Emotional Speech Transfer on Demand Based on Contextual Information and Generative Models: A Case Study

Andrea Veronica Porco
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

Kang Dongshik
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

***Abstract*—The automated generation of speech audio that closely resembles human emotional speech has garnered significant attention from the society and the engineering academia. This attention is due to its diverse applications, including audiobooks, podcasts, and the development of empathetic home assistants. In the scope of this study, it is introduced a novel approach to emotional speech transfer utilizing generative models and a selected emotional target desired for the output speech. The natural speech has been extended with contextual information data related with emotional speech cues. The generative models used for pursuing this task are a variational autoencoder model and a conditional generative adversarial network model. In this case study, an input voice audio, a desired utterance, and user-selected emotional cues, are used to produce emotionally expressive speech audio, transferring an ordinary speech audio with added contextual cues, into a happy emotional speech audio by a variational autoencoder model. The model try to reproduce in the ordinary speech, the emotion present in the emotional contextual cues used for training. The results show that, the proposed unsupervised VAE model with custom dataset for generating emotional data reach an MSE lower than 0.010 and an SSIM almost reaching the 0.70, while most of the values are greater than 0.60, respect to the input data and the generated data. CGAN and VAE models when generating new emotional data on demand, show a certain degree of success in the evaluation of an emotion classifier that determines the similarity with real emotional audios.**

***Keywords*—Emotion transfer; contextual information; speech processing; generative models; variational autoencoder; conditional generative adversarial networks; empathetic systems**

I. Introduction

The creation of empathetic systems, capable of understanding and responding to human emotions, marks a significant advancement in artificial intelligence. Empathetic systems hold the promise of transforming human-machine interactions, offering not just responses but genuine understanding. However, this task presents a profound challenge. During speech processing, the subtle nuances of human emotions often dissipate, making it a complex endeavor to imbue artificial intelligence with empathy. Preserving these emotional characteristics during speech processing remains an open issue in the field of AI.

It is expected in the near future to have home assistants capable of not only recognizing when people surrounded are feeling sad, happy, or anxious but also responding with appropriate empathy [1]–[3]. Such a system could offer invaluable support especially to the vulnerable population, providing comfort to the lonely, reassurance to the anxious, joy to the despondent and safety to children and seniors. The potential applications are vast, extending beyond homes to healthcare, customer service, and mental health support. Achieving this level of artificial empathy stands at the frontier of AI research, requiring innovative solutions to bridge the gap between raw data and the rich emotional tapestry of human speech.

There are several models proposed to achieve this target, including generative and non generative models.

Generative models have emerged as valuable tools for the synthesis of speech audios. These models offer the unique ability to create audio data that captures the nuanced patterns and complexities of human speech. Variational Auto-Encoders (VAEs) [4], [5] for instance, provide a structured approach to encoding and decoding data, allowing for the generation of diverse, high-quality audio samples. Conditional Generative Adversarial Networks (cGANs), on the other hand, introduce conditional factors, enabling the generation of speech data with specific attributes, such as different emotional states or gender-specific characteristics. These generative models excel in creating natural-sounding speech and have the potential to revolutionize applications like voice assistants, speech synthesis, and emotional speech generation.

Despite their promising capabilities, generative models also come with inherent challenges. One notable drawback is the risk of generating audio samples that, while coherent, may lack the nuanced emotional expressiveness present in natural human speech. The delicate interplay of pitch, rhythm, and intensity that defines emotional speech can be challenging to replicate accurately. Furthermore, generative models may struggle with gender-specific patterns, such as pitch variations and resonance differences between male and female voices. Additionally, ensuring the generated audio remains consistent with the intended emotional state or gender identity presents a significant challenge. These difficulties underscore the need for continued research and development in the field of generative speech modeling, particularly in the context of emotion, gender, and natural speech synthesis [6], [7].

There are several non-generative models used for speech synthesis, including Hidden Markov Models (HMMs) [8],

Long Short-Term Memory (LSTM) networks [9], and deep neural networks (DNNs) [10], [11]. While these models have proven effective in certain aspects of speech processing, they exhibit limitations that generative models, like VAEs and cGANs, can address.

One limitation of traditional non-generative models, such as HMMs and LSTMs, is their reliance on a fixed set of acoustic features or linguistic representations. These models often struggle to capture the rich nuances of natural speech, including emotional variations and gender-specific characteristics. The adaptability of non-generative models to generate highly expressive and contextually rich speech remains limited. Furthermore, these models may require extensive data preprocessing and manual feature engineering, making them less flexible and more labor-intensive in comparison to generative models.

Generative models, on the other hand, have the potential to overcome these limitations [12]–[14]. They can operate in an end-to-end fashion, learning complex patterns without the need for extensive feature engineering. By leveraging latent spaces, conditional information, and adversarial training, generative models can synthesize speech that better resembles natural human communication, including emotional variations and gender-specific traits. This adaptability and capacity to capture nuanced characteristics make generative models an attractive choice for applications where high-fidelity and emotionally expressive speech synthesis is crucial, such as the creation of diverse empathetic systems. Nonetheless, it is essential to recognize that both generative and non-generative models have their own strengths and limitations, and the choice between them depends on specific task requirements and constraints [15]–[21].

In this work, two generative models such as variational auto-encoder and conditional adversarial network were utilized, to train speech audio data, contextual audio data and an emotional selected target on demand, to generate an emotional speech audio with the target emotion. In this proposed case study, we also show how a neutral speech audio with a specific gender and a specific contextual data input (laughing by giggling, angry by shouting, crying sound, etc.) is converted into a happy speech audio automatically by the variational auto-encoder model. For the creation of the testing data, a TTS system is used by selecting a gender specific voice and an utterance close in pronunciation to the trained data.

To the best of our knowledge previous research works did not propose an emotional speech transfer on demand that train extended generative models such as VAE and CGAN with speech data and contextual related cues in gender and emotion. Furthermore, a TTS system is utilized to generate testing data for the proposed case study with a trained variational auto-encoder model and additional contextual cues.

The paper is structured as follows. Subsequent sections of the introduction section, sequentially detail the proposed approach and associated experiments, incorporating comprehensive information on data preprocessing, experimental methodologies, and results. Following this, the case study section is introduced to illustrate a practical application, featuring the integration of real Text-to-Speech (TTS) samples with the utilization of the proposed model and custom data. The following discussion section serves to expound on the evaluations and results pertaining to both models, delineating inherent limitations and identifying potential avenues for further research or enhancements. Lastly, the conclusion section encapsulates final remarks on the presented works and outlines prospective future endeavors.

II. Proposed Approach

In this proposed work, the functionality of a variational autoencoder model and a conditional adversarial network model were extended, for learning emotional patterns that have associated emotional contextual audio data, such as crying, shouting and laughing by giggling sounds. The sad emotion is associated with the crying sound, the angry emotion with the shouting sound, the happy emotion with the laughing by giggling sound and a normal emotion has a simple whisper sound of 1 second pattern.

The dataset used is an extended Ravdess dataset. The Ravdess database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent, “Kids are talking by the door” and “Dogs are sitting by the door” for speech and sing. Speech includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions. The selected data from Ravdess dataset is audio-only with 1 second recoding for 24 actors, equally gender balanced, with additional 24 equally balanced contextual audios that match by gender and by emotion.

The selection of contextual data associated is random, forming a total of a 2 second recording for input data. However, the model cannot distinguish this associated data, since is using it directly as a complete input data. The contextual audio data is firstly divided by gender, into female and male data. Even though the speech audio used correspond to a female or male specific voice, we did not associate the same exact voice by gender with the same pattern male/female voice. This is due to the generalisation of the voice we do have while laughing, crying or shouting and singing, where the voice is difficult to be perfectly recognised.

The generative models used benefit in different ways. The variational autoencoder model will keep training in an unsupervised manner, because after training, once we input a data with a specific associated contextual audio, and an additional gender specific voice speech input, the variational autoencoder will try to reconstruct the emotion present in similar audios. Therefore, trying to reconstruct the weak emotional speech part into a stronger emotional speech. For example, when people speech is happy, people use to laugh by giggling, and these two correlated actions we expect that our models will capture the essence and associate both action into one emotional concept. Certainly, other external noises from the context could be learnt by the models, however there will be not correlation with the emotional speech we emphasised.

The generative adversarial network in counterpart, will use the same input data but to feed the discriminator that during training passes information to the generator loss, and therefore to the generator itself. In this case, we do not expect the generator of CGAN model to be as good as the VAE encoder and decoder. This assumption is based on the CGAN model specifications, where after training, we ask the generator to

generate a happy audio but we cannot ensure which samples it will create in relation to gender and utterance given. This issue is improved when we train data by gender (male or female, but not both of them), or by specific utterance which is very restricted in terms of the usage of TTS normal emotion generated input audios. Therefore, for the case study we will show an example based on a variational auto-encoder model instead, when giving a TTS neutral audio voice as an input, setting the gender and the utterance that is expected to be reconstructed.

The VAE and CGAN models will try to transfer the emotion in a new audio file. It is important to mention the limitations we feat with this approach. The first limitation is the reconstruction of the input and the generation of the target data. The input data origin was initially separated, which causes that the output received by each model should also be treated separately at the end. This will cause the data to be more noisy and more difficult to reconstruct as an audio file.

Another important point is that these models produce noisy results with complex data, and speech data enter in that category. Times series data cannot be manipulated such as the image data because rotating, flipping, augmenting or shifting the data for images will not affect the final position or structure of the objects in an image, but it will completely corrupt our times series data. Therefore, we made a great effort while passing through these models to leave the data without manipulating it, whenever was possible during the emotional transfer process.

The details of the architecture of the proposed models are shown in Fig. 1, 2 and 3, respectively.

The input data of VAE and CGAN architectures, is a combined audio data between a 1 second speech audio from Ravdess dataset audios, with a connected 1 second associated emotional contextual information in a form of audio. Therefore, the two audios are concatenated into one audio with 2 seconds of total duration and the same sample frequency, 44100Hz. As previously mentioned, the relation between the speech data and the contextual information is the weakly emotional pattern present in Ravdess dataset, and specifically, the gender present in each speech audio data. As it was mentioned, it significantly differs from using any other unrelated noise that would not be useful in terms of our target, which is the emotional transfer. For our case study we extracted the log Mel spectrogram data and converted them into images.

Our proposed model architecture for VAE for training and testing can be observed in Fig. 1 and 2, respectively. The basic architecture of the VAE model has two associated networks, called encoder and decoder, which are connected by the latent space representing the probabilistic part of the model.

The proposed CGAN model's architecture can be observed in Fig. 3. In this figure we can identify two main parts, that serve as networks, the generator and the discriminator. The discriminator will output the fake or real classification affecting the discriminator loss or the generator loss being mutually exclusive results. The generator cannot see directly the input images which make it more difficult to learn the emotional patterns in the input audios, but while receiving the generator loss information it will learn to generate them as closer as possible to the original input data. In this case we assumed

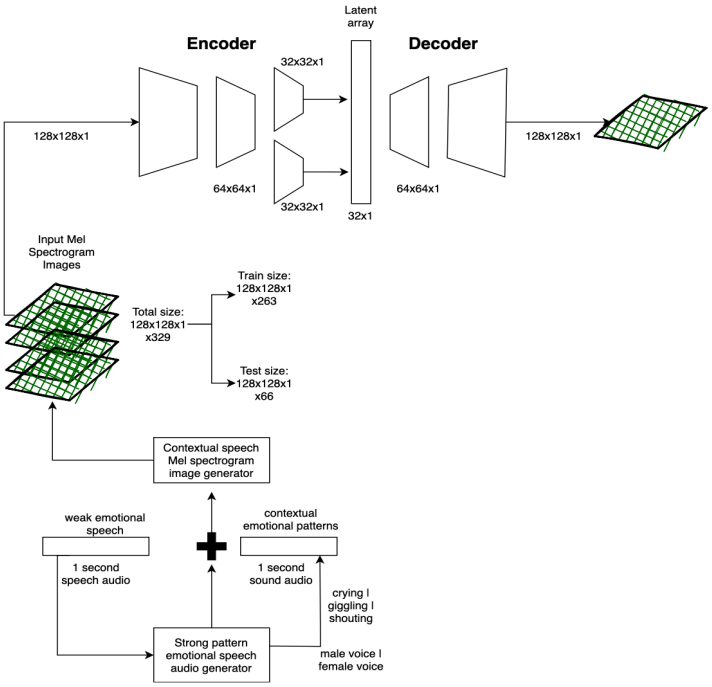

Fig. 1. Proposed architecture of the VAE model for training.

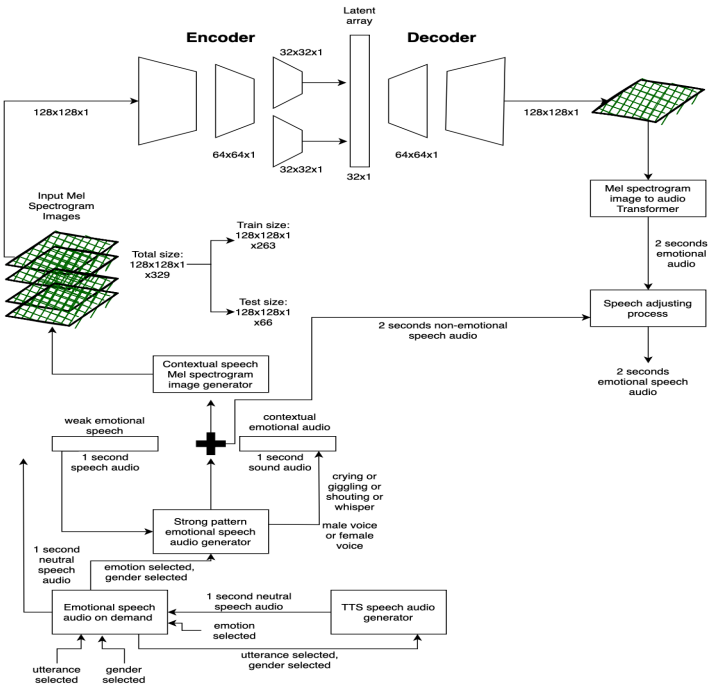

Fig. 2. Proposed architecture of the VAE model for testing.

some restrictions in the gender training to be male or female and not both of them, because CGAN is not good in terms of managing multiple conditions at the same time. In the testing of the CGAN model, the generator of the model is directly asked to produce the emotional data by emotion selected label.

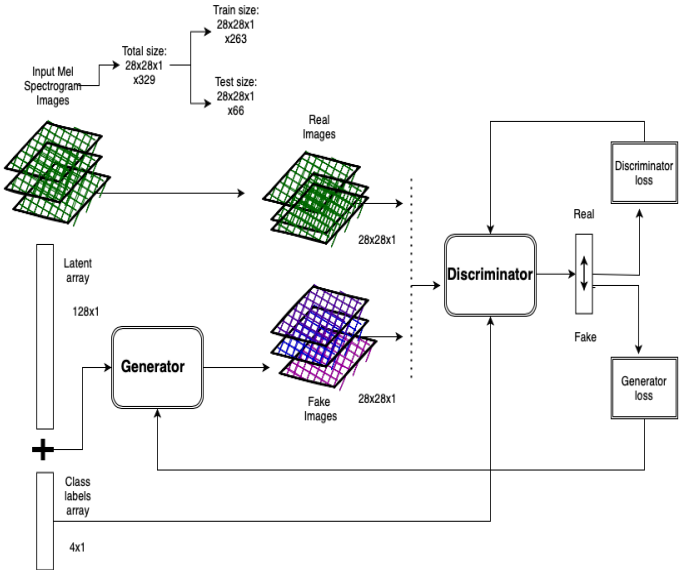

Fig. 3. Proposed architecture of the CGAN model for training.

In each model, the loss function is affected by the phonetic regularization with the extraction of the formants F1 to F5, presents in the Mel spectrogram data.

III. Experiments

A. Data Preprocessing

The input audio data before transforming into log Mel spectrograms can be observed in Fig. 4. This data were used for training, for the VAE and CGAN model to recognizse the happy speech and the contextual data that are usually present in a normal conversation. We limited our research to a one side speaker to be analysed.

The first audio A represents the speech of a male speaker from the Ravdess dataset that has an original duration of 3 seconds. The utterance in this happy speech is saying "Dogs are sitting by the door". The identification number in the original Ravdess dataset is "03-01-03-01-02-01-09" and is open to the public. This happy audio does not sound happy to our ears, since it is part of a weak emotional dataset, which is one of the reasons why the emotional transfer is complex to achieve in time series data. As can be observed in the image, the audio A contains zero data in the beginning, and at the end of the speech. With the aim to eliminate the non-useful data, the audio was reduced into a 1 second audio, while remaining the speech content.

The second audio B remains in 1 second, since originally each contextual data has 1 second of duration. This represents naturally what happens with human beings, since usually our laughing takes about the same time in being produced. The combination results in a total duration of 2 seconds for the concatenated audio C. These preprocessing tasks provide the desired adjustment, while maintaining a consistent audio quality.

The contextual information related to emotion is added as follows. We selected 6 audios per gender and per weakly emotion to add to any weakly emotional audio that matches by gender and emotion class, and chosen randomly. This is possible, given that, while singing or making emotional noises, we can not perfectly distinguishing these pattern belonging gender and voice. The duration of each audio is 1 second, its sampling rate is 44100Hz and its format extension is WAV. For happy emotion, we selected 6 female giggling patterns and 6 male giggling patterns, in total 12 audios for constructing the emotional extended audios. In the case of sad emotion, we selected 6 female crying pattern audios and 6 male crying pattern audios. When selecting the angry audios, we picked 6 female and 6 male shouting audios. However for neutral pattern audios we selected 6 female and 6 male whisper and natural English speaking pauses, such as, "Emmmm....", "AHA....", "Ahhhh....", "cause.....", and so on, where the utterance is almost not listened.

All the training and testing audios were denoised, trimmed and adjusted by volume, especially the audio B, since many artefacts were present. This is because it is custom data, downloaded freely from the Freesound site. Generally, custom data has many artefacts content present in the audios.

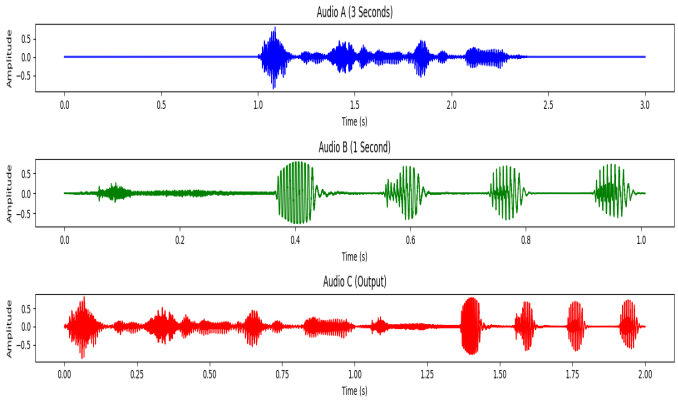

Fig. 4. Original audios A and B representing a happy speech, preprocessed and combined into one audio output file C.

The following step after gathering the clean concatenated audio is to transform the audio into frequency domain. The log Mel spectrogram image data generated for testing, as a case study can be observed in Fig. 5. The sampling rate is 44100Hz, the number of FFTs is 2048, with a hop length of 512. For the VAE model the target size is a 128 by 128 image. For the CGAN we selected the target 28 by 28, because CGAN generates better results with smaller sizes of images.

Before training, the pixel values were normalized to a (-1, 1) interval instead of using RGB values ranging from 0 to 255. Firstly, larger input values may slow down or disrupt the learning process of neural networks and setting them to smaller values is good practice [21]. Secondly, this normalization was required since the generator outputs tanh activations within the same interval.

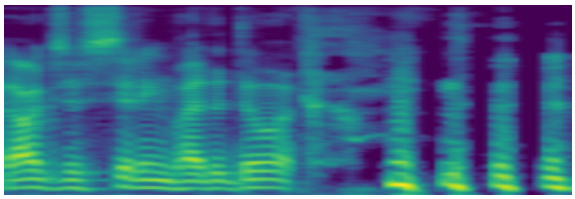

Fig. 5. Input log Mel spectrogram data example for the VAE and CGAN model. Happy emotion male speaker with original Ravdess dataset extended with male giggling sound.

B. Experimental Details

Two distinct methodologies were introduced, involving the VAE and CGAN models. The initial approach utilizes the VAE model in an entirely unsupervised manner, whereas the second approach employs the CGAN model in a supervised capacity.

Through the experiments, it was demonstrated that enhancing the input dataset with contextual attributes yields superior outcomes compared to explicit labeling or enforced supervision. Despite CGAN being a supervised model, it also benefits the core task of emotional transfer. This advantage is attributed to the data complexity being reduced when well-patterned contextual data is linked with the original input. Such simplification is only achievable with contextual information related to emotions and speaker gender.

In CGAN's labeled input-guided framework, we must assign appropriate labels to guide the generation of new samples. The model needs to be trained with labels that are familiar to it during the training phase. In contrast, the VAE utilizes extended input data without the explicit management of labels; instead, it separates classes by observing the data's intrinsic patterns during training.

To generate data from a specific class using the aforementioned CGAN model, several structured steps need to be followed. Firstly, a one-hot encoded label vector representing the desired class is created. This label must align with the format of labels in our training dataset. Secondly, random latent vectors are generated as input for the generator, sampled from a normal distribution. Thirdly, the one-hot encoded label vector is concatenated with the random latent vectors to form the conditional input for the generator. In the final step, the generator is employed to create data samples based on the prepared conditional input.

C. Experimental Results

The training dataset comprises 263 samples, while the test dataset consists of 66 samples. The distribution of training data is as follows: 37 samples for neutral audios, 72 for happy audios, 75 for sad audios, and 79 for angry audios. In the testing set, there are 11 neutral audios, 20 happy audios, 18 sad audios, and 17 angry audios. The limited number of "neutral" samples is due to the smaller quantity available in the Ravdess dataset for each actor, requiring additional steps to balance the dataset, such as cloning voices and creating neutral audios with different utterances and intensities. This task, although valuable, falls beyond the scope of our current research and could be explored in future studies.

The computational environment used for testing included a RAM occupancy of 5.38GB out of 51GB and a disk space usage of 26.83GB out of 166.77GB. The experiments utilized a NVIDIA T4 GPU provided by Google Compute Engine. The execution time for the CGAN model training over 10,000 iterations was 20 minutes. For the VAE models, the training process involved 10,000 epochs and took approximately 20 minutes. The programming language employed for these tasks was Python version 3.

To assess the VAE model, an unsupervised approach lacking predefined target classes, two metrics were utilized, mean square error (MSE) and structural similarity index (SSIM) between the original and generated data. In MSE, lower values indicate higher similarity between images, with 0 representing a perfect match, although realistically, a small value signifies good similarity. SSIM values range from -1 to 1, where 1 signifies a perfect match. Values closer to 1 indicate strong similarity, and a value above 0.9 is generally considered a robust match. It is essential to consider that ideal values can vary based on the specific domain and image quality.

VAE model results of training data after 10000 epochs, for original Ravdess dataset can be observed in Fig. 6, 7, and 8. The training and validation loss of the variational autoencoder trained with the original Ravdess dataset for 10000 epochs can be observed in Fig. 6. The training loss starts in a high value 10^5 from epoch 1 and the validation loss starts in 10 in the epoch 1. The values are steadily decreasing over the epochs as expected. The data generated by VAE model in 10000 epochs

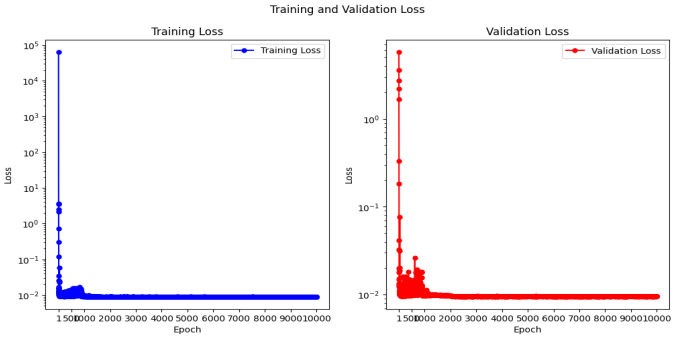

Fig. 6. Training and validation loss of the variational autoencoder after 10000 epochs of training with Ravdess dataset.

of training can be observed in Fig. 7. The generated results

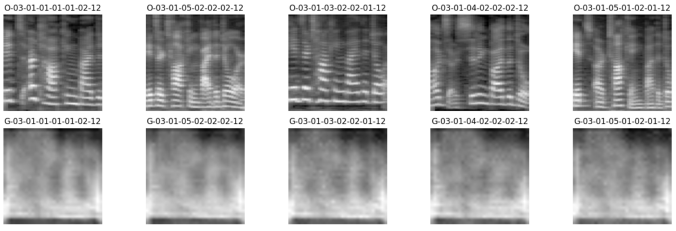

Fig. 7. VAE model random generation of different classes of data after 10000 epochs of training with Ravdess dataset. "O" stands for original, "G" for generated, and the emotion class is the 3rd number from left to right reading (01:Neutral, 03:Happy, 04:Sad, 05:Angry).

comparison with Ravdess dataset can be seen in Fig. 8.

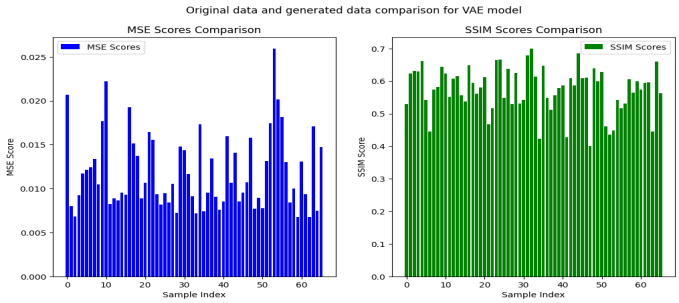

Fig. 8. VAE results comparison with Ravdess dataset. MSE and SSIM measured data results.

CGAN model results of training data after 10000 epochs, for original Ravdess dataset can be observed in Fig. 9, 10, 11, 12 and 13. CGAN loss for training data over 10000 epochs can be observed in Fig. 9. The angry, sad, happy and

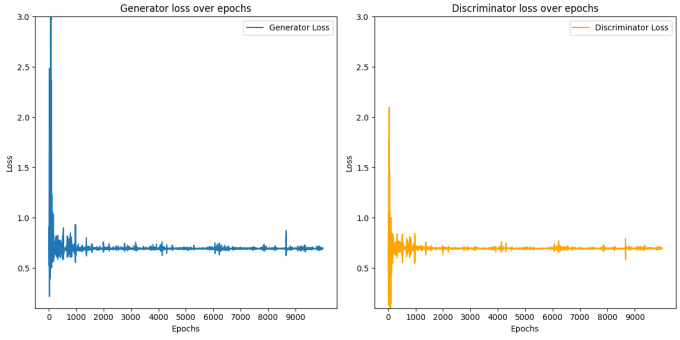

Fig. 9. CGAN loss over 10000 epochs of training.

neutral emotional data generated by CGAN model after 10000 epochs of training can be observed in Fig. 10, 11, 12 and 13, respectively.

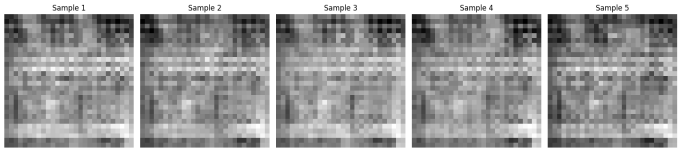

Fig. 10. CGAN model generation of data class 1 (neutral emotion) after 10000 epochs of training.

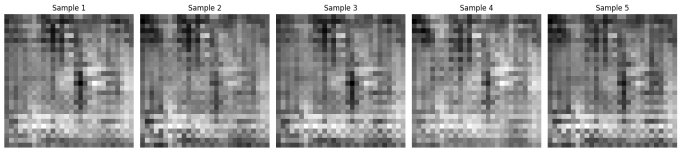

Fig. 11. CGAN model generation of data class 3 (happy emotion) after 10000 epochs of training.

VAE proposed model results after 10000 epochs of training, with Ravdess dataset extended with contextual information data, can be observed in Fig. 14, 15 and 16. The training

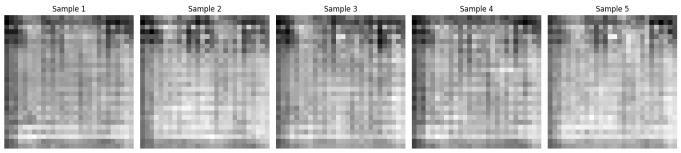

Fig. 12. CGAN model generation of data class 4 (sad emotion) after 10000 epochs of training.

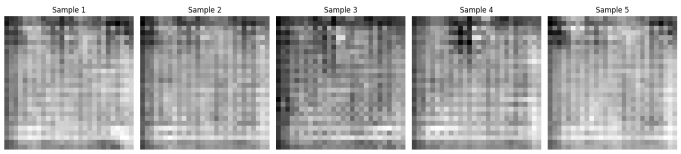

Fig. 13. CGAN model generation of data class 5 (angry emotion) after 10000 epochs of training.

and validation loss of the variational autoencoder trained with Ravdess dataset extended with contextual information data for 10000 epochs can be observed in Fig. 14. The training loss starts in a lower value above 10^1 from epoch 1 and the validation loss starts also in a lower value of 10^0 in the epoch 1, in comparison with the training and validation loss of original Ravdess dataset in Fig. 6. The values are steadily decreasing over the epochs as expected. The data generated

Fig. 14. Training and validation loss of the variational autoencoder after 10000 epochs of training with the Ravdess dataset extended with contextual information data.

by VAE model in 10000 epochs of training can be observed in Fig. 15. The generated data improved the previous original Ravdess training with VAE, since the quality of each image increased significantly while remaining the same training and testing conditions. The generated results comparison after

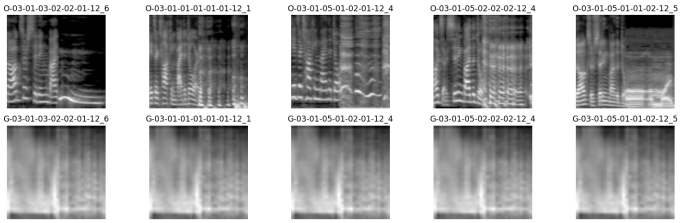

Fig. 15. VAE model random generation of different classes of data after 10000 epochs of training with contextual information extended data. “O” stands for original, “G” for generated, and the emotion class is the 3rd number from left to right reading (01:Neutral, 03:Happy, 04:Sad, 05:Angry).

10000 epochs of training with Ravdess dataset extended with contextual information data, can be seen in Fig. 16. CGAN

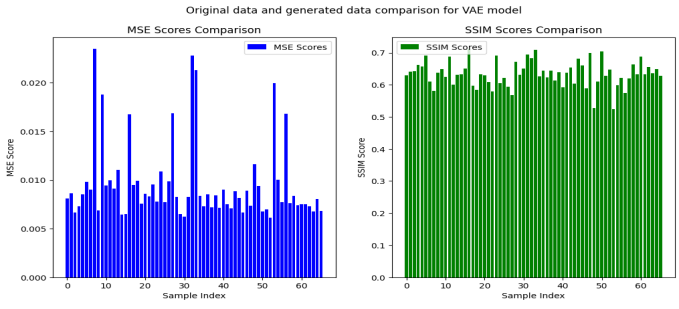

Fig. 16. VAE results comparison with the proposed custom dataset. MSE and SSIM measured data results.

proposed model results after 10000 epochs of training, for Ravdess dataset extended with contextual information data can be observed in Fig. 17, 18, 19, 20 and 21. CGAN loss for training data over 10000 epochs can be observed in Fig. 17. It shows that the loss starts with a higher value and decreased accordingly, as it is expected for loss functions in both sides, with no strange jumpings or increasing values from both sides. It also shows a convergence and a estabilization point. The

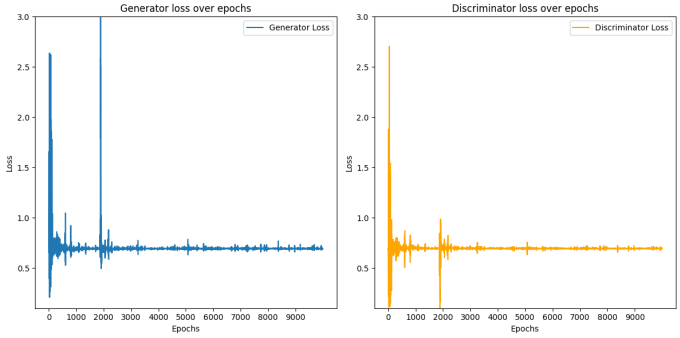

Fig. 17. CGAN loss over 10000 epochs of training.

angry, sad, happy and neutral emotional data generated by CGAN model after 10000 epochs of training can be observed in Fig. 18, 19, 20 and 21, respectively. The generated data improved the previous original Ravdess training with CGAN, since the quality of each image increased significantly.

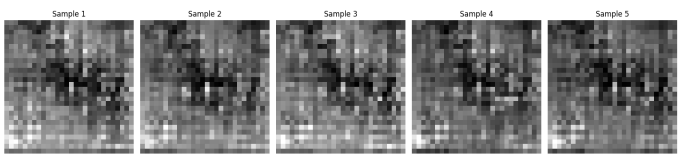

Fig. 18. CGAN model generation of data class 1 (neutral emotion) after 10000 epochs of training.

To validate the accuracy of the generated data, we developed an emotional classifier program, trained on the same input data utilized in both the VAE and CGAN models. The classifier's accuracy indicates the models' ability to generate emotional data that can be correctly classified into specific

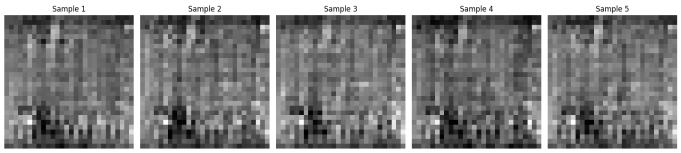

Fig. 19. CGAN model generation of data class 3 (happy emotion) after 10000 epochs of training.

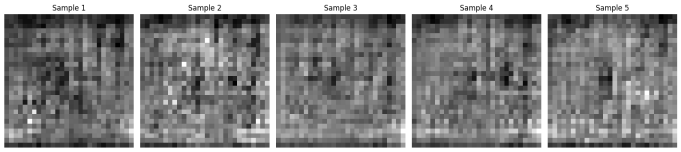

Fig. 20. CGAN model generation of data class 4 (sad emotion) after 10000 epochs of training.

emotion classes. It is important to remark that this measure can not obtain a 100 percent of accuracy due to the lack of train and validation over this fresh generated data. Therefore, the generated data is non-trained and unseen data for the classifier, however it will give us a notion of how much we should adjust the classifier and the models to improve the results. Most importantly, it will result in all classes zero classified, if it is not able to detect any emotional generated data.

When testing both models with the original Ravdess input data, it is anticipated that the generated results to be weakly classified by the classifier. This expectation arises because the Ravdess data, as previously discussed, is inherently weakly emotional. Moreover, it is important to acknowledge the inherent noise in the outcomes produced by generative models, particularly concerning audio data features such as log Mel spectrogram values.

The VAE generated data was sent to the classifier for testing generation accuracy. The results can be observed in Fig. 22. The results of the classifier after training with original Ravdess data shows a 20.51 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as explained above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data. This is the evidence that our CGAN model in Fig. 23, is trying to reconstruct the data getting a better emotional generated data's result, in comparison with VAE model. Additionally, the classifier needs to be adjusted for future works to measure more precisely our data.

The CGAN generated data was sent to the classifier for testing generation accuracy. The results can be observed in

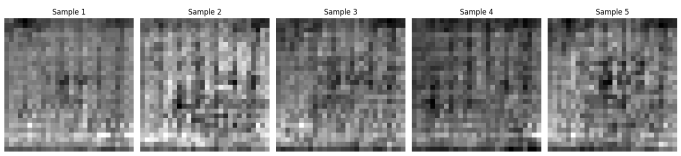

Fig. 21. CGAN model generation of data class 5 (angry emotion) after 10000 epochs of training.

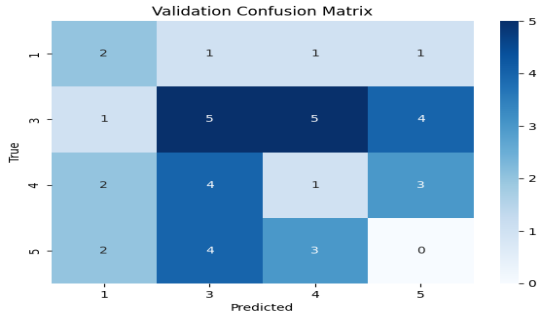

Fig. 22. Validation confusion matrix for VAE model generated data after classified by an emotional classifier.

Fig. 23. The results of the classifier after training CGAN with extended data shows a 25.641 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as discussed above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data. This is the evidence that our model is trying to gradually reconstruct the data and the classifier needs to be adjusted for future works to measure more precisely our generated data.

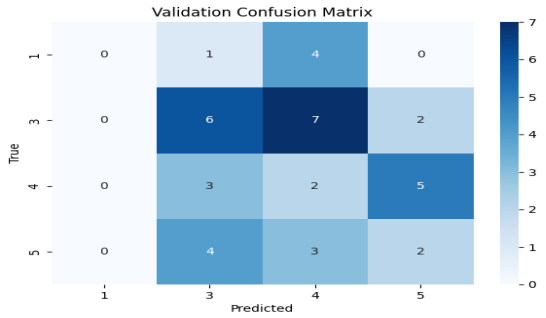

Fig. 23. Validation confusion matrix for CGAN model generated data after classified by an emotional classifier.

During this evaluation, we anticipate the input data to be classified in closer proximity to the generated data. However, we cannot definitively confirm how the generative models interpret the essential features in the presence of additional contextual information. Nevertheless, the generated data is expected to align more closely with its input data due to the utilization of paired information. For instance, the weak speech emotion in the Ravdess dataset, when associated with the extended contextual information like a female voice's giggling pattern linked to a happy emotion, should be more accurately classified as happy. In comparison with the vanilla VAE model and standard CGAN model trained with original Ravdess dataset, the results are clearly improved. The results shows also that more efforts are required for future works in order to better represent the reconstructed emotional audios.

D. Case Study

As an extension of the emotional data generation capabilities with the proposed VAE and CGAN models, in this case study we showcase the specific emotion transfer with additional conditions such as specific gender, specific voice, and specific utterance, with an additional change in emotions, from neutral to happy, sad, angry respectively. Since the variational auto-encoder model is more flexible in terms of receiving new input data to regenerate, we reutilized the pre-trained proposed variational auto-encoder model.

Furthermore, the utterances in this study were generated using a Text-to-Speech (TTS) system, which inherently produces a "Neutral" emotion speech audio since it lacks emotional variation. Converting "Neutral" audio to another "Neutral" audio is not necessary for our evaluation; it would not yield any change and only indicates the models' understanding of elements like whispering or speaking pauses. Although our training data includes neutral sounds, such as whispers or slight delays in speech, they are not utilized in our TTS systems for testing purposes.

For specific test scenarios, assuming the model is trained, we paired our TTS-generated voice with additional emotional audio, creating samples as follows:

1) 'Bob is by the door' with female giggling.
2) 'Bob is by the door' with male giggling.
3) 'Bob is by the door' with female shouting.
4) 'Bob is by the door' with male shouting.
5) 'Bob is by the door' with female crying.
6) 'Bob is by the door' with male crying.

To demonstrate the practicality of this approach, a "Neutral" emotional speech is converted into a "Happy" emotional speech on demand through a TTS system, by using the proposed variational auto-encoder model with custom data for the reconstruction. A selected raw audio data generated for testing can be seen in Fig. 24. The TTS system utilizes a neutral MAC PC male voice "Alex", uttering "Bob is by the door". The duration was condensed to one second for consistency across speech data. The TTS voice is neutral without any emotional inflection. For this use case, the utterance was varied from the trained sets u_1 ("Kids are talking by the door") to u_2 ("Dogs are sitting by the door"). This variation illustrates the model's ability to transfer emotion even with slight differences in the trained words.

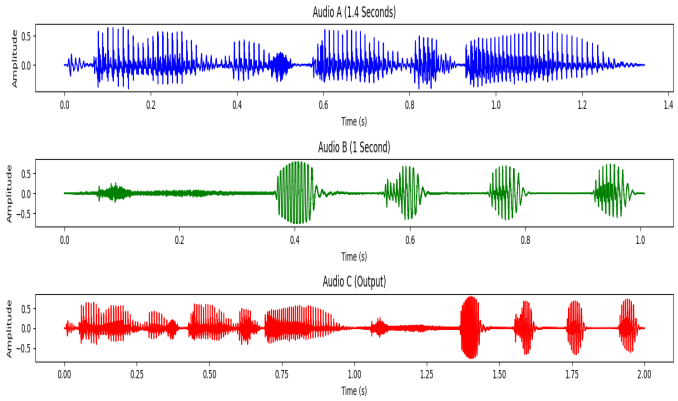

Fig. 24. Raw audio A is a Neutral speech audio generated by a TTS MAC OS male voice system. Original audio B representing a happy emotion by giggling. Preprocessed audios were combined into one audio output file C.

The following step after gathering the clean concatenated audio is to transform the audio into frequency domain. The log Mel spectrogram image data generated for testing, as a case study can be observed in Fig. 25. The sampling rate is 44100Hz, the number of FFTs is 2048, with a hop length of 512. For the VAE model the target size is a 128 by 128 image.

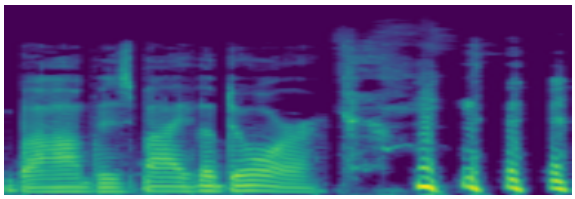

Fig. 25. Input log Mel spectrogram example for the VAE. Neutral speech with additional happy contextual data.

The result of the VAE model emotion transfer can be observed in Fig. 26, when given the happy bob spectrogram generated, the proposed VAE model try to reconstruct it with happy emotion.

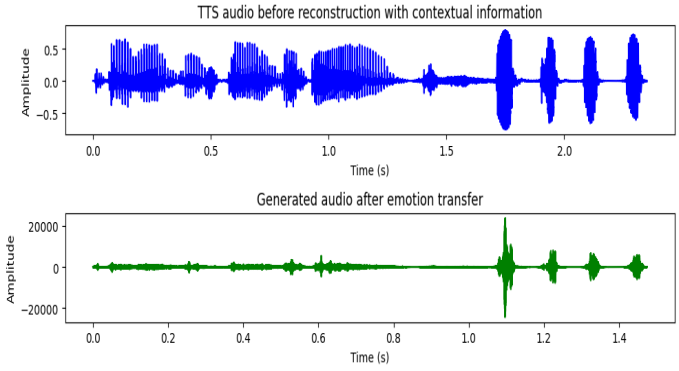

Fig. 26. Result of the VAE model emotion transfer from neutral to happy speech sample.

In this implementation, an unsupervised model was used to infer the emotion that should be reconstructed by contextual information. This task was developed with the proposed VAE model. The TTS speech was created with 1 second plus a related audio extension of 1 second, depending of the gender and the emotion desired. These samples show how the emotion is reconstructed among the trained utterances and among similar utterances with words composition variability. The utterance selected for this testing is as follows. u_3 : "Bob is by the door".

The total data size used are 329 audios, the train data size is 263 audios, transformed into log Mel spectrogram images. The test data size is 66. This case study showcase one sample among 6 samples presented. The size of each image is 128 by 128. The initial VAE training is performed with grayscale images. Each audio has a sampling rate of 44100Hz, which makes it more difficult to recover the audio after passing to Mel spectrogram.

Some pending tasks are required such as volume regularization and denoising tasks, however it can be seen how the contextual information such as giggling in a male speaker is greatly represented and trying to recover the left side of the speech with the same accent and pauses that the male speaker represents.

IV. Discussion

To assess the VAE model, an unsupervised approach lacking predefined target classes, two metrics were utilized, mean square error (MSE) and structural similarity index (SSIM) between the original and generated data. In 66 reconstructed audios for the proposed VAE model, when testing the custom data, the MSE is lower than 0.010 and the SSIM is almost reaching the 0.70, while most of the values are greater than 0.60. For the Ravdess dataset, the MSE of the reconstruction is lower than 0.020 and the SSIM is almost reaching the 0.70, while most of the values are higher than 0.50. This signify that the proposed data along with the proposed model is getting better similarity in the reconstruction of tested data, while the mean squared error is lower which emphasizes the improvement in the reconstruction by using this proposed unsupervised model.

An external tool was created to measure the level of reconstruction in terms of emotions. This tool is an emotional classifier, trained with the same emotion classes and the same input data size and characteristics as the proposed models. Certainly the reconstructed data from CGAN by setting a specified emotion label and the TTS generated data used as input for the VAE make them different than the trained data of the created tool. However, this tool can state if the generated data by both models is not considered emotional at all by the classifier showing near zero values in the accuracy. This tool help to test the research results while avoiding the subjective measures by human resources. The VAE and the CGAN generated data were sent to the classifier for testing generation accuracy. The results of the classifier after training with custom data shows a 20.51 percent of correct classification of emotions in a total of 39 tested generated data for VAE. The results of the classifier after training CGAN with custom data shows a 25.641 percent of correct classification of emotions in a total of 39 tested generated data. Even though the result is low in comparison with a perfect accuracy, as explained above, it is expected that the classifier cannot easily extends its classification with these non-trained unseen data.

The results of the classifier also show that there is a need of a better vocoder function in future improvements of this work, and that the output should be better adjusted by a volume regularizer and an extra denoise function. This is due to the weak points present in generative models, where a better vocoder could probably increase the accuracy in the emotional classifier results, after generating new unseen and non-trained data. The new generated data is still identified as different for the emotional classifier, because the data just created posses different characteristics such as moved characteristics from the original input data that the classifier usually consumes.

The treatment of the utterances could be improved in future works, such as injecting the words spoken and its characteristics right after the generative models resampling. This is because the generative models tend to loose the accents of each spoken word during the transition from original data

to resampled data. However, a strong point of this research to mention is the extension of the trained utterances to similar utterances that have some words in common but differ in others. As it is known, with models such as HMM, RNN and LSTM, the utterances are totally fixed without any possibility of extension. This also make the models fail in the presence of utterances that contain different words, even when the general pronunciation of the words is similar and the sentence has the same duration. In comparison with other datasets, we used the Toronto emotional speech set (TESS) with two female speakers and a variation in utterances such as "Say the word book", "Say the word bought", or other variations from 200 target words for the emotions happy, sad, angry and neutral. SSIM and MSE results with our proposed model and extension with contextual cues show, for female speaker's emotions, similarity to the evaluated Ravdess dataset with 0.7 and 0.020, respectively. Therefore, demonstrating the ability to be robust and scalable to other utterances and voice characteristics. For future works, the evaluation with other datasets that include the male counterpart audios is also required for better determine the level of scalability by gender.

V. Conclusion

In this research, a method for transferring emotional speech utilizing generative models and specific emotional targets for the output was presented. The generative models employed in this task include a variational autoencoder model and a conditional generative adversarial network model. Although further refinement is needed to enhance the accuracy of generated emotional speech, both proposed models have demonstrated the ability to reconstruct emotional speech with a certain degree of quality.

In the presented case study, it was utilized an input voice audio, a desired utterance, and user-selected emotional cues to automatically transform ordinary speech into emotionally expressive speech audio using a variational autoencoder model. Remarkably, the proposed VAE model achieved this task without requiring specific labels to control the generated output, highlighting the efficiency of this approach with unsupervised learning. The model attempts to replicate the emotion inherent in the emotional contextual cues used for training directly into the ordinary speech. The findings reveal that the suggested unsupervised VAE model for generating emotional data achieves an MSE below 0.010 and an SSIM nearly reaching 0.70. The majority of values surpass 0.60 in comparison to both the input and generated data. When generating new emotional data on demand, CGAN and VAE models demonstrate a discernible level of success in the evaluation of an emotion classifier, determining its similarity to real emotional audios used for training.

In future works, the primary focus will be on refining the generative models further and implementing additional strategies to achieve a better balance between real contextual information and emotionally rich speech. This ongoing effort aims to bridge the gap between human understanding and artificial agents' comprehension in the essence of a desirable authentic speech.

References

[1] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in Interspeech, Sep 2022.

[2] HSU, Jia-Hao, et al. Empathetic Response Generation based on Plug-and-Play Mechanism with Empathy Perturbation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[3] PAIVA, Ana, et al. Caring for agents and agents that care: Building empathic relations with synthetic agents. In Autonomous Agents and Multiagent Systems, International Joint Conference on IEEE Computer Society, 2004. p. 194-201.

[4] 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., pp. 3391– 3395, 2020.

[5] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 6284–6288.

[6] K.Zhou, B. Sisman, R. Rana, B.W.Schuller and H.Li, "Emotion intensity and its control for emotional voice conversion," IEEE Tran. on Affective Computing, 2023.

[7] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in Proc. Odyssey Speaker Lang. Recognit. Workshop, 2020, pp. 230–237.

[8] DENG, Jun, et al. Recognizing emotions from whispered speech based on acoustic feature transfer learning. IEEE Access, 2017, vol. 5, p. 5235-5246.

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

[10] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in Proc. IEEE Int. Conf. Acoust., Speech Signal Pro- cess., 2017, pp. 4905–4909.

[11] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2015, pp. 4475–4479.

[12] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," 2020, arXiv:2007.12932.

[13] 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.

[14] K. Zhou, B. Sisman, and H. Li, "Vaw-GAN for disentanglement and recomposition of emotional elements in speech," in Proc. IEEE Spoken Lang. Technol. Workshop, 2021, pp. 415–422.

[15] D. P.. Kingma, M. Welling, "Auto-encoding variational Bayes", in Proc. 2nd International Conference on Learning Representations, 2014.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 10, 2014. arXiv: 1406.2661 [cs, stat]. URL: http://arxiv.org/abs/ 1406.2661

[17] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. Nov. 6, 2014. arXiv: 1411.1784 [cs, stat]. URL: http://arxiv.org/abs/1411.1784

[18] Ali Borji. Pros and Cons of GAN Evaluation Measures. Oct. 23, 2018. arXiv: 1802.03446 [cs]. URL: http://arxiv.org/abs/1802.03446

[19] Pegah Salehi, Abdolah Chalechale, and Maryam Taghizadeh. Generative Adversarial Networks (GANs): An Overview of Theoretical Model, Evaluation Metrics, and Recent Developments. May 27, 2020. arXiv: 2005.13178 [cs, eess]. URL: http://arxiv.org/ abs/2005.13178

[20] Sudarshan Adiga, Mohamed Adel Attia, Wei-Ting Chang, and Ravi Tandon. "On the tradeoff between mode collapse and sample quality in generative adversarial networks". In: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Anaheim, CA, USA: IEEE, Nov. 2018, pp. 1184–1188. I SBN: 978-1-72811-295-4. DOI: 10.1109/GlobalSIP.2018.8646478. URL: https:// ieeexplore.ieee.org/document/8646478/

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. June 10, 2016. arXiv: 1606.03498 [cs]. URL: http://arxiv.org/abs/1606.03498