Non-Parallel Training Approach for Emotional Voice Conversion Using CycleGAN

Mohamed Elsayed[1], Sama Hadhoud[1], Alaa Elsetohy[1], Menna Osman[1] and Walid Gomaa[1,2]
[1]*Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt*
[2]*Faculty of Engineering, Alexandria University, Alexandria, Egypt*

Keywords: Emotional Voice Conversion, CycleGAN, World Vocoder, Non-Parallel.

Abstract: The focus of this research is proposing a nonparallel emotional voice conversion for Egyptian Arabic speech. This method aims to change emotion-related features of a speech signal without changing its lexical content or speaker identity. We relied on the assumption that any speech signal can be divided into content and style code and the conversion between different emotion domains is done by combining the target style code with the content code of the input speech signal. We evaluated the model using an Egyptian Arabic dataset covering two emotion domains and the conversion results were successful depending on a survey conducted on random people. Our purpose is to produce a state-of-the-art pre-trained model as it will be an unprecedented model in the Egyptian Arabic language as far as we are concerned.

1 INTRODUCTION

Voice conversion (VC) focuses on extracting the acoustic features of the source voice and then, mapping them to those of the target voice. After that, the waveforms are synthesized from the generated acoustic features. Pre-processing is done before training the mapping function. Such pre-processing includes using dynamic time warping (DTW) which is used to time align between the source and target voice features under study. This is done when working on parallel data. Emotional voice conversion approach is similar to that of the VC, with the important role of prosody to express emotions in speech (Choi and Hahn, 2021). Emotional voice conversion focuses on converting the emotion-related features from the source emotion domain to that in the target domain while preserving the linguistic content and speaker identity. Such emotion-related features include prosodic and spectrum-related features.

Emotional voice conversion has various applications in conversational agents, intelligent dialogue systems, and other expressive speech synthesis applications (Luo et al., 2017). Additionally, it is promising for applications in human-machine interaction, such as enabling robots to respond to people with emotional intelligence (Olaronke and Ikono, 2017). Speech not only conveys information but also shows one's emotional state.

Given the significance of emotions in communication, we focused on emotion-voice transformation in this work. Prosodic features like pitch, intensity, and speaking rate can be used to help identify emotions (Scherer et al., 1991). Emotional voice conversion aims to change emotion-related features of a speech signal without changing its lexical content or speaker identity (Choi and Hahn, 2021).

It is proposed in (Huang and Akagi, 2008) that the perception of emotion is multi-layered. Thus, from top to bottom, the layers are represented by emotion categories, semantic primitives, and acoustic features. It is further suggested that emotion production and perception are inverse processes. An inverse three-layered model for speech emotion production was proposed in (Xue et al., 2018). Studies on speech emotion generally utilize prosodic features concerning voice quality, speech rate, fundamental frequency (F_0), spectral features, duration, F_0 contour, and energy envelope (Schröder, 2006).

Early studies on emotional voice conversion mostly relied on parallel training data, or a pair of utterances that contain the same content but with different emotions from the same speaker. Through the paired feature vectors, the conversion model learns mapping from the source emotion A to the target emotion B during training. The authors in (Tao et al., 2006) mainly addressed prosody conversion by decomposing the pitch contour of the source speech using classification and regression trees, then utilizing Gaussian Mixture Model (GMM) and regression-based clustering techniques.

Recent works proposed deep learning approaches

Elsayed, M., Hadhoud, S., Elsetohy, A., Osman, M. and Gomaa, W.
Non-Parallel Training Approach for Emotional Voice Conversion Using CycleGAN.
DOI: 10.5220/0012156000003543
In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2023) - Volume 2*, pages 17-24
ISBN: 978-989-758-670-5; ISSN: 2184-2809

achieving remarkable performance. In (Luo et al., 2016), an emotional voice conversion model is proposed. This model is divided into two parts. In the first part, Deep Belief Networks (DBNs) are used to modify spectral features, while in the second one Neural Networks (NNs) are used to modify the fundamental frequency (F_0). The STRAIGHT-based approach was used for extracting features from the source voice signal and the destination speech signal while introducing the spectral conversion part and F_0 conversion part. This model successfully adjusts both the acoustic voice and the prosody for the emotional voice simultaneously when compared to traditional approaches (NNs and GMMs). However, this work and other recent emotional voice conversion techniques require temporally aligned parallel samples, which is very difficult to attain in practical applications. Additionally, accurate time alignment requires manual segmentation of the speech signal, which is also time-consuming.

Beyond the parallel training data, new methods for learning the translation across emotion domains utilizing CycleGAN and StarGAN have been developed (Gao et al., 2019). In this paper*, we used the CycleGAN architecture, which switches between two emotion domains rather than learning a one-to-one mapping between pairs of emotional utterances. The training approach for the model is speaker-dependent. It depends on extracting the emotion-related speech features using WORLD vocoder. Such features include the fundamental frequency (F_0) and spectral envelope. Gaussian normalization is applied to the F_0 whereas, spectral envelope is introduced to the auto-encoder model. Disentanglement is applied to separate the content code from the style code. This is very beneficial to preserve the lexical and speaker identity during the learning process. A survey was conducted on random people in which people were exposed to original and converted samples. The survey included two emotions, the neutral and the angry, accuracy of convergence to neutral domain is about 63.42% whereas it is 56.19% for convergence to the angry domain.

The paper is organized as follows. Section 1 is an introduction that gives a general overview of the subject under research. Section 2 covers related work. In Section 3, we introduce the structure of our model, the loss function used in the training, and the experimental setup followed by a description on the training dataset. In Section 4, we discuss the results obtained from the training. Dashboard graphs are used to demonstrate these outputs. Section 5 shows the survey outputs and an analysis of the results. Section 6 summarizes our work and gives general ideas about future work.

*The conversion system code is uploaded to the following repository: https://github.com/MohamedElsayed-22/non-parallel-training-for-emotion-conversion-of-arabic-speech-using-cycleGAN-and-WORLD-Vocoder.

2 RELATED WORK

Earlier conversion models changed prosody-related emotional features directly. According to (Xue et al., 2018), the acoustic features of spectral sequence and F_0 have the highest effects in converting emotions, with duration and the power envelope having the least impact. Traditional approaches to conversion include modeling spectral mapping using statistical techniques such as partial least squares regression (Helander et al., 2010), sparse representation (Sisman et al., 2019), and Gaussian Mixture models (GMMs) (Toda et al., 2007).

2.1 Emotion in Speech

Emotion is introduced in both the spoken linguistic content and acoustic features (Zhou et al., 2021). Acoustic features play the important role in human interactions. Emotions can be characterized by either the categorical or the dimensional representation.

The categorical approach is easier and more straightforward where emotions are labeled and the model architecture is built based on that. Ekman's six basic emotions theory (Ekman, 1992) is one of the most famous categorical approaches, where emotions are categorized into 6 categories which are anger, disgust, fear, happiness, sadness, and surprise. The downside of the categorical approach is that it ignores the minor differences in human emotions. However, the dimension-representation approach models the physical properties of emotion-related features. Russell's circumplex model is one of the models that represent emotion in terms of arousal, valence, and dominance. In this paper, we use the categorical way of representing emotions. The process of generating emotion from the speaker is opposite to that of the emotion perception of the listener. Consequently, we assume that the encoding of the speaker's emotion is the opposite process of the listener's decoding of that emotion.

In (Gao et al., 2019), an approach was introduced to learn a mapping between the distributions of of two utterances from distinct emotion categories, $x_1 \in X_1$ and $x_2 \in X_2$. However, the joint distribution $p(x_1, x_2)$ cannot be directly determined for nonparallel data. Therefore, they assume that the speech signal can be broken down into an *emotion-invariant* content component and an *emotion-dependent* style component and that the encoder E and the decoder D are inverse functions. These assumptions allow estimating $p(x_1|x_2)$ and $p(x_2|x_1)$ conversion models.

2.2 Disentanglement

Based on the results obtained from *disentangled representation learning* in image style transfer (Gatys et al., 2016), (Huang et al., 2018) , Such approach can be utilized similarly in speech. Each speech signal taken from either domain X_1 or X_2 can be divided into a content code (C) and a style code (S), as proposed in (Gao et al., 2019). The content code carries emotion-independent information, while the style code represents emotion-dependent information.

The content code is shared across domains and contains the data that we wish to retain. The style code is domain-specific and includes the data we wish to alter. As shown in fig. 1, we take the content code of the source speech and merge it with the style code of the target emotion during the conversion stage.

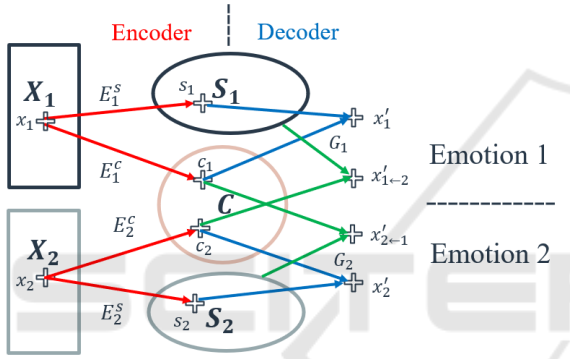

Figure 1: Nonparallel training inspired by disentanglement (Gao et al., 2019).

2.3 WORLD Vocoder

WORLD (Morise et al., 2016), fig. 2, a vocoder-based system, is a high-quality real-time speech synthesis system. It consists of three algorithms for analysis through retrieving three speech parameters, in addition to a synthesis algorithm that takes these parameters as inputs. The fundamental frequency contour (F_0), spectral envelope, and excitation signal are the parameters obtained. The F_0 estimation algorithm is DIO (Morise et al., 2009). The spectral envelope is determined by Cepstrum and linear predictive coding algorithms (LPC) (Atal and Hanauer, 1971). Those two parameters, as mentioned before, are emotion-related speech features. They are taken as input to the auto-encoder and the Gaussian normalization for the process of conversion. Furthermore, the outputs are taken to the synthesis part of the WORLD to retrieve the resulted converted voice.

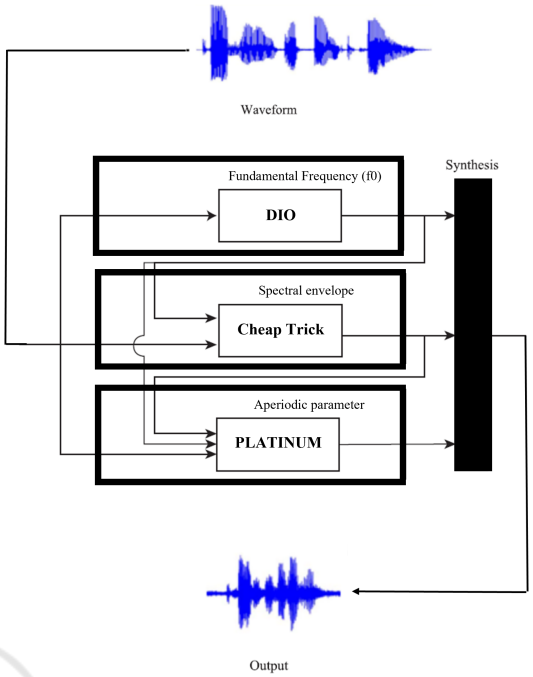

Figure 2: WORLD Vocoder Working Mechanism (Morise et al., 2016).

2.4 CycleGAN

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have produced remarkable results in various fields like image processing, computer vision, and sequential data (Gui et al., 2021). The main goal is to generate new data based on the training data, which is done through concurrently training two models: a generative model, the generator G, which captures the data distribution, and a discriminative model, the discriminator D, which classifies/decides whether a given sample is real or fake. The learning behavior of G is designed to maximize the probability of D making a mistake. This typically produces a two-player *minimax* game. In this paper, we are concerned with a particular GAN-based network architecture called CycleGAN (Zhu et al., 2017). CycleGAN learns mapping $G_{X\to Y}$ from a source domain X to a target domain Y without the need for parallel data. It is also combined with an inverse mapping $G_{Y\to X}$.

CycleGAN is based on two losses: the first one is adversarial loss, which is defined as follows.

$$L_{ADV}(G_{X\to Y}, D_Y, X, Y) = E_{y\sim P(y)}[D_Y(y)] + E_{x\sim P(x)}[log(1 - D_Y(G_{X\to Y}(x)))] \quad (1)$$

It shows how matching the distribution of generated images $G_{X\to Y}(x)$ is to the distribution in the target domain y.Thus, as long as the distribution of the generated images $P(x)$ to that of the target domain $P(y)$

becomes closer, the loss will consequently become smaller. The generator $G_{X\to Y}$ works on generating images while trying to maximize the error of the discriminator D_Y, as mentioned before, whereas, D_Y works in the opposite direction. The second loss is cycle consistency loss, which is defined using the $L1$ norm, is as follows.

$$\begin{aligned} &L_{CYC}(G_{X\to Y}, G_{Y\to X}) \\ &= E_{y\sim P(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1] \\ &+E_{x\sim P(x)}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1] \end{aligned} \tag{2}$$

It focuses on preventing the learned mappings $G_{X\to Y}$ and $G_{Y\to X}$ from contradicting each other. As explained in (Kaneko and Kameoka, 2018) working on the adversarial loss will not guarantee that the core information of X and $G_{X\to Y}(x)$ are preserved.

X, Y are for the domain, while x, y are for the samples, as explained earlier in section 2.1: $x_1 \in X_1$ and $x_2 \in X_2$ This is because the main target of the adversarial loss function is to guarantee whether $G_{X\to Y}(x)$ follows the distribution of the target domain. Cycle-consistency loss function focuses on making $G_{X\to Y}$ and $G_{Y\to X}$ find a pair (x, y) having the same core information.

An identity mapping loss is also introduced in (Kaneko and Kameoka, 2018).

$$\begin{aligned} L_{ID}(G_{X\to Y}, G_{Y\to X}) &= E_{y\sim P(y)}[\|G_{X\to Y}(y) - y\|] \\ &+E_{x\sim P(x)}[\|G_{Y\to X}(x) - x\|] \end{aligned} \tag{3}$$

To achieve identity transfer, the features we are concerned about must be transferred without modification from the source domain to the target domain. $G_{X\to Y}$ and its inverse generator $G_{Y\to X}$ are directed to find a mapping to achieve this target.

3 EXPERIMENTAL WORK

3.1 Methodology

In this paper, we propose a nonparallel training approach for the Egyptian Arabic Language emotional voice conversion, since using the traditional parallel approach is inefficient and infeasible to create parallel datasets. Thus, we concentrated on learning the spectral sequence and the fundamental frequency F_0 conversion using the CycleGAN and Gaussian Normalization models. Moreover, the use of cycleGAN architecture is due to its breakthrough in style transfer in images, a scope similar to that of our proposed problem, mentioned in Section 2.4. The concept of disentanglement, Section 2.2, is used to separate the emotion-related features, style code (S), from the speech content (C) thus, training is undergone on these features separately without affecting the content (C). We used World Vocoder, Section 2.3, which extract the emotion-related features that are fed to the training architecture, described in detail in Section 3.3.

3.2 Non-Parallel Training

In a nonparallel training method, we train a conversion model between two partially shared emotion domains ($X_1 : X_2$). Unlike the parallel training approach which depends on the mapping between two utterances which are the same for all features except for the aspects under study. In this way, nonparallel training is more practical and of less cost making it feasible for industrial applications. Parallel training requires time alignment of the samples. This can be done manually, which is difficult for large datasets, or using dynamic time warping (DTW) (Berndt and Clifford, 1994) which depends on pattern detection in time series to match word templates against the waveform of discrete time series of the continuous-waveform voice samples.

The approximate word matching that DTW relies on can be a drawback for their work, as it might include a wide range of pronunciations and map them to the same word, even though each pronunciation may carry a different emotion. This can lead to inaccuracies in emotional conversion.

3.3 Model

The conversion system, shown in fig. 3, takes the emotion-related features, more specifically fundamental frequency F_0 and spectral sequence, of both the source and target domains which are extracted by WORLD vocoder. The F_0 of the source emotion domain is then transformed by a linear transformation using log Gaussian normalization to match the F_0 of the target emotion domain.

Aperiodicity, which is analyzed from the input sample, is mapped directly since it is not one of the features under study. Low-dimensional representation of spectral sequence in mel-cepstrum domain is introduced to the auto-encoder. Gated CNN is used to implement the used encoders and decoders. A GAN module is used to produce realistic spectral frames. In this way, each feature is converted separately without being dependent on other features under study. Lastly, the converted emotion-related features are introduced to the World vocoder for recombining them back with the content code of the speech signal.

The introduced neural network architecture consists of an autoencoder, content encoder, style encoder,

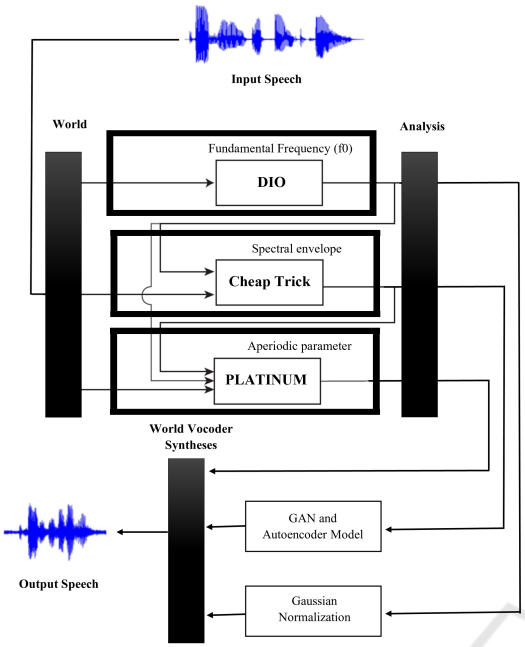

Figure 3: Voice emotion conversion mechanism.

decoder, and GAN discriminator. They work sequentially to generate the output as shown in fig. 4.

For an emotional speech signal x_i that is mapped between two partially shared emotion domains (X_1, X_2), Instance normalization (Ulyanov et al., 2016) was used to remove the emotional style mean and variance. The content encoder E_i^c is responsible for extracting the content code c_i of signal sample x_i:

$$E_i^c(x_i) = c_i \tag{4}$$

A 3-layer MLP (Huang and Akagi, 2008) was used to encode the emotional characteristics, The style encoders E_i^s is responsible for extracting the style code S_i of signal sample x_i:

$$E_i^s(x_i) = s_i \tag{5}$$

The decoder $G_i(c_i, s_j)$ is responsible for recombining the content code from one emotion with the style code of another, for example, it uses c_1 and s_2 to get $x'_{2\leftarrow 1}$:

$$G_i(c_i, s_j) = x'_{j\leftarrow i} \tag{6}$$

Note that the style code is learned from the entire emotion domain. This was accomplished by adding an adaptive instance normalization layer (Huang and Belongie, 2017).

Finally, The GAN discriminator is responsible for distinguishing real samples from machine-synthesized samples.

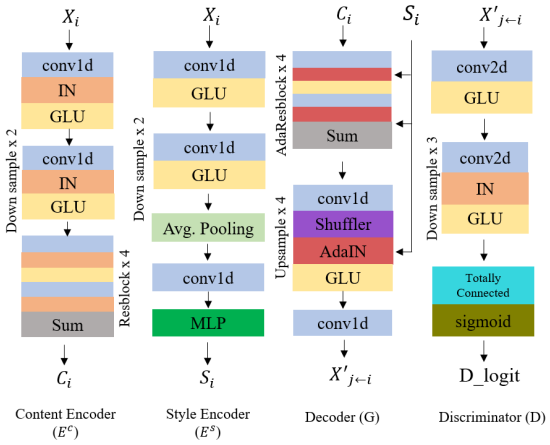

Figure 4: Network Structure.

3.4 Loss Function

The auto-encoder, fig. 1, consists of an encoder and a decoder thus, to keep them inverse operations to each other (Gao et al., 2019). Reconstruction loss is applied in the direction of $x_i \rightarrow (c_i, s_i) \rightarrow x'_i$. The reconstruction loss, as calculated in 7 where $\mathbb{E}$ represents the expected value, focuses on quantifying the model ability to regenerate the original sample x_i from the synthesized sample x'_i in the same domain.

$$L_{rec}^{x_i} = \mathbb{E}_{x_i}(\|x_i - x'_i\|_1) \tag{7}$$

The spectral sequence ought to remain unchanged following the process of encoding and decoding:

$$x'_i = G_i(E_i^c(x_i), E_i^s(x_i)) \tag{8}$$

The latent space is partially shared between the two emotions, specifically the content code is the shared space, so semi-cycle consistency loss is preferred which is applied in the direction of encoding. In content code, the content of an arbitrary sample $x_1 \in X_1$ represented as c_1 is coded to that of the equivalent sample $x'_{2\leftarrow 1}$ in the target domain X_2, this is done as follows: $c_1 \rightarrow x'_{2\leftarrow 1} \rightarrow c'_{2\leftarrow 1}$. The coding direction can not be represented as $x''_{1\leftarrow 2\leftarrow 1} = x_1$ because we take the semi-cycle consistency loss approach as the latent space of the content code is shared and such coding direction will lead to changes to the content of the speech. In the style code, the following coding direction is implemented as well: $s_1 \rightarrow x'_{2\leftarrow 1} \rightarrow s'_{2\leftarrow 1}$. Consequently, we can construct the loss functions for the content and style codes separately equations 11 and 10 respectively similar to that in the reconstruction loss 7.

$$\begin{aligned} L_{cycle}^{c_1} &= \mathbb{E}_{c_1, s_2}(\|c_1 - c'_{2\leftarrow 1}\|_1), \\ c'_{2\leftarrow 1} &= E_2^c(x'_{2\leftarrow 1}) \end{aligned} \tag{9}$$

$$L_{cycle}^{s_2} = \mathbb{E}_{c_1,s_2}(\|s_2 - s_{2\leftarrow 1}^{'}\|_1), \quad s_{2\leftarrow 1}^{'} = E_2^s(x_{2\leftarrow 1}^{'}) \tag{10}$$

A GAN module is used to keep the converted samples indistinguishable from that in the target emotion. Thus, we improve the quality of the synthesized samples. The GAN loss is computed on $x_{i\leftarrow j}^{'}$ where $i \neq j$

$$L_{GAN}^{i} = \mathbb{E}_{c_j,s_i}[\log(1 - D_i(x_{i\leftarrow j}^{'}))] + \mathbb{E}_{x_i}[\log D_i(x_i)] \tag{11}$$

Thus the used loss function is the weighted sum of the three loss functions L_{rec}, L_{cycle}, and L_{GAN}.

3.5 Experimental Setup

Although there are several speech emotion databases for different European and Asian languages, there are very few Arabic speech emotion databases in literature. Moreover, those datasets were recorded in Arabic spoken by Syrian, Saudi, and Yemeni native speakers, so there is a deficiency in Egyptian Arabic datasets and research. The Egyptian Arabic dataset that we used consists of a total of 0.9 hr of audio recordings from 1 native Egyptian speaker covering 2 different emotion categories (neutral and angry). It was recorded using Audacity sound engineering software in a silent room with a sampling rate of $48KHz$. The Dataset contains 1000 utterances for each emotion. The training and testing sets were randomly selected (80% for training, 20% for testing). The training set was sampled with a fixed length of 128 frames 5ms each.

4 RESULTS

The focus is mainly on the results of the loss functions discussed earlier in 3.4. In fig. 5, the discriminator loss curve has significant oscillations. This is because of the small batch size chosen which is 1. We tested various values including 8, 16, etc., however, a batch of size 1 was the most suitable as far as we are concerned to keep a trade-off with the generator loss. The generator loss keeps converging to a fair extent, indicating how the generation of utterances to the target domain is improving.

The semi-cycle loss convergence shown in fig. 6 is a reflection of the naturalness of the reconstructed utterances. It causes the utterance generation in both emotion domains to demonstrate low generation losses which are explicitly clarified in fig. 7. Discriminator losses in both emotion domain directions demonstrated in fig. 8 provide robust evidence on the naturalness of converted samples since they represent the model's ability to distinguish between the real and synthesized samples.

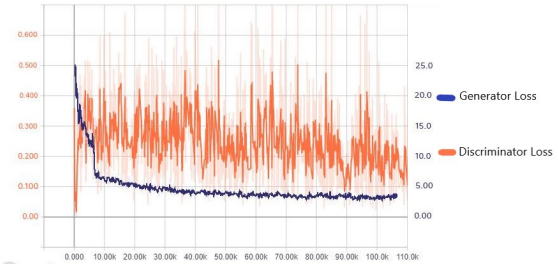

Figure 5: Generator Loss and Discriminator Loss.

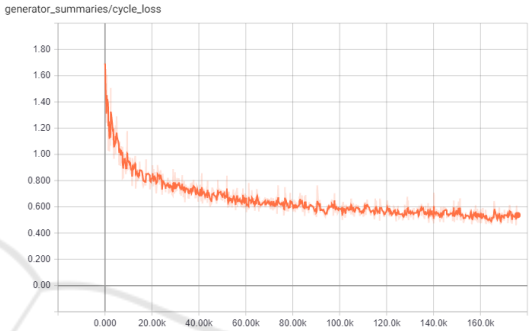

Figure 6: Semi-cycle Loss.

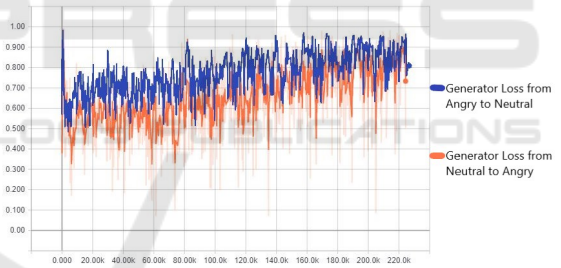

Figure 7: Generator Loss from Angry to Neutral and Generator Loss from Neutral to Angry.

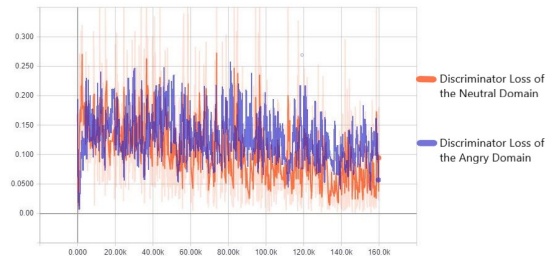

Figure 8: Discriminator Loss of the Angry Domain and Discriminator Loss of the Neutral Domain.

The speaker's identity was almost steady and no much change happened to it as shown in fig. 9. Moreover, alongside the training steps, the identity loss decreased significantly at the beginning and then kept steady at a relatively low value. Since the model train-

ing approach is speaker-dependent, the pre-trained model will be for one speaker. Consequently, the training dataset for each pre-trained model will be of one speaker. Thus, the speaker's identity is preserved.

Generally, the generation loss decreased significantly at the beginning, it rose up a little bit, nevertheless, it ended at a relatively sufficient value to generate the target emotion. The discriminator loss is oscillating, however, keeping the cycle loss at low values and generating utterances in both domain directions efficiently.

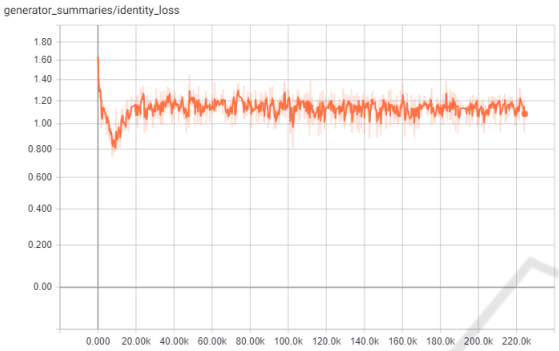

Figure 9: Speaker Identity Change.

5 DISCUSSION

The assessment of the results is based on a survey through which participants rated the clarity of the voice, emotion, and speaker identity out of 10. About 105 participants took place in the assessment. Each one listened to 5 different utterances for both the angry and neutral emotions. The mean accuracies of convergence to neutral and angry are 63.42% and 56.19% respectively based on the survey results. The assessment results are shown in fig. 10. The results might not be up to our expectations regarding the model architecture. This is due to the fact that angry emotion in the dataset is more shouting than expressing anger, however, our model performed better than VC- StarGAN (Kameoka et al., 2018) in terms of conversion ability (average 59.8% vs 44%). Regarding the newly proposed model EVC-USEP (Shah et al., 2023), our model performed better (average 59.8% vs 41.5%). Further modifications to the dataset might introduce better results. In addition to taking into consideration the other emotion-related features that affect the delivery of emotions. They might not be of much significance to affect the output, however, they will surely introduce improvements to the model results.

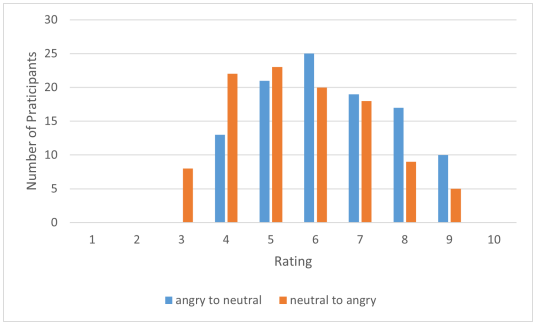

Figure 10: Survey Results.

6 CONCLUSION

This research proposes a nonparallel speaker-dependent emotional voice conversion approach for Egyptian Arabic speech using CycleGAN. The proposed method successfully changes emotion-related features of a speech signal without altering the lexical content or speaker identity. However, the results might not be up to expectations due to the nature of the dataset. Further modifications to the dataset and considering other emotion-related features are likely to introduce improvements to the model results. Future work includes using continuous wavelet transform (CWT) to decompose F_0 into 10 different scales so it can observe abrupt changes, modifying the model to be speaker independent. Overall, this study provides a significant contribution to the development of emotional voice conversion for Egyptian Arabic speech and can pave the way for further research in this area.

REFERENCES

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 359–370. AAAI Press.

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

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Gao, J., Chakraborty, D., Tembine, H., and Olaleye, O. (2019). Nonparallel emotional speech conversion. In *Interspeech 2019*. ISCA.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., and ..., Y. B. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, volume 27.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*.

Helander, E., Virtanen, T., Nurminen, J., and Gabbouj, M. (2010). Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*.

Huang, C.-F. and Akagi, M. (2008). A three-layered model for expressive speech perception. *Speech Communication*, 50(10):810–828.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510. IEEE.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *The European Conference on Computer Vision (ECCV)*.

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273.

Kaneko, T. and Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE.

Luo, Z., Takiguchi, T., and Ariki, Y. (2016). Emotional voice conversion using deep neural networks with mcc and f0 features. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*.

Luo, Z.-H., Chen, J., Takiguchi, T., and Sakurai, T. (2017). Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *Journal of Audio, Speech, and Music Processing*, 18.

Morise, M., Kawahara, H., and Katayose, H. (2009). Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. Paper 11.

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99-D(7):1877–1884.

Olaronke, I. and Ikono, R. (2017). A systematic review of emotional intelligence in social robots.

Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15(2):123–148.

Schröder, M. (2006). Expressing degree of activation in synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1128–1136.

Shah, N., Singh, M. K., Takahashi, N., and Onoe, N. (2023). Nonparallel emotional voice conversion for unseen speaker-emotion pairs using dual domain adversarial network & virtual domain pairing.

Sisman, B., Zhang, M., and Li, H. (2019). Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6):1085–1097.

Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1145–1154.

Toda, T., Black, A. W., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2016). Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022.

Xue, Y., Hamada, Y., and Akagi, M. (2018). Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*.

Zhou, K., Sisman, B., Liu, R., and Li, H. (2021). Emotional voice conversion: Theory, databases and esd. *arXiv*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232. IEEE.