Intelligible Emotional Voice Conversion with StarGAN assisted by DTW and Speaker Classifier

by

Gökçe İymen

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Data Science

KOÇ ÜNİVERSİTESİ

September 8, 2022

Intelligible Emotional Voice Conversion with StarGAN assisted by DTW and Speaker Classifier

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Gökçe İymen

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Metin Sezgin (Advisor)

_____

Prof. Björn Schuller

_____

Prof. Yücel Yemez

Date: _____

ABSTRACT

Intelligible Emotional Voice Conversion with StarGAN assisted by DTW and Speaker Classifier
Gökçe İymen
Master of Science in Data Science
September 8, 2022

Human speech carries not only linguistic content and speaker identity but also emotional content. The ability to alter emotional colouring of speech has the potential to enable a variety of tasks such as producing affective speech for intelligent dialogue systems and guiding people in their emotional expression abilities. While performing emotional voice conversion, special focus needs to be given to both the quality of speech (e.g., naturalness, intelligibility) and the perceived emotion of the generated speech. In this study, we propose a method for converting the emotion of speech across multiple emotion categories with a single trained model. Our StarGAN-based model, enhanced by the DTW algorithm and auxiliary speaker classifier, can change the emotion of a given speech signal into 3 emotion classes: angry, happy and sad. When building our model, we determine the loss functions targeting distinct attributes of speech including the authenticity, linguistic information, speaker identity, and emotional expression. We evaluate the performance of our model through objective and subjective evaluation criteria for both audio quality and emotional content of the converted speech. The results show that our method compares favourably with the state-of-the-art method in terms of both speech quality and emotional articulateness.

ÖZETÇE

Dinamik Zaman Bükmesi ve Konuşmacı Sınıflandırıcı Destekli StarGAN ile Anlaşılır Duygusal Ses Dönüşümü
Gökçe İymen
Veri Bilimi, Yüksek Lisans
8 Eylül 2022

İnsan konuşması sadece dilsel içerik ve konuşmacı kimliği değil aynı zamanda duygusal içerik de taşır. Konuşmanın duygusal rengini değiştirme yeteneği, akıllı diyalog sistemleri için duygusal konuşma üretme ve insanlara duygusal ifade yeteneklerinde rehberlik etme gibi çeşitli görevleri mümkün kılma potansiyeline sahiptir. Duygusal ses dönüşümü gerçekleştirilirken, hem konuşma kalitesine (örneğin doğallık, anlaşılabilirlik) hem de üretilen konuşmanın algılanan duygusuna özel olarak odaklanılması gerekir. Bu çalışmada, tek bir eğitimli model ile konuşmanın duygusunu çoklu duygu kategorisine dönüştürebilen bir yöntem öneriyoruz. DTW algoritması ve yardımcı konuşmacı sınıflandırıcısı ile geliştirilmiş StarGAN tabanlı modelimiz, verilen bir konuşma sinyalinin duygusunu kızgın, mutlu ve üzgün olmak üzere 3 duygu sınıfına dönüştürebilir. Modelimizi oluştururken, kayıp fonksiyonlarını özgünlük, dilsel içerik, konuşmacı kimliği ve duygusal ifade gibi konuşmanın farklı niteliklerini hedefleyecek şekilde belirliyoruz. Modelimizin performansını, dönüştürülen konuşmanın hem ses kalitesi hem de duygusal içeriği için nesnel ve öznel değerlendirme kriterleri aracılığıyla değerlendiriyoruz. Sonuçlar, yöntemimizin hem konuşma kalitesi hem de duygusal ifade açısından son teknoloji yöntem ile avantajlı kalacak şekilde kıyaslanabilir olduğunu göstermektedir.

ACKNOWLEDGMENTS

I would like to thank my advisor Assoc. Prof. Metin Sezgin for his unwavering guidance, support and encouragement throughout this thesis study. Working with him has not only advanced me in affective computing, deep learning, and research experience, but also in communication, organization, and social skills. I am so grateful to him for the opportunity to work in the Intelligent User Interfaces Lab with high-calibre collaborators. Moreover, I would also like to thank my thesis committee members Prof. Björn Schuller and Prof. Yücel Yemez for their interest and valuable time while evaluating this thesis.

I would like to acknowledge the Scientific and Technological Research Council of Turkey (TÜBİTAK) and the Graduate School of Sciences and Engineering of Koç University for their financial support during my studies.

I am very fortunate to have made great friendships at Koç University. I am in debt to my lab mates Alara, Ezgi, Barış and Alpay for always being supportive for me. I especially want to thank Gizem for our friendship that will last forever. What we learned and experienced together will always be unforgettable to me. I am delighted to know my other friends Baturhan, Ecem, Buse, Çelen, Kerim, Bengi, Mohammed, and Waris. I owe the biggest thanks to Karim for his motivational support during my work and the great times we spent together.

I would also like to thank my friends Beril, Ayşemine, Murat, Nazemin, Aslı, Semih, and Şafak, who were with me on this journey as always. I am grateful to Fatih, Yamaç, Yağmur, Enver, Sencer, Yağız, Emirhan, and Ali for their guidance and support in my hard times. Last but not least, I would like to express my indescribable gratitude to my family. Without their constant support and belief in me, it would not have been possible to complete this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANOVA	Analysis of Variance
Ap	Aperiodicity Parameter
ASR	Automatic Speech Recognition
AS-CWT	Arbitrary Scales Continuous Wavelet Transform
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
EVC	Emotional Voice Conversion
F0	Fundamental Frequency
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LGNT	Logarithm Gaussian Normalized Transformation
LSTM	Long Short-Term Memory
MCC	Mel-Cepstral Coefficient
MCD	Mel-Cepstral Distortion
MOS	Mean Opinion Score
NMF	Non-negative Matrix Factorization
NSF0	Normalized-Segment-F0
RNN	Recurrent Neural Network
SC	Speaker Classifier
SER	Speech Emotion Recognition
TSC	Time Series Classification
TTS	Text-to-Speech
VC	Voice Conversion

Chapter 1

INTRODUCTION

Human speech carries para-linguistic information such as emotional state and speaker identity in addition to the verbal messages to be conveyed. Conversion studies in speech can target any of these non-lexical properties. Although speech conversion is more focused on changing speaker identity (i.e., voice conversion (VC)) [Sisman et al., 2020], [Lorenzo-Trueba et al., 2018], [Zhao et al., 2020], there are also efforts towards emotional voice conversion (EVC) [Zhou et al., 2022], [Luo et al., 2016], [Gao et al., 2019], [Rizos et al., 2020]. Emotional voice conversion aims to alter the emotional state of speech while preserving the speaker identity and linguistic content. Changing the emotional content of an existing speech enables intelligent dialogue systems to produce more affective, and hence more natural speech. It can also help to guide people in their emotional expression abilities through the formative feedback on emotional speech, in which emotions are accurately expressed.

Generally inspired by developments in the VC field [Toda et al., 2007], [Aihara et al., 2014], statistical approaches such as Gaussian Mixture Model (GMM) [Aihara et al., 2012] and Non-negative Matrix Factorization (NMF) [Aihara et al., 2014] were utilized in emotional voice conversion before recent advances in deep learning. Although these works have yielded promising results, the most notable advances in EVC have been achieved through the use of deep learning models.

Previous deep learning-based studies on emotional speech alteration can be grouped into two approaches: models using parallel data [Luo et al., 2016], [Ming et al., 2016] and models using non-parallel data [Gao et al., 2019], [Rizos et al., 2020], where parallel data refers to recordings collected from each speaker with the same linguistic content but different emotions. Collecting parallel speech data is not

practical. Relying, therefore, on parallel data greatly limits the data that can be used. This is why recent efforts are mostly on the usage of non-parallel data.

In models using non-parallel data, we can see inspirations from the image domain (e.g., image style transfer [Gatys et al., 2016] and image-to-image translation [Zhu et al., 2017]). In one of the first non-parallel emotional speech conversion attempts, [Gao et al., 2019] postulate that a speech signal with a disentangled representation can be decomposed into an emotion-independent content code and an emotion-informed style code. Their model design has a distinct style encoder for each emotion, and they perform speaker-dependent model training. Furthermore, inspired by the pioneering method of unpaired image-to-image translation methods, [Bao et al., 2019], [Zhou et al., 2020a], and [Liu et al., 2020] adopt CycleGAN [Zhu et al., 2017] for conversion from neutral speech to each emotion. One impact of CycleGAN is that as many new models have to be trained as the number of emotions, but speaker-independent training remains another challenge [Zhou et al., 2020a, Liu et al., 2020].

In the later-proposed StarGAN-based emotional voice conversion model (StarGAN-EVC) [Rizos et al., 2020], the issues of multi-emotion non-scalability and speaker dependency in previous models are resolved. In that work, the performance of StarGAN-EVC in carrying emotion information is validated by both quantitative and qualitative evaluations, yet speech quality of the generated samples is not analyzed in detail. Since the speeches converted by StarGAN-EVC have artifacts, and mostly suffer from low intelligibility, an improvement on StarGAN-EVC has been studied by [He et al., 2021], where they change the structure of StarGAN from a monolithic feature transformation model to one that learns to disentangle speech features using two encoders; the emotion-independent encoder and the emotion encoder.

Our work adopts the state-of-the-art non-parallel EVC model unified for multi-emotion, i.e., StarGAN-EVC, and aims to improve the intelligibility of the converted speeches without compromising the quality of emotional alteration. For this purpose, we propose to implement dynamic time warping (DTW) loss which takes into account the temporal relation in speech, and to extend the model with an additional

speaker classifier to emphasize speaker features during conversion.

1.1 Contributions

To address the speech quality issue, we modify the distance measure of cycle consistency and identity mapping losses, whose role is to preserve the linguistic content in StarGAN-EVC. Instead of L1 loss, we exploit dynamic time warping (DTW) loss for its better tolerance of time-axis fluctuations in time-dependent data such as speech. Moreover, we discover that attaching an auxiliary speaker classifier - with the same logic as the auxiliary emotion classifier in StarGAN-EVC - improves speech quality and contributes to the emotional change. Furthermore, in order to achieve more refined speech outputs, we combine these improvements in a three-stage training process consisting of reconstruction, emotional conversion and fine-tuning, and we propose to train the model in a progressive manner. We design the training procedure in such a way that different speaking qualities are targeted sequentially, so that authenticity and linguistic content are targeted first, followed by emotional expression and finally speaker identity. Lastly, while comparing the performance of the proposed model with that of the baseline model, we report automatic speech recognition (ASR)-based loss using pre-trained ASR model in addition to mel-cepstral distortion (MCD) and data augmentation results, and present the results of our customized experiments for subjective evaluations. To summarize, our main contributions are: *i)* improving the intelligibility of the audio outputs by switching from L1 loss to DTW loss, *ii)* adding an auxiliary speaker classifier so as to explicitly control the speaker identity feature, *iii)* combining these improvements within a three-stage training process that delivers better quality results, and *iv)* providing novel objective and subjective evaluation approaches for audio quality and emotional content of converted speech outputs.

1.2 Organization

The organization of the following parts is as follows: In chapter 2, we cover the background information on emotional state of speech, emotional speech datasets,

and audio feature extractors. To put our contribution in context, we review the state-of-the-art in emotional voice conversion in chapter 3 while introducing the StarGAN architecture and DTW algorithm. Next, we present the methodology of our contribution in chapter 4 with details of our motivation and model architecture. In chapter 5, we explain the experiment conditions such as the dataset, feature extractor and training details. The results of our evaluations are explained in chapter 6. We conclude in chapter 7 with some remarks and future directions.

Chapter 2

BACKGROUND

2.1 Emotional State of Speech

Human speech expresses explicit and implicit emotions both through acoustic features and linguistic content. Since the emotion in speech is formed by the sophisticated interaction of semantic choices of spoken words and vocal attributes such as intonation, loudness and speech pace, it is not easy to fully define the *emotional state of speech*. Therefore, we see different emotion representations in the literature [Banse and Scherer, 1996], [Cowie and Cornelius, 2003], [Zhou et al., 2022], [Ekman, 1992], [Russell, 1980]. The most widely accepted representations used in the affective community are the categorical and dimensional models. While Ekman's framework [Ekman, 1992] groups emotions into six basic categories as anger, fear, sadness, enjoyment, disgust, and surprise, the affective states are not considered independent of each other in Russell's circumplex model [Russell, 1980] where emotions are projected on arousal (active/passive) and valence (pleasure/displeasure) dimensions.

Beyond the lexical words used in speech, acoustic features are the most important factors that create emotion in speech [Zhou et al., 2022]. Analyzing the significance of different factors on emotional state of speech, [Xue et al., 2018] concentrate on four types of acoustic features: duration, F0 contour, spectral sequence and power envelope. With the procedure of modifying only one factor at a time, they conclude that replacement of F0 contour and spectral sequence alter the emotional expressiveness of neutral speech in the direction of the desired emotion, while change in duration and power envelope do not have much effect. This makes F0 contour and spectral sequence the features in focus for emotional speech synthesis and conversion [Luo et al., 2016], [Gao et al., 2019], [Rizos et al., 2020].

2.2 Emotional Speech Datasets

The characteristics of datasets used in training of deep learning models are as significant as the model architectures or training parameters. Yet, the variety and quality of datasets in the field of emotional speech have not progressed remarkably until recent times. Fortunately, with the increasing interest in emotional speech studies, suitable academic datasets have emerged recently. In this context, Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [Busso et al., 2008] can be considered as a groundbreaking English-language collection consisting of audio-visual recordings of 10 actors. The total size of the dataset (∼12 hours) and its focus on the authenticity of emotions make it invaluable for the affective computing community.

Open-source emotional speech datasets vary in different aspects, such as their recording composition (i.e., acted, induced, and spontaneous), their emotion model choice (i.e., sentimental, categorical, and dimensional), or their data annotation method (i.e., actor intention and perceptual evaluation). These aspects should be taken into account when choosing among these datasets according to the intended use of the models. For instance, when building a text-to-speech (TTS) model that attempts to have more natural speeches with different emotions, the dataset to be used should be as clear as possible without overlapping speech, and as diverse as possible in terms of sentences used in recordings [Adigwe et al., 2018]. On the other hand, if a model is designed to be used on real-life sounds, as in the EVC model, it would be more accurate to continue with a dataset which includes some background noise. Otherwise, the model will perform poorly when it encounters real-life sounds in service.

Based on their emotional expression setup, these datasets can be classified as acted, induced, or spontaneous, with the majority being acted or induced. While in acted ones [Adigwe et al., 2018], [Zhou et al., 2021], the same utterance is performed in different emotions that are directly tied to the actor's performance, the speakers are directed prior to recording in induced ones [Cao et al., 2014], [Busso et al., 2017], [James et al., 2018]. Moreover, it is also possible to find spontaneous speeches in

datasets picked from audio-video sharing platforms [Zadeh et al., 2018], [Lotfian and Busso, 2019]. On the other hand, the selection of the linguistic content may vary in different datasets. For example, [Cao et al., 2014], [Adigwe et al., 2018], [James et al., 2018] have paid great attention to choosing semantically-neutral sentences. How the emotion labels are annotated in a dataset is as crucial as its collection design. While perceptual evaluations have been conducted subsequent to dataset collection in [Busso et al., 2008], [Busso et al., 2017], and [Lotfian and Busso, 2019], [Adigwe et al., 2018] and [Zhou et al., 2021] assign emotion labels to samples based on actor intention during recording.

According to the designs of deep learning models, another feature that is investigated in datasets is whether they have parallel or non-parallel data. For the EVC field, we see that models using parallel data were initially more common [Luo et al., 2016], [Ming et al., 2016], [Robinson et al., 2019]. Since sequence-to-sequence models (e.g., LSTM) are trained to learn mappings between two records of the same sentence in different emotions, parallel training data is needed for these models. However, following the improvements in image-domain for unpaired translation [Zhu et al., 2017], [Choi et al., 2018] and their reflections to voice-domain [Kaneko and Kameoka, 2018], [Kameoka et al., 2018], EVC models have started to eliminate the requirement for parallel data. Overall, we see that both parallel [Adigwe et al., 2018], [Zhou et al., 2021] and non-parallel [Busso et al., 2008] emotional speech datasets are used in emotional speech synthesis [Tits et al., 2019] and conversion models [Zhou et al., 2022], [Rizos et al., 2020]. However, the IEMOCAP dataset [Busso et al., 2008] appears to be the most favoured in EVC models [Gao et al., 2019], [Bao et al., 2019], [Rizos et al., 2020] due to its superiority in simulating real-life conversations and emotions.

2.3 Audio Feature Extractors

Nowadays, deep learning models have become increasingly popular in speech-related applications such as automatic speech recognition [Dong et al., 2018], [Baevski et al., 2020], speech synthesis [Oord et al., 2016], [Wang et al., 2017], voice conversion

[Kaneko and Kameoka, 2018], [Kameoka et al., 2018], and emotional voice conversion [Luo et al., 2016], [Rizos et al., 2020]. These models are expected to learn various speech characteristics (intonation, loudness variations, pitch, etc.) from the speech data they use. Thus, the data preprocessing step that extracts useful information from speech is of paramount importance in the audio field.

When corpora containing speech data are examined, we see that speech is mostly collected and stored in waveform. Waveform, which is the time-domain representation of a sound, shows the air pressure variations over time in terms of amplitude. Although raw audio waveforms can be used in deep learning models [Baevski et al., 2020], [Oord et al., 2016], mostly audio in waveform is transformed to feature vector sequences by applying audio signal processing methods [Dong et al., 2018], [Wang et al., 2017], [Kameoka et al., 2018], [Rizos et al., 2020]. There are two main motivations to implement feature extraction on audios. First, using audio feature extractors reduces the data dimension while preserving significant characteristics and removing irrelevant and noisy data. Second, converting audio data into 2D representation such as spectrogram and mel-cepstral coefficients (MCCs) enables the acquisition of image-like data from audio, and this gives the opportunity to use deep learning architectures originally developed for the image domain.

Reviewing feature extraction for audio, we should definitely mention vocoders that have been employed for audio analysis and synthesis for decades [Schroeder, 1966]. Originally developed for the reconstruction of human speech [Dudley, 1939], they have since been used as an important tool in fields such as speech transmission, audio storage, and voice encryption. Although unidirectional vocoders exist [Griffin and Lim, 1984], they typically include both speech encoding and decoding, the former allowing to extract interpretable features from the audio.

In addition to conventional vocoders which use digital signal processing (DSP) for audio analysis, neural network-based vocoders are becoming more common through easier access to data and recent advances in deep learning. From conventional vocoders, STRAIGHT [Kawahara et al., 1999] and WORLD [Morise et al., 2016] are the most preferred in deep learning models for speech [Hsu et al., 2016], [Luo et al., 2016], [Ming et al., 2016], [Kameoka et al., 2018], [Gao et al., 2019], [Rizos

et al., 2020]. Recently, however, neural speech synthesizers such as WaveNet [Oord et al., 2016] and WaveRNN [Kalchbrenner et al., 2018] have also started to be used [Shen et al., 2018], [Paul et al., 2020].

Chapter 3

RELATED WORKS

3.1 Emotional Voice Conversion

Voice conversion is a research topic that has been active for decades [Abe et al., 1988] [Sisman et al., 2020], and its aim is to change the speaker identity in speech while preserving linguistic structure. With increasing interest in affective computing [Schuller and Schuller, 2018], the methods used for voice conversion have also started to be applied in emotional voice conversion to transition between different emotional states of speech. In this regard, statistical approaches from VC domain such as Gaussian Mixture Model (GMM) [Toda et al., 2007] and Non-negative Matrix Factorization (NMF) [Aihara et al., 2014] were initially proposed for emotional voice conversion [Aihara et al., 2012] [Ming et al., 2015]; however, the major breakthroughs were made by later works using deep learning models. Deep learning-based emotional voice conversion studies can be examined under two groups: models with parallel data and models with non-parallel data.

3.1.1 Deep Learning-Based Models Using Parallel Data

The use of parallel data was primarily more common as early studies learned the mapping between paired speech data of two different emotions. To model the relationship between source and target speech, neural network and deep belief network were employed by [Luo et al., 2016], [Luo et al., 2017] for the conversion of F0 and MCC, respectively. They investigated the use of normalized-segment-F0 (NSF0) [Luo et al., 2016] and arbitrary scales continuous wavelet transform (AS-CWT) [Luo et al., 2017] representations of F0 in their studies. On the other hand, [Ming et al., 2016] and [Robinson et al., 2019] proposed the usage of sequence-to-sequence models for feature mapping between speech sequences.

3.1.2 Deep Learning-Based Models Using Non-Parallel Data

Although the aforementioned studies rely upon the availability of parallel training data, inspired by advances in image-domain for unpaired translation, models using non-parallel data have been developed in the VC domain, and subsequently in the EVC domain. The first attempt for non-parallel EVC [Gao et al., 2019] considered the speech signal as having a disentangled representation and separately extracted emotion code and emotion-independent content code from speech. Using emotion-specific style encoders and adversarial principles, they altered one speaker's voice from the source emotion to the target one. However, this approach lacks scalability for translation into multiple emotions since it requires as many style encoders as emotions. Later, CycleGAN [Zhu et al., 2017] became the most popular architecture for EVC with non-parallel data [Bao et al., 2019], [Zhou et al., 2020a], [Liu et al., 2020], but this approach also requires training a separate model for each pair of emotions. In addition to scalability, speaker dependency, i.e., the fact that models can be trained with data from a single speaker, remains another limitation in these prior studies [Gao et al., 2019], [Zhou et al., 2020a], [Liu et al., 2020]. As a solution for speaker-independent EVC, VAW-GAN-based EVC [Zhou et al., 2020b] and StarGAN-based EVC [Rizos et al., 2020] frameworks have been proposed. StarGAN-EVC, the baseline used in our study, also enables many-to-many translation through a unified model using the auxiliary classifier. Nevertheless, the StarGAN-EVC suffers from poor speech quality and linguistic distortions.

3.2 StarGAN

The StarGAN framework, which allows multi-domain translation in a unified model, has become highly favoured in both the image and voice domains. Our work, based on the StarGAN-EVC model, can perform many-to-many domain (i.e., emotion) translation with a single generator, and it does not depend on the presence of parallel data thanks to its model and loss structures.

StarGAN [Choi et al., 2018] is originally proposed to tackle the scalability and effectiveness issues of cross-domain translation models in computer vision. There-

after, StarGAN implementations in the voice domain begin with the StarGAN-VC [Kameoka et al., 2018] model which enables non-parallel many-to-many VC. Unlike StarGAN, StarGAN-VC employs an auxiliary domain classifier (i.e., speaker classifier) different from the discriminator, and adopts identity mapping loss in CycleGAN [Zhu et al., 2017] to strengthen the preservation of linguistic content. Although StarGAN-VC is capable of performing non-parallel many-to-many VC, the audios generated by StarGAN-VC are not of great quality; thus, models with enhancements to StarGAN-VC were proposed by [Li et al., 2020] and [Sakamoto et al., 2021].

Utilizing a similar network architecture to StarGAN-VC, StarGAN-EVC [Rizos et al., 2020] aims to modify the emotion in speech rather than the speaker identity. The components of StarGAN-EVC, namely the generator, discriminator and emotion classifier are illustrated in Figure 3.1 and briefly explained in the following sections.

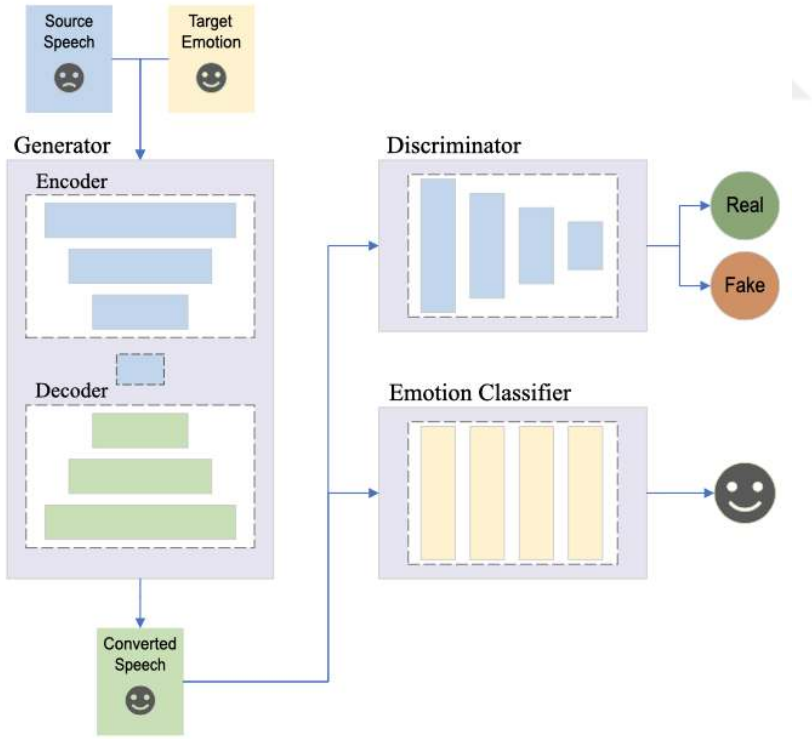


Figure 3.1: Components of StarGAN-EVC: the generator, discriminator and emotion classifier.

3.2.1 The Generator

The generator consists of fully convolutional layers and takes the input acoustic feature as a 3D array of size (F x T x 1), where F is the number of features and T is the number of time frames, similar to a single-channel image. Having an encoder-decoder structure, the generator first downsamples the input in order to obtain an emotion-independent latent code, then upsamples the latent code to generate an output acoustic feature of the same size as the input; (F x T x 1). While upsampling the latent code one-hot encoded emotion labels are channel-wise concatenated to the output of each layer.

During training, the generator attempts to deceive the discriminator by producing real-like speeches using adversarial loss while also trying to preserve linguistic content through cycle consistency and identity mapping losses. In addition, encouraged by the classification loss provided by the emotion classifier, it aims to ensure that the generated output is in the target emotion. The losses from which the generator learns can be formulated as follows:

Adversarial loss:

$$\mathcal{L}^{G}_{adv} = -\mathbb{E}_{x,c}\left[\log D(G(x,c),c)\right] \tag{3.1}$$

Cycle consistency loss:

$$\mathcal{L}^{G}_{cyc} = \mathbb{E}_{x,c}\left[\|x - G(G(x,c),c_x)\|_1\right] \tag{3.2}$$

Identity mapping loss:

$$\mathcal{L}^{G}_{id} = \mathbb{E}_{x}\left[\|x - G(x,c_x)\|_1\right] \tag{3.3}$$

Emotion classification loss:

$$\mathcal{L}^{G}_{emo\_cls} = -\mathbb{E}_{x,c}\left[\log C_{emo}(c \mid G(x,c))\right] \tag{3.4}$$

3.2.2 The Discriminator

Following the convolutional downsampling layers and average pooling, the discriminator tries to estimate the realness/fakeness of the samples with the tanh function.

Since the training objective of the discriminator is to differentiate fake samples from real samples, it uses the following adversarial loss:

$$\mathcal{L}^{D} = -\mathbb{E}_{x}\left[\log D(x, c_x)\right] - \mathbb{E}_{x,c}\left[\log(1 - D(G(x, c), c))\right] \tag{3.5}$$

3.2.3 The Emotion Classifier

Inspired by the attention-augmented speech emotion recognition (SER) model [Zhang et al., 2019], [Rizos et al., 2020] employ an emotion classifier which includes two-dimensional convolutional layers and LSTMs, followed by weighted pooling with attention. Full architecture of the emotion classifier can be seen in Figure 3.2. This model is pre-trained with weighted cross-entropy loss before being combined with the entire model, and then fixed while training the generator and discriminator. The loss of emotion classifier is denoted as follows:

$$\mathcal{L}^{C_{emo}} = -\mathbb{E}_{x}\left[\log C_{emo}(c_x \mid x)\right] \tag{3.6}$$

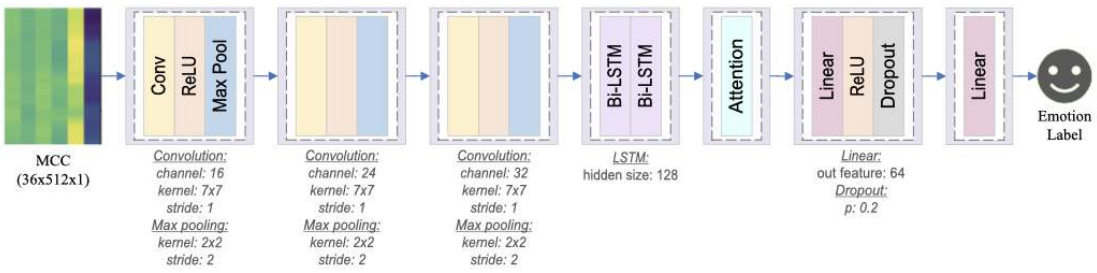


Figure 3.2: Emotion classifier model used in StarGAN-EVC.

3.3 Dynamic Time Warping and Soft-DTW

Dynamic time warping (illustrated in the lower part of Figure 3.3) is a technique used to find an alignment or to compare the similarity between two time series, i.e., two sets of samples collected at equal time intervals [Müller, 2007]. Although its applications were initially focused on speech processing [Sakoe and Chiba, 1978], [Rabiner et al., 1978], [Myers et al., 1980], it is currently used in various fields

such as time series classification (TSC) [Bagnall et al., 2017], [Abdelmadjid and Boucheham, 2021], gesture generation [Maghoumi, 2020], [Maghoumi et al., 2021], and handwriting recognition [Rath and Manmatha, 2003], [Efrat et al., 2007]. In EVC, DTW has so far been used only to find the alignment between parallel source and target data prior to model training [Ming et al., 2015], [Luo et al., 2016], [Luo et al., 2017], [Adigwe et al., 2018].

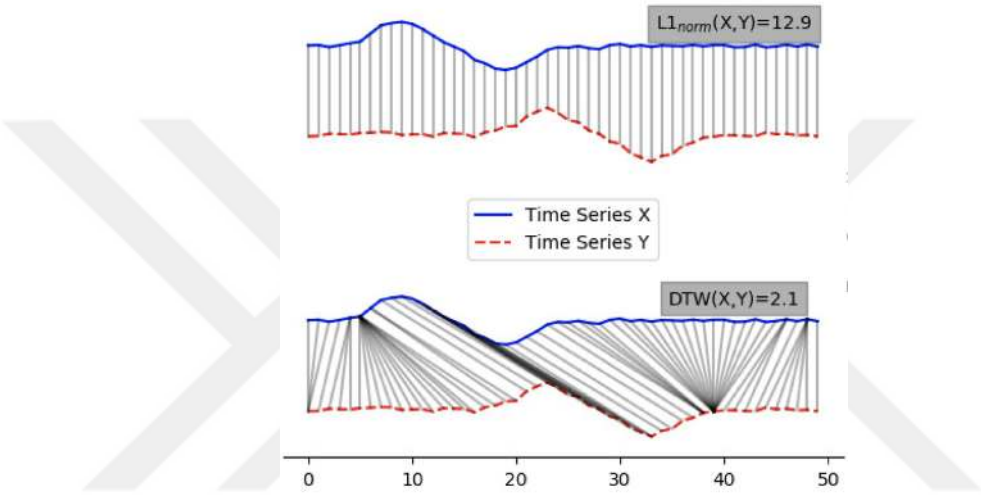


Figure 3.3: Illustration of L1-norm distance and DTW distance between two time series, X and Y [Cai et al., 2021].

3.3.1 DTW Algorithm

Given two time series $X = (x_1, x_2, \ldots, x_N), N \in \mathbb{N}$ and $Y = (y_1, y_2, \ldots, y_M), M \in \mathbb{N}$, the dynamic time warping algorithm first generates the local distance matrix:

$$D \in \mathbb{R}^{N*M} : \quad D_{n,m} = d(x_n, y_m) \quad n \in [1, N], \quad m \in [1, M] \quad (see\ Figure\ 3.4) \tag{3.7}$$

where local distance measure $d(x_n, y_m)$ is any distance function which typically yields low distance when x_n and y_m are similar, and higher distance otherwise. Using the local distance matrix, which contains the pairwise distances of time frames in X and Y, the algorithm aims to find the optimal alignment path p^* between two time series. An alignment path p (also called a warping path) consisting of a sequence

$p = (p_1, p_2, \ldots, p_K), K \in \mathbb{N}$ where $p_k = (n_k, m_k)$ must adhere to the following rules [Müller, 2007], [Senin, 2009]:

Boundary condition: The starting and ending pairs of warping path should match the first and last time frames of X and Y.

$$p_1 = (1, 1) \quad and \quad p_K = (N, M)$$

Monotonicity condition: Alignment indices of time series should monotonically increase on the alignment path.

$$n_1 <= n_2 <= \ldots <= n_K$$
$$m_1 <= m_2 <= \ldots <= m_K$$

Step size condition: Alignment path should not jump over any index of any time series so that each index from one time sequence can match with one or more indices from the other sequence.

$$p_{k+1} - p_k \in \{(0, 1), (1, 0), (1, 1)\}$$

All alignment paths (examples shown in Figure 3.4) satisfying these three conditions form the set of alignment paths P.

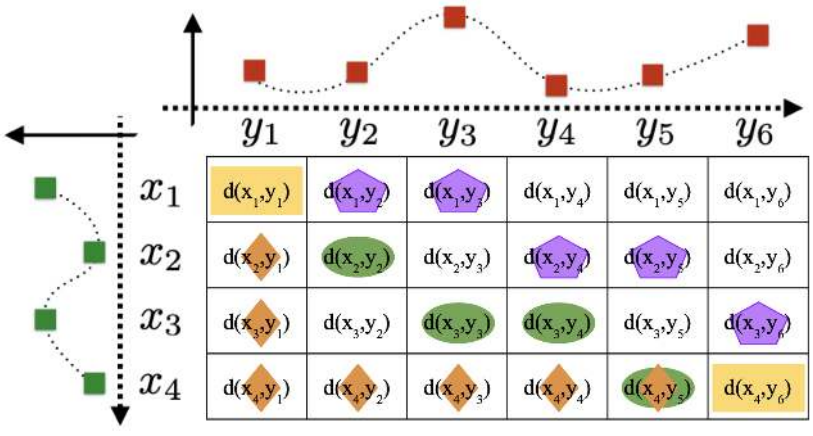


Figure 3.4: Local distance matrix where N = 4, M = 6. Three examples of alignment paths shown in orange, green and purple, with the start and end points of all alignment paths shown in yellow (Figure adapted from [Cuturi and Blondel, 2017]).

Alignment cost $c_p(X, Y)$ is defined as the total distance of alignment path p on local distance matrix D, and the minimum of all alignment costs (c_{p^*}) gives the DTW distance of two time series:

$$c_p(X, Y) = \sum_{k=1}^{K} D_{p_k} = \sum_{k=1}^{K} d(x_{n_k}, y_{m_k}) \tag{3.8}$$

$$DTW(X, Y) = min\ \{c_p(X, Y) \mid p \in P\} \tag{3.9}$$

Instead of calculating the cost of all possible alignment paths and finding the optimal one among them, DTW proposes to utilize dynamic programming (DP) to build the accumulated cost matrix C, thereby obtaining $DTW(X, Y)$:

$$C \in \mathbb{R}^{N*M}: \quad C_{n,m} = \begin{cases} \sum_{k=1}^{m} d(x_1, y_k) & \text{if } n = 1 \\ \sum_{k=1}^{n} d(x_k, y_1) & \text{if } m = 1 \\ d(x_n, y_m) + min\ \{C_{n,m-1}, C_{n-1,m}, C_{n-1,m-1}\} & \text{otherwise} \end{cases} \tag{3.10}$$

$$DTW(X, Y) = C_{N,M} \tag{3.11}$$

3.3.2 Soft-DTW

DTW can find the optimal alignment path p^* and the corresponding minimum alignment cost between two time series; $c_{p^*}(X, Y)$. However, the fact that the DTW loss is only locally differentiable on the optimal alignment path makes it difficult to optimize, hence preventing us from using it to train a neural network with a gradient-based optimization technique. Therefore, [Cuturi and Blondel, 2017] present a method for differentiable DTW called *soft-DTW* using a smoothing parameter γ. They replace the *min* operation in equation 3.9 with the *sum* operation. Since this formulation is differentiable everywhere, this loss can be exploited for backpropagation in neural networks [Maghoumi, 2020], [Maghoumi et al., 2021].

$$soft\text{-}DTW(X,\ Y) = -\gamma\ log(\sum_{p \in P} e^{\frac{-c_p(X,Y)}{\gamma}}), \quad \gamma > 0 \tag{3.12}$$

Chapter 4

METHODOLOGY

4.1 Motivation

StarGAN-EVC uses a single generator for many-to-many emotional voice conversion while allowing speaker-independent training on non-parallel data. These capabilities make it one of the most advanced methods for EVC; however, its outputs need further improvement in terms of speech quality in order to be used in real-world applications such as intelligent dialogue systems. Hence, we have focused on techniques to enhance its speech quality without diminishing its effectiveness in emotional change.

To encourage better intelligibility, we concentrate on training losses responsible for maintaining linguistic information: cycle consistency and identity mapping losses. Considering that DTW loss is a more robust similarity measure against small time variations in time-series than the pixel-by-pixel L1 loss, we explore the use of DTW loss in StarGAN-EVC. Indeed, [Donahue et al., 2021] have shown that DTW loss which relaxes the exact alignment for target and generated speech features provides significant performance increase in text-to-speech.

In StarGAN-EVC, the cycle consistency and identity mapping losses implicitly attempt to keep the speaker characteristics of the converted speech the same as the source speech. However, since an additional loss dedicated specifically to this task can further assist the model training in obtaining better outcomes, we have attached another component to the StarGAN-EVC network architecture, namely the speaker classifier.

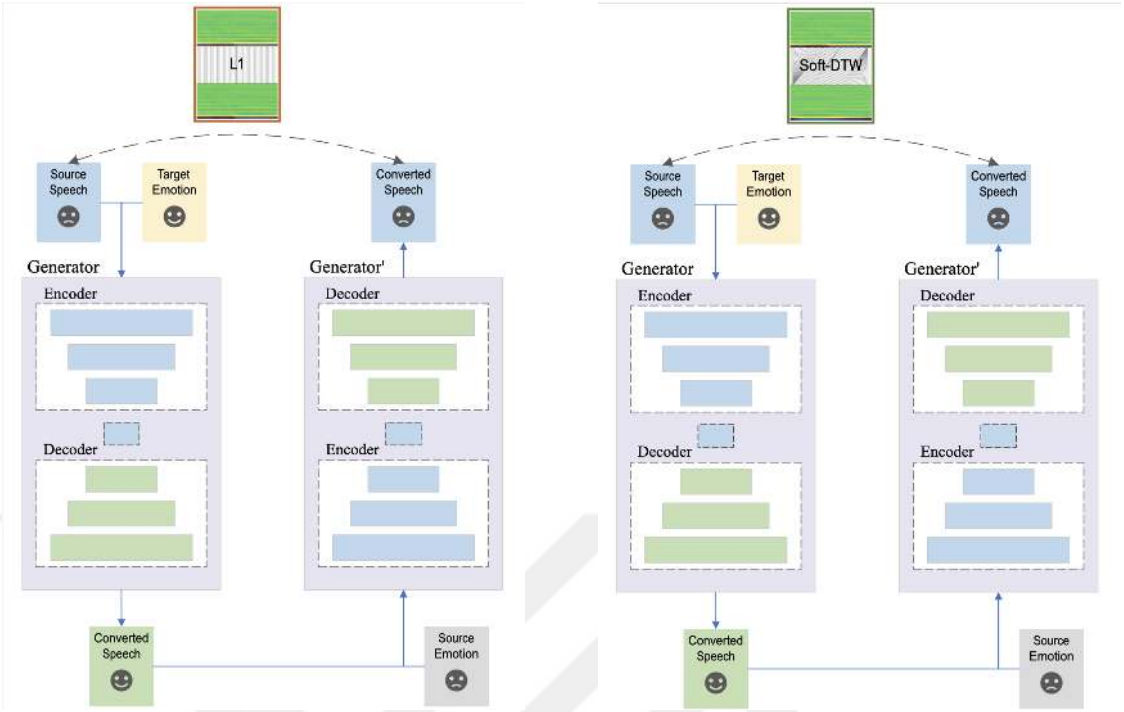


Figure 4.1: Changing cycle loss from L1 loss to DTW loss: (a) cycle loss in StarGAN-EVC, (b) cycle loss in our model.

4.2 DTW Loss

Relying on L1 loss, the cycle consistency and identity mapping losses in StarGAN-EVC are intended to ensure that the linguistic information of speech remains unchanged during the conversion process. However, the simple L1 loss cannot tolerate any time shift in model outputs, making it too restrictive when comparing two speech feature sequences, thus misleading for the model. Considering the temporal characteristics of speech, instead of L1 loss, we utilize a DTW-based loss which is a well-accepted practice to compare the similarity between two time series data. To the best of our knowledge, this is the first work that utilizes DTW-based training loss in the non-parallel EVC domain. The DTW-based loss relaxes the time-dimensional matching of the speech features, allowing the model to focus more on mismatches in the frequency dimension. This helps improve the sound quality of speech by eliminating artifacts such as crackle and buzz.

Since vanilla DTW computation is not differentiable everywhere, it is not practical to use it for the gradient flow in neural networks training. At this point, the use of Soft-DTW having a soft minimum instead of a minimum has become functional. Therefore, as depicted for cycle consistency loss in Figure 4.1, we replace the equations 3.2 and 3.3 with the following equations:

Cycle consistency loss:

$$\mathcal{L}_{cyc}^{G} = \mathbb{E}_{x,c}\left[soft\text{-}DTW(x,\ G(G(x,c),c_x))\right] \tag{4.1}$$

Identity mapping loss:

$$\mathcal{L}_{id}^{G} = \mathbb{E}_{x}\left[soft\text{-}DTW(x,\ G(x,c_x))\right] \tag{4.2}$$

4.3 Auxiliary Speaker Classifier

The generator in StarGAN-VC is driven by an auxiliary speaker recognizer to modify the speaker characteristics of the voice, while the generator in StarGAN-EVC likewise uses an auxiliary emotion recognizer to alter the emotional state. In StarGAN-EVC, other features of speech including linguistic content as well as speaker identity are protected from destruction by the cycle consistency and identity mapping losses. Since there is no direct control of the speaker class, some information about this feature may be missed. In fact, we think that an auxiliary speaker classifier could be employed to preserve the speaker feature of speech this time, rather than modifying it. As a result, we extend the StarGAN-EVC model with an auxiliary speaker classifier. In this way, the model explicitly learns to produce the speech in such a way that the speaker class of converted speech is the same as that of the source speech.

In the speaker recognizer, reusing the architecture of the emotion classifier, we change the output dimension of the last fully connected layer in order to adapt it to the number of speaker classes. The loss of speaker classifier is defined as follows:

$$\mathcal{L}^{C_{spk}} = -\mathbb{E}_{x}\left[\log C_{spk}(s_x \mid x)\right] \tag{4.3}$$

After pre-training the speaker classifier, we use it to further fine-tune the generator and discriminator. During fine-tuning, we freeze the weights of the pre-trained

speaker classifier, and train only the generator and discriminator. The speaker classifier adds the following loss to the training objectives of the generator.

Speaker classification loss:

$$\mathcal{L}^{G}_{spk\_cls} = -\mathbb{E}_{x,c}\left[\log C_{spk}(s_x \mid G(x,c))\right] \tag{4.4}$$

4.4 DTWSC-StarGAN-EVC

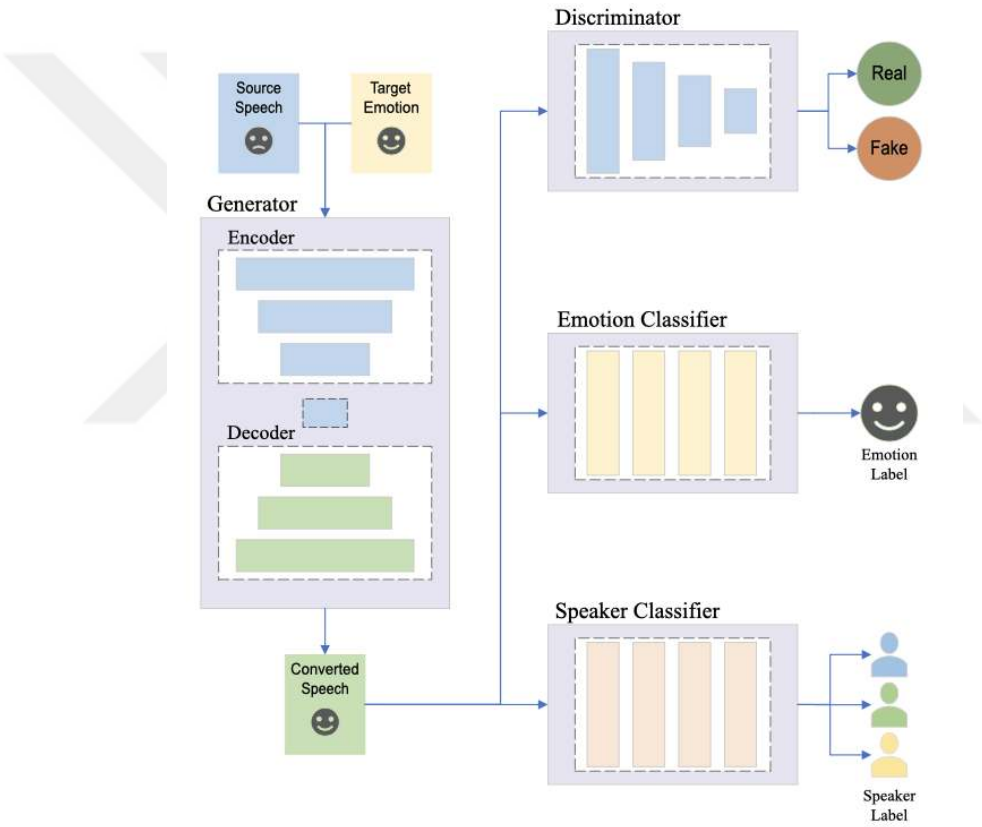


Figure 4.2: Components of DTWSC-StarGAN-EVC: the generator, discriminator, emotion classifier, and speaker classifier.

Our proposed model, *DTWSC-StarGAN-EVC*, is a further development of StarGAN-EVC with DTW loss and auxiliary speaker classifier enhancements. The full architecture of DTWSC-StarGAN-EVC is shown in Figure 4.2. In addition to the generator G, discriminator D, and emotion classifier C_{emo}, we add a speaker classifier

C_{spk} to the network architecture in the proposed model. Moreover, instead of L1 loss, we implement soft-DTW for linguistic content preservation losses.

With a special focus on speech intelligibility as well as the perceived emotion of the converted speech, our contributions to the model mainly affect the training objectives of generator G. As a result, the full training objective of the generator G becomes:

$$\begin{aligned}\mathcal{L}^G = &- \mathbb{E}_{x,c}\left[\log D(G(x,c),c)\right] \\ &+ \lambda_{cyc}\ soft\text{-}DTW(x,\ G(G(x,c),c_x)) + \lambda_{id}\ soft\text{-}DTW(x,\ G(x,c_x)) \\ &- \lambda_{emo\_cls}\ \mathbb{E}_{x,c}\left[\log C_{emo}(c \mid G(x,c))\right] - \lambda_{spk\_cls}\ \mathbb{E}_{x,c}\left[\log C_{spk}(s_x \mid G(x,c))\right]\end{aligned} \tag{4.5}$$

Chapter 5

IMPLEMENTATION DETAILS

5.1 Emotional Speech Dataset - IEMOCAP

In our work, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [Busso et al., 2008], which has approximately 12 hours of non-parallel emotional speech collected in 5 sessions. Each session of this dataset contains audiovisual recordings of a male and female actor pair, so there are 10 speakers in total. Focusing on the authenticity of emotional expression, two different approaches are adopted in the recordings: scripted and improvised. In the first, the actors adhere to the pre-given emotional script, while in the second, they improvise on hypothetical situations specifically designed to evoke different emotions. The effort to reflect real-life emotions makes it an invaluable asset in the affective speech community, as well as a widely used dataset in EVC.

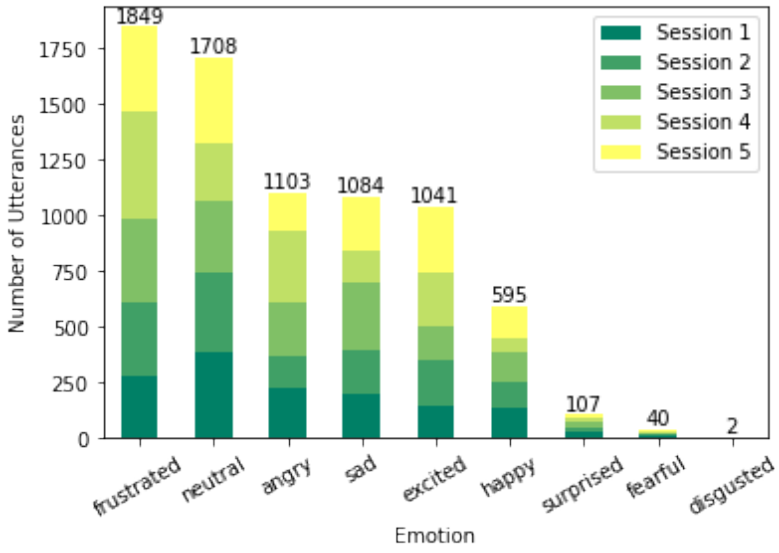


Figure 5.1: Number of utterances distributions in IEMOCAP for emotions and sessions.

We also used this dataset in our experiments to ensure compatibility with previous studies. The dataset has a total of 10039 utterances with both categorical and dimensional emotion annotations. From 9 emotion categories (see Figure 5.1), we chose 4490 utterances, including angry, happy, sad and neutral data. We used sessions 1, 2 and 3 for training, session 4 for quantitative and qualitative validation, and session 5 for testing. Hence, all numerical data reported in the objective and subjective evaluation sections are results obtained on the test set, unless otherwise stated.

5.2 Audio Feature Extractor - WORLD

To manipulate the emotional state of speech, the acoustic features that reveal emotional expression, such as the F0 contour and spectral sequence [Xue et al., 2018], must first be acquired from the raw waveform of speech. For this purpose, we used the vocoder-based WORLD [Morise et al., 2016] as a speech analyzer. Since it has better sound quality and faster processing time than other speech synthesis systems, it is more preferred in real-time speech applications such as VC and EVC [Kameoka et al., 2018], [Rizos et al., 2020]. WORLD extracts aperiodicity parameter (Ap), fundamental frequency (F0), and spectral envelope features from speech, and can resynthesize speech using them.

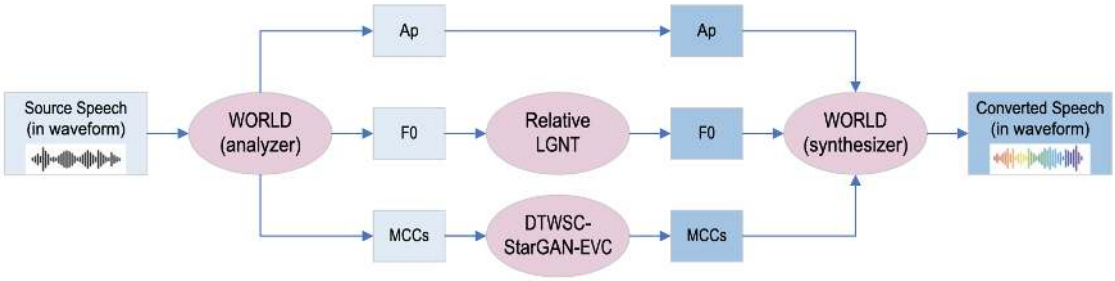


Figure 5.2: Our EVC system.

We employed the Python implementation of WORLD, PyWORLD [Jeremy Hsu, 2022], in order to derive *Ap*, *F0*, and the reduced-size spectral envelope having 36 mel-cepstral coefficients (*MCCs*) from each speech sample, as well as to resynthesize

speech from modified versions of these features. In our EVC system (see Figure 5.2), we transformed the *MCCs* and *F0* using the StarGAN-based model and the relative logarithm Gaussian normalized transformation (LGNT) proposed in [Rizos et al., 2020], respectively, keeping *Ap* the same.

5.3 Baseline Models

To evaluate the performance of the proposed model DTWSC-StarGAN-EVC, we train models with only one of the varying factors in addition to the vanilla StarGAN-EVC model. In our experiments, the model with DTW loss but no auxiliary speaker classifier is referred to as DTW-StarGAN-EVC. In the same manner, the model with auxiliary speaker classifier but no DTW loss is referred to as SC-StarGAN-EVC.

The training of the vanilla StarGAN-EVC was originally optimized for 3 emotions: angry, happy and sad. However, when performing conversion between these 3 strong emotions, we realized that the linguistic content of speech also affected the subjective perception of emotion; therefore, we believed that it would be more proper to use speech samples converted from neutral to any emotion in subjective evaluation.

As a consequence, we perform two separate trainings of the baseline models and DTWSC-StarGAN-EVC, once with 3 emotions and another with 4 emotions (including the neutral samples). In total, we have trained 8 models: StarGAN-EVC, DTW-StarGAN-EVC, SC-StarGAN-EVC and DTWSC-StarGAN-EVC, once with each of the two modifications for the number of emotions.

5.4 Training Details

The role of the generator G in DTWSC-StarGAN-EVC is to be able to change the emotion of speech while maintaining speech quality. To this end, the model follows a three-stage training process:

Reconstruction: In the first stage, only the generator and discriminator are included in the training. The goal is that the generator can learn to reconstruct speech from the low-dimensional latent code (the output of the encoder part) while

retaining the realism and linguistic content of the input speech. Therefore, the generator uses a certain part of its loss function as follows:

$$\begin{aligned}\mathcal{L}^G = &- \mathbb{E}_{x,c}\left[\log D(G(x,c),c)\right] \\ &+ \lambda_{cyc}\ soft\text{-}DTW(x,\ G(G(x,c),c_x)) + \lambda_{id}\ soft\text{-}DTW(x,\ G(x,c_x))\end{aligned} \tag{5.1}$$

Emotional conversion: In the second stage, the pre-trained emotion classifier is included in the model, and it is aimed that the generator can change the emotion of the input voice towards the target emotion. While training the generator and discriminator, the weights of the emotion classifier are kept constant. At this stage, the generator uses the following loss function:

$$\begin{aligned}\mathcal{L}^G = &- \mathbb{E}_{x,c}\left[\log D(G(x,c),c)\right] \\ &+ \lambda_{cyc}\ soft\text{-}DTW(x,\ G(G(x,c),c_x)) + \lambda_{id}\ soft\text{-}DTW(x,\ G(x,c_x)) \\ &- \lambda_{emo\_cls}\ \mathbb{E}_{x,c}\left[\log C_{emo}(c \mid G(x,c))\right]\end{aligned} \tag{5.2}$$

Fine-tuning: The last stage of model training uses all components of the model, the generator G, discriminator D, emotion classifier C_{emo}, and speaker classifier C_{spk}. The speaker classifier is also pre-trained before being combined with the entire model, and the weights of both of the auxiliary classifiers are fixed while training G and D. At this stage, we try to boost both the intelligibility and emotional content of the converted speech with feedback from the emotion and speaker classifiers. The generator uses the full loss function in equation 4.5.

The training hyperparameters are listed in Table 5.1. For different variations of the model, we replace the soft-DTW functions in the loss equations with L1 loss if the model does not include the DTW loss, or we do not continue with the last fine-tuning stage of the training if there is no speaker classifier.

Table 5.1: Model hyperparameters.

Batch size	4
Number of iterations	1^{st} stage: 200000 2^{nd} stage: 100000 3^{rd} stage: 100000
λ_{cyc}, λ_{id}, λ_{emo_cls}, λ_{spk_cls}	3, 2, 1, 1

Chapter 6

EVALUATION RESULTS

In this chapter, we report the results of our experiments to objectively and subjectively evaluate the performance of 4 different frameworks, namely StarGAN-EVC, DTW-StarGAN-EVC, SC-StarGAN-EVC and DTWSC-StarGAN-EVC. In our evaluations, we focus on assessing two main features of human speech: intelligibility and emotion. In all of our evaluation experiments, we perform our metric calculations on the test set which is samples from session 5 of IEMOCAP.

6.1 Objective Evaluation

While comparing the performance of the proposed model with the vanilla StarGAN-EVC model objectively, we also examine the DTW-StarGAN-EVC and SC-StarGAN-EVC to determine the impact of each factor. In general, we find that the DTW loss leads to an improvement in speech quality, whereas the auxiliary SC contributes to a greater elicitation of the emotion feature by distinctly directing the speaker feature of speech. However, the combination of these two factors in the DTWSC-StarGAN-EVC model provides us to obtain the relatively best model in terms of both intelligibility and emotion. The details of these investigations are analyzed in the relevant parts.

6.1.1 Intelligibility

To objectively evaluate the intelligibility performance of the models, we employ the widely used mel-cepstral distortion (MCD), and also propose using a novel approach to measure the recognizability of speech in voice, ASR-based loss. In the following, the MCD and ASR metrics are briefly explained.

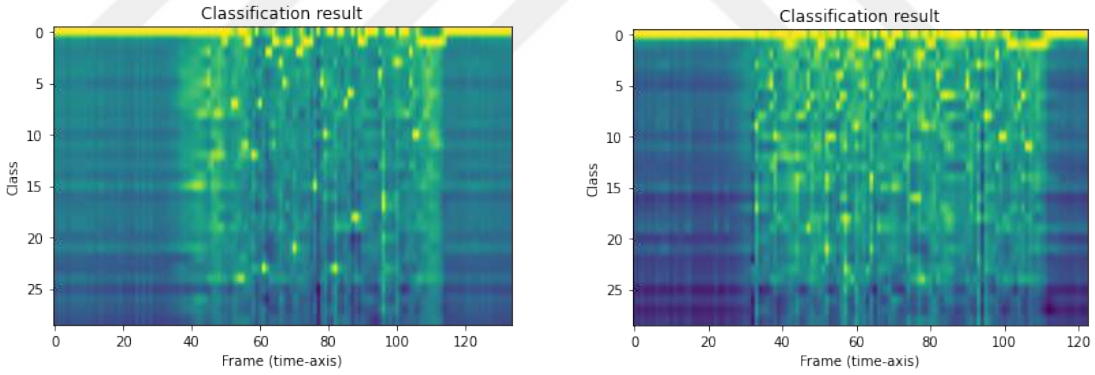
MCD loss: Mel-cepstral distortion (MCD) calculates the distance between two

sequences of cepstral coefficients on the mel-frequency scale [Kubichek, 1993]. It uses the following formulation to compute the difference in t-th time frame between two MCCs having F coefficients:

$$MCD(t) = \frac{10}{\log 10}\sqrt{2\sum_{f=1}^{F}(MCC_{org}(f,t) - MCC_{conv}(f,t))^2} \tag{6.1}$$

In our case, it is used to measure the distances between the original and converted speeches of our test set samples. A lower MCD indicates less distortion during conversion, meaning the model performs better in terms of speech quality.

ASR loss: In addition to the fact that the voice is not distorted, the level of recognizability of the spoken words also indicates the quality of the speech. Considering the increasing success of deep learning-based speech recognition models, we propose to utilize deep ASR models to measure the speech recognizability on the generated samples.



(a) (b)

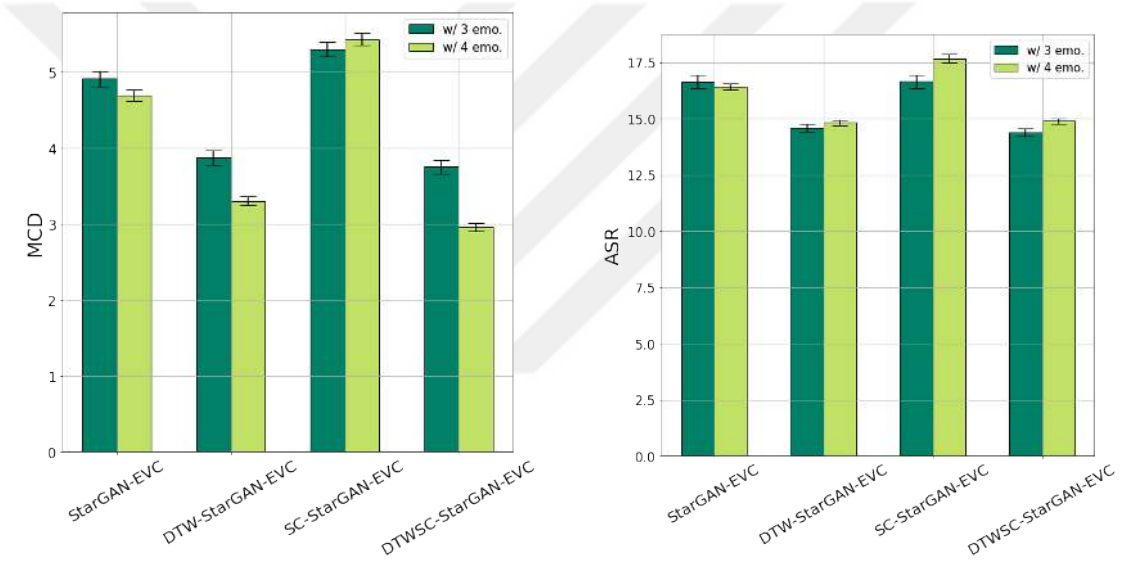
Figure 6.1: Examples of Wav2Vec2 class probability outputs: (a) original speech, (b) converted speech.

Given the input speech in the waveform, Wac2Vec2 [Baevski et al., 2020], one of the state-of-the-art ASR models, outputs the frame-by-frame probability predictions for class labels containing the letters of the English alphabet. Using the pre-trained Wav2Vec2 [Paszke et al., 2019], [PyTorch, 2022], we obtain the classification probabilities of speech samples as shown in the examples in Figure 6.1. Taking the

class label predictions of the original speech as ground truth, we compute the difference between the t-th time frame predictions of the original speech and the speech samples converted by different frameworks as follows:

$$ASR(t) = \sum_{c=1}^{C} (prob_{org}(c,t) - prob_{conv}(c,t)) \tag{6.2}$$

where C denotes the total number of class labels, and prob represents the probability output of the ASR model. A lower ASR is better as it means less inconsistency with the original linguistic content.



(a) (b)

Figure 6.2: Results of intelligibility metrics in our experiments: (a) MCD loss, (b) ASR loss.

Figure 6.2 illustrates the overall mean values of the MCD and ASR losses calculated over the converted speeches. In these graphs, the effects of using only DTW loss and adding only the speaker classifier are also shown to better understand their individual effects on speech intelligibility. To obtain these statistics, for models trained with 3 emotions, all samples of 3 emotions in the test set are transformed into 3 emotions, and for models trained with 4 emotions, all samples of 4 emotions in the test set are transformed into 4 emotions. The error bar on each column represents the 95% confidence interval for mean values. Additionally, when these mean values

are analyzed to see whether the difference between the different model frameworks is significant, it is found that the difference is significant for all pairs of different model frameworks ($p < 0.01$) except for the mean ASR losses of StarGAN-EVC and SC-StarGAN-EVC trained with 3 emotions ($p = 0.98$).

In Figure 6.2a, we see that while adding the speaker classifier to the StarGAN-EVC model increases the MCD value by around 7.9% and 15.7% for models with 3 and 4 emotions, respectively, it affects the combined DTWSC-StarGAN-EVC model with 3.2% and 10.4% reduction in MCD when compared with DTW-StarGAN-EVC. In total, DTWSC-StarGAN-EVC models provide 23.5% and 37.0% improvement in MCD compared to StarGAN-EVC models for 3-emotion and 4-emotion models.

Figure 6.2b shows similar patterns for the ASR metric. For the model trained with three emotions, adding the speaker classifier causes almost no increase in ASR while combining it with the DTW loss in the DTWSC-StarGAN-EVC model achieves about 13.5% reduction in ASR, compared with the vanilla StarGAN-EVC model. Training the model with four emotions, ASR loss increases by 7.7% with the speaker classifier but again falls by around 9.3% with the combined model trained under the same conditions.

Table 6.1: Two-way ANOVA for MCD of 3-emotion models.

	F Value	**P Value**
DTW	1409.837	< 0.01
SC	69.531	< 0.01
DTW:SC	275.006	< 0.01

Table 6.2: Two-way ANOVA for MCD of 4-emotion models.

	F Value	**P Value**
DTW	5211.613	< 0.01
SC	280.373	< 0.01
DTW:SC	1981.295	< 0.01

Table 6.3: Two-way ANOVA for ASR of 3-emotion models.

	F Value	**P Value**
DTW	390.302	< 0.01
SC	1.754	0.186
DTW:SC	1.928	0.165

Table 6.4: Two-way ANOVA for ASR of 4-emotion models.

	F Value	**P Value**
DTW	2147.318	< 0.01
SC	431.121	< 0.01
DTW:SC	334.057	< 0.01

Moreover, we further analyze the effect of DTW and speaker classifier on MCD and ASR losses, and their interaction by implementing two-way repeated measures ANOVA tests (see Tables 6.1, 6.2, 6.3, and 6.4). In these tests, simple main effects analysis shows that DTW has a statistically significant effect on both MCD and ASR. However, while SC has a statistically significant effect on MCD, its effect on ASR may not be significant ($p > 0.10$ for 3-emotion model). The two-way ANOVA also reveals that there is a statistically significant interaction between the effects of DTW and speaker classifier for MCD loss, but this is not always the case for ASR ($p > 0.10$ for 3-emotion model).

Table 6.5: MCD losses of models with 3 emotions for pairwise conversion between emotions.

	AtoH	**AtoS**	**HtoA**	**HtoS**	**StoA**	**StoH**	**Overall**
StarGAN-EVC	6.368	5.878	5.796	4.544	4.299	3.791	4.913
DTW-StarGAN-EVC	5.408	5.748	3.995	3.942	2.805	2.675	3.879
SC-StarGAN-EVC	6.409	6.032	5.970	4.989	4.834	4.567	5.302
DTWSC-StarGAN-EVC	**5.272**	**5.539**	**3.967**	**3.804**	**2.714**	**2.514**	**3.756**

Table 6.6: ASR losses of models with 3 emotions for pairwise conversion between emotions.

	AtoH	**AtoS**	**HtoA**	**HtoS**	**StoA**	**StoH**	**Overall**
StarGAN-EVC	17.573	16.911	16.142	15.013	17.558	16.717	16.635
DTW-StarGAN-EVC	16.179	16.301	14.452	**14.561**	13.634	13.667	14.588
SC-StarGAN-EVC	18.338	17.749	15.270	15.405	16.556	16.422	16.638
DTWSC-StarGAN-EVC	**15.844**	**16.257**	**14.226**	14.596	**13.298**	**13.328**	**14.382**

Table 6.7: MCD and ASR losses of models with 4 emotions for conversion from neutral to one of A/H/S emotions.

	NtoA		**NtoH**		**NtoS**	
	MCD	ASR	MCD	ASR	MCD	ASR
StarGAN-EVC	5.281	16.553	4.610	16.250	3.828	15.905
DTWSC-StarGAN-EVC	**2.881**	**14.784**	**2.644**	**14.908**	**2.661**	**15.059**

In addition to overall comparison, we report the MCD and ASR losses for pair-wise conversion between emotions. Tables 6.5 and 6.6 show the results of models trained with 3 emotions, while Table 6.7 includes the results of models with 4 emotions. Between StarGAN-EVC and DTWSC-StarGAN-EVC models, the highest gain is observed in the conversion of sad-to-angry and neutral-to-angry for the 3-emotion and 4-emotion models, respectively. The DTWSC-StarGAN-EVC model with 3 emotions shows the best performance in conversions from sad emotion, while the conversion from neutral using the 4-emotion DTWSC-StarGAN-EVC model does not differ much in performance between different emotions.

6.1.2 Emotion

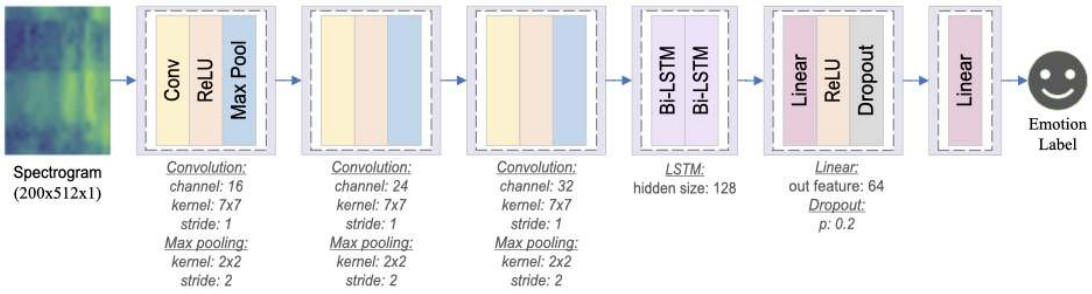


Figure 6.3: Speech emotion recognition model used in data augmentation experiments.

In this part, we try to answer the question of whether improving the audio quality with our novel DTWSC-StarGAN-EVC model has a negative effect on the emotion conversion ability, as compared with the vanilla StarGAN-EVC model. For this purpose, we study the usefulness of the generated audio samples in data augmentation for a SER model to verify the integrity of the emotional content.

For our experiments, we train the SER model in Figure 6.3 using both the original training data only (sessions 1, 2, and 3) and the augmented training data. In this way, we investigate whether we can achieve a performance improvement on the test set with the augmented data. The samples used in data augmentation are generated by several variations of StarGAN model; namely, the StarGAN-EVC, the

StarGAN-EVC with the DWT loss, the StarGAN-EVC with the speaker classifier, and the combined DTWSC-StarGAN-EVC.

Table 6.8: F1-scores of data augmentation experiments.

	with 3 emotions		**with 4 emotions**	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Ses. 1-3	57.780%	61.828%	45.280%	48.089%
Ses. 1-3 + StarGAN-EVC	56.590%	60.215%	48.969%	**55.414**%
Ses. 1-3 + DTW-StarGAN-EVC	47.460%	46.416%	44.222%	49.469%
Ses. 1-3 + SC-StarGAN-EVC	60.825%	62.007%	**51.107%**	55.202%
Ses. 1-3 + DTWSC-StarGAN-EVC	**62.926%**	**65.950%**	47.765%	51.592%

In Table 6.8, we show the macro and micro F1 scores on the test set for each of the mentioned models. Here we perform our experiments with both 3-emotion and 4-emotion SER models. For data augmentation, we use 3-emotion and 4-emotion StarGAN frameworks accordingly. Our dataset is unbalanced as seen in Figure 5.1, yet we want to give equal importance to each emotion class. Thus, we will focus on the macro F1 values, but we also include the micro F1 values for reference.

Table 6.8 shows that 3-emotion and 4-emotion combined DTWSC-StarGAN-EVC models provide 5.146% and 2.485% improvement in Macro-F1 performance of the SER model on the test set. Therefore, we can conclude that adding the DTW loss and the speaker classifier to the vanilla StarGAN-EVC model not only improves speech intelligibility as evidenced by improvements in MCD and ASR metrics but also retains the effect on the alteration of emotional content in the voice; such that the use of the generated audios in the training of the SER model enables the model to learn emotion differences better and increases its performance in the test set. On the other hand, when we compare the data augmentation impact of DTWSC-StarGAN-EVC model with StarGAN-EVC, we see that DTWSC-StarGAN-EVC performs 6.336% better in SER with 3 emotions, while slightly worse (1,204%) in SER with 4 emotions. The best performing model appears to be SC-StarGAN-EVC for the 4-emotion SER model; however, if we take into account intelligibility performance, DTWSC-StarGAN-EVC is still more advantageous.

6.2 Subjective Evaluation

The outputs of the models are subjectively assessed using 3 different evaluation experiments, 2 on intelligibility and 1 on emotion, with a total of 10 participants. The details of the experiments and discussion of the results are described in the relevant subsections, but briefly, for intelligibility we perform a *perception test* measuring the comprehensibility level of each word, and a *preference test* comparing the results of 2 different models in terms of speech clarity, while for emotion we carry out a *classification test* on converted samples to observe whether the outputs are classifiable by humans.

To perform these experiments, we select a subset of 4 samples from the neutral labeled samples in the test set. Since our evaluations will also include the subjective perception of emotion, we use speech samples that have been converted from neutral to any of the emotions (angry, happy or sad) in all our subjective evaluation test cases, concerning that otherwise linguistic content may influence emotion perception. Moreover, we limit the number of words in the experiment samples to between 5 and 10 words in order to ensure that the audio samples are not too long or too short. Apart from these restrictions, we randomly select 4 audios from the remaining samples, and use them in all 3 experiments.

6.2.1 Intelligibility

In both evaluation experiments for intelligibility, we include only neutral samples that have been converted to a happy state, in order to avoid introducing any other variable than the model structure into the model comparisons and to keep the number of samples to be listened to within a certain limit.

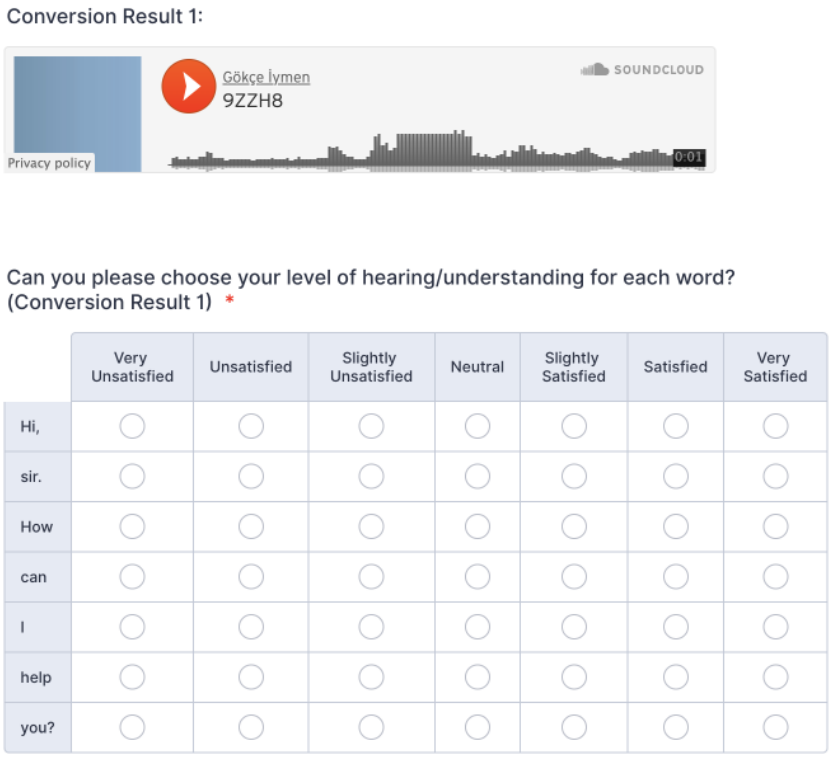
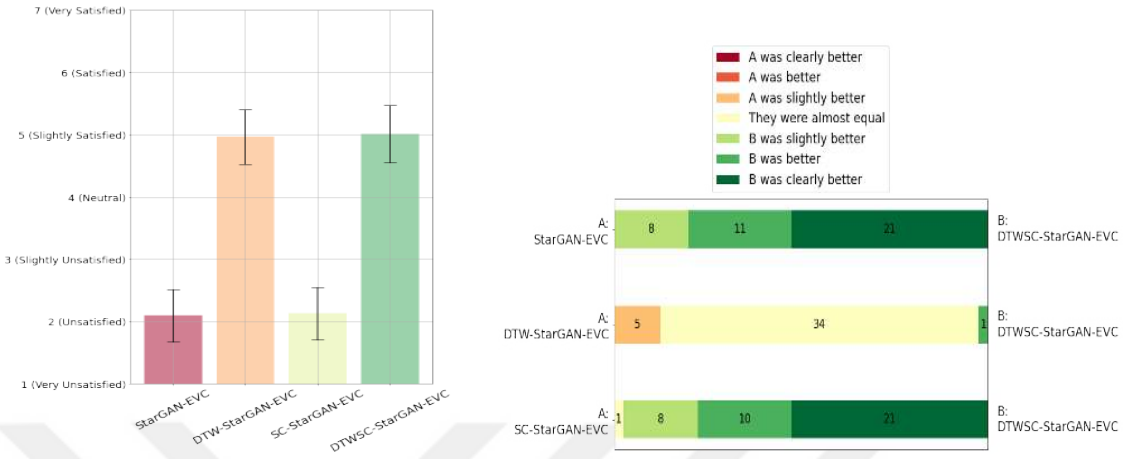


Figure 6.4: Example showing perception test for intelligibility of words.

In the first of our experiments measuring intelligibility, we conduct the mean opinion score (MOS) test on a 7-point Likert scale to investigate how much the 4 different models affect the audibility of words when converting from one emotion to another. In this *perception test* for intelligibility of words, participants are asked to rate their level of hearing/understanding for each word in the samples converted by the 4 different STARGAN-based EVC frameworks, as exemplified by the screenshot in Figure 6.4. The scale used for this assessment ranges from 1 to 7, with 1 meaning *Very Unsatisfied*, 4 meaning *Neutral*, and 7 meaning *Very Satisfied*.

Figure 6.5a shows the average of the MOS evaluations for each framework. These results show that the proposed method, DTWSC-StarGAN-EVC, outperforms the others in terms of speech intelligibility, which is actually in line with the objective evaluation results. However, we should also note that, according to T-test, DTWSC-StarGAN-EVC is significantly better than StarGAN-EVC and SC-StarGAN-EVC ($p < 0.01$), while the difference between DTW-StarGAN-EVC and DTWSC-StarGAN-

EVC is not found to be as significant (p = 0.71).



(a) (b)

Figure 6.5: Results of subjective evaluation experiments for intelligibility: (a) perception test, (b) preference test.

For the second intelligibility experiment, i.e., the *preference test*, participants are expected to compare 2 listening samples in terms of clarity of speech, defined as being clear and less disturbing. Each of the samples in the pairs is generated using a different model, i.e., there are 6 pairs in total where 4 models are compared for each original speech sample. Participants listen to a total of 24 test cases for 4 samples, as shown in Figure 6.6, and in each test case, they select the one they prefer as better in terms of speech clarity. In doing so, they are expected to make a choice within a scale of -3 "A was clearly better" and 3 "B was clearly better", where 0 means "They were almost equal".

Figure 6.5b shows the preference results of paired combinations between one of the baseline models and the proposed model. According to these results (T-test is performed using the scale from -3 to 3), DTWSC-StarGAN-EVC is found to be significantly favourable over StarGAN-EVC and SC-StarGAN-EVC ($p < 0.01$). Furthermore, although DTW-StarGAN-EVC seems to be more preferred, according to p value (p=0.32), it cannot be said that the difference is significant.

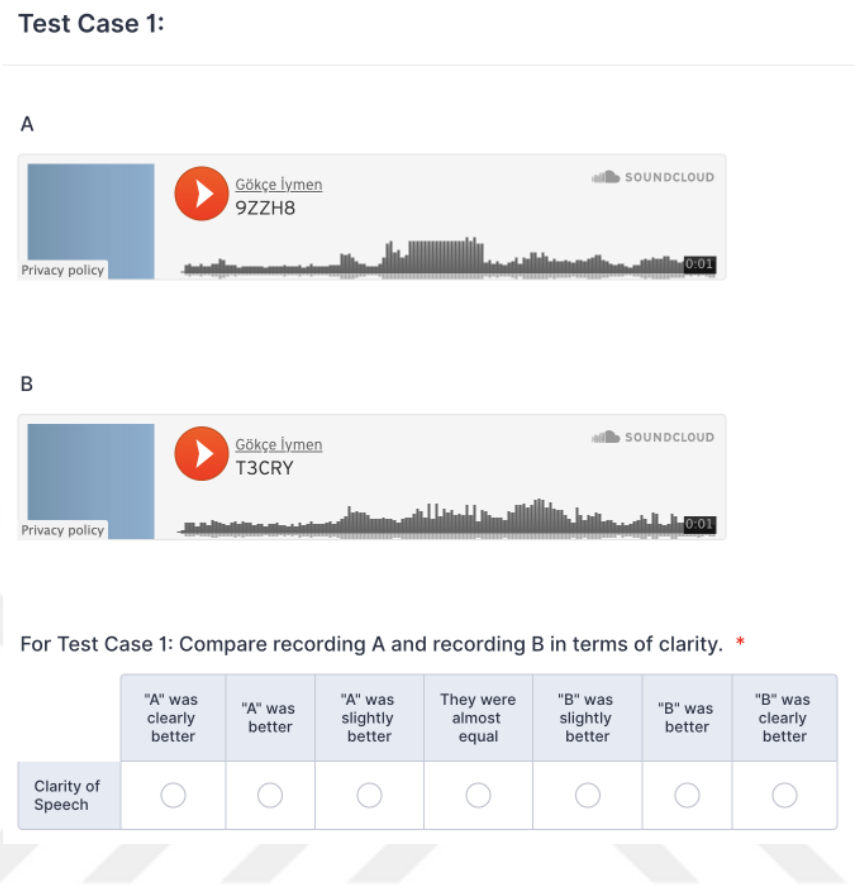


Figure 6.6: Example showing preference test between 2 model outputs.

6.2.2 Emotion

The results of objective evaluations investigating the usefulness of the generated audio samples to improve the prediction performance of the SER model through data augmentation indicate that the converted audios carry similar emotional content to the original dataset. However, we want to further explore the emotion classes perceived by humans in the model outputs via *classification test*. For this purpose, we prepare an experiment in which participants perform one-to-one matching between 3 recordings converted to angry, happy and sad emotions and 3 corresponding emotion labels (see Figure 6.7). Participants perform 16 matching in this experiment, corresponding to 4 model outputs of 4 original samples.

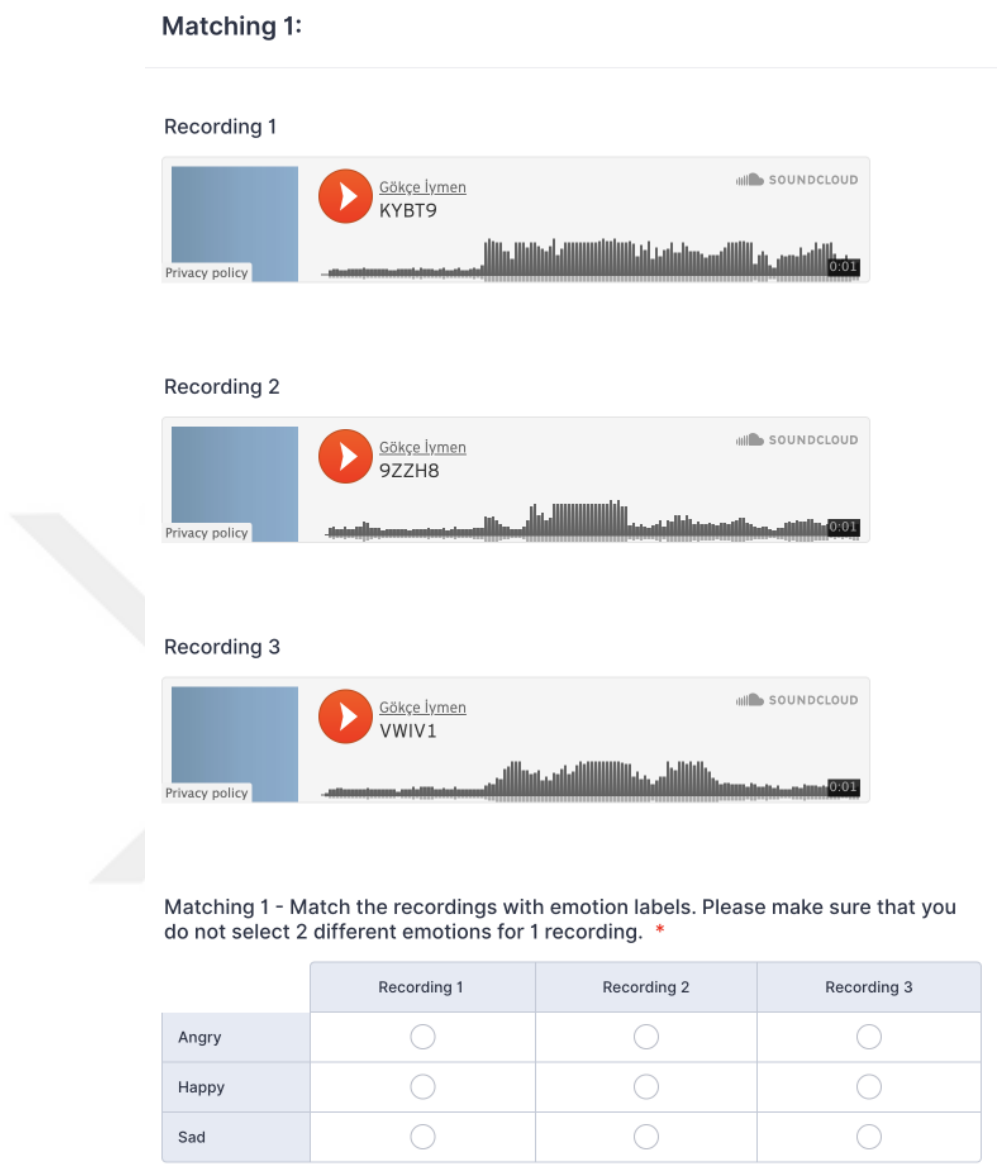


Figure 6.7: Example showing classification test for perceived emotion class of converted samples.

Graph in Figure 6.8 shows the accuracy performance of this experiment for 4 different models. Although the mean values of the accuracy scores differ slightly for the 4 frameworks, T-test indicates that the difference between any paired combination of the 4 models is not found to be as significant ($p > 0.10$ for all pairs). It should also be noted that in this experiment, where the random prediction accuracy will be 33.33%, the accuracy performance of any model is not good enough; therefore, it can be concluded that the outputs are not yet human classifiable.

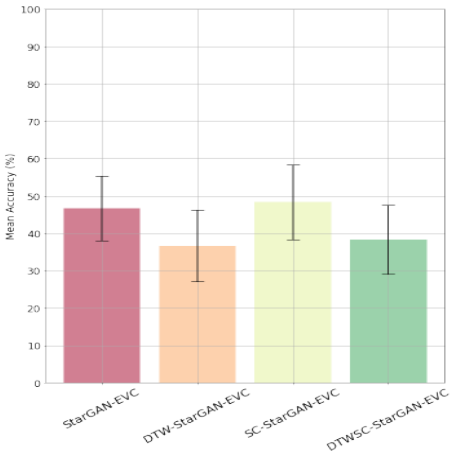


Figure 6.8: Result of subjective classification test.

Chapter 7

CONCLUSION

In this thesis, we proposed the DTWSC-StarGAN-EVC model, which enables the alteration of emotion in human speech. Our StarGAN-based model, which is speaker-independent in both training and inference phases, has a single generator for multi-emotion conversion. In general, we aimed to achieve more intelligible speech outputs by improving the baseline model.

7.1 Remarks

We investigated DTW-based training loss in the non-parallel EVC field, which is the first such study to the best of our knowledge. Furthermore, we experimented with adding an auxiliary speaker classifier to the model for a better conversion process, which is also guided in terms of speaker identity. Building these improvements on a three-stage training process of reconstruction, emotional conversion and fine-tuning, we have obtained a model that is more intelligible than the baseline model but also has comparable success in terms of emotional conversion performance.

In the evaluation step, we tried to assess the performance of the model outputs by targeting both intelligibility and emotional content features. To this end, we proposed a novel ASR-based objective evaluation metric in addition to MCD and data augmentation results, and presented our purpose-designed subjective evaluation setups. The objective and subjective evaluation results showed that the DTW-based training loss improved the model mainly in terms of speech quality, while the auxiliary speaker classifier revealed the emotion feature of the speech by specifically isolating the speaker feature. Overall, we found that the combination of these two factors in the DTWSC-StarGAN-EVC model yielded the relatively best model when considering the relationship between intelligibility and emotion.

7.2 Future Directions

Despite the progress made with the model in terms of intelligibility and emotional content, we acknowledge the limitations that the model still has. The output speech converted by the model has not yet reached the quality of speech that can be used in real-life applications. It may be possible to work with larger models, such as Transformer [Vaswani et al., 2017] and BERT [Devlin et al., 2018], to better preserve the linguistic information of the voice, but at this point, the real-time conversion capability may have to be sacrificed. Nevertheless, as future work, we think that the benefits of using larger models could be examined. Furthermore, our overall impression on the converted speeches and the evaluators' feedbacks in the subjective evaluation suggest that the model tends to increase the amplitude of the whole speech when making a speech angry and similarly decrease the amplitude of the whole speech when making it sad, without paying attention to accents or the spoken word. This is mostly due to the fact that the speech samples in the dataset for angry and sad emotions generally have these features. As a future work, we think that a dataset of larger size and with more divergent emotional features could provide improvements for the model's ability to convert emotion. Alternatively, the model could also be trained to learn how to pay attention to parts of speech according to the spoken words using a text-assisted approach.

BIBLIOGRAPHY

[Abdelmadjid and Boucheham, 2021] Abdelmadjid, L. and Boucheham, B. (2021). A comparison study of dynamic time warping's variants for time series classification. 4:56–71.

[Abe et al., 1988] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 655–658 vol.1.

[Adigwe et al., 2018] Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., and Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

[Aihara et al., 2012] Aihara, R., Takashima, R., Takiguchi, T., and Ariki, Y. (2012). Gmm-based emotional voice conversion using spectrum and prosody features. *American Journal of Signal Processing*, 2:134–138.

[Aihara et al., 2014] Aihara, R., Ueda, R., Takiguchi, T., and Ariki, Y. (2014). Exemplar-based emotional voice conversion using non-negative matrix factorization. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–7.

[Baevski et al., 2020] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

[Bagnall et al., 2017] Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660.

[Banse and Scherer, 1996] Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70:614–36.

[Bao et al., 2019] Bao, F., Neumann, M., and Vu, N. T. (2019). CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition. In *Proc. Interspeech 2019*, pages 2828–2832.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

[Busso et al., 2017] Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2017). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

[Cai et al., 2021] Cai, B., Huang, G., Samadiani, N., Li, G., and Chi, C.-H. (2021). Efficient time series clustering by minimizing dynamic time warping utilization. *IEEE Access*, 9:46589–46599.

[Cao et al., 2014] Cao, H., Cooper, D., Keutmann, M., Gur, R., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390.

[Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8789–8797.

[Cowie and Cornelius, 2003] Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1):5–32.

[Cuturi and Blondel, 2017] Cuturi, M. and Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Donahue et al., 2021] Donahue, J., Dieleman, S., Binkowski, M., Elsen, E., and Simonyan, K. (2021). End-to-end adversarial text-to-speech. In *International Conference on Learning Representations*.

[Dong et al., 2018] Dong, L., Xu, S., and Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

[Dudley, 1939] Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177.

[Efrat et al., 2007] Efrat, A., Fan, Q., and Venkatasubramanian, S. (2007). Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. *Journal of Mathematical Imaging and Vision*, 27(3):203–216.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

[Gao et al., 2019] Gao, J., Chakraborty, D., Tembine, H., and Olaleye, O. (2019). Nonparallel emotional speech conversion. In Kubin, G. and Kacic, Z., editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2858–2862. ISCA.

[Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style

transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423.

[Griffin and Lim, 1984] Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

[He et al., 2021] He, X., Chen, J., Rizos, G., and Schuller, B. W. (2021). An improved stargan for emotional voice conversion: Enhancing voice quality and data augmentation. *arXiv preprint arXiv:2107.08361*.

[Hsu et al., 2016] Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., and Wang, H.-M. (2016). Voice conversion from non-parallel corpora using variational auto-encoder. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. IEEE.

[James et al., 2018] James, J., Tian, L., and Watson, C. I. (2018). An open source emotional speech corpus for human robot interaction applications. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2768–2772.

[Jeremy Hsu, 2022] Jeremy Hsu (accessed 2022). Python-wrapper-for-world-vocoder. `https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder`.

[Kalchbrenner et al., 2018] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. (2018). Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

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

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

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3):187–207.

[Kubichek, 1993] Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

[Li et al., 2020] Li, Y., Xu, D., Zhang, Y., Wang, Y., and Chen, B. (2020). Non-parallel many-to-many voice conversion with psr-stargan. pages 781–785.

[Liu et al., 2020] Liu, S., Cao, Y., and Meng, H. (2020). Emotional voice conversion with cycle-consistent adversarial network.

[Lorenzo-Trueba et al., 2018] Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., and Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.

[Lotfian and Busso, 2019] Lotfian, R. and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

[Luo et al., 2017] Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2017). Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.

[Luo et al., 2016] Luo, Z., Takiguchi, T., and Ariki, Y. (2016). Emotional voice conversion using deep neural networks with mcc and f0 features. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5.

[Maghoumi, 2020] Maghoumi, M. (2020). *Deep Recurrent Networks for Gesture Recognition and Synthesis*. PhD thesis, University of Central Florida Orlando, Florida.

[Maghoumi et al., 2021] Maghoumi, M., Taranta, E. M., and LaViola, J. (2021). Deepnag: Deep non-adversarial gesture generation. In *26th International Conference on Intelligent User Interfaces*, pages 213–223.

[Ming et al., 2015] Ming, H., Huang, D., Dong, M., Li, H., Xie, L., and Zhang, S. (2015). Fundamental frequency modeling using wavelets for emotional voice conversion. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 804–809.

[Ming et al., 2016] Ming, H., Huang, D.-Y., Xie, L., Wu, J., Dong, M., and Li, H. (2016). Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion. In *Interspeech*, pages 2453–2457.

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

[Myers et al., 1980] Myers, C., Rabiner, L., and Rosenberg, A. (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):623–635.

[Müller, 2007] Müller, M. (2007). Dynamic time warping. *Information Retrieval for Music and Motion*, 2:69–84.

[Oord et al., 2016] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Paul et al., 2020] Paul, D., Pantazis, Y., and Stylianou, Y. (2020). Speaker conditional wavernn: Towards universal neural vocoder for unseen speaker and recording conditions. In *INTERSPEECH*.

[PyTorch, 2022] PyTorch (accessed 2022). Wav2vec2 model weights. `https://download.pytorch.org/torchaudio/models/wav2vec2_fairseq_base_ls960_asr_ls960.pth`.

[Rabiner et al., 1978] Rabiner, L., Rosenberg, A., and Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(6):575–582.

[Rath and Manmatha, 2003] Rath, T. and Manmatha, R. (2003). Word image matching using dynamic time warping. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II.

[Rizos et al., 2020] Rizos, G., Baird, A., Elliott, M., and Schuller, B. (2020). 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)*, pages 3502–3506.

[Robinson et al., 2019] Robinson, C., Obin, N., and Roebel, A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6830–6834.

[Russell, 1980] Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

[Sakamoto et al., 2021] Sakamoto, S., Taniguchi, A., Taniguchi, T., and Kameoka, H. (2021). Stargan-vc+asr: Stargan-based non-parallel voice conversion regularized by automatic speech recognition. *arXiv preprint arXiv:2108.04395*.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

[Schroeder, 1966] Schroeder, M. (1966). Vocoders: Analysis and synthesis of speech. *Proceedings of the IEEE*, 54(5):720–734.

[Schuller and Schuller, 2018] Schuller, D. and Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9):38–46.

[Senin, 2009] Senin, P. (2009). Dynamic time warping algorithm review.

[Shen et al., 2018] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

[Sisman et al., 2020] Sisman, B., Yamagishi, J., King, S., and Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 29:132–157.

[Tits et al., 2019] Tits, N., El Haddad, K., and Dutoit, T. (2019). Exploring transfer learning for low resource emotional tts. In *Proceedings of SAI Intelligent Systems Conference*, pages 52–60. Springer.

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

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Wang et al., 2017] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[Xue et al., 2018] Xue, Y., Hamada, Y., and Akagi, M. (2018). Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, 102:54–67.

[Zadeh et al., 2018] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

[Zhang et al., 2019] Zhang, Z., Wu, B., and Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709. IEEE.

[Zhao et al., 2020] Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., and Toda, T. (2020). Voice conversion challenge 2020:

Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*.

[Zhou et al., 2020a] Zhou, K., Sisman, B., and Li, H. (2020a). Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv preprint arXiv:2002.00198*.

[Zhou et al., 2021] Zhou, K., Sisman, B., Liu, R., and Li, H. (2021). Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE.

[Zhou et al., 2022] Zhou, K., Sisman, B., Liu, R., and Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

[Zhou et al., 2020b] Zhou, K., Sisman, B., Zhang, M., and Li, H. (2020b). Converting anyone's emotion: Towards speaker-independent emotional voice conversion. pages 3416–3420.

[Zhu et al., 2017] Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251.