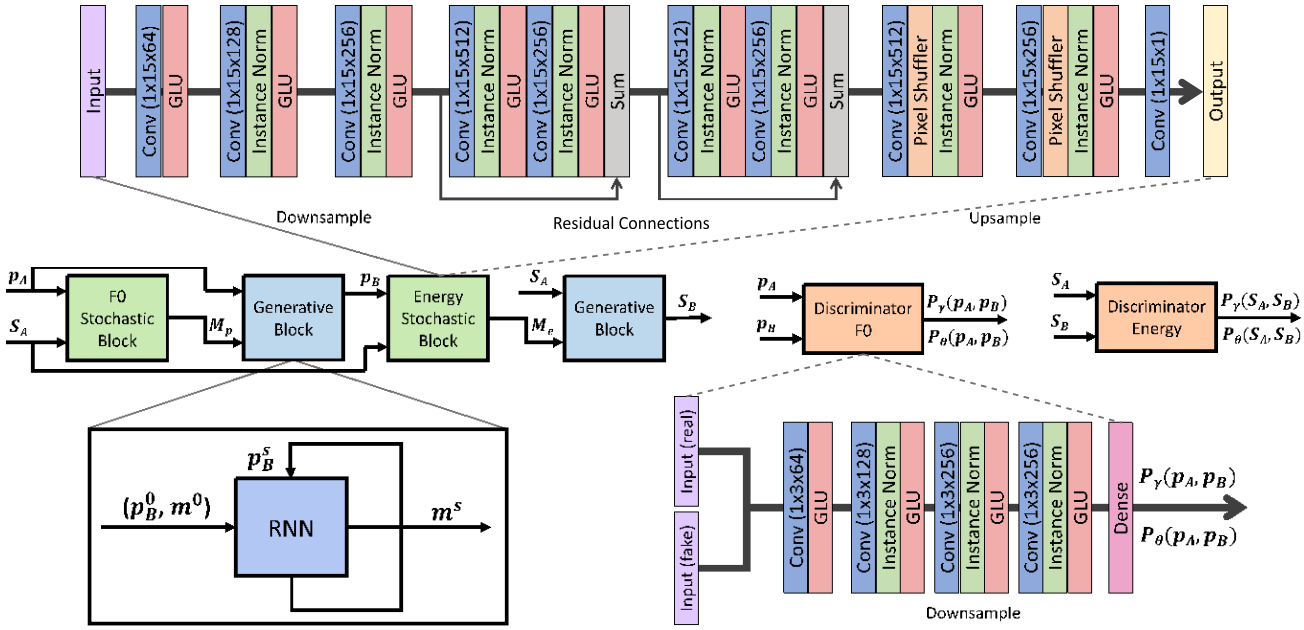

Fig. 3. Architecture of the neural network for F0 and energy prediction. The output of F0 prediction is fed as input for energy estimation. Each generator has two blocks: a stochastic block for sampling momenta and a generative/deterministic block for curve warping (represented as an RNN).

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.

C. 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 [31] for the stochastic block of the generators as shown in Fig. 3. It has been experimentally verified that fully convolutional networks are more stable in a GAN framework than including fully-connected layers [32]. 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 $L_m = E[k\nabla m_p k^2] + E[k\nabla m_e k^2]$ 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

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 loss of generator G_γ is as follows:

$$\begin{aligned} L_{G_\gamma} &= \lambda_{c_1} E\,[k p_A - p_{\phi} k] + \lambda_m E\, k \mathbb{m}_p k_2 + k \mathbb{m}_e k_2 \;{}_2 \\ &+ \lambda_i E\;\; k e_A - e_{\hat{A}} k \;\; + \lambda_{c_2} E\,[k e_A - e_{\hat{\phi}} k] \\ &+ \lambda_d E_{(S_A, p_A)}\;\; E_{(S_B, p_B) \mathbb{P} P_\gamma}\,[\log(D_\gamma(S_A, p_A, S_B, p_B))] \end{aligned} \tag{11}$$

In the case of energy contour modification, we add an identity loss to 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 [33] 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.

D. Discriminator Loss and Architecture

We model the probability ratio term in (9) by a discriminator denoted by D_γ. Conceptually, this discriminator distinguishes between the joint distributions of (S_A, p_A) and (S_B, p_B) learned by generators G_γ and G_ϑ, respectively.

During training of the discriminator D_γ, we minimize:

$$\begin{aligned} L_{D_\gamma} &= -E_{(S_A, p_A)} E_{(S_B, p_B) \mathbb{P} P_\gamma}\,[\log(D_\gamma(S_A, p_A, S_B, p_B))] \\ &- E_{(S_B, p_B)}\;\; E_{(S_A, p_A) \mathbb{P} P_\vartheta}\,[\log(1 - D_\gamma(S_A, p_A, S_B, p_B))] \end{aligned} \tag{12}$$

A similar discriminator has been proposed to train autoencoders in an adversarial setting [35], [36]. 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 (12) 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 $P_\gamma(S_A, p_A, S_B, p_B)$ and $P_\vartheta(S_A, p_A, S_B, p_B)$ (derived in Appendix B).

We split each discriminator in two partial discriminators which separately provide the feedback for F0 and energy conversion task (see Appendix B). 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 (see Appendix C). 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).

E. 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:

$$e_A = \sum_{f=0}^{\frac{F_s}{2}} [S_A]_f^t \tag{13}$$

where, f corresponds to the frequency and t is the time. Once the energy contour has been modified through the VCGAN, denoted as e_B, then the converted spectrum S_B is given by:

$$S_B = S_A \times \frac{e_B}{e_A} \tag{14}$$

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 5 ms 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

Algorithm 2: Training Procedure of VCGAN model.

Result: Parameters $(\theta_G, \theta_D, \gamma_G, \text{and } \gamma_D)$

while *not converged* **do**

  for $i = 1..n$ **do**

    Draw $(\mathbf{S}_A^i, \mathbf{p}_A^i)$ and $(\mathbf{S}_B^i, \mathbf{p}_B^i)$ from emotion classes A and B;

    Compute $\mathbf{e}_A^i$ and $\mathbf{e}_B^i$ from $\mathbf{S}_A^i$ and $\mathbf{S}_B^i$;

    Sample $\mathbf{m}_{p,AB}^i$ and $\mathbf{m}_{p,BA}^i$ using F0 momenta samplers ;

    Generate F0 contours $\tilde{\mathbf{p}}_B^i$ and $\tilde{\mathbf{p}}_A^i$;

    Sample $\mathbf{m}_{e,AB}^i$ and $\mathbf{m}_{e,BA}^i$ using energy momenta samplers ;

    Generate energy contours $\tilde{\mathbf{e}}_B^i$ and $\tilde{\mathbf{e}}_A^i$ and corresponding $\tilde{\mathbf{S}}_B^i$ and $\tilde{\mathbf{S}}_A^i$;

    Sample $\hat{\mathbf{m}}_{p,AB}^i$ and $\hat{\mathbf{m}}_{p,BA}^i$ using the F0 momenta samplers for cyclic conversion;

    Generate cycle converted F0 contours, $\hat{\mathbf{p}}_A^i$ and $\hat{\mathbf{p}}_B^i$;

    Sample $\hat{\mathbf{m}}_{e,AB}^i$ and $\hat{\mathbf{m}}_{e,BA}^i$ using the energy momenta samplers for cyclic conversion;

    Generate cycle converted energy contours $\bar{\mathbf{e}}_A^i$ and $\bar{\mathbf{e}}_B^i$;

    Sample identity converted $\bar{\mathbf{m}}_{e,AB}^i$ and $\bar{\mathbf{m}}_{e,BA}^i$ using original spectrum and F0 contours;

    Generate identity converted energy contours $\bar{\mathbf{e}}_A^i$ and $\bar{\mathbf{e}}_B^i$;

  end

  $\nabla_{G_\gamma} \leftarrow \frac{1}{n}\sum_{i=1}^{n}[\lambda_d \nabla \log(D_\gamma(\mathbf{S}_A^i, \mathbf{p}_A^i, \tilde{\mathbf{S}}_B^i, \tilde{\mathbf{p}}_B^i)) + \lambda_{c_1}\|\mathbf{p}_A^i - \hat{\mathbf{p}}^i\|_1 + \lambda_m(\,|\nabla \mathbf{m}_{AB}^i|^2 + \|\nabla \hat{\mathbf{m}}_{AB}^i\,|^2 + \|\nabla \bar{\mathbf{m}}_{AB}^i\|^2) + \lambda_i\|\mathbf{e}_A^i - \bar{\mathbf{e}}_A^i\|_1 + \lambda_{c_2}\|\mathbf{e}_A^i - \hat{\mathbf{e}}_A^i\|_1]$;

  $\nabla_{G_\theta} \leftarrow \frac{1}{n}\sum_{i=1}^{n}[\lambda_d \nabla \log(D_\theta(\mathbf{S}_B^i, \mathbf{p}_B^i, \tilde{\mathbf{S}}_A^i, \tilde{\mathbf{p}}_A^i)) + \lambda_{c_1}\|\mathbf{p}_B^i - \hat{\mathbf{p}}_B^i\|_1 + \lambda_m(\,|\nabla \mathbf{m}_{BA}^i\,|^2 + \|\nabla \hat{\mathbf{m}}_{BA}^i\,|^2 + \|\nabla \bar{\mathbf{m}}_{BA}^i\,|^2) + \lambda_i\|\mathbf{e}_B^i - \bar{\mathbf{e}}_B^i\|_1 + \lambda_{c_2}\|\mathbf{e}_B^i - \hat{\mathbf{e}}_B^i\|_1]$;

  $\nabla_{D_\gamma} \leftarrow \frac{1}{n}\sum_{i=1}^{n}[-\nabla \log(D_\gamma(\mathbf{S}_A^i, \mathbf{p}_A^i, \tilde{\mathbf{S}}_B^i, \tilde{\mathbf{p}}_B^i)) - \nabla \log(1 - D_\gamma(\tilde{\mathbf{S}}_A^i, \tilde{\mathbf{p}}_A^i, \mathbf{S}_B^i, \mathbf{p}_B^i))]$;

  $\nabla_{D_\theta} \leftarrow \frac{1}{n}\sum_{i=1}^{n}[-\nabla \log(D_\theta(\mathbf{S}_B^i, \mathbf{p}_B^i, \tilde{\mathbf{S}}_A^i, \tilde{\mathbf{p}}_A^i)) - \nabla \log(1 - D_\theta(\tilde{\mathbf{S}}_B^i, \tilde{\mathbf{p}}_B^i, \mathbf{S}_A^i, \mathbf{p}_A^i))]$;

  $\gamma_G \leftarrow \gamma_G - \eta_g \nabla_{G_\gamma}$;

  $\theta_G \leftarrow \theta_G - \eta_g \nabla_{G_\theta}$;

  $\gamma_D \leftarrow \gamma_D - \eta_d \nabla_{D_\gamma}$;

  $\theta_D \leftarrow \theta_D - \eta_d \nabla_{D_\theta}$;

end

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 [37] with an exponential decay of 0.5 for the first moment. Sampling process in the generators is implemented via dropout [38] rate of 0.3 during both training and testing.

III. 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 [23]. 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.

A. VESUS Dataset

We evaluate our algorithms on the VESUS dataset [21] 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) [39]. These ratings represent the ratio of workers who correctly identify the intended 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.

B. 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 [40]. 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 generated samples. Fig. 4 shows the difference between the generator and discriminator losses for the Cycle-GAN (orange) and our proposed VCGAN (blue). We note

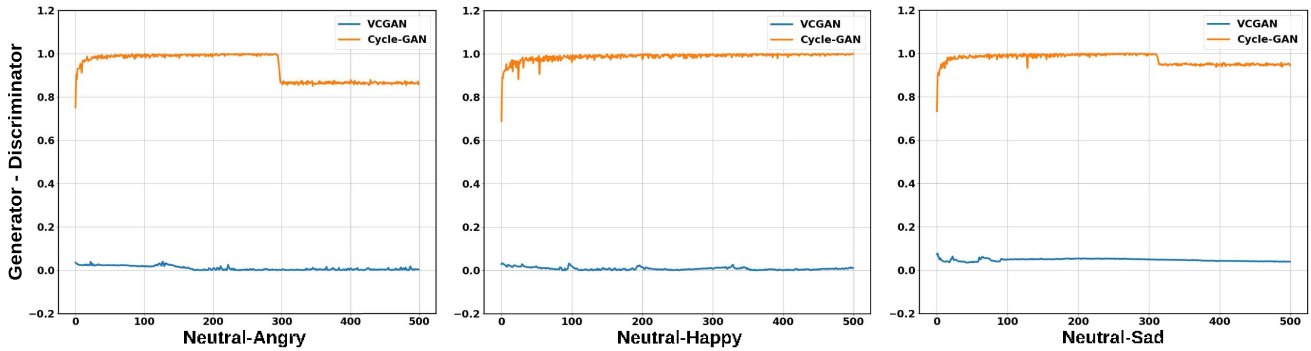

Fig. 4. Comparing Cycle-GAN with its variational counterpart. On Y-axis, we denote the difference between the generator and discriminator loss. On X-axis, we denote the number of epochs. The plots represent the mismatch between the adversarial losses which is an indicator of instability in training [34].

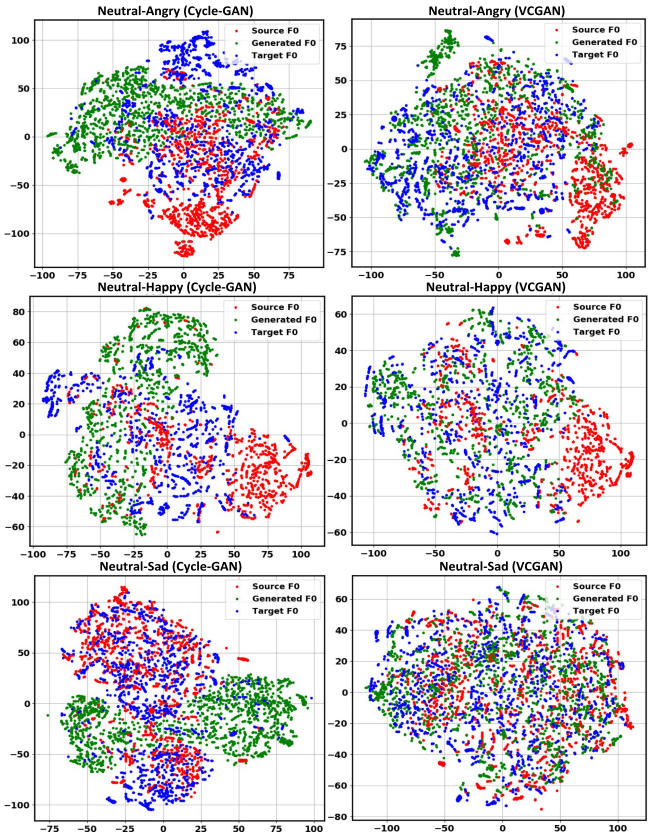

Fig. 5. Visualizing t-SNE embeddings of source, converted and target F0 contours. The left column shows the embeddings generated using Cycle-GAN and the right column shows the same for variational model.

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 [41].

To illustrate the improved generator calibration, Fig. 5 shows the tSNE plots [42] of the source, generated, and target emotion F0 values extracted over 640 ms 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.

C. *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 [16], [24] use a continuous wavelet transform to parameterize the F0 contour to stabilize its generative process. Empirically, we observe that our proposed momenta-based warping allows more flexible transformations and better scale matching between the generated and target contours. Fig. 6 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. 6).

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. 7) and energy (Fig. 8) contours. As seen, our momenta-based warping is significantly better than wavelet based regularization used

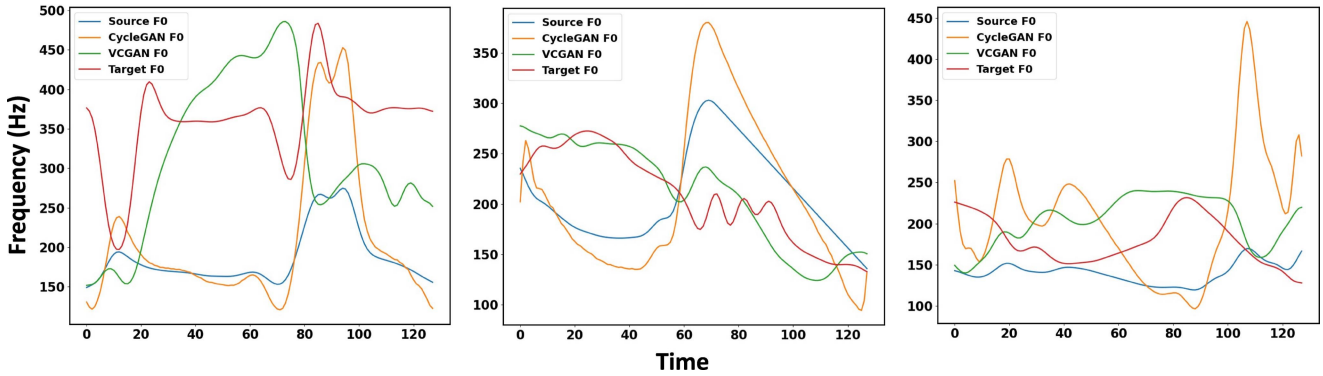

Fig. 6. Comparing the F0 contours generated by Cycle-GAN and our momenta regularized variational model. Using diffeomorphic warping as a regularizer leads to more stable F0 contour generation in comparison to wavelet based regularization.

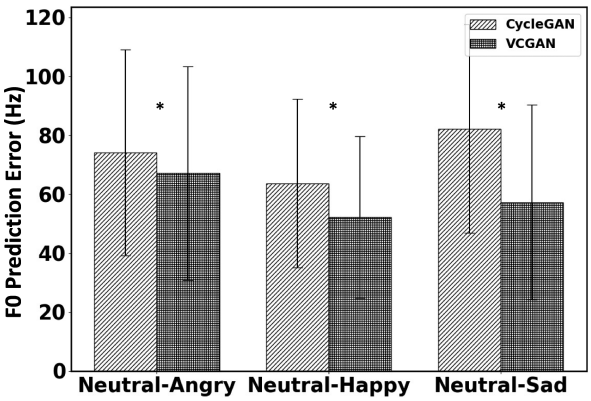

Fig. 7. F0 RMSE comparison between Cycle-GAN and VCGAN. The results are statistically significant at level 0.05 (* denote p-value $\leq 1e-10$).

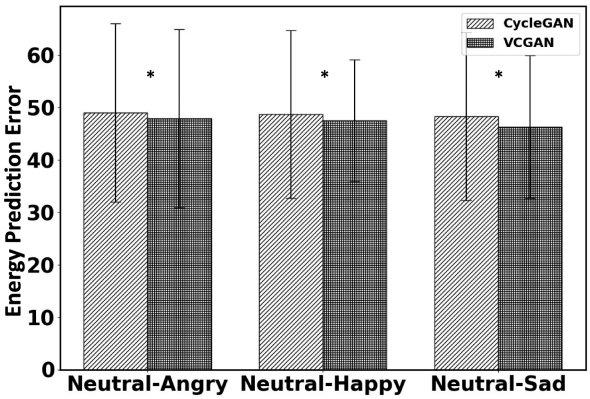

Fig. 8. Energy RMSE comparison between Cycle-GAN and VCGAN. Results are statistically significant at level 0.05 (* denote p-value $\leq 1e-10$).

in [16], [23] 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.

IV. 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.

A. 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 [11]. The second baseline uses a Bi-LSTM model [16] 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

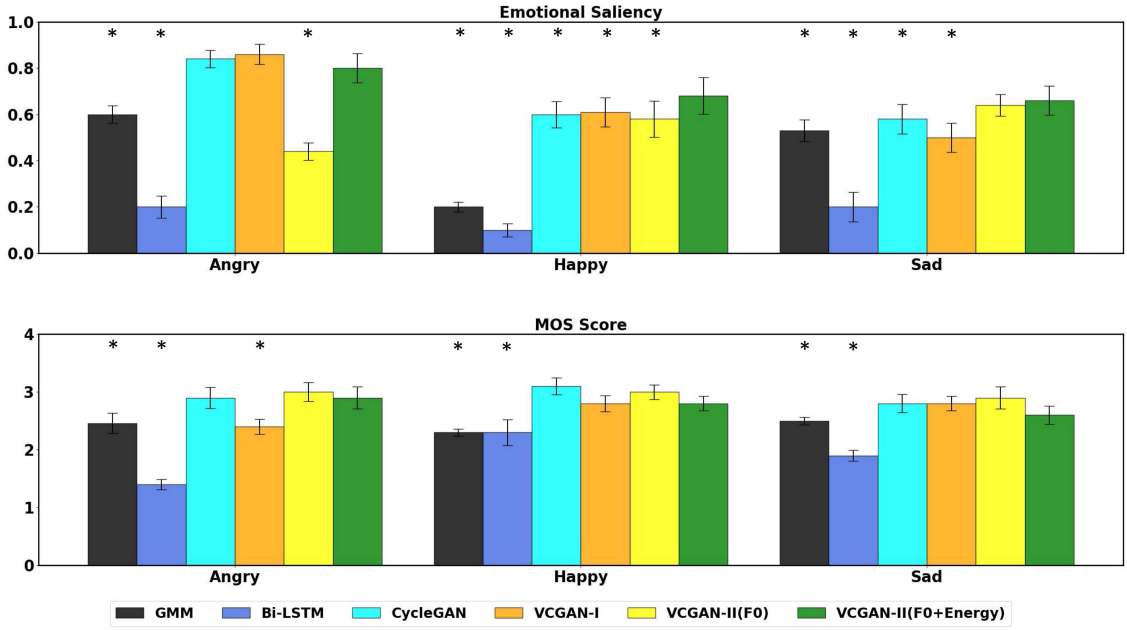

Fig. 9. **Single-Speaker Evaluation:** we create the training, validation and testing sets for each emotion pair based on the speaker from VESUS with the highest number of emotionally salient utterances. The asterisk (*) denotes statistical significance for the test (VCGAN-II (F0+Energy) > Method) at $p < 0.05$.

prosody. The third technique is a recently proposed Cycle-GAN framework [24] 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 [25] (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. We refer the reader to the Appendix C for detail descriptions of each baseline architecture, including size and number of parameters.

B. 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. 9 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 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.

C. 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. 10 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.

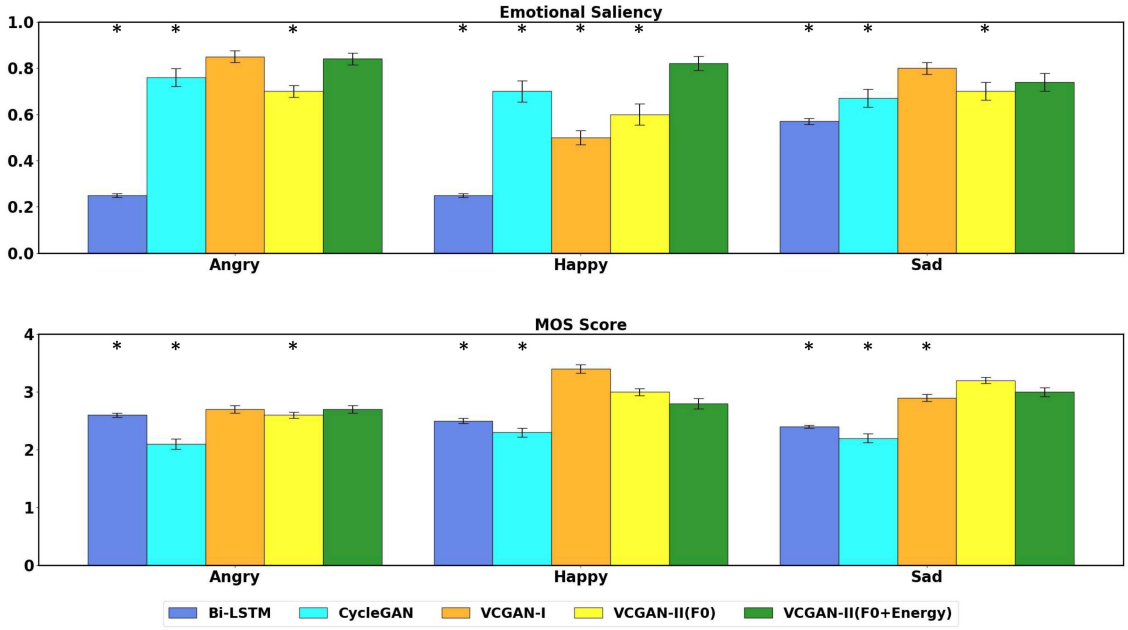

Fig. 10. **Mixed Speaker Evaluation:** we create the training, validation and testing sets by randomly sampling utterances from VESUS across all speakers. The asterisk (*) denotes statistical significance for the one-tailed t-test (VCGAN (F0+Energy) > Method) at $p < 0.05$.

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 map-ping, meaning it does not perform any emotion conversion, but simply reconstructs the (already high quality) input utterance. The Cycle-GAN [23] 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 least consistent.

TABLE I
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

D. 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 I.

We sample 10 utterances for each fold and each emotion pair to collect the final ratings. Fig. 11 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 mod-els (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

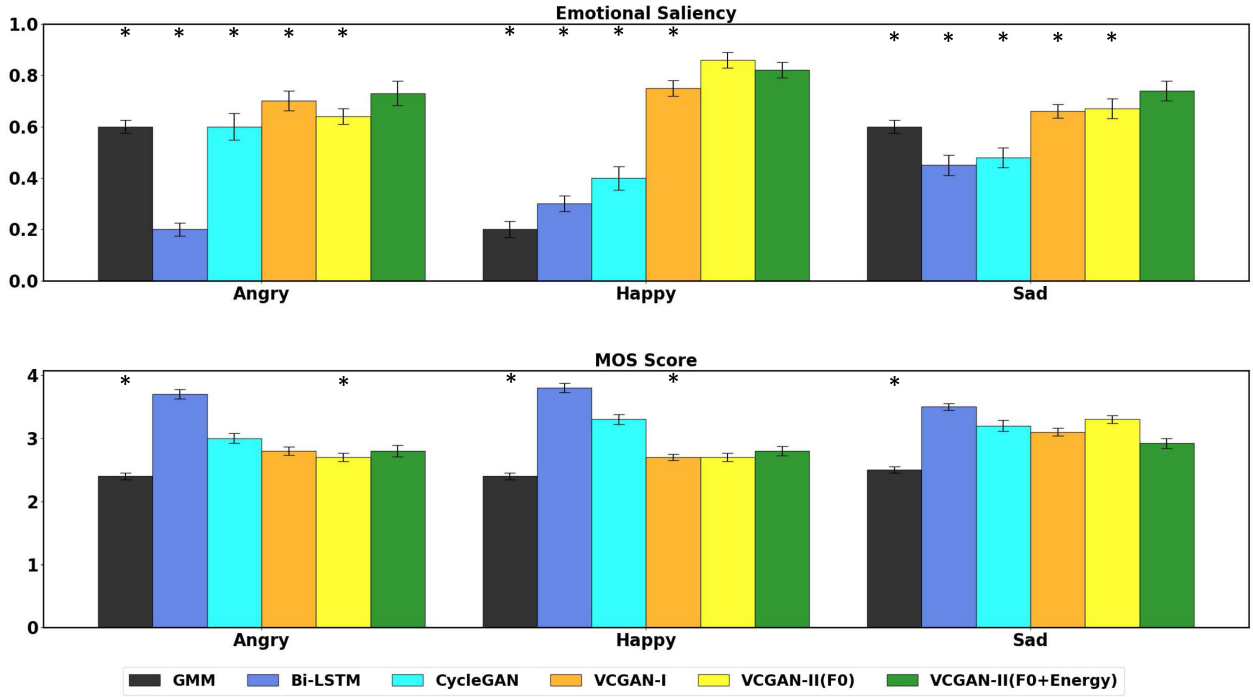

Fig. 11. **Out-of-Speaker Evaluation:** we create 5 folds from VESUS, each comprising of a male and a female speaker. We train the VCGAN model on four folds and evaluate its performance on the fifth. The asterisk (*) denotes statistical significance for the test (VCGAN-II (F0+Energy) > Method) at $p < 0.05$.

"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 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.

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

Fig. 12 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).

F. Summary of Results

Table II summarizes 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,

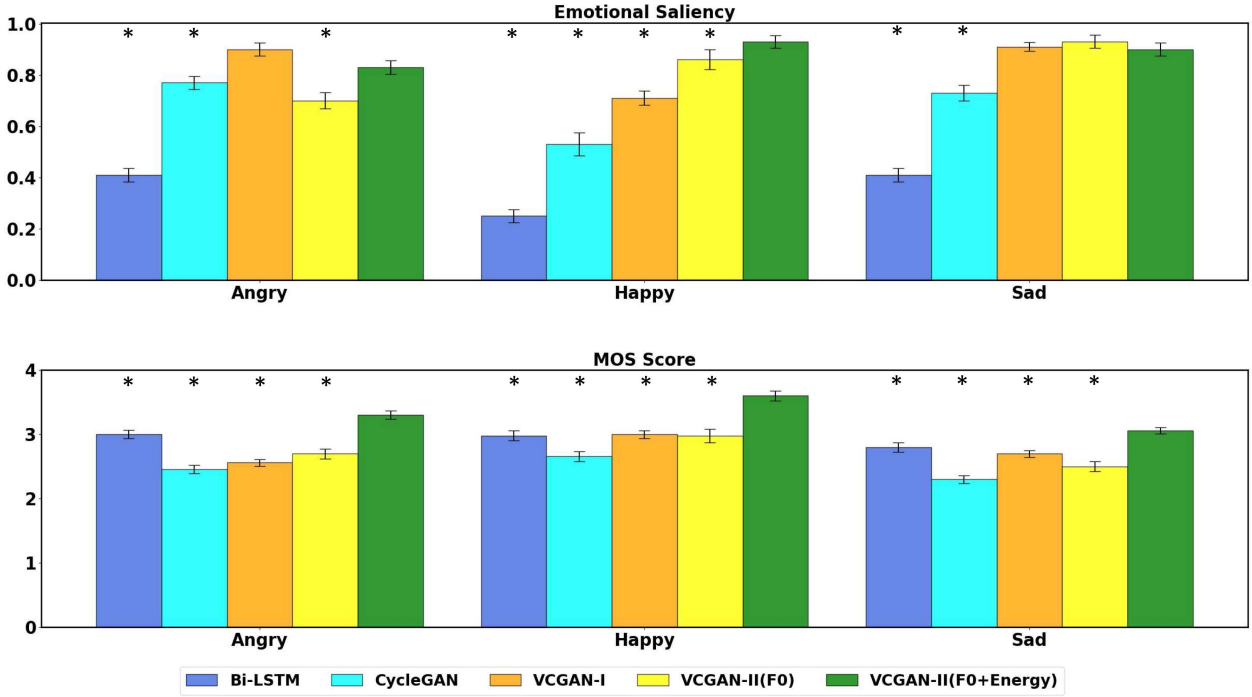

Fig. 12. **Wavenet Evaluation:** We apply our mixed speaker models (without fine-tuning) to modify speech generated by the Wavenet model. The asterisk (*) denotes statistical significance for the one-tailed t-test (VCGAN (F0+Energy) > Method) at $p < 0.05$.

TABLE II
PERFORMANCE ACROSS THE FOUR EVALUATION PARADIGMS: SINGLE-SPEAKER, MIXED-SPEAKER, OUT-OF-SPEAKER, AND WAVENET

Evaluation	Algorithm	Neutral-angry		Neutral-happy		Neutral-sad	
		Acc.	MOS	Acc.	MOS	Acc.	MOS
Single Speaker	GMM [11]	0.6±0.1	2.5±0.6	0.2±0.1	2.3±0.2	0.53±0.2	2.5±0.4
	Bi-LSTM [16]	0.2±0.2	1.4±0.3	0.1±0.1	2.3±0.7	0.2±0.2	1.9±0.3
	Cycle-GAN [23]	0.84±0.1	2.9±0.6	0.6±0.2	3.1±0.5	0.58±0.2	2.8±0.5
	VCGAN-I [25]	**0.86±0.1**	2.4±0.4	0.6±0.2	2.8±0.4	0.5±0.2	2.8±0.4
	VCGAN-II(F0)	0.44±0.1	**3.0±0.5**	0.58±0.2	**3.0±0.4**	0.64±0.2	**2.9±0.6**
	VCGAN-II(F0+Energy)	0.8±0.2	2.9±0.6	**0.68±0.2**	2.8±0.4	**0.66±0.2**	2.6±0.5
Mixed Speaker	Bi-LSTM [16]	0.25±0.1	2.6±0.3	0.25±0.1	2.5±0.3	0.57±0.1	2.4±0.2
	Cycle-GAN [23]	0.76±0.3	2.1±0.7	0.7±0.3	2.3±0.5	0.67±0.3	2.2±0.6
	VCGAN-I [25]	**0.85±0.2**	**2.7±0.5**	0.5±0.2	**3.4±0.5**	0.8±0.2	2.9±0.5
	VCGAN-II(F0)	0.7±0.2	2.6±0.4	0.6±0.3	3.0±0.4	0.7±0.3	**3.2±0.4**
	VCGAN-II(F0+Energy)	0.84±0.2	**2.7±0.5**	**0.82±0.2**	2.8±0.6	**0.74±0.3**	3.0±0.6
Out-of-Speaker	GMM [11]	0.6±0.2	2.4±0.4	0.2±0.2	2.4±0.4	0.6±0.2	2.5±0.4
	Bi-LSTM [16]	0.2±0.2	**3.7±0.6**	0.3±0.2	**3.8±0.6**	0.45±0.3	**3.5±0.4**
	Cycle-GAN [23]	0.6±0.4	3.0±0.6	0.4±0.3	3.3±0.6	0.48±0.3	3.2±0.7
	VCGAN-I [25]	0.7±0.3	2.8±0.5	0.75±0.2	2.7±0.4	0.66±0.2	3.1±0.5
	VCGAN-II(F0)	0.64±0.2	2.7±0.5	**0.86±0.2**	2.7±0.5	0.67±0.3	3.3±0.5
	VCGAN-II(F0+Energy)	**0.73±0.3**	2.8±0.7	0.82±0.2	2.8±0.6	**0.74±0.3**	2.9±0.6
Wavenet	Bi-LSTM [16]	0.41±0.2	3.0±0.5	0.25±0.2	2.98±0.5	0.41±0.2	2.8±0.6
	Cycle-GAN [23]	0.77±0.2	2.46±0.5	0.53±0.3	2.66±0.5	0.73±0.2	2.3±0.5
	VCGAN-I [25]	**0.9±0.2**	2.56±0.4	0.71±0.2	3.0±0.4	0.9±0.1	2.7±0.4
	VCGAN-II(F0)	0.7±0.24	2.7±0.6	0.86±0.3	2.98±0.7	**0.93±0.2**	2.5±0.6
	VCGAN-II(F0+Energy)	0.83±0.2	**3.3±0.5**	**0.93±0.2**	**3.6±0.5**	0.9±0.2	**3.1±0.4**

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. 9–12 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 II, 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. 1) is another notable aspect of our approach; empirically, it helps preserve the naturalness of the modified speech.

V. Conclusion

In this paper, 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] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernandez-Dols, "Facial and vocal expressions of emotion," *Annu. Rev. Psychol.*, vol. 54, pp. 329–349, Nov. 2003.

[2] D. Schacter, D. T. Gilbert, and D. M. Wegner, *Psychology*, 2nd ed. Duffield, U.K.: Worth Publications, 2011.

[3] M. Swerts and E. J. Krahmer, "On the use of prosody for on-line evaluation spoken dialogue systems," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2000.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.

[5] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3502–3506.

[6] J. Schmidt, E. Janse, and O. Scharenborg, "Perception of emotion in conversational speech by younger and older listeners," *Front. Psychol.*, vol. 7, 2016, Art. no. 781.

[7] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 490–493.

[8] Y. Hashizawa, S. 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. 2nd Int. Conf. Speech Prosody*, 2004, pp. 655–658.

[9] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[11] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *Amer. J. Signal Process.*, vol. 2, pp. 134–138, Dec. 2012.

[12] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *Proc. IEEE Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2014, pp. 1–7.

[13] T. Virtanen, B. Raj, J. F. Gemmeke, and H. Van Hamme, "Active-set newton algorithm for non-negative sparse coding of audio," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 3092–3096.

[14] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*,vol. 5, pp. 1457–1469, 2004. [Online]. Available: http://arxiv.org/abs/cs.LG/0408058

[15] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[16] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2453–2457.

[17] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*.
[18] R. Shankar, J. Sager, and A. Venkataraman, "A multi-speaker emo-tion morphing model using highway networks and maximum likelihood objective," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2848–2852.
[19] R. Shankar, H.-W. Hsieh, N. Charon, and A. Venkataraman, "Automated emotion morphing in speech based on diffeomorphic curve registration and highway networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4499–4503.
[20] R. Shankar, H.-W. Hsieh, N. Charon, and A. Venkataraman, "Multi-speaker emotion conversion via latent variable regularization and a chained encoder-decoder-predictor network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3391–3395.
[21] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "VESUS: A crowd-annotated database to study emotion production and perception in spoken english," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 316–320.
[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.
[23] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2858–2862.
[24] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Odyssey 2020 Speaker Lang. Recognit. Workshop*, 2020, pp. 230–237.
[25] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3396–3400.
[26] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, pp. 1877–1884, Jul. 2016.
[27] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical im-age registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.
[28] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, pp. 139–157, 2005.
[29] S. C. Joshi and M. I. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000.
[30] H.-W. Hsieh and N. Charon, "Diffeomorphic registration of discrete geometric distributions," Lecture Notes Series, Institute for Mathematica Sciences, National University of Singapore, 2019.
[31] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," 2017, *arXiv:1711.11293*.
[32] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Representations*, 2016.
[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[34] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
[35] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. A. Sutton, "VEEGAN: Reducing mode collapse in GANs using implicit variational learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3308–3318.
[36] V. Dumoulin et al., "Adversarially learned inference," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ElR4cgg
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.
[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html
[39] Amazon, "Amazon mechanical turk," 2022. [Online]. Available: https://www.mturk.com
[40] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: GANs do not need to decrease a divergence at every step," in *Proc. Int. Conf. Learn. Representations*, 2018.
[41] R. Durall, A. Chatzimichailidis, P. Labus, and J. Keuper, "Combating mode collapse in GAN training: An empirical analysis using Hessian eigenvalues," 2020, *arXiv:2012.09673*. [Online]. Available: https://arxiv.org/abs/2012.09673
[42] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Ravi Shankar (Student Member, IEEE) received the B.Tech. degree from the Indian Institute of Technology, Guwahati, India, in 2015. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA. His research interests include domain of generative modeling for expressive speech synthesis and emotion conversion in speech. He was the recepient of ISCA Travel Award (2020) and MINDS research fellowship for data science (2020). He has also was a Reviewer for Interspeech, NeuRips, ICLR and CISS in the past.

Hsi-Wei Hsieh received the Ph.D. degree in applied mathematics statistics from Johns Hopkins University, Baltimore, MD, USA, under the supervision of Dr. Nicolas Charon. He is currently a Peter ODonnell, Jr. Postdoctoral Fellow with Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX, USA. He was with the Scientific AI Research Group led by Prof. Yen-Hsi Richard Tsai. Prior to this position,

Nicolas Charon is currently an Assistant Professor of applied mathematics and statistics, he develops mathematical and numerical models for the representation, registration, and statistical analysis of geometric structures occurring in the anatomy and in biology. His research has innovative applications to the modeling and analysis of the shape of curves and surfaces extracted from medical images, in particular. He is also interested in the interplay between anatomy and function in datasets that combine geometry with additional measured signals, such as response to stimuli or tissue thickness. He is creating new analysis tools to adequately model and estimate joint variabilities in shape and function. He also works in developing computational methods to compress and accelerate the vast amounts of data generated by diffusion MRI (dMRI), an imaging protocol that retrieves the diffusion patterns within the brain and is key to understanding its connectivity. His work provides researchers and clinicians new computational tools for fast reconstruction and analysis of diffusion data.

Archana Venkataraman received the B.S., M.Eng., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology Cambridge, MA, USA, in 2006, 2007, and 2012, respectively. She is currently a John C. Malone Assistant Professor with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA. She directs the Neural Systems Analysis Laboratory, and is a core Faculty Member of the Malone Center for Engineering with Healthcare and the Mathematical Institute for Data Science. Her research interests include intersection of artificial intelligence, network modeling and clinical neuroscience. Her work has yielded novel insights in to debilitating neurological disorders, such as autism, schizophrenia, and epilepsy, with the long-term goal of improving patient care. She was the recipient of the MIT Provost Presidential Fellowship, the Siebel Scholarship, the National Defense Science and Engineering Graduate Fellowship, the NIH Advanced Multimodal Neuroimaging Training Grant, the CHDI Grant on network models for Huntington's Disease, numerous Best Paper awards, and the National Science Foundation CAREER award. Dr. Venkataraman was also named by MIT Technology Review as one of 35 Innovators Under 35 in 2019.