Zewail City for Science and Technology
University of Science and Technology
Communications and Information Engineering

Arabic Emotionally Intelligent Voice-Enabled Chatbot

A Graduation Project
Submitted in Partial Fulfillment of
B.Sc. Degree Requirements in
Communications and Information Engineering

Prepared By

Moustafa Abada	201601212
Abdelmoez Elsaadany	201500438
Mohamed Ellebody	201601854

Supervised By
Dr. Moustafa Elshafei Ahmed

Signature

Table of contents

Acknowledgments

We would love to express our sincere gratitude to the Late Dr. Ahmed Zewail for his efforts in establishing such a great and developed University for science and technology and letting us have the chance to join this beautiful memorial place. We would also love to thank our supervisor Dr. Moustafa Elshafei for his consistent support and guidance throughout our senior year. We would also like to thank our professors in the major, without whom we would not have been able to complete this senior project, and without whom we would not have made it through our bachelor's degree. Thanks to our beloved families for their endless love, support, and encouragement. Without them, we would not have been able to finish our undergraduate studies. Thanks also to our friend Ahmed Hassan for his collaborative effort during data collection and debugging the code.

Declaration

This is to certify that the work done in this thesis is our original work done by the names found on the title page. The work done is not submitted to any other degree or institution. It’s intended for our senior project only. This project is not part of any other project. This is a new field in Arabic chatbots and Arabic emotion recognition and conversion.

List of Tables

List of figures

List of Acronyms

SER	Speech Emotion Recognition
EVC	Emotion Voice Conversion
VC	Voice Conversion
MFCC	Mel-frequency cepstral coefficients
MS	Mel spectrogram
E-Learning	Electronic learning
MLP	Multilayer Perceptron
SVM	Support Vector Machine
CNN	Convolutional Neural Network
TESS	Toronto emotional speech set
EmoDB	Berlin emotional speech set
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
PCA	Principal component analysis
API	Application Programming Interface
ESD	Emotional Speech Dataset
IEMOCAP	Interactive Emotional Dyadic Motion Capture
F0	Pitch contour/Fundamental Frequency
SP	Spectral Prosody
AP	Aperiodicity
GAN	Generative Adversarial Network
CycleGAN	Cycle Consistency Loss-Generative Adversarial Network
Sec2Sec	Sequence to Sequence
VAE-GAN	Variational Autoencoder-Generative Adversarial Network
CWT	Continuous Wavelet Transform
GMM	Gaussian mixture model
NMF	Non-negative Matrix Factorization
DNN	Deep Neural Network
VAW-GAN	Variational Autoencoding Wasserstein Generative Adversarial Networks
MCEP	Mel-Cepstral Coefficients

Abstract

Emotions are the underlying hidden mental states that people feel every time they speak. Emotions can be easily recognized by human brains. However, in machine learning, Emotions are hard to detect, but previous research has shown that we can use facial expressions to reveal people's emotions [1]. New research has shown that voices also can be used to recognize different emotions such as being happy, sad, surprised, angry, nature, fear, and disgust [2]. This field is called "Speech Emotion Recognition (SER)". We have built many models that recognize emotions from speech signals and we extended our work and decided to work on "Emotion Voice Conversion (EVC)" in which we can change the source emotion of a given sentence for a given speaker to another target emotion preserving the speaker identity and the sentence or the content spoken.

Emotions can be detected from each other by several features. For example, sadness and happiness are considered as calm speech, so that they have low average amplitude. While anger and surprise have high average amplitude. Also, Happiness emotion has a higher pitch variance compared to other emotions. This leads to the conclusion that emotions are the same in every language. For example, being angry in English is similar to being angry in Arabic or Chinese. Thus, we propose that emotion recognition models work for all languages. Also, emotion voice conversion models work for all languages. However, we might add one layer in our convolutional neural network structure to adjust the model to a unique language. This research is independent of the text, instead, it cares about the signal speech or voices. Lastly, we have tried to mimic Arabic, voice-enabled, conversation, and emotional chatbots and we proposed using it in E-learning and online classes.

This field has a lot of applications in real life such as E-learning, call centers, commercial applications, and computer games. For example, in E-learning or online classes, information about the underlying emotional state of students can be revealed and some improvements in the educational systems or equipment may occur. Also, we may know what subjects the students find interesting and others that the students find boring so that we develop more fun strategies to help students. SER systems and Emotional Chatbots also can be used in mobile services, and call centers. The experience of clients or customers can be evaluated using SER Systems and Emotional Chatbots and further improvements can be made. In psychological assessment, the emotions of the patients can be evaluated and medication is assigned to each patient according to his medical state.

Introduction

Emotions are hidden within humans. We took a new and different path to recognize emotions by analysing voices. We extended the project and worked on Emotion Voice Conversion (EMV). Moreover, we decided to work on Arabic language as the research in this area is very limited. We also thought about giving a product that serves the community. That's why our objective is to make an intelligent emotional conversational chatbot that can detect emotions and understand the speaker's speech and reply to them with appropriate replies and emotions. This project consists of three layers: Emotion Speech Recognition Layer, Chatbot Layer, and Emotion Voice Conversion Layer. The first layer in the project is the recognition layer where emotions are recognized by extracting some features from speech such as Mel-frequency cepstral coefficients (MFCCs), chroma, Mel spectrogram, contrast, tonnez, spectrum centroid, and zero-crossing rate. These features are used in machine learning and deep learning models to classify and recognize emotions. We made a comparison between different machine learning classification algorithms such as Decision Tree, Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN). Also, we are testing different combinations of the features extracted. We used English, Arabic, and Berlin datasets. The datasets used are the Toronto emotional speech set (TESS) [3], the Berlin emotional speech set (EmoDB)[4], and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). For the Arabic speech emotions dataset, we used the Principal component analysis (PCA) algorithm which is considered as a feature selection technique to avoid redundancy in features and select the most important features that well represent the dataset. In the second layer, the chatbot is a simple chatbot that gets the input voice message from the user and converts it from speech-to-text using Google API and then it defines the intent of the user and using the appropriate emotion and response, it answers the user after converting it from text-to-speech using Google API. In the third layer, we do emotion voice conversion. We convert the emotion of a spoken speech sentence from one source emotion to another target emotion without changing the speaker's identity and the content or sentence said. We have used two datasets: Emotional Speech Dataset (ESD) and The Interactive Emotional Dyadic Motion Capture (IEMOCAP). We extracted some features from both source emotion and the target emotion like Frequency F0, spectral prosody (SP), and aperiodicity(AP) and learned how to map from the source emotion to the target emotion using deep learning models such as GANs. We tested different types of GANS such as CycleGan, Sec2Sec model, and VAE-GAN and compared their performance.

Literature Review

In [5], the authors have used continuous wavelet transform (CWT) decomposition for F0 modeling. CWT provides a way to decompose a signal into different temporal scales that explain prosody in different time resolutions. The authors trained two CycleGAN pipelines for spectrum and prosody mapping respectively. In this way, they eliminated the need for parallel data of any two languages

Zongyang Du et al were motivated by the success of CycleGAN in spectral conversion. Thier proposed Spectrum-Prosody-CycleGAN framework outperforms the Spectrum-CycleGAN baseline in subjective evaluation. They extracted features such as fundamental frequency (F0) that represents the variation of the vocal pitch over the whole time domain. Since F0 is hierarchical in nature and affected by short-term as well as long-term dependencies, they used CWT decomposition to analyse F0 in different time scales, and find a mapping for each time scale[5]. The CWT decomposition describes a frame-based F0 value with a set of CWT coefficients that represent prosodic features.

The structure in figure (1, 2)shows the model used in [6]. The first picture is the training model in which we have a source speech and target speech. Firstly, the paper did world encoder to the source speech and extracted F0 features and MCEPs coefficients and did CWT Analysis to the F0 and then fed the result and the MCEP to the CycleGAN training model. The same happened in the target speech, and F0 with CWT components and MCEPs are fed into the CycleGAN that learns the mapping technique to convert the Source speech to the Target speech.

In the Second picture, also, the F0 with CWT components and MCEPs are extracted from source speech and fed into the CycleGAN which convert the features to the features of the target speech.

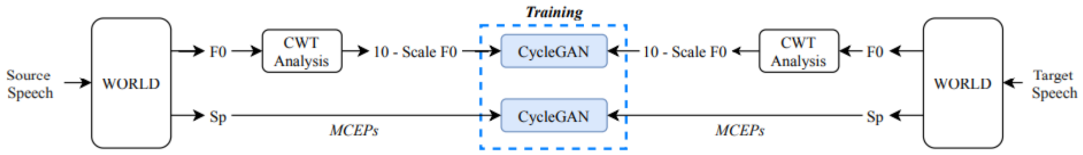


Fig. (1) Flowchart of the training process[5].

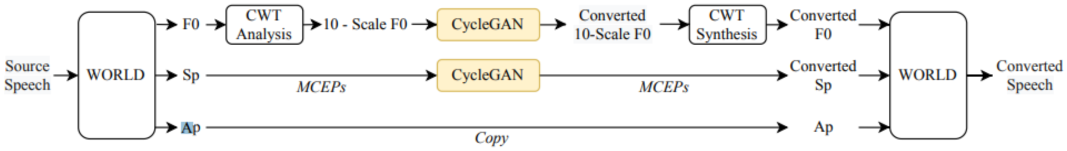


Fig. (2) Flowchart of the Conversion process[5].

Zongyang Du et al had some limitations. For example, prosody conversion for cross-lingual voice conversion has not been well studied. Also, the differences between two languages lie not only in phonetic systems, but also in linguistic prosody and speaking style, that are characterized by the F0 contour of speech[5].

Another structure used is in 2020, where Kun Zhou et al used the model VAW-GAN that’s based on encoder-decoder structure to learn the spectrum and prosody mapping and learn the emotion-independent representations. They performed prosody conversion by using continuous wavelet transform (CWT) to model the temporal dependencies. They made use of F0 as an additional input to the decoder to improve emotion conversion performance.[7]

Kun Zhou et al extracted Spectral and F0 features using the WORLD vocoder. F0 contains segmental information so it will need to be decomposed using CWT decomposition. The speaker-dependent and independent components that were extracted will be dealt with differently[7]. The extracted components out of F0 will allow the encoder to learn and differentiate between speakers. CWT is sensitive to the discontinuities in F0 so linear interpolation will be done. Then the scale of F0 will be transformed into a logarithmic scale and normalized. It was found that training of spectrum and prosody separately is better than training collectively. Two networks were constructed for the purpose of training and each consists of an encoder, decoder, and discriminator. The encoder is exposed to multiple inputs from different speakers with different emotions and output the emotion-independent features such as speaker identity and phonetic content as latent code z. a one-hot vector is used as an emotion ID and is fed into the decoder. Figure 1 contains the training process[7].

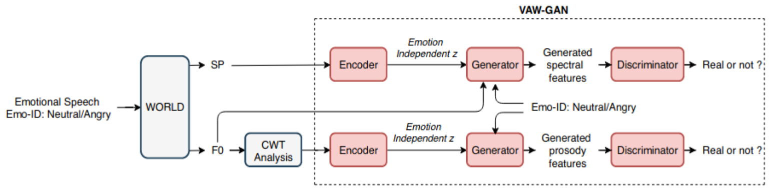


Fig (3) comparison between different frameworks in their ability to convert different pairs of emotions[7].

The generative model is trained based on adversarial learning to find the best solution through a min-max game. During run time conversion the steps are quite similar to the training procedure. F0 is decomposed into coefficients of different time scales using CWT analysis then the coefficients are fed into the encoder to generate the emotion ID. The generator is conditioned on emotion ID with and CWT-based F0 coefficients. The converted spectral feature is obtained through the trained VAW-GAN for spectrum then WORD vocoder is used to synthesize the speech. Figure 2 illustrates the run-time conversion.

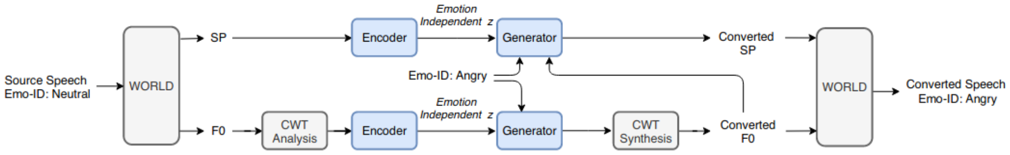


Fig (4) Run-time conversion[7].

In Table 1, The authors of [8] reported the performance of spectrum conversion in terms of Mel-cepstral distortion (MCD) and log-spectral distortion (LSD) [9–11, 12];

and that of prosody conversion in terms of Pearson correlation coefficient (PCC) and root mean square error (RMSE) of F0 contours.

Table 1: MCD [dB], LSD [dB], PCC and RMSE [Hz] results for seen and unseen speakers fo CycleGAN-based EVC [22]

Speaker	MCD	LSD	PCC	RMSE
Seen	4.948	7.028	0.721	54.043
Unseen	5.131	7.298	0.594	62.826
Seen(Zero effort)	5.210	7.383	0.571	62.242
Unseen(Zero effort)	5.296	7.400	0.440	66.646

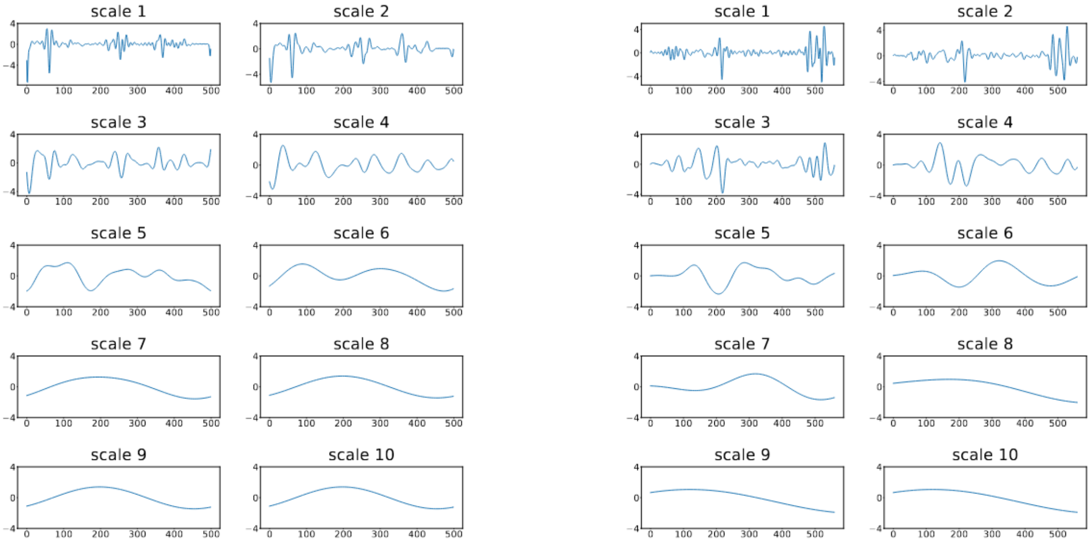
The results for the unseen speaker are clearly better than those for Zero Effort in terms of MCD and LSD for spectrum, and PCC and RMSE for prosody.

Emotions are important in human communication. These emotions can be represented by spectral features and prosodic features. Prosodic features include pitch, intensity, and speaking rate. Inheriting the emotions in speech is important in applications like conversational agents and intelligent dialogue systems. Emotion conversion is a technique for converting the emotion of a specific utterance into different emotions while preserving the speaker's identity and linguistic information. This technique shares some similarities with conventional voice conversion like converting the non-linguistic information from the source to target. They differ as conventional techniques consider the characteristics of voice is dependent on the physical features only. On the other hand, emotional voice conversion takes the physical features(spectral features) and prosodic features into consideration. Some machine learning approaches were proposed but they were not efficient[14]. The advancement in deep learning led to some advancement in VC techniques. Neural Network-based methods, such as Restricted Boltzmann Machine (RBM)[15], Feed Forward NN [16], Deep Neural Network (DNN)[17], and Recurrent Neural Network (RNN) [18] have helped in achieving a higher level in terms of modeling the relationship between source and target feature. Training these models using parallel data approaches was not practical. As a result, some approaches were proposed such as Deep Bidirectional Long short-term memory, Variational autoencoder, Deep Bidirectional long short term memory with model using phonetic posterior grams, and GANs. Earlier, prosody conversion was based only on classification and regression trees then the output was fed into the Gaussian mixture model or regression-based clustering method. Another proposed prosody conversion was the combination of the hidden Markov model, Gaussian mixture model, and F0 segment. Some other techniques like the Deep belief network and exemplar-based emotional VC approach based on NMF were proposed to convert prosodic features. Sequence to sequence encoder-decoder structure was proposed. A style transfer autoencoder was proposed that could learn from non-parallel training data using autoencoder. Prosody has some linguistic features which are dependent on the F0 prosodic factor. This factor was decomposed earlier by the Logarithm gaussian-based linear transformation method. This was not sufficient as a single pitch of F0 does not represent the signal well so continuous wavelet transform was applied to know the frequency component of the signal with different scales. The contributions of this paper are

1- proposing a parallel-data-free emotional voice conversion framework.
2- showing the effect of prosody for emotional voice conversion.
3- Effectively converting spectral and prosodic features withCycleGAN.
4- Investigating different training strategies for spectrum and prosody conversion such as
separate training and joint training.
5- Outperforming the baseline approaches, and achieving quality converted voice.

Cycle GANs were used to train the model. They are based on the concept of adversarial training. They have some losses like adversarial loss, cycle-consistency loss, and identity-mapping loss, leaning forward and inverse mapping between source and target. Adversarial loss tells about the difference in the distribution between the data in target and source. The cycle consistency loss is for knowing whether the data is consistent or not. The identity mapping loss is used for preserving linguistic information without any external processes.

Continuous Wavelet transform provides an easily interpretable visual representation of signals. Using CWT, a signal can be decomposed into different temporal scales. Also, CWT helps in overcoming the problem of discontinuity in F0 which is challenging in modeling it.



(a) ’It is well never to know an author’ in a neutral tone.

(b) ’It is well never to know an author’ in an angry tone.

Fig (5) different scales of the two emotions neutral and angry [14].

The prosody of emotion is expressed differently at different time scales. With multi-scale representations, lower scales capture the short-term variations, and higher scales capture the long-term variations. Thus F0 can be modeled and transferred. The above figure is an example to compare two utterances with the same content but different emotions across time scales.

Spectrum and prosody conversion are done using cycle consistent adversarial networks. F0 is converted into 10-time scales using CWT. As CWT is sensitive to discontinuities in F0 linear interpolation is done on F0 then it is converted to logarithmic scale then normalized. These F0 features are extracted from both source and target utterance using the WORLD vocoder. CycleGAN is trained for spectrum conversion using 24-dimensional Mel cepstral coefficients and prosody conversion is trained using 10-dimensional F0 features for each speech frame. Source and target training data are from the same speaker but consist of different linguistic content and with different emotions. When the model learns how to inverse and forward map between different emotions using adversarial and cycle consistent losses, Cycle GANs are encouraged to find the optimal mapping between source and target spectrum and prosody features.
The below figure shows the training process of CycleGAN.

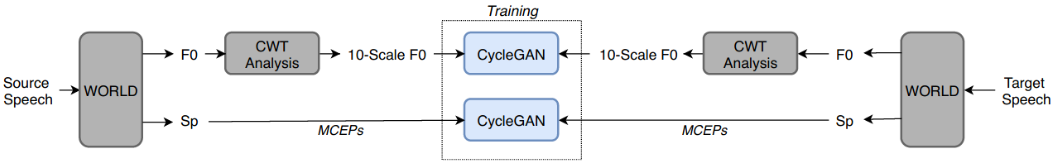


Fig. (6) Training process of the CycleGANs [14].

During runtime conversion, the spectral prosody and F0 features are extracted using WORLD vocoder. The spectral prosody is 24 dimensional and the output of CWT analysis is 10 dimensions. Both of the outputs are used separately by two CycleGANs. The output of the two CycleGANs is mapped to the output utterance using the inverse of processes used. The below figure shows the pipeline of operation.

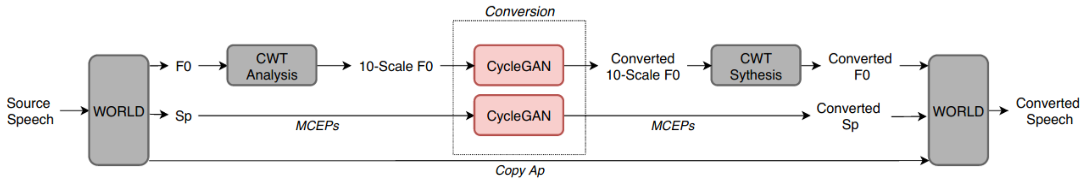


Fig.(7) Conversion process of the CycleGAN [14].

Training of CycleGANs is done on four emotions which are natural, angry, sad, and surprise. CWT is used to decompose F0 into 10-dimensional features and is used with 24 MCEPs separately or jointly to train the CycleGANs. The below table shows the difference between separate and joint CycleGANs.

Framework	**Spectrum Conversion**	**Prosody Conversion (F0)**
Baseline	Spectrum CycleGAN	LG-based F0 linear transformation
CycleGAN-Joint	Joint Spectrum-Prosody CycleGAN	
CycleGAN-Separate (proposed)	Spectrum CycleGAN	Prosody CycleGAN

Table (2) illustrates the training methods used in each framework [14].

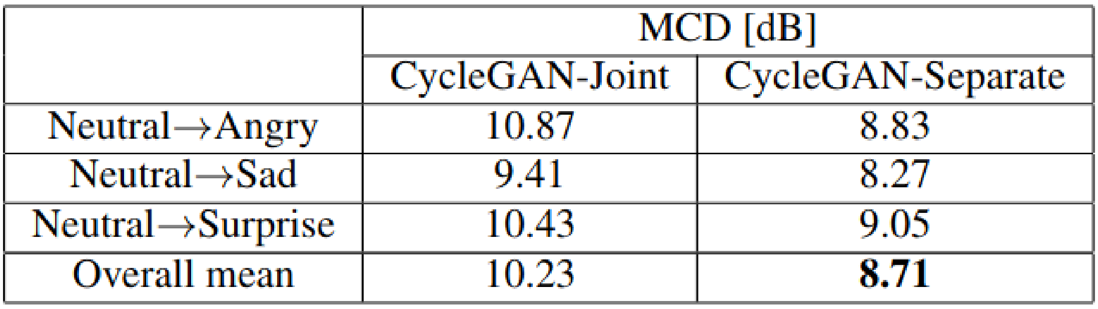
The specifications of the experiment are as follows. The data is sampled at 16kHz. The audio files of each emotion are divided into 100 sentences(3 minutes) . 90 sentences are for training and the 10 are for testing. The first 45 utterances of the 90 are used for the source and the other 45 are used for the target. The features are extracted every 5 ms using the WORLD vocoder. Three frameworks using CycleGANs are used for evaluation. They are baseline, CycleGAN-Joint, and CycleGAN separate. In the baseline, 24-dimensional MCEPs are extracted besides one dimensional F0 features for each frame. For separate and joint cycleGANs 10 dimensional F0 features are used. The generators are constructed using 1D CNN besides downsampling, residual and up-sampling layers. The discriminator is constructed using 2D CNN. The networks are trained using Adam optimizer. The initial learning rate is set to 0.0002 for the generators and 0.0001 for the discriminators.

The performance of the voice conversion is based on spectrum and prosody conversion. For the spectrum conversion, the performance is evaluated using Mel-cepstral distortion between target and converted MCEPs sequences. The below equation shows the calculator of Mel-cepstral distortion(MCD).

$$MCD[dB] = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}(mceps_i^t - mceps_i^c)^2}$$

Eq (5) Calculation of MCD loss[14].

MCD values were calculated for both separate and joint CycleGANs and the below table shows the results.



	MCD [dB]	
	CycleGAN-Joint	CycleGAN-Separate
Neutral→Angry	10.87	8.83
Neutral→Sad	9.41	8.27
Neutral→Surprise	10.43	9.05
Overall mean	10.23	**8.71**

Table. (3) Comparison between separate and joint training in CycleGANs[14].

From the results, CycleGAN-separate is better in the case of spectrum conversion. CycleGAN-separate results are similar to the baseline as both are trained using spectral features. So the spectral distortion is supposed to be the same in both of them.

For prosody conversion, the Pearson correlation coefficient(PCC) and Root Mean Squared Error(RMSE) are used to evaluate the performance. RMSE and PCC are shown below.

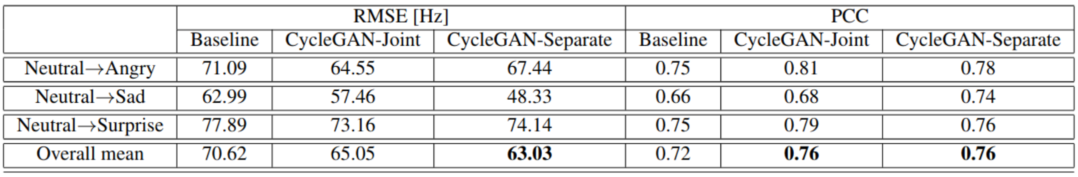
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F0_i^c - F0_i^t)^2}$$

Eq (5) Calculation of RMSE loss[14].

$$\rho(F0^c, F0^t) = \frac{cov(F0^c, F0^t)}{\sigma_{F0^c}\sigma_{F0^t}}$$

Eq (5) PCC loss[14].

Where F0c and F0t denote the converted and target interpolated features. And the denominator in PCC denotes the multiplication of standard deviations of Converted and targeted F0 sequences.



	RMSE [Hz]			PCC		
	Baseline	CycleGAN-Joint	CycleGAN-Separate	Baseline	CycleGAN-Joint	CycleGAN-Separate
Neutral→Angry	71.09	64.55	67.44	0.75	0.81	0.78
Neutral→Sad	62.99	57.46	48.33	0.66	0.68	0.74
Neutral→Surprise	77.89	73.16	74.14	0.75	0.79	0.76
Overall mean	70.62	65.05	**63.03**	0.72	**0.76**	**0.76**

Table. (4) Comparison between the different frameworks with two metrics[14].

Table. (4) shows the results of using RMSE and PCC as evaluation metrics. The table concludes that separate training of CycleGAN for prosody conversion is better than the other methods.

Another comparison between the three frameworks(CycleGAN-separate, CycleGAN-joint, and CycleGAN baseline) was done by humans to evaluate the emotions produced by the frameworks. The below results show that CycleGAN-separate outperforms the other two frameworks despite the small size of data. The joint CycleGAN can not generalize as both the F0 coefficients and prosody are trained jointly and an implicit assumption is made which is that emotions are dependent on the spectrum and prosody, not the prosody only.

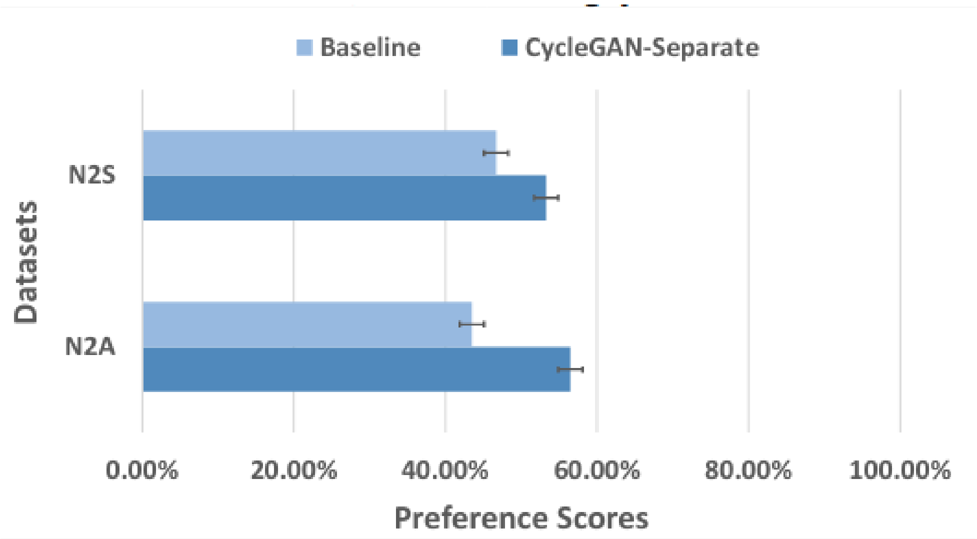


Fig. (8) comparing the performance of CycleGAN-separate and baseline[14].
N2S: Neutral to sad. N2A: Neutral to angry.

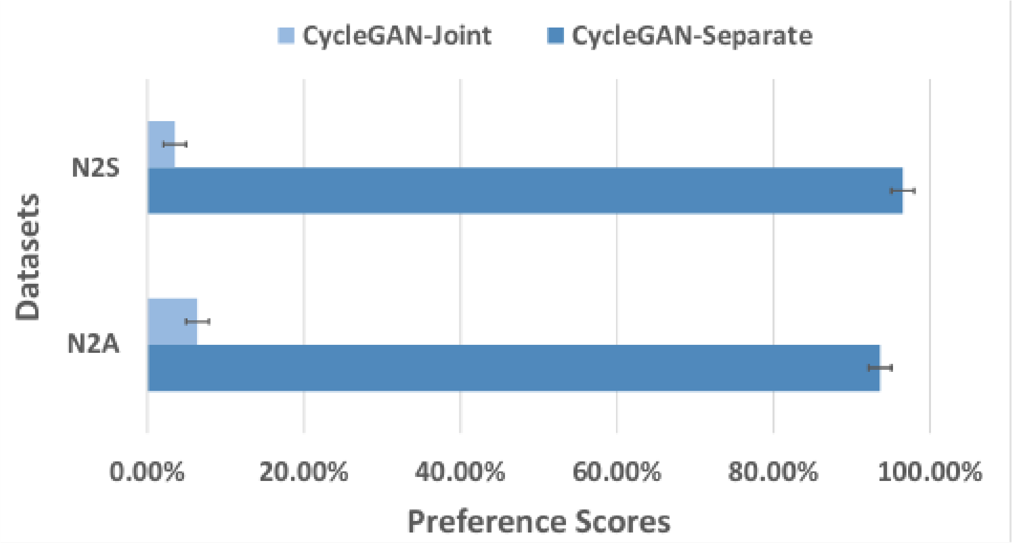


Fig. (9) Comparing the performance of CycleGAN separate and joint [14].
N2S: Neutral to sad. N2A: Neutral to angry.

In paper [19], a 2-stage training strategy for seq2seq was used. The first stage performs style initialization in which the speaking style and the linguistic content are disentangled. The second stage includes emotion training by limited speech data. The paper also has an emotion encoder that learns to disentangle emotional style from the speech. Also, an emotion classifier is obtained so any emotion content is eliminated from the linguistic data.

In 2021, Kun Zhou et al said that Seq2seq models performed well in the field of machine translation and speech synthesis[19]. The attention mechanism allows the network to focus only on the regions with emotions in the text. Seq2seq helped in eliminating the frame-based techniques. The proposed frameworks till now that use seq2seq require a large amount of data. Kun Zhou et al proposed seq2seq EVC model uses a small amount of data while maintaining a good performance[19]. The model consists of 5 components: text decoder, seq2seq automatic speech recognition(ASR) encoder, style encoder, classifier, and seq2seq decoder. The training happens in two stages: the first one includes disentangling the linguistic content from the speaker style(emotion). The second stage includes the training of the components in stage one. During run-time conversion, the emotion is added to the utterance with inference.

Stage 1: the framework takes the acoustic features and the phoneme sequences as input. The linguistic embeddings are predicted from the audio input and text input. The acoustic features are embedded in the style embeddings using the style encoder. In the end, the seq2seq decoder recovers the acoustic features with the style and the linguistic embeddings. The style decoder detects speaker-dependent information and excludes the linguistic data from the acoustic features. To make the text neutral, an adversarial training with the classifier is made to eliminate speaker-dependent information from the linguistic space[19]. Stage 1 is considered emotion initialization as the encoder can not learn how to make sense of any emotion. However, the style encoder with emotional corpus has a lot of information about the speaker style and speaker information thus it can learn about emotional style with small data.

Stage 2: Emotion training
The framework is restrained with the limited amount of speech data and is supposed to have learned some of the basic functions of voice conversion from the first stage. The style encoder then will learn the emotional style of the text from the data. Also, the classifier will eliminate the emotions from the data and will leave the linguistic space only. Here the style encoder will act as an emotional encoder as it will put the emotions of the text in the emotions vector[19]. The emotional encoder learns the emotional representation through the cross-entropy of the embedding sequences. For the emotion classifier, it learns through adversarial loss. The figures below show the two stages of training.

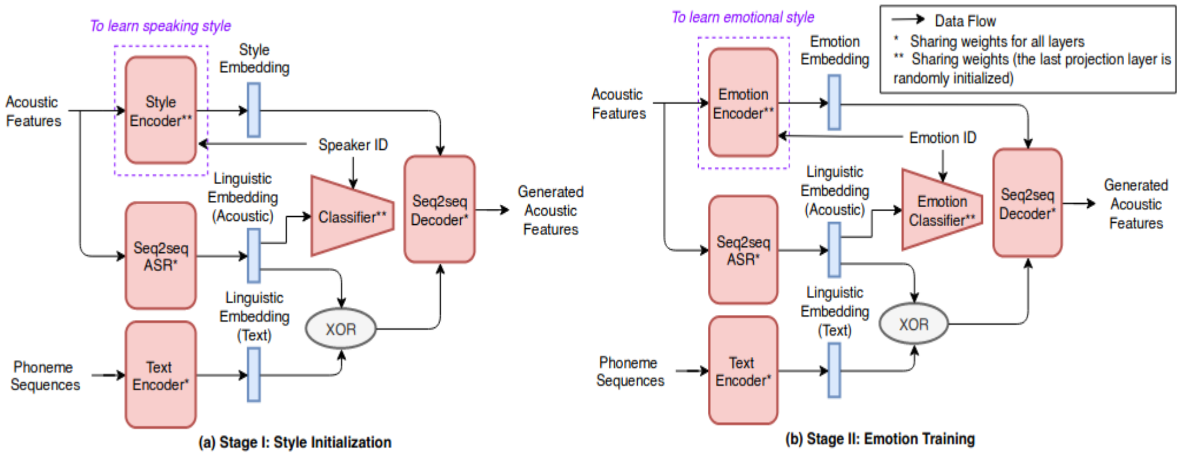

To learn speaking style
Style Embedding
Acoustic Features
Style Encoder**
Speaker ID
Linguistic Embedding (Acoustic)
Classifier**
Seq2seq Decoder*
Generated Acoustic Features
Seq2seq ASR*
Linguistic Embedding (Text)
XOR
Phoneme Sequences
Text Encoder*
(a) Stage I: Style Initialization
To learn emotional style
Emotion Embedding
Data Flow
* Sharing weights for all layers
** Sharing weights (the last projection layer is randomly initialized)
Acoustic Features
Emotion Encoder**
Emotion ID
Linguistic Embedding (Acoustic)
Emotion Classifier**
Seq2seq Decoder*
Generated Acoustic Features
Seq2seq ASR*
Linguistic Embedding (Text)
XOR
Phoneme Sequences
Text Encoder*
(b) Stage II: Emotion Training

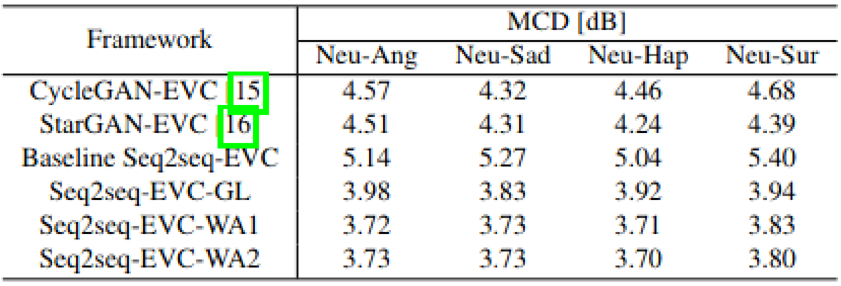
Fig (10) the two stages of training the seq2seq model [19].

In Fig(10), the style encoder is used in paper [19] as an emotion encoder and the speaker ID is replaced with emotion ID. At run time the emotion encoder is used to generate emotion embeddings from utterances from the same emotion ID category. Then given a source utterance and intended emotion, the seq2seq automatic speech recognition encoder derives the linguistic embedding of the source and gives them as input to the decoder.

The following models were implemented:

1. CycleGAN-EVC [20] (baseline): CycleGAN-based emotional voice conversion with WORLD vocoder.
2. StarGAN-EVC [21] (baseline): StarGAN-based emotional voice conversion with WORLD vocoder.
3. Baseline Seq2seq-EVC (baseline): Seq2seq-EVC trained directly with limited ESD data without any pre-training, and followed by a Griffin-Lim vocoder [22].
4. Seq2seq-EVC-GL (proposed): Seq2seq-EVC followed by a Griffin-Lim vocoder.
5. Seq2seq-EVC-WA1 (proposed): Seq2seq-EVC followed by a WaveRNN vocoder [23] that is pre-trained on VCTK corpus.
6. Seq2seq-EVC-WA2 (proposed): Seq2seq-EVC followed by a WaveRNN vocoder that is pre-trained onVCTK corpus, and fine-tuned with limited ESD data.
7. CycleGan-EVC can map only one pair of emotions so it's trained separately for each emotion. This paper[19] proposed the two models StarGAN-EVC and seq2seq-EVC

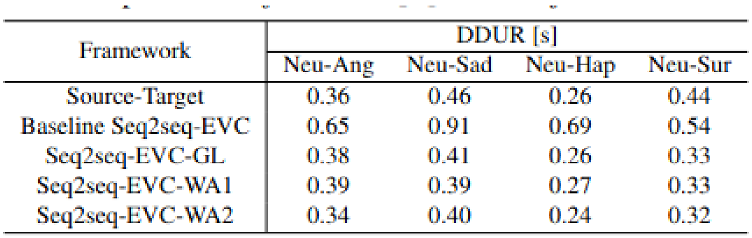
For the evaluation, it was done using Mel-cepstral distortion(MCD) and the average absolute differences of utterance duration(DDUR)[24]. When the model is trained only with ESD limited data it has the worst performance which ensures the need for the 2-stage approach. The table below shows the MCD values for the models



Framework	MCD [dB]			
	Neu-Ang	Neu-Sad	Neu-Hap	Neu-Sur
CycleGAN-EVC [15]	4.57	4.32	4.46	4.68
StarGAN-EVC [16]	4.51	4.31	4.24	4.39
Baseline Seq2seq-EVC	5.14	5.27	5.04	5.40
Seq2seq-EVC-GL	3.98	3.83	3.92	3.94
Seq2seq-EVC-WA1	3.72	3.73	3.71	3.83
Seq2seq-EVC-WA2	3.73	3.73	3.70	3.80

Table (5) comparison between different frameworks in their ability to convert different pairs of emotions[19].

The acoustic features in seq2seq models were Mel-spectrograms and Mel-cepstral coefficients. seq2seq-EVC-WA1 and seq2seq-EVC-WA2 outperformed the rest of the models. The below table shows the average duration of the utterances that get out of the models.



Framework	DDUR [s]			
	Neu-Ang	Neu-Sad	Neu-Hap	Neu-Sur
Source-Target	0.36	0.46	0.26	0.44
Baseline Seq2seq-EVC	0.65	0.91	0.69	0.54
Seq2seq-EVC-GL	0.38	0.41	0.26	0.33
Seq2seq-EVC-WA1	0.39	0.39	0.27	0.33
Seq2seq-EVC-WA2	0.34	0.40	0.24	0.32

Table (6) comparison between different frameworks in their ability to convert different pairs of emotions[19].

Recognizing emotions from the speech is important in human-computer interactions. Some features are needed for this task. The explored features include energy, pitch, linear predictive spectrum coding (LPCC)[25], Mel-frequency spectrum coefficients (MFCC), and mel-energy spectrum dynamic coefficients (MEDC). The paper only classifies three emotions: happiness, sadness, and neutral. The Berlin Database of Emotional Speech is used. SVM is used to classify data. The combination of both features MFCC and MEDC has shown an accuracy in the Chinese emotional database(91.3%) and Berlin emotional database (95.1%)[25]. SER Systems have a lot of applications, for example, in E-learning, the teacher can know if the students like the subject or not by knowing their emotions. If they are interested and happy, then they like the subject. Many datasets are tested before like the Berlin Database of Emotional and Speech Danish Emotional Speech. Features lie energy, pitch frequency, formant frequency, Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) are extracted in this paper. The classification techniques used in previous papers are DNN, GMM, HMM, Maximum Likelihood Bayesian classifier (MLC), KNN, and SVM. This paper uses both the Berlin dataset and a Chinese dataset. It also extracts all the features mentioned previously. Only 3 emotions are classified. Many models have made as shown in the table below:

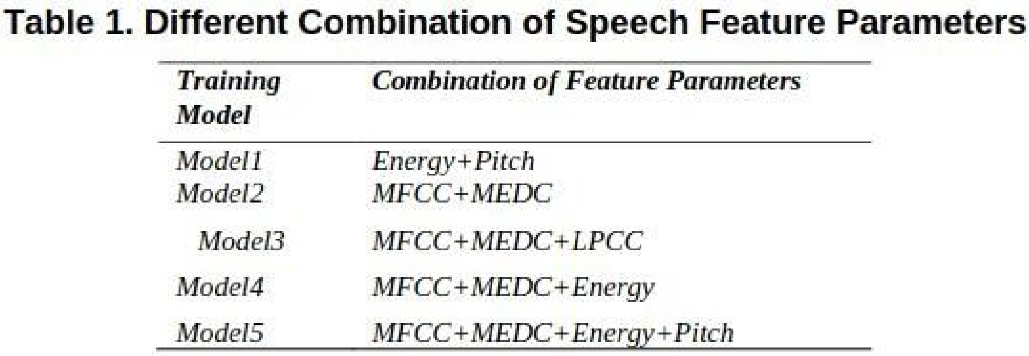


Table 1. Different Combination of Speech Feature Parameters

Training Model	*Combination of Feature Parameters*
Model1	*Energy+Pitch*
Model2	*MFCC+MEDC*
Model3	*MFCC+MEDC+LPCC*
Model4	*MFCC+MEDC+Energy*
Model5	*MFCC+MEDC+Energy+Pitch*

Table (7) Combinations of feature parameters[25].

The results of this paper are shown in the following tables:

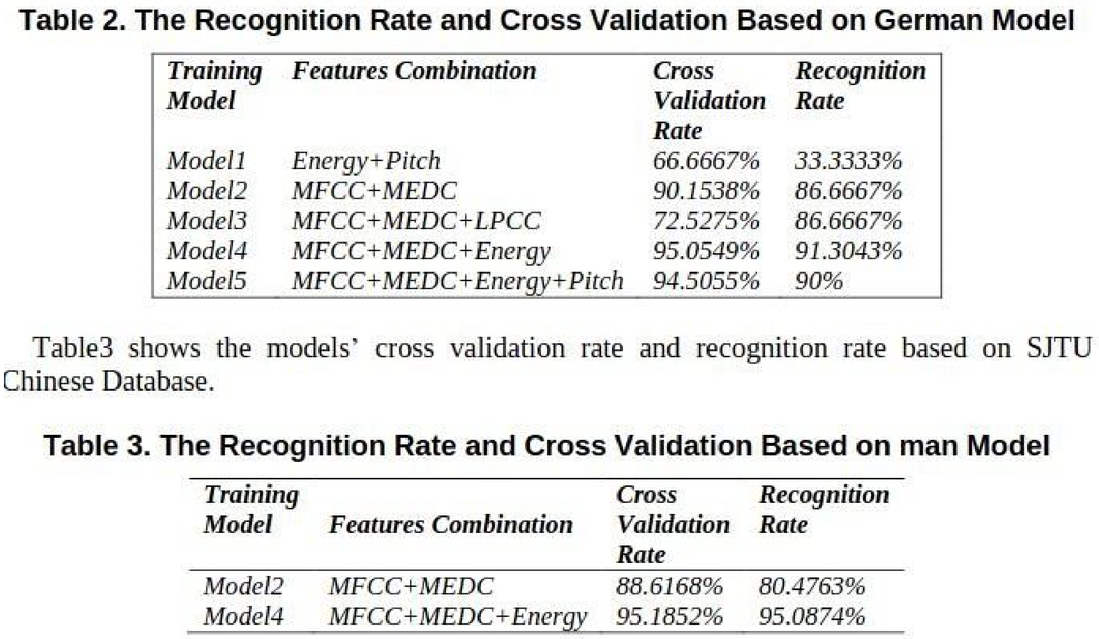


Table 2. The Recognition Rate and Cross Validation Based on German Model

Training Model	*Features Combination*	*Cross Validation Rate*	*Recognition Rate*
Model1	*Energy+Pitch*	*66.6667%*	*33.3333%*
Model2	*MFCC+MEDC*	*90.1538%*	*86.6667%*
Model3	*MFCC+MEDC+LPCC*	*72.5275%*	*86.6667%*
Model4	*MFCC+MEDC+Energy*	*95.0549%*	*91.3043%*
Model5	*MFCC+MEDC+Energy+Pitch*	*94.5055%*	*90%*

Table3 shows the models' cross validation rate and recognition rate based on SJTU Chinese Database.

Table 3. The Recognition Rate and Cross Validation Based on man Model

Training Model	*Features Combination*	*Cross Validation Rate*	*Recognition Rate*
Model2	*MFCC+MEDC*	*88.6168%*	*80.4763%*
Model4	*MFCC+MEDC+Energy*	*95.1852%*	*95.0874%*

Table (8) Recognition rate and cross-validation[25].

Here the importance of Speech emotion recognition systems and its importance in the field of computer science[26]. It explained that these systems consist of three stages: dataset processing, feature extraction, and feature recognition. The quality of the features extracted directly influences the accuracy of the recognition. The dataset used is the ChineseHigh-Performance for analysis of the analysis dataset. It consists of 7 males and 8 females and it has 7 different basic emotions. We can distinguish emotions from each other by several features. For example, sadness and happiness are considered Markov for the analysis of August 30ing the calm speech, so that they have low average amplitude. While anger and surprise have high average amplitude. Also when the speaker is happy, the fundamental frequency curve is going upward. The paper used a combination of Deep Belief Network (DBN) and SVM for the classification task.

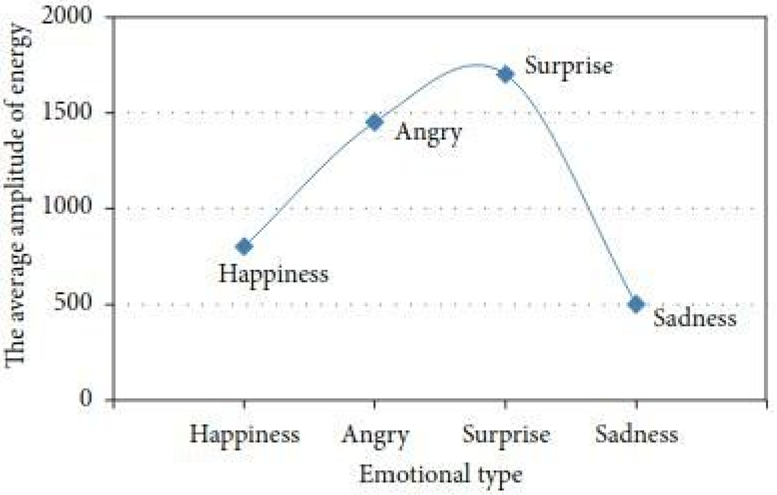


FIGURE 2: The distribution of emotional speech amplitude energy.

Fig (11) Distribution of emotional speech amplitude energy[26].

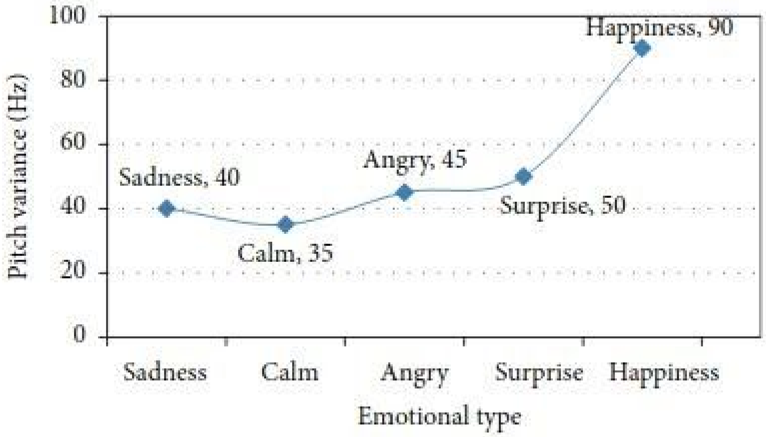


FIGURE 3: Curves of different emotional variance.

Fig (12) Curves of different emotional variance[26].

The DBN layer consists of the following units:

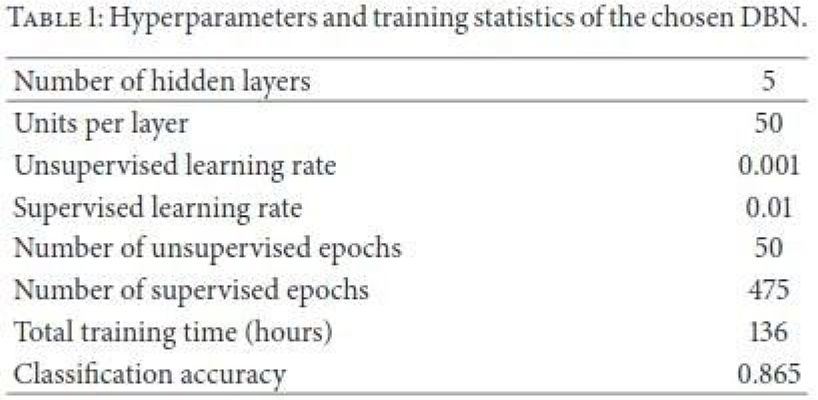


TABLE 1: Hyperparameters and training statistics of the chosen DBN.

Number of hidden layers	5
Units per layer	50
Unsupervised learning rate	0.001
Supervised learning rate	0.01
Number of unsupervised epochs	50
Number of supervised epochs	475
Total training time (hours)	136
Classification accuracy	0.865

Table (9) Hyperparameters and training statistics of the following units[26].

The dataset is spilled into 40% for training and 60% for testing. The accuracy was 7% higher than the traditional classifier.

To conclude, the combinations of features and techniques used to extract emotions were not diverse so we used different sets of combinations(Different algorithms and different features). This led to better emotion recognition. The authors used CycleGANs networks as the dataset used in training is not parallel. Also, to avoid the shortcut that may happen in the GAN when it disregards the input and outputs some limited samples that beat the discriminator.

Project Design

We used multiple datasets in our implementation for both emotion recognition and emotion conversion. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset contains 7356 files. It has 24 professional actors (12 female, 12 male). Emotions are calm, happy, sad, angry, fearful, surprise, and disgusted expressions. The song contains calm, happy, sad, angry, and fearful emotions [5]. In the Toronto emotional speech set, a set of 200 target words were spoken by two actresses (aged 26 and 64 years) and recordings were made of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality) [6]. The EMODB database comprises seven emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral. The dataset consists of 535 utterances from actors who say, one word per utterance [7]. The Arabic Natural Audio dataset(ANAD) consists of 1384 records with 7 different emotions. Each record is 1 second long and is evaluated using 18 listeners who decide which emotion appears in this record. Laughs, silence, and noisy parts were eliminated from the records. The whole dataset was obtained from 8 videos between an anchor and a human outside of the studio. The features of the utterance were extracted then (Principal component analysis)PCA was applied to reduce the number of features to 10%. This means every 190 features were reduced to nineteen features. then the emotions of each utterance were detected [28]. ESD(Emotional Speech Database)[37] is a dataset for emotional voice conversion. It consists of 350 parallel utterances by 10 native English speakers and 10 native Chinese speakers. It covers five emotions(neutral, happy, angry, sad, and surprise). Its duration is 29 hours and can be used in multispeaker and cross-lingual conversion tasks.

- **Features used in emotion recognition layer:**

1. MFCC is a small set of features that can represent the shape of the signal’s spectral envelope.
2. MEDC is similar to MFCC, however, the main difference is, in MEDC, the logarithmic mean of energies is taken while in MFCC logarithmic is taken. Also, the 1st and 2nd differences are computed.
3. The spectral centroid is a measure of how low or high the voice signal is. It is calculated by finding the centroid of the signal using the following formula:-

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

Eq. (1) Centroid Calculation

The result of the formula is the weighted mean of the frequencies of the signal. These frequencies are obtained using a Fourier transform. In the above equation f(n) is the center frequency at bin number n and x(n) is the value or magnitude of the signal at bin number n
The signal regions with a higher magnitude of the spectral centroid mean that the signal is brighter there.

4. Chroma Is a condensed representation of the content of the musical signal. It has an important role in preparation for high-level semantic analysis. It can be used in the analysis of voices whose pitches can be categorized into twelve categories.
5. Contrast is the difference between different speech sounds that makes each sound different from other speech sounds.

- **Features used in emotion conversion layer:**

1. Converted F0 fundamental frequency is the fundamental frequency or F0 is the frequency at which vocal cords vibrate in voiced sounds. This frequency can be identified in the sound produced, which presents quasi-periodicity, the pitch period being the fundamental period of the signal (the inverse of the fundamental frequency).
2. In sound processing, the Mel-frequency cepstrum (MCEPs) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients are coefficients that collectively make up an MFC
3. Aperiodicity measures the aperiodicity of the input utterance and maps it to the output utterance.

Our project consists of three main layers:

I. The recognition layer:

As shown in figure 1, we used Arabic, English, and German emotional datasets and extracted the same features in all three languages. The features that are extracted from the speech are Mel-frequency cepstral coefficients (MFCCs), chroma, Mel spectrogram, contrast, tonnez, spectrum centroid, and zero-crossing rate. These features are used in the machine learning classification to classify emotions. More than one classification algorithm was used on each dataset and the performance was compared to obtain the best classification algorithm. The used algorithms are multilayer perceptron, support vector machine, and decision tree.

A. Multilayer perceptron(MLP):

A multilayer perceptron with 300 neurons and a batch size of 256. MLP was used as it can solve problems stochastically with a good approximate solution like the one of detecting the emotion from voice.

B. Support Vector Machine(SVM):

In this project, SVM was used to classify the emotions obtained from voice as its main use is in classification problems.

C. Decision tree:

It was used as it has an advantage: the ability to classify the input by dividing it into smaller subsets of the dataset.

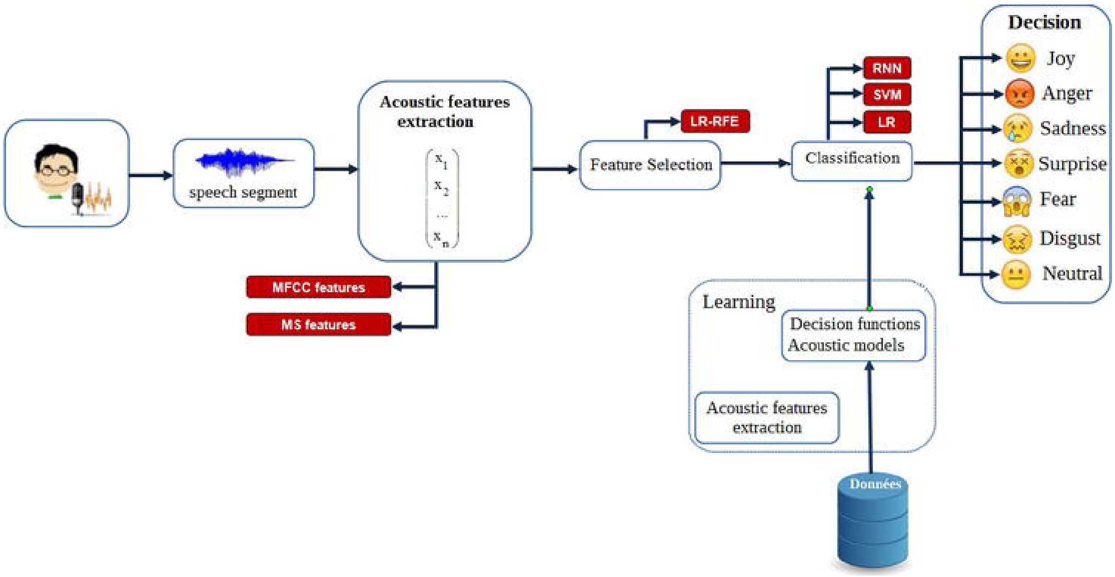


Fig. (13) Speech Emotion Recognition System Flow-Chart
Picture is taken from the journal: Automatic Speech Emotion Recognition Using Machine Learning

II. The chatbot layer:

A chatbot is a software that uses artificial intelligence to conduct a conversation with any human using natural language to perform a specific task or ask a question about anything in different fields[29]. As shown in the figure below, The chatbot gets input from the user and analyzes it. Then the intent of the user is evaluated and the appropriate response is composed.

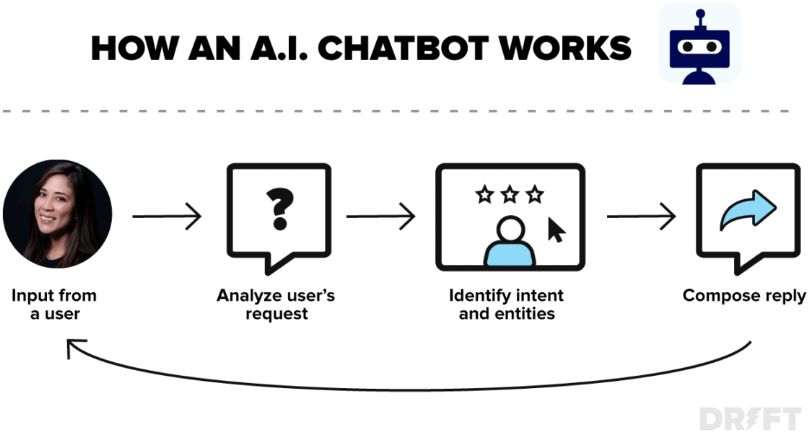


Fig. (14) Operation of the chatbot
The picture is taken from: drift.com An Introduction to AI Chatbots

There are many types of chatbots as they can be classified according to their domain of use as open domain chatbots which do not have a specific field and can talk about general topics, and closed domain chatbots-which is going to be used which has a specific field and focus on a particular knowledge domain[30]. The chatbot that we built is just a proof of idea. It's a simple chatbot. As shown below, it gets the input voice message from the user and converts it from speech to text using Google API and the chatbot knows the intent of the user and finds the appropriate response with emotion and sends it to Google API text-to-speech and the message is said with the appropriate emotion to the user.

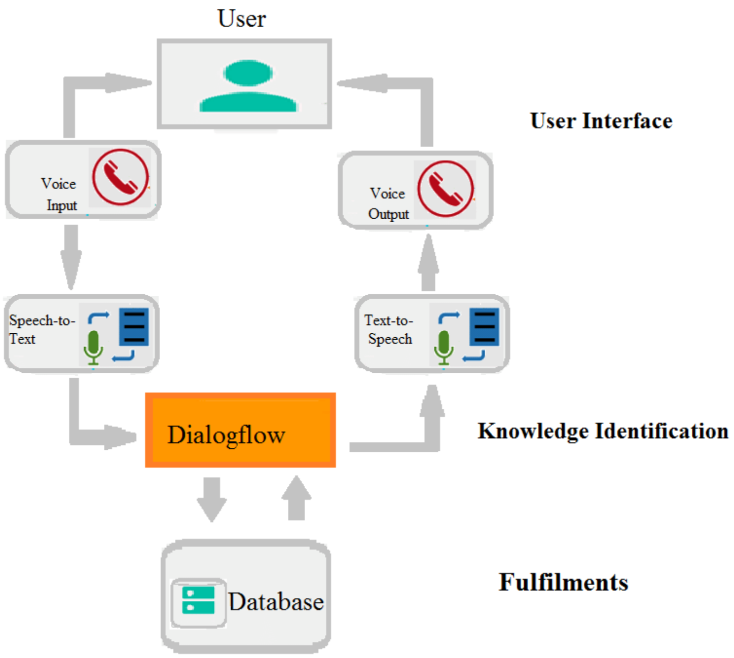


Fig. (15) Architecture of voice based chatbot[31].

III. The emotional voice conversion layer:

This layer converts the speaker's emotion from one emotion to another. Keep in mind that the speaker's identity and the sentence said are not changed. Unlike the voice conversion, the identity of the speaker is changed but the sentence said is the same, as shown below:

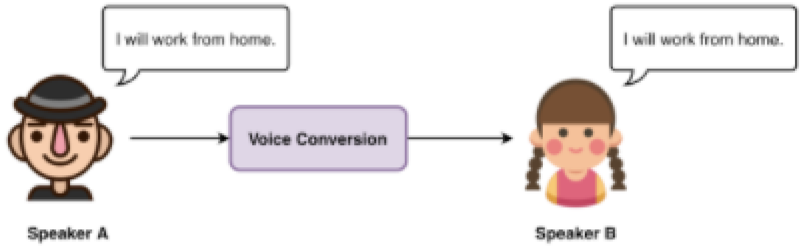


Fig. (16) Voice Conversion (VC) Example with Different Speaker Identities
Source: https://paperswithcode.com/dataset/esd

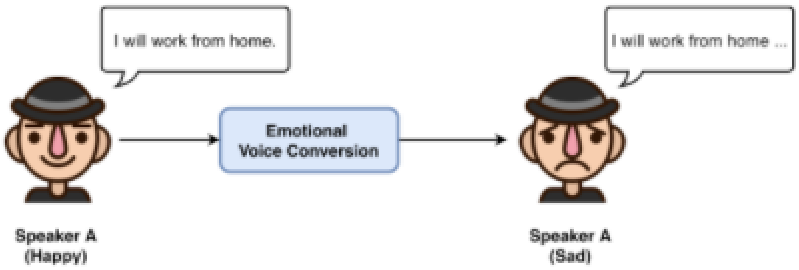


Fig. (17) Emotion Voice Conversion (EVC) Example with Same Speaker Identity
Source: https://paperswithcode.com/dataset/esd

The datasets used in the emotion conversion process are the Emotional Speech Dataset (ESD) and The Interactive Emotional Dyadic Motion Capture (IEMOCAP)[38] full release without videos. We have tried the conversion of Neutral-to-Angry emotion, Neutral-to-Sad emotion, Neutral-To-Happiness emotion.

Below is the training procedure used in our project. We extracted fundamental frequency F0, spectral prosody (SP) using python library "WORLD" from both the source speech and the target speech. We did a log-gaussian normalized transform. We train CycleGAN for spectrum conversion with Mel-cepstral coefficients (MCEPs), and with F0 features for each speech frame. We note that the source and target training data are from the same speaker, but consist of different linguistic content and different emotions. The CycleGAN finds an optimal mapping between source and target spectrum.

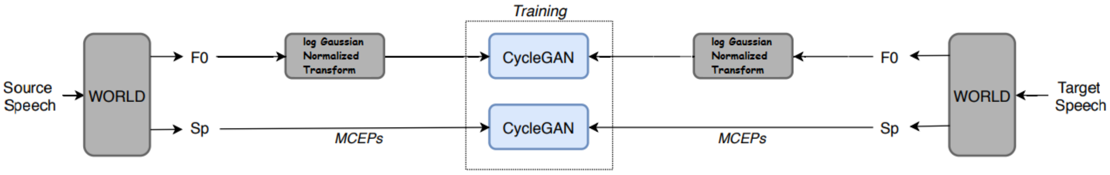


Fig. (18) Flowchart of Training Process[14].

In the conversion process, Frequency F0, spectral prosody (SP), and aperiodicity(AP) are extracted. AP is copied to the target emotion directly without conversion. However, F0 normalized and SP with MCEPs are fed into the CycleGAN that converted them to another F0 and SP as shown below:

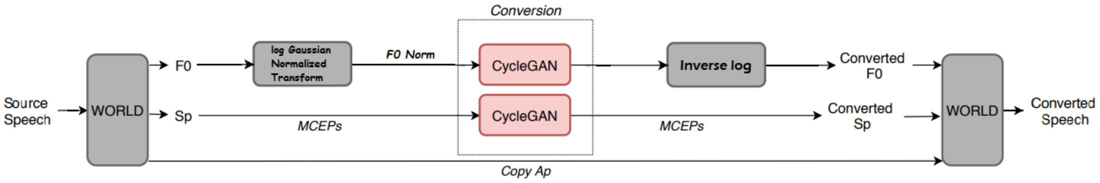


Fig. (19) Flowchart of Conversion Process[14].

The model that we used is cycle-consistent adversarial network Voice Conversion (CycleGAN-VC) with gated convolutional neural networks (CNNs) [4] and an identity-mapping loss [5]. It's used to learn the mapping from one source emotion voice to another target emotion voice. Since speech has sequential and hierarchical structures such as voiced or unvoiced segments and phonemes or morphemes. An effective way to represent such structures would be to build a CycleGAN using gated CNNs that allow parallelization in data. The architecture of the Generator and Discriminator is shown below:

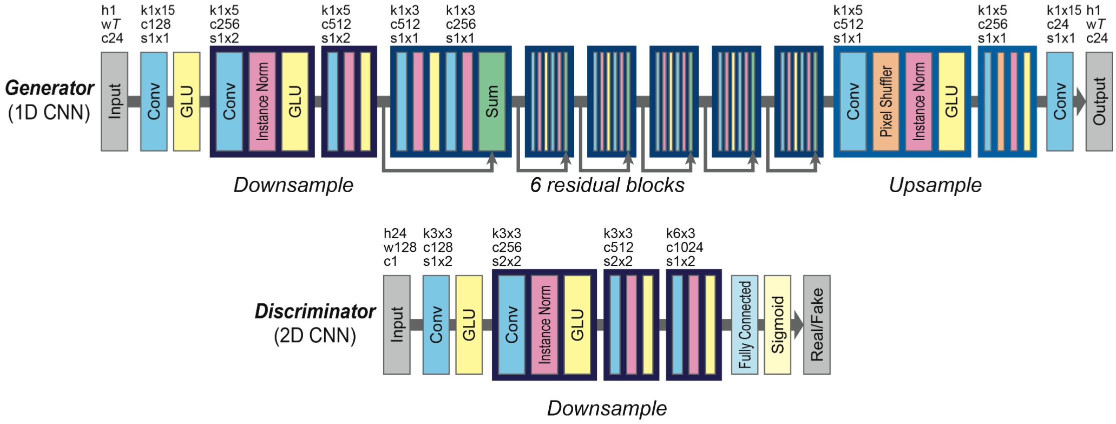


Fig. (20) Architecture of the Generator and Discriminator[32].

Project Execution

The project execution was divided into simple steps as follows:

1. Using the mentioned algorithms and datasets to create an emotion recognition model.
2. Comparing the results of the algorithms and saving the best model.
3. Use Google's speech to text API to convert the audio to text.
4. As we implemented the first and the last layer only by ourselves the second layer will be a simple chatbot just to demonstrate the idea.
5. Then we use Google's text to speech API to convert the text from the chatbot to audio (which has a neutral voice).
6. Then we used the Cyclegan and trained it using both the ESD and the ANAD sets.
7. Finally, we combine all the layers to create a program that takes the audio from the user as an input, recognizes its emotion, converts it into text then the emotion and the text are inputs to the chatbot which should output the text and emotion of the response which is converted to neutral audio using Google's API and finally converted to the desired emotion of the chatbot using the final layer of the EVC.

Results:

The results are divided into two sections for the emotion recognition and conversion layers as follows:

1. The emotion recognition layer:

a. RAVDESS Dataset[33]

Only six emotions are tested which are: Neutral Calm Happy Sad Angry Fearful The first part compares the different classification

High-Performance Realize Algorithms and only MFCC features are extracted. The results are shown below:

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	64.5%
Multilayer perceptron classifier	MFCC	86%
Support vector machine classifier	MFCC	82.21%
Convolutional Neural Network	MFCC	68%

Table (10) Comparison between different classification algorithms with MFCC feature.

The second part is the result of the comparison between different combinations of features and only an MLP classification algorithm is used:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	86%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	85%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	73.52%

Table (11) Comparison between different combination of features with MLP Classifier

b. TESS Dataset[34]

This dataset contains 7 emotions and was generated by 2 females both of them generated 200 audio files for each emotion making a total number of samples in this dataset 2800 samples, the emotions found in this dataset are: Fear Pleasant surprise Sad Angry Disgust Happy Neutral.

The same procedure is done for the TESS dataset; different classification algorithms are tested and only MFCC features are extracted. The results are shown below:

Table. (3) Comparison of different used algorithms and different extracted features

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	91.43%
Multilayer perceptron classifier	MFCC	98.81%
Support vector machine classifier	MFCC	99%
Convolutional Neural Network	MFCC	98.81%

Table (12) Comparison between different classification algorithms with MFCC feature

Furthermore, different features combinations are tested and only MLP classification algorithm is used as shown below:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	98.81%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	99.76%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	99.52%

Table (13) Comparison between different combination of features with MLP Classifier

c. EmoDB Dataset[35]

This dataset contains 7 emotions and was generated by professional speakers (five males and five females) who participated in data recording. The database contains a total of 535 utterances.

As shown below different classification algorithms are tested and only the MFCC feature is extracted:

Algorithm	Features Used	Validation Accuracy
Decision tree classifier	MFCC	36.7%
Multilayer perceptron classifier	MFCC	50%
Support vector machine classifier	MFCC	23%
Convolutional Neural Network	MFCC	NA

Table (14) Comparison between different classification algorithms with MFCC feature

Also, different features combinations and only MLP classification algorithm are used:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC	50%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz	99.76%
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Tonnetz & Spectrum centroid & Zero crossing rate	66.7%

Table (15) Comparison between different combination of features with MLP Classifier

d. The Arabic dataset (ANAD)[36]

The Arabic dataset accuracy without PCA:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	79.5%
Convolutional Neural Network	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	92.6%

Table (16) Comparison between different classification algorithms with ALL features without PCA

The Arabic dataset accuracy with PCA:

Algorithm	Features Used	Validation Accuracy
Multilayer perceptron classifier	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	95.8%
Convolutional Neural Network	MFCC & Mel & Chroma & Contrast & Spectrum centroid & Zero crossing rate	96.5%

Table (17) Comparison between different classification algorithms with ALL features with PCA

2. The emotion Conversion layer:

For the emotion conversion layer, we have used the University's High-realized performance Computing Lab to build and train a CycleGAN Neutral to Angry model. This means we can convert from neutral emotion to angry emotion and vice versa.

- From Neutral to Angry

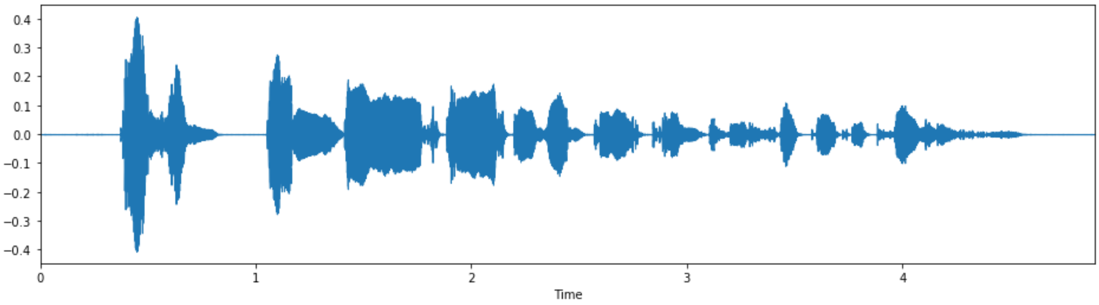


Fig. (21). Neutral Speech Voice

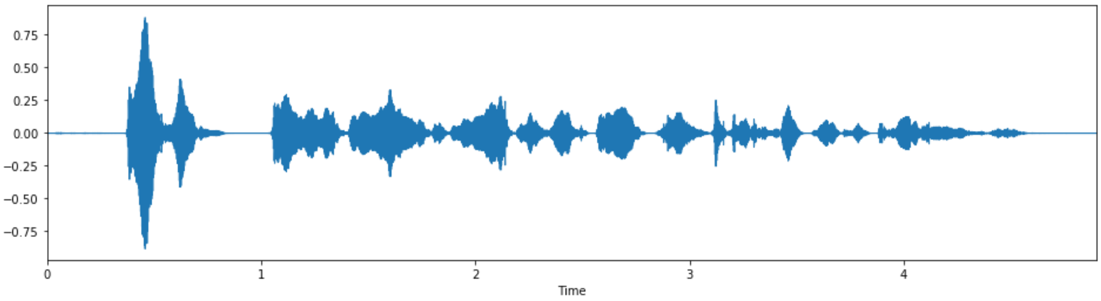


Fig. (22). Converted Angry Tone

- From Angry to Neutral

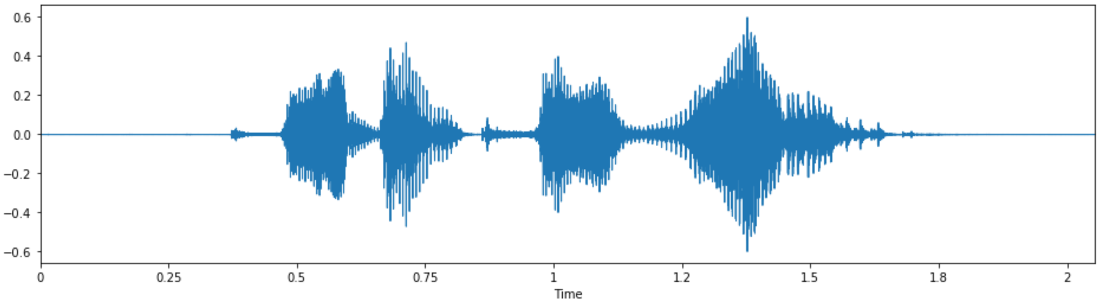


Fig. (23). Angry Speech Voice

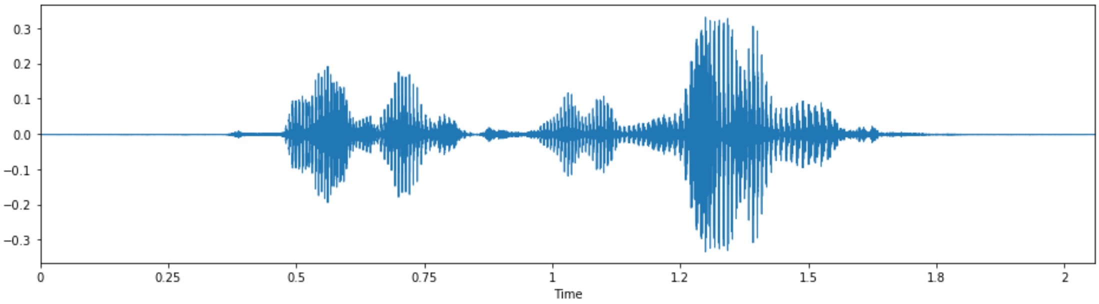


Fig. (24). Converted Neutral Tone

If one listens to both the original audio file and the converted audio file, he will realize that the model produced a good quality audio file with the target emotion although we trained only in 85 epochs and half of the ESD dataset due to the limited time and resources. Note also the Neutral-to-Angry model is exactly similar to Neutral-to-Sad and Neutral-to-Surprise. We, currently, are training the models on IBM Cloud and testing Arabic audio files.

Cost Analysis

This project has no negative effects on the environmental sector. It has a positive impact on the health sector and social impact as this field can be used in the psychological assessment to help the patients and provide emotionally intelligent machines that reply to them appropriately. At the social impact, this project can also be used in call centers to provide the best help to the customers and also can be used in education to help the students and fully understand their questions and give them the best answers they need. Lastly, there are no ethical impacts in our project as there is no use of personal information of any clients or customers, only their emotions are known for better support and response.

Conclusion and Future work

As shown in the previous work it's proven that emotions can be detected and recognized from speech using machine learning and deep learning algorithms with extracting voice features such as MFCC, Mel Contrast, Tonnetz, spectrum centroid, and zero-crossing rate. It's proven also that emotions can be converted from one source emotion to another target emotion using CycleGAN and by extracting features such as F0, spectral prosody, and aperiodicity. We have tried three models: Neutral-to-Angry emotion, Neutral-to-Sad emotion, Neutral-To-Happiness emotion. We used both English and Arabic datasets in the recognition and conversion layer. Moreover, it's proven that the MLP classifier showed better performance if the dataset is small while the neural network overfits. If the dataset is large, the Neural Network will be better. Also, we showed that using a combination of all features and then doing PCA for feature reduction gives greater accuracy and performance than without PCA.

Due to the lack of resources, we couldn't try different types of GANs. We also didn't try the whole dataset; we tried only half of it because it takes days to train. As a result, in the future, we will compare the types of GANs: VAE-GANs, and Sec2Sec models and our CycleGANs model. We will train the whole dataset. We will use a large number of epochs. We will try to build an advanced Arabic chatbot as we were limited in time this semester.

References

[1] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. Social Media and Machine Learning. doi:10.5772/intechopen.84856

[2] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, [1] databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76. doi:10.1016/j.specom.2019.12.001

[3] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS). Toronto: University of Toronto, Psychology Department.

[4] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[5]Z. Du, K. Zhou, B. Sisman, and H. Li, "Spectrum and Prosody Conversion for Cross-lingual Voice Conversion with CycleGAN," *arXiv.org*, 03-Nov-2020. [Online]. Available: https://arxiv.org/abs/2008.04562. [Accessed: 11-Jul-2021].

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization,"ICASSP-88., International Conference onAcoustics, Speech, and Signal Processing, pp. 655–658 vol.1, 1988

[7] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion," *Interspeech 2020*, 2020.

[8] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-basedemotional voice conversion using spectrum and prosody features,"American Journal of Signal Processing, 2012

[9] W.-C. Huang, Y.-C. Wu, C.-C. Lo, P. L. Tobing, T. Hayashi,K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Investi-gation of f0 conditioning and fully convolutional networks in variational autoencoder based voice conversion,"arXiv preprintarXiv:1905.00615, 2019.

[10] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Zero-shot voice style transfer with only autoencoder loss,"arXiv preprint arXiv:1905.05879, 2019.

[11] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via con-ditional autoencoder,"2020 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP), May 2020.

[12] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody con-version,"IEEE/ACM Transactions on Audio, Speech, and Lan-guage Processing, vol. 27, no. 6, pp. 1085–1097, 2019

[13] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel train-ing data," arXiv preprint arXiv:2002.00198, 2020.

[14]K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020.

[15] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 12, pp. 1859–1872, 2014.

[16] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad "Spectral mapping using artificial neural networks for voice conversion," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 954–964, 2010.

[17] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learninJ algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "High- order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in Fifteenth annual conference of the international speech communication association, 2014.

[19]K. Zhou, B. Sisman, and H. Li, "Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training," *arXiv.org*, 09-Jun-2021. [Online]. Available: https://arxiv.org/abs/2103.16809. [Accessed: 12-Jul-2021].

[20] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 2020, pp. 230–237.

[21] 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 ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3502–3506.

[22] D. Griffin and J. Lim, "Signal estimation from modified short time fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.

[23] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in International Conference on Machine Learning, 2018, pp. 2410–2419.

[24] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence- to-sequence voice conversion with disentangled linguistic and speaker representations," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 540–552, 2019.

[25]Pan, Y., Shen, P., & Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine.

[26]Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM. Mathematical Problems in Engineering, 2014 , 1-7. doi:10.1155/2014/749604

[27] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody For Emotional Voice Conversion with Non-Parallel Training Data," inProc. Odyssey 2020 The Speaker and Language Recognition Workshop,2020, pp. 230–237.

[28] B. Sisman and H. Li, “Wavelet analysis of speaker-dependent and independent prosody for voice conversion.” in Interspeech, 2018, pp.52–56

[29] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, “Cross-lingual voice conversion with bilingual phonetic posterior gray and average modeling,” inICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp.6790–6794

[30] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “Gmm-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, vol. 2, no. 5, pp. 134–138,2012.

[31]K. Zhou, B. Sisman, and H. Li, “Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training,” *arXiv.org*, 09-Jun-2021. [Online]. Available: https://arxiv.org/abs/2103.16809. [Accessed: 12-Jul-2021].

[32]K. Zhou, B. Sisman, and H. Li, “Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training,” *arXiv.org*, 09-Jun-2021. [Online]. Available: https://arxiv.org/abs/2103.16809. [Accessed: 12-Jul-2021].

[33]K. Zhou, B. Sisman, and H. Li, “Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training,” *arXiv.org*, 09-Jun-2021. [Online]. Available: https://arxiv.org/abs/2103.16809. [Accessed: 12-Jul-2021].

[34]E. J. Lok, “Toronto emotional speech set (TESS),” *Kaggle*, 24-Aug-2019. [Online]. Available: https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess. [Accessed: 12-Jul-2021].

[35]E. J. Lok, “Toronto emotional speech set (TESS),” *Kaggle*, 24-Aug-2019. [Online]. Available: https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess. [Accessed: 12-Jul-2021].

[36]SamiraKlaylat, “Arabic Natural Audio Dataset,” *Kaggle*, 01-Dec-2017. [Online]. Available: https://www.kaggle.com/suso172/arabic-natural-audio-dataset. [Accessed: 12-Jul-2021].

[37]“ESD dataset 2020 for the time period 2005-2019,” *European Environment Agency*, 16-Dec-2020. [Online]. Available: https://www.eea.europa.eu/data-and-maps/data/esd-2/esd-dataset-2020-for-the. [Accessed: 12-Jul-2021].

[38]IEMOCAP- Home. [Online]. Available: https://sail.usc.edu/iemocap/. [Accessed: 12-Jul-2021].