Choosing Only the Best Voice Imitators: Top-K Many-to-Many Voice Conversion with StarGAN-VC

KTH Thesis Report

Claudio Fernandez Martín
**Author**

Claudio Fernandez Martín <clferma1@i3b.upv.es>
Double Master’s Degree in Human-Computer Interaction and Design, and ICT Innovation - EIT Digital HCID
Information and Communication Technology
KTH Royal Institute of Technology

**Date and Place for Project**

December 2021
Stockholm, Sweden
Valencia, Spain

**Examiner**

Dr. Haibo Li
KTH Royal Institute of Technology in Stockholm, Sweden

**Supervisors**

Claudio Panariello
KTH Royal Institute of Technology in Stockholm, Sweden
Adrián Colomer Granero
CVBLAB | Computer Vision and Behaviour Analysis Lab in Valencia, Spain
Abstract

Voice conversion systems are becoming more relevant as the popularity of voice technologies is growing with the increased adoption of voice assistants and the increased demand for speech-based interfaces in recent years. This scenario would not have been possible without the latest developments in the generative deep learning field, where novel neural networks architectures such as generative adversarial networks (GANs) are providing researchers with previously unimaginable possibilities in the creation of synthetic media. In the field of speech synthesis, voice conversion through deep learning has shown to be the most promising approach, especially for real-life scenarios where it is not possible to have the same sentence uttered by both the source and target speaker. The most recent example that has outperformed the previous approaches for non-parallel voice conversion is StarGAN-VC, which makes use of a single pair of generator and discriminator that allows conversions between multiple speakers. Nonetheless, some limitations of GANs, in general, are the long time they take to train, due to their adversarial nature, and the instability during the training, which often leads to problems such as overfitting. Recently, a big focus from the deep learning community has been placed on trying to solve these issues during training. This is the example of the Top-K methodology, which consists of training the generator only with the best $K$ generated fake samples that achieved to fool the discriminator, throwing away the bad ones. However, this methodology has only been applied to image-related deep learning tasks and to simple GAN architectures. In the present work, we show that applying the Top-K methodology to a state-of-the-art voice conversion system such as StarGAN-VC can significantly improve the quality of the converted voices, obtaining a quicker convergence of the model and improving the stability during training. We also perform a novel study about the most optimal moment for starting to apply the Top-K methodology and how to reduce the value of $K$ during the training. For achieving this purpose, we conduct a quantitative and qualitative study of the results due to the lack of a golden standard metric for the voice conversion evaluation and because of the inherently subjective nature of the speech perception.
Sammanfattning

Acknowledgements

First of all, I would like to thank the CVBLab in Valencia, and specifically to Sandra Morales Martínez and Valery Naranjo Ornedo, for believing in me and offering me the chance to work in such an inspiring environment where I am learning every day how to become a great researcher in the field of Deep Learning.

Particularly, I would like to thank my supervisor at the CVBLab, Adrián Colomer Granero for his kindness and support during all this process. Thank you especially for having the patience and being always happy to answer my never-ending questions and for all the laughs that we shared trying to understand the extraction of the Mel-Cepstral Coefficients together.

Also, I would like to appreciate the job that my KTH supervisor, Claudio Panariello, has done giving me always great advice on how to proceed with the project always with a smile on his face.

This project has been a very important milestone in my life, because it represents the last step to finish my life as a University student. It all started in Valencia, my hometown, where I became a Telecommunications Engineer in the Polytechnic University of Valencia (UPV). Then, I did the last year of my Bachelor’s and my Bachelor Thesis in the Technical University of Munich, where I made my first contact with Machine Learning and with the international student life. After this moment, thanks to the European Institute of Technology (EIT) I had the chance to start my journey living and studying abroad permanently, as I got accepted in the EIT Digital Master’s in Human-Computer Interaction and Design. This way, I got the chance to study my master’s in two of the best European universities, the first year at the University of Twente in Enschede, The Netherlands; and the second and final year at KTH Royal Institute of Technology in Stockholm, Sweden.

During these six years, I have had the chance to meet plenty of people from everywhere in the world and I believe that I am very lucky for being able to consider many of them my friends. From my first year in Twente, I would like to particularly thank Barry O’Sullivan and Saumya Singh for bringing me endless afternoons of laughter and also hard work during my time in the Netherlands. Regarding my time in Stockholm I want to mention specially my friend Juan Alvarez Del Vallado for being my confident during hard times, and for our long coffee talks by the window of his room.

This last year of my University life would not have been the same without Karolina Siemieniuk, which has been my life partner during very tough moments this year. Together we have shared unforgettable moments like our
trip to the north of Sweden to see the Northern Lights. This was just the first of many more trips together. I am very glad to have met you and very keen to keep sharing adventures with you.

Last but not least, I would like to thank my mother, Maria Antonia Fernandez Martín, for always pushing me to become a better version of myself, for always being there when I needed her and for feeling my accomplishments as hers. I will always be grateful for how much you keep believing in me every day.

Stockholm, November 2021

Claudio Fernández Martín
# Contents

1 Introduction .................................. 1
   1.1 Voice Conversion .......................... 4
   1.2 Research Questions ....................... 5
   1.3 Thesis Outline .......................... 6

2 Background: Theoretical Framework and Evolution of Voice Conversion Techniques 9
   2.0.1 The General Voice Conversion Pipeline ............. 11
   2.0.2 Meaning of *Domain* in a Voice Conversion Task .... 12
   2.0.3 Classifications of Voice Conversion Systems ........ 13
   2.0.4 Acoustic Features and Conversion Process .......... 16
   2.1 Deep Learning and Generative Deep Learning .......... 19
       2.1.1 Generative Adversarial Networks (GANs) .......... 21
       2.1.2 CycleGAN and CycleGAN-VC ....................... 24
       2.1.3 StarGAN and StarGAN-VC ......................... 25
       2.1.4 Top-K GAN Approach .......................... 27

3 Methodology .................................. 31
   3.1 Material and Development Environment ................ 32
       3.1.1 Dataset .................................. 32
       3.1.2 Development Environment ....................... 33
   3.2 Background: StarGAN Voice Conversion (StarGAN-VC) as a Baseline Model .......................... 34
       3.2.1 Model Pipeline .......................... 34
       3.2.2 Baseline Model Training ....................... 36
   3.3 Proposed Model Architecture: Top-K StarGAN-VC ........ 37
       3.3.1 The *Critical Point* of GAN StarGAN-VC Training .... 38
       3.3.2 The Optimal Value of $\gamma$, the Decay of K ....... 40
4 Results Analysis and Evaluation 41
   4.1 Quantitative or Objective Evaluation 42
      4.1.1 Mel-Cepstral Distance (MCD) 42
      4.1.2 Most Optimal Top-K Parameters Study 42
      4.1.3 Convergence Study of the Top-K Methodology 44
   4.2 Qualitative or Subjective Evaluation 45
      4.2.1 Subjective Listeners’ Evaluation of the Obtained Voice Conversions 46

5 Discussion and Conclusions 51
   5.1 Conclusions from the Research Questions 51
   5.2 Ethical Considerations and Implications of Speech Synthesis and Voice Conversion 53
   5.3 Future Work 56

References 57
List of Figures

2.1 The sub-field of voice conversion inside the field of speech synthesis and the super-field of digital speech processing [1]. . 10
2.2 Typical flow of a voice conversion system. The pink box represents the training of the mapping function, while the blue box applies the mapping function at run-time [2]. . . . . . . . 12
2.3 Parallel vs. Non-parallel VC system representation [1]. . . . 14
2.4 Mono-lingual vs. Cross-lingual VC system representation [1]. 14
2.5 Intra-gender vs. Cross-gender VC system representation [1]. . 15
2.6 One-to-One vs. Many-to-Many VC system representation [1]. . 16
2.7 Mel scale [3]. . . . . . . . . . . . . . . . . . . . . . . . . . . 17
2.8 MFCC (mel-frequency cepstral coefficients) characteristic vectors extraction flow [4]. . . . . . . . . . . . . . . . . . . . . . . . . . . 18
2.9 Representation of the scope of AI, ML and DL [5]. . . . . . 20
2.10 Generator and Discriminator as GAN building blocks [6]. . . 22
2.11 Objective function in GAN formulation [6]. . . . . . . . . . . 23
2.12 (a) The CycleGAN model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators $D_Y$ and $D_X$. $D_Y$ encourages $G$ to translate $X$ into outputs indistinguishable from domain $Y$, and vice versa for $D_X$ and $F$. To further regularise the mappings, they introduced two cycle consistency losses that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ [7]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
2.13 Training procedure of CycleGAN-VC. $G_{X \rightarrow Y}$ is forward generator that transforms $X$ to $Y$. $G_{Y \rightarrow X}$ is inverse generator that transforms $Y$ to $X$. $D_X$ and $D_Y$ are discriminators in $X$ and $Y$ domains, respectively. (a)(b) We use adversarial losses and cycle-consistency losses to find optimal pseudo pair from unpaired data. (c)(d) Furthermore, we use identity-mapping losses to preserve linguistic information [8].

2.14 Comparison between cross-domain models and our proposed model, StarGAN. (a) To handle multiple domains, cross-domain models should be built for every pair of image domains. (b) StarGAN is capable of learning mappings among multiple domains using a single generator. The figure represents a star topology connecting multi-domains [9].

2.15 Overview of StarGAN, consisting of two modules, a discriminator $D$ and a generator $G$. (a) $D$ learns to distinguish between real and fake images and classify the real images to its corresponding domain. (b) $G$ takes in as input both the image and target domain label and generates a fake image. The target domain label is spatially replicated and concatenated with the input image. (c) $G$ tries to reconstruct the original image from the fake image given the original domain label. (d) $G$ tries to generate images indistinguishable from real images and classifiable as target domain by $D$ [9].

2.16 Illustration of A-StarGAN training from [10]. The $A$ network is designed to produce $2K$ probabilities, where the first and second $K$ probabilities correspond to real and fake classes, and simultaneously play the roles of the real/fake discriminator and domain classifier.

2.17 Diagram of top-$k$ training of a GAN. The generator generates a batch of samples, which are scored by the critic. Only the $K$ samples with the highest scores are used to update the generator [11].

3.1 Illustration of A-StarGAN training. The $A$ network is designed to produce $2K$ probabilities, where the first and second $K$ probabilities correspond to real and fake classes, and simultaneously play the roles of the real/fake discriminator and domain classifier [10].

3.2 Loss curves using the vanilla StarGAN-VC training.
4.1 Loss curves using the proposed Top-K A StarGAN-VC training. 46
xii | LIST OF FIGURES
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Experimental Parameters for the Top-K Variants</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Ray Tune Results for the Different Top-K Variants compared to the Vanilla version</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>MCD Comparison of the Top-K Variants with the Baseline on the VCTK Corpus Dataset</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td>Convergence Study Results Comparing the MCDs at Critical Steps Between Vanilla and the Best Top-K Approach</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>Results of the realness study performed in the qualitative study, comparing</td>
<td>48</td>
</tr>
<tr>
<td>4.5</td>
<td>Results of the Convergence Study Comparing the Perceived Speech Naturalness at Critical Steps Between Vanilla and the Best Top-K Approach by the Participants in the Qualitative study</td>
<td>49</td>
</tr>
</tbody>
</table>
List of acronyms and abbreviations

AI  Artificial Intelligence

DL  Deep Learning

GAN  Generative Adversarial Network

GPU  Graphics Processing Unit

MCD  Mel-cepstral Distortion or Mel-cepstral Distance

MFCC  Mel Frequency Cepstral Coefficients

MOS  Mean Opinion Score

VC  Voice Conversion
Chapter 1

Introduction

Speech corresponds to an essential part of communication among individuals. In fact, it is one of the traits that distinguishes humans from the rest of the animal species. Its importance relies on the different purposes that speaking can have, such as informing, convincing, persuading, stimulating, entertaining or even consoling. For this reason, the development of computational speech processing systems, as part of the field of signal processing, has posed a challenge in the last decades. The popularity of voice technology began to emerge in 2011 with the introduction of Siri, a voice assistant present in some Apple products. However, foreseeing the enormous impact this technology has had when developing new products was almost impossible at that time.

Nowadays, roughly one out of four U.S. adults own a smart speaker, such as Google Home, Alexa or Amazon Echo, which means that there are now an estimate of 157 million of these devices in American homes [12]. Recent reports forecast that by 2023, nearly 92.3 percent of smartphone users will be using voice assistants [13]. This shift towards speech-based interfaces can be explained by the change in the user demands, which now focuses on speed, efficiency and convenience. Another factor that explains the shift towards voice applications is the mass adoption of artificial intelligence (AI) in the user’s everyday lives with the raise of use of smart devices. Industry experts even predict that nearly every application will integrate voice technology in some way in the next five years [14].

Progress in Machine Learning (ML) algorithms and the integration of GPU hardware into computing systems have allowed for commoditisation of custom voice creation and speech enhancing allowing for more emotion-like features, making a computer-generated voice almost indistinguishable from the real one. Therefore, in the present but mostly in the future, voice cloning will become an indispensable tool for different applications and fields like
medical patients, education, game developers, advertisers, content creators, filmmakers and dubbing actors.

In this sense, in the medical field patients that suffer from conditions that leave them unable to use their voices anymore such as amyotrophic lateral sclerosis (ALS), apraxia, Huntington’s disease, autism, strokes, or traumatic brain injuries, could have their voices recreated via voice cloning if they bank them beforehand. Moreover, positive commercial implications can arise as voice cloning with AI reduces the need for additional recordings for interactive voice response systems, and thus allows actors to make better, more creative use of their time. Also, the importance of voice conversion with AI has also become more salient for providing interactive content for online learning courses during the times of lockdown due to COVID-19, as voice conversion technology makes it easier to record audio notes because it makes it unnecessary to do so for every new session, or to address the mistakes in previous sessions. This way the operational costs of professionally recorded lectures can be dramatically reduced, and students could benefit from the educational materials as if they were in a regular classroom. Voice conversion also represents a great opportunity for the gaming industry, because voice actors spend large amounts of time in the recording studios, therefore, if their cloned voice could take over, this would highly increase the time efficiency of this process. Furthermore, voice cloning brings the opportunity to give old voices a new life as game developers could add historical voices to the game script, or also be able to finish the game with an actor who unfortunately passed away. Advertisements can also be benefited from the possible applications of voice conversion as it could allow them to use voices which otherwise would be very difficult to record as unavailable actors, kids or historical figures, significantly contributing to the reduction of production time and costs. Nonetheless, one of the sectors where the impact of voice conversion would be more notable would be dubbing. Not only voice conversion would reduce the time and cost of the actors voiceover work during post-production, but also if producers would choose a language agnostic voice conversion technology, they would be able to record the needed voices in any language and then simply translate it in an automated way for the different world regions where a movie would be released. Furthermore, film and advertisement producers could be more adaptable to their target audience by using precisely the kind of accent that is presumably best received in a particular region.

It is clear that the positive socio-economic impact that voice conversion technologies can provide is very substantial. However, the risks that they might pose to societies and their ethical implications must also be considered.
These could be more realistic deepfakes with the combination of audio and video, that could lead to complete imitations of people, which in the case of politicians it could potentially lead to a very dangerous manipulation of societies. Also, phone phishing attacks could be a threat to companies and people. For these reasons, in the last section of this work, we provide an exhaustive review of the ethical considerations of speech synthesis and voice conversion technologies.
1.1 Voice Conversion

Voice Conversion (VC) consists in transforming the voice characteristics from a source speaker into those of a desired target speaker while keeping the linguistic content intact. In other words, a voice conversion system modifies a particular person’s voice (source speaker) to make it sound like a second person (target speaker) is speaking the same sentences. In this sense, the non-linguistic information is what is attempted to be replicated in the output of the system. These correspond to features such as the speaker identity, accent, pronunciation and speaking style which includes expression of emotions, emphasis, intentions, etc.

Therefore, the main goal of automatic VC is to generate synthetic voices very similar to the original ones. Based on Deep Learning (DL) and Artificial Intelligence (AI) techniques, the system takes advantage of a set of audio features of the original voice to form a model capable of generating new voices that aim to be difficult to distinguish from the original ones. Several examples of the applications of VC techniques include speaker identity modification, speaking assistance, speech enhancement, bandwidth extension, accent conversion, voice translation or human-like voice synthesis.

In literature, we find successful VC frameworks involving approaches that utilise acoustic models represented by Gaussian mixture models (GMMs) [15, 16] for feature mapping. Recently, frameworks based on neural networks (NNs) [8, 17] and non-negative matrix factorization (NMF) [18] have also shown satisfactory results. Indeed, most of these conventional VC methods require accurately aligned parallel source and target speech data: this is called parallel VC. In parallel VC, the VC system makes use of parallel voice data which consists of recordings of parallel (same) sentences or utterances read by both the source and the target speakers. Therefore, the correspondence between paired utterances can be directly achieved by performing time-alignment (i.e., Dynamic Time-Wrapping). However, collecting parallel voice data requires an enormous effort compared to that for non-parallel voice data, as this one does not require parallel utterances, transcriptions or time alignments. Hence, the use of non-parallel voice data becomes really convenient in practical applications, especially in real-time, where approaches that can collect data with ease are desirable.

In non-parallel VC, it is not trivial to measure the correspondence between the source and the target spectral features due to the non-existence of paired utterances. For this reason, VC systems that use parallel data usually provide a higher quality outcome compared to non-parallel ones. Therefore, in the last
couple of years, it has become a challenge to develop competitive non-parallel VC systems. Examples of techniques that have been implemented to achieve non-parallel VC are automatic speech recognition, variational autoencoders (VAE) [19] and, more recently, generative adversarial networks (GANs) [20, 10]. The latter ones have shown very promising results in implementations like CycleGAN-VC or StarGAN-VC. GANs consist of a general framework for training a generator network with the objective of generating fake data samples, in the case of this project fake voices, that are able to mislead a real/fake discriminator in the form of a mini-max game.

Hence, the objective of this dissertation will be to develop a non-parallel VC system using GANs which tries to obtain competitive results compared to existing implementations by changing the way the StarGAN-VC model is trained, using only the best samples created by the Generator for the training. This is the Top-K methodology, which aims to train the Generator throwing away the bad samples with the objective of obtaining better quality samples in less training time. Also, we aim to perform a study in order to determine what is the most optimal moment for starting to apply the Top-K methodology and also how the reduction of the value of $K$ should be carried out.

1.2 Research Questions

In this section, we introduce the main research question that we aim to answer by both our quantitative and our qualitative study.

RQ1. Does the application of the Top-K training to a state-of-the-art voice conversion system like StarGAN-VC improves the quality of the converted voices both quantitatively and qualitatively?

Sub-questions are also derived from this central research question. The more technical ones will be addressed in the Quantitative Evaluation and the ones corresponding to the actual listener’s perception of the converted voices using our model will be addressed through Qualitative Evaluation with human participants.

Questions addressed in the quantitative study:

1. When is the most optimal moment during a GAN training to start applying the Top-K methodology?

2. What is the most optimal way of reducing the value of $K$ after starting to apply the Top-K methodology?
3. Does the learning loss of the Generator becomes smaller using the Top-K methodology compared to the Vanilla version of StarGAN-VC?

4. Does the learning loss of the Generator reaches the convergence point quicker using the Top-K methodology compared to the Vanilla version of StarGAN-VC?

5. Does the learning loss of the Generator becomes more stable once it reaches the convergence point using the Top-K methodology compared to the Vanilla version of StarGAN-VC?

6. Are the state-of-the-art metrics for evaluating a Voice Conversion system improved using the Top-K methodology compared to the Vanilla version of StarGAN-VC?

Questions addressed in the qualitative study:

1. Is it possible for the human listener to mistake a cloned voice generated by an AI with the voice of a real human speaker?

2. Which cloned voices created by an AI are more mistaken by the listeners with the voices of a real human speaker, the ones generated using the Top-K methodology or the ones generated using the Vanilla version of the StarGAN-VC?

3. Which cloned voices have a higher quality, the ones generated using the Top-K methodology or the ones generated using the Vanilla version of the StarGAN-VC?

1.3 Thesis Outline

This thesis is divided into six chapters. In Chapter 2, a theoretical framework and an explanation about the evolution of voice conversion techniques will be presented, as well as several basic concepts that are essential to understanding the general voice conversion system framework. Additionally, some fundamental concepts about Deep Learning and more specifically, about Generative Adversarial Networks (GANs) are explained in detail along with a description of the state-of-the-art GAN architectures, including CycleGAN and StarGAN.
In Chapter 3, the chosen dataset used for training and testing, together with the experimental setup and development environment are presented to the reader.

In Chapter 4, first, we describe the chosen state-of-the-art voice conversion baseline for this thesis, StarGAN-VC. This includes an in-depth description of the architecture of its generator and discriminator along with the used loss functions for training them. Then, the three variants of the Top-K methodology that were applied to the baseline StarGAN-VC are described in detail. Additionally, the convergence study of the Top-K training that was carried out is also presented.

Chapter 5 includes the descriptions of the quantitative and qualitative studies, together with the analysis of the results obtained from them.

In Chapter 6 the obtained results in both studies are discussed and several conclusions are derived from the experimental results. Also, possible future work directions are discussed.
8 | Introduction
Chapter 2

Background: Theoretical Framework and Evolution of Voice Conversion Techniques

This chapter covers the theoretical background needed for understanding this thesis. Firstly, an introduction to the field of voice conversion is presented, where the general framework of a voice conversion system is explained together with the different categories or types of voice conversion, with a special focus on the parallel and non-parallel voice conversion distinction. Also, the concept of "domain" in a voice conversion task is introduced. The rest of the chapter introduces the theory about deep learning, and more specifically about generative deep learning (DL) and generative adversarial networks (GANs) which are the models used in this work. After this, we will introduce the StarGAN architecture which is the principal paradigm of the baseline method chosen for this work, the StarGAN-VC. Please note that in this section, several analogies with image-related tasks will be used in order to make the understanding easier for the reader.
Voice conversion (VC) is the study of the transformation of the voice characteristics of a source speaker into those of a desired target speaker, while keeping the linguistic content intact. Voice conversion belongs to the general technical field of speech synthesis (Figure 2.1), which consists in the artificial production of human voices, changing the properties or the non-linguistic information of the speech.

The main non-linguistic information from the human voice are: the speaker identity, which involves all the features that make a listener recognise someone’s voice. These are the accent, the pronunciation and the speaking style, including the expression of emotions, the emphasis or the intentions of a voice utterance. In this sense, voice conversion is focused on the manipulation of the voice identity in speech, which represents one of the main challenging research problems in speech processing in the last years.

Since the beginning of computer-based speech synthesis in the 1950s, there has been a constant effort in the search of an effective manipulation of speech properties. The strong development of digital signal processing in the 1970s translated to an easier control of the parameters for speech manipulation. Despite the fact that the original motivation of voice conversion was merely curiosity and novelty, these technological advancements from statistical modelling to deep learning have made a considerable impact on several real-life applications which benefited its users. These benefits range from personalised speech synthesis, communication aids and speaking assistance for the speech-impaired to speech enhancement, voice translation services or voice dubbing for movies [2].

Generally, a human speaker’s voice can be characterised by three kinds of factors:

1. Linguistic factors: correspond to the features that are reflected in the sentence structure, lexical choice and idiolect, which is an individual’s distinctive and unique use of the language. These include vocabulary, grammar and pronunciation.
2. Supra-segmental factors: such as the prosodic characteristics of a speech signal, these consist in the elements of speech that are not individual phonetic segments but are properties of syllables and larger units of speech. These correspond to linguistic functions such as intonation, stress, and rhythm.

3. Segmental factors: are the ones related to short-term features such as the spectrum and formants, which are the broad peaks or local maximums in the spectrum resulting from the acoustic resonance of the human vocal tract. These include vowels and consonants.

As previously mentioned, the idea in voice conversion is that the linguistic content of the input speech remains untouched. Therefore, the segmental and supra-segmental factors become the most relevant factors in a speaker individuality. For this reason, an effective voice conversion system is expected to convert both of these factors. Nonetheless, nowadays, voice conversion is still far from perfect but the recent application of deep learning techniques is helping its refinement achieving satisfactory results in this task.

### 2.0.1 The General Voice Conversion Pipeline

In this subsection, we will introduce the more general voice conversion pipeline. The whole pipeline of the proposed voice conversion system will be then explained in detail.

As shown in Figure 2.2, a typical voice conversion pipeline includes, fundamentally, a speech analysis and feature extraction, mapping and reconstruction modules. The speech analyser and feature extractor module decomposes the speech signal of a source speaker into features that represent the supra-segmental and segmental information of the speech. These features will be further described in Section 3.1.1 where the preprocessing and feature extraction of an audio file will be explained in detail. Then, the mapping module is essential, as it is in charge of mapping the speech characteristics from the source speaker to the ones of the target speaker. Finally, the speech reconstruction module re-synthesises the feature mappings into time-domain speech signals generating a reconstructed audible speech from the converted speech features.

Due to the recent advances of deep learning in the last years, VC has emerged as an active research area for the deep learning community. The VC systems based on deep learning, are generally composed by two phases: the training phase and the conversion phase. During the training phase, a model
gets trained using the speech features extracted in the speech feature analysis stage, and through many iterations, it learns the feature mappings between speakers. Then, during the conversion phase, the model is given a certain speech audio input from a source speaker and using the knowledge from the training phase, it modifies the voice features to sound and resemblance the target speakers’ voice.

2.0.2 Meaning of Domain in a Voice Conversion Task

Generally, in generative deep learning tasks the goal is to learn a mapping between different *domains*. For example, in image-to-image translation, these *domains* are visual [21]. In this case, *domain* implies a set of images that can be grouped as a visually distinctive category, and each image has a unique appearance, which is called *style*. For example, two separate image domains could be based on the gender of a person, in which case the style would include hairstyle, makeup or beard [22].

However, when we address a voice conversion task, *domain* means a set of recorded voices that can be grouped as auditory distinctive, thus meaning that different domains relate to different speakers. In this sense, two different and distinctive domains in a voice conversion task could be male and female speakers, in which the style would include a lower or higher pitch, for instance.

Domains can be far from each other (e.g., conversion between male and female speaker and vice-versa) or close (e.g., conversion between two male speakers). Since a voice conversion between closer domains is not as challenging for a deep learning model as a conversion between further domains, a voice conversion system is usually tested with conversion experiments between two male and two female speakers.
2.0.3 Classifications of Voice Conversion Systems

Depending on the approach of the voice conversion task, VC systems can be divided into eight different categories: parallel VC, non-parallel VC, mono-lingual speech data based VC, cross-lingual speech data based VC, intra-gender VC, cross-gender VC, one-to-one VC, and many-to-many VC.

Parallel vs. Non-Parallel Voice Conversion

The parallel and non-parallel VC systems are categorised based on the linguistic contents of the speech dataset with which the VC modules are trained.

The parallel voice data consists of recordings of the same sentences read by the source and target speakers. This means that finding the correspondence between paired utterances can be directly achieved by performing time-alignment. Therefore, the speech feature mapping component can easily map the source speakers’ vocal features to the target speakers’ vocal features due to the presence of frame-wise alignment of the linguistic contents. However, collecting parallel voice data requires a very large effort compared to that for non-parallel voice data.

On the other hand, in non-parallel VC systems, training data consists of samples that contain misalignment of the linguistic contents of the source and target speakers. The advantage of non-parallel voice data is that it does not require parallel utterances, transcriptions or time alignment procedures. Also, it is more suitable for practical applications, specially in real-time situations, where approaches that are able to collect data with ease are desirable.

The main issue with non-parallel voice conversion is that it is not straightforward to measure the correspondence between the source speaker spectral features and the target ones, due to the non-existence of paired utterances. For this reason, in general the quality and the conversion effect obtained with non-parallel methods is normally limited compared with the ones that use parallel data, because of the disadvantage related to the training condition (Figure 2.3). Therefore, developing non-parallel voice conversion methods whose synthesised speech quality is comparable to the one of parallel ones, represents an important challenge for the speech synthesis field.

Mono-lingual vs. Cross-lingual Voice Conversion

The division of voice conversion systems between mono-lingual and cross-lingual is based on the language of the speech dataset that is used for the
conversion process.

In mono-lingual VC, the training of the VC module is performed using the speech dataset with utterances in the same language for both the source and the target speakers. Whereas for cross-lingual VC, this module is trained with speech data from different languages (Figure 2.4). This difference in the language might seem trivial, however it corresponds to a considerable challenge as the differences in the basic speech features such as syllable, prosody, intonation, etc. change from language to language making the mappings of from the source to the target speaker harder to learn, compared to mono-lingual VC.
Intra-gender vs. Cross-gender Voice Conversion

VC systems also happen to be divided based on the genders of the speakers involved in the voice conversion process.

The gender of the speakers plays a very important role in the voice conversion process, as male and female speakers’ speech features are considerably different. In intra-gender VC both source and target speaker belong to the same gender, in this sense, the conversion domains are closer, and therefore, the voice conversion task becomes easier as their speech features are more similar between each other.

On the other hand, in cross-gender VC the source and the target speaker belong to the opposite gender, making the voice conversion task not as straightforward as in intra-gender VC (Figure 2.5). Nonetheless, cross-gender VC systems are considered more robust as they can be used for both intra and cross-gender voice conversion tasks.

One-to-One vs. Many-to-Many Voice Conversion

The last categorisation of VC systems is based on the number of speakers among which the system is able to perform the conversion task.

In one-to-one VC systems, the model can convert the speech features from only one source speaker to a determined target speaker. Whereas in many-to-many, the VC model can convert the vocal features of more than one source speaker to more than one target speaker during the voice conversion process (Figure 2.6).

The recent advances in deep learning have focused their efforts in creating
many-to-many VC systems, as they are more robust, and even though they are more complex they are also capable of performing one-to-one VC tasks.

Figure 2.6: One-to-One vs. Many-to-Many VC system representation [1].

2.0.4 Acoustic Features and Conversion Process

In the same way as in image-related tasks, the features that are input into the deep learning model and that describe the image numerically usually correspond with the pixel values from the RGB channels, in audio-related tasks these features are different.

The first step in any voice conversion or automatic speech recognition system is to extract the features from the speech utterances. These audio features represent the components of the audio signal that are relevant for identifying the different characteristics of a speech signal. However, this task with voice data is not as trivial as in image-related tasks.

Sounds generated by a human are filtered by the shape of the vocal tract including the tongue, teeth, etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. In this sense, Mel Frequency Cepstral Coefficients (MFCCs) are features widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980’s and have been state-of-the-art ever since.
The Mel-scale and Mel Frequency Cepstral Coefficients (MFCCs)

For understanding what MFCCs are, it is necessary to know what is the Mel scale. Generally, we are used to interpreting audio signals using the Hertz (Hz) scale. The difference in frequency in Hz is linear, however when you compare for example, the jump between 65-262 Hz and from 1567-1760 Hz, even though the frequency jump is numerically the same, to the human ear, the two notes in the second jump are much closer in pitch between each other than the first jump. The way humans perceive pitch is non-linear, this means in a logarithmic way. Therefore, there was a need for a non-linear scale, and that was the reason for the introduction of the Mel-scale.

The Mel scale relates the perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely with the differences that humans hear compared to the Hertz scale. For reference, in the Mel-scale $1000\text{Hz} = 1000\text{Mel}$, see Figure 2.7.

For extracting the MFCCs from an audio signal the next steps should be followed (see Figure 2.8) [23]:

1. Frame the signal into short frames.

2. For each frame, calculate the periodogram estimate of the power spectrum.
3. Apply the mel filter-bank to the power spectra, sum the energy in each filter.

4. Take the logarithm of all filter-bank energies.

5. Take the discrete cosine function (DCT) of the log filter-bank energies.

6. Keep DCT coefficients 2-13, discard the rest.

The advantages of the MFCCs include:

- Quantifies the gross-shape of the spectrum (the spectral envelope), which is important in, for example, identification of vowels. At the same time, it removes fine spectral structure (micro-level structure), which is often less important. It thus focuses on that part of the signal which is typically most informative.

- Straight-forward and computationally reasonably efficient calculation.

- Their performance is well-tested and understood.

Some of the issues with the MFCC include:

- The choice of perceptual scale is not well-motivated. Scales such as the equivalent rectangular bandwidth (ERB) or gamma-tone filterbanks might be better suited. However, these alternative filterbanks have not demonstrated consistent benefit, whereby the mel-scale has persisted.
• MFCCs are not robust to noise. That is, the performance of MFCCs in presence of additive noise, in comparison to other features, has not always been good.

• The MFCCs work well in analysis but for synthesis, they are problematic. Namely, it is difficult to find an inverse transform (from MFCCs to power spectra) which is simultaneously unbiased (accurate) and congruent with its physical representation (power spectrum must be positive) [3].

• The calculation of the MFCCs has a high computational cost which translates in longer times during the pre-processing of the audio files to extract the features.

These MFCC of voice data correspond to the audio features that we will use for training and testing the deep learning models developed in this work. The way for extracting them is described in more detail in Section 3.1.1.

2.1 Deep Learning and Generative Deep Learning

The objective of this section is to provide the reader with the theoretical background behind the basic principles about deep learning (DL), and more specifically about generative deep learning and generative adversarial networks (GANs).

As a result of the technological advances from the last years, the field of artificial intelligence (AI) has become really popular within and out the scientific community. This recent rise on its popularity has made it easy to forget that the AI field has been around for decades. During this time, its popularity and the confidence in this kind of technology has experienced several fluctuations. One of the main problems in the past of the field, was related with the low computational capacity of the graphic processing units (GPUs) which made the training of the models with large amount of data an arduous task. However, nowadays, the field has become more accessible, allowing one of its disciplines, machine learning (ML), to be applied to almost every aspect in science.

With the purpose of a better comprehension from the reader, it is essential to differentiate the terms of artificial intelligence, machine learning and deep learning (Figure 2.9). The field of AI is, in essence, when machines (robots or
computers) are able to perform tasks that typically require human intelligence. Machine learning is the application of AI that provides systems the capacity to automatically learn and improve their from experience, similarly to the way human beings learn. There are many branches under the field of machine learning, however, artificial neural networks (ANN or simply NN), have shown outstanding potential and have been a big part of the scientific research development during the last years. These, consist in computational algorithms inspired by the functioning of the neurons in the human brain, that are capable to learn from large amounts of data.

On the other hand, deep learning is a branch of ML which uses algorithms with multiple layers to progressively extract higher level features from the raw input. The main idea is to implement multiple layers of features so the model will increasingly learn certain features with a high abstraction level. Furthermore, each of these high-level features are defined in terms of their relationship with simpler features. In other words, the algorithms learn complicated concepts from simpler ones, creating a hierarchical structure that creates large concatenations (deep) of features, which provide the name to these kind of techniques. The main building blocks used for that purpose are Convolutional Neural Networks (CNN).

![Figure 2.9: Representation of the scope of AI, ML and DL](image-url)
2.1.1 Generative Adversarial Networks (GANs)

In this subsection we aim to introduce the reader to the field of generative deep learning, focusing on one of its most popular types of networks, generative adversarial networks (GANs), which were chosen as the main part of the voice conversion system developed in this thesis.

GANs were firstly introduced in 2014 by Goodfellow et al. [24]. In this article, they proposed the vanilla version of GANs, a novel framework for estimating generative models through an adversarial process. In this process, two models are trained simultaneously: a generative model or generator $G$ that is in charge of capturing the data distribution and a discriminative model or discriminator $D$, that estimates the probability that a given sample corresponds to a real element from the training data rather than from $G$. This simultaneous training corresponds to a mini-max like game, where the training objective of $G$ is to maximise the probability of $D$ making a mistake.

When first introduced, the vast potential of GANs’ adversarial training was considered as "the most interesting idea in the last ten years in ML" [25]. This is because GANs could be trained to generate new fake samples that were significantly realistic in different domains, such as images, music, speech or prose.

In order to understand GANs, it is necessary to understand how generative algorithms work in contrast with the vastly used discriminative algorithms. Discriminative algorithms (supervised learning) aim to classify input data, which means that given the features of a data sample, they predict or estimate a label or category to which that data belongs. In discriminative modelling, each observation in the training data is associated with a label (i.e. 0 or 1), and the purpose of the model is to learn how to differentiate between these two groups, and therefore it outputs the probability that a new observation has label 1, it maps features to labels. In other words, discriminative modeling attempts to estimate the probability that an observation $x$ belongs to category $y$, $p(y|x)$. On the other hand, generative modelling is usually performed with an unlabelled dataset (unsupervised learning). It could be understood as instead of predicting a label given certain features, it attempts to predict features given a certain label, this is $p(x)$ or $p(x|y)$ if the data is labelled [26].

**Vanilla GAN**

In [24], Goodfellow et al. introduced the basic or vanilla version of the GAN architecture. This is based in a model where two antagonistic networks are trained and learn against each other: the Generator $G$ and the Discriminator...
\[ D. \] They are considered antagonistic between each other, as \( G \) generates a fake image while \( D \) tries to determine whether this is a real image or not as represented in Figure 2.10. This way the Generator \( G \) aims to fool the Discriminator and learns through backpropagation when \( D \) makes a wrong decision. However, the Discriminator \( D \) is simultaneously learning and becoming better at discerning if images are real or not, but as \( G \) becomes better at learning how to produce fake images, it will become a harder task for \( D \) to determine whether they are real.

![Figure 2.10: Generator and Discriminator as GAN building blocks [6].](image)

In Figure 2.11 the objective function that aims to be optimised is depicted. The Discriminator function is termed as \( D \) and the Generator function is termed as \( G \). \( P_z \) is the probability distribution of the latent space which is usually a random Gaussian distribution. \( P_{\text{data}} \) is the probability distribution of the training dataset. When \( x \) is sampled from \( P_{\text{data}} \), the Discriminator wants to classify it as a real sample. \( G(z) \) is a generated sample when \( G(z) \) is given as input to the Discriminator, it wants to classify it as a fake one. The Discriminator intends to drive the likelihood of \( D(G(z)) \) to 0. Hence it wants to maximise \( (1 - D(G(z))) \) whereas the Generator wants to force the likelihood of \( D(G(z)) \) to 1 so that Discriminator makes a mistake in calling out a generated sample as real. Hence Generator wants to minimise \( (1 - D(G(z))) \).

This vanilla version of the GANs showed several limitations related to the training of the networks. Firstly, the mode collapse, which happens as, normally, the generative goal is to produce a wide range of samples that are somehow different between each other and that all are capable of fooling the discriminator. However, if a generator produces an especially plausible output, the generator may learn to produce only that output. In fact, the generator is always trying to find the one output that seems most plausible
to the discriminator. Therefore, if the generator starts producing the same output (or a small set of outputs) over and over again, the discriminator’s best strategy is to learn to always reject that output. But if the next generation of discriminator gets stuck in a local minimum and doesn’t find the best strategy, then it becomes too easy for the next generator iteration to find the most plausible output for the current discriminator. Each iteration of generator over-optimises for a particular discriminator, and the discriminator never manages to learn its way out of the trap. As a result the generators rotate through a small set of output types, losing the variability on the output samples and hence, losing creativity [26].

Moreover, a common problem that is faced during the GANs training is that the losses of the generator and the discriminator might never reach a convergence point. This phenomenon is called the convergence failure and it is due to the high sensitivity of these kind of networks to the hyper-parameters.

Finally, one of the main inconvenient of this basic GAN architecture is that there is not a real control about the content of the generated samples, meaning that they could be fooling the discriminator but they might be of low quality or not reaching the requirements of the user or the output domain.

With the objective of solving some of the limitations of the GANs stated above new modifications to the original vanilla version were proposed, such as Wasserstein, Conditional, Cycle and StarGAN.
Wasserstein GAN (WGAN), Wasserstein GAN - Gradient Penalty (WGAN-GP) and Conditional GAN (CGAN)

In the first place, the one known as Wasserstein GAN (WGAN) [27] and its successor, the Wasserstein GAN - Gradient Penalty (WGAN-GP) [28] [22], constitute improvements of the vanilla version that tried to increase stability, when training the model and reduce the problem of mode collapse. To do this, both replace the binary cross-entropy loss function with a loss function that correlates with the quality of the images generated.

On the other hand, the Conditional GAN (CGAN) [29] addressed the problem of poor control of the content synthesis of the vanilla version. To do this, they use external information that guides the imaging process. This way, a CGAN model always combines a basic GAN with an external information source, such as discrete class labels, text descriptions, semantic maps, conditional images, object masks or attention maps, to better control the content of the generated image.

2.1.2 CycleGAN and CycleGAN-VC

Until 2017, the use of the vanilla version of GANs for image-to-image translation presented one major limitation, which was the need of paired training data between two domains. This means that two sets of the exact same images in different domains were needed in order for the model to learn the mapping but this is not available in many tasks. However, the publication of [7] by Zhu et al. was considered a game changer, as they presented an approach for learning to translate an image from a source domain $X$ to a target domain $Y$ in the absence of paired examples, $G : X \rightarrow Y$. For doing so, the system only needed different images that belonged to the two domains and that were tagged, without the need of having paired images from both domains. Moreover, the success of their approach was that it also allowed inverse mapping $F : Y \rightarrow X$ and the introduction of the cycle consistency loss that enforced $F(G(X)) \approx X$ and vice-versa.

The architecture of the CycleGAN is shown on Figure 2.12. The model consists of four neural networks: two generators ($G$ and $F$) and two discriminators ($D_X$ and $D_Y$). $G$ is in charge of transforming images from domain $X$ to domain $Y$. Analogously, $F$ transforms the images from domain $Y$ to domain $X$. Then, $D_Y$ determines if the input image is a real image of domain $Y$ or a fictitious image that tries to resemble domain $Y$. For its part, $D_X$ tries to discriminate if an input image is a real image of the domain $X$ or a fake image that tries to resemble the domain $X$. 
Figure 2.12: (a) The CycleGAN model contains two mapping functions $G : X \to Y$ and $F : Y \to X$, and associated adversarial discriminators $D_Y$ and $D_X$. $D_Y$ encourages $G$ to translate $X$ into outputs indistinguishable from domain $Y$, and vice versa for $D_X$ and $F$. To further regularise the mappings, they introduced two cycle consistency losses that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \to G(x) \to F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \to F(y) \to G(F(y)) \approx y$.

Following the same idea and architecture, in 2018 Kaneko and Kameoka presented CycleGAN-VC [8] which consisted in a non-parallel voice conversion method that was able to learn a mapping from source to target speech without the need of parallel data. This way, they were able to generate high quality conversions without the need of extra data, modules or time alignment procedure. Similarly than in [7] they used cycle-consistency losses an an identity-mapping loss (see Figure 2.13), which allowed the mapping function to capture sequential and hierarchical structures while preserving linguistic information.

### 2.1.3 StarGAN and StarGAN-VC

Despite the good quality of the conversions generated by CycleGAN and CycleGAN-VC, one major limitation was that these models were only able to learn mappings between a single pair of domains, or speakers in the case of the latter one.

In the most of real-life VC application scenarios, it is desirable to be able to convert speech into multiple domains or speakers, not just one. Therefore, a naive way of applying CycleGAN to multi-domain VC tasks would be to prepare and train a different mapping pair for each domain pair, see (a) in Figure 2.14. However, this methodology is very ineffective since each
mapping pair fails to use the training data of the other domains for learning, even though there must be a common set of latent features that can be shared across different domains.

In order to overcome this issue in the image-related tasks, in 2018 Choi et al. published [9] where they presented StarGAN, a new GAN architecture that solved the limitation of the two-domain conversion. In this work, they presented a novel and scalable approach which was capable of performing image-to-image translations for multiple domains using only a single generator model. The StarGAN architecture allowed for simultaneous training of multiple domains with different domains within a single network, one pair of generator and discriminator (Figure 2.15), leading to a superior quality of translated images compared to existing models as well as the novel capability of flexibly translating an input image to any desired target domain.

Similarly to what happened with CycleGAN and CycleGAN-VC, the StarGAN architecture originally used for image-related tasks was adapted to a voice conversion system in 2018 by Kameoka et al. in [20] and in 2020 in [10]. Their main features of StarGAN-VC were the following: first, it did not require utterances, transcriptions, or time alignment procedures for
Figure 2.14: Comparison between cross-domain models and our proposed model, StarGAN. (a) To handle multiple domains, cross-domain models should be built for every pair of image domains. (b) StarGAN is capable of learning mappings among multiple domains using a single generator. The figure represents a star topology connecting multi-domains [9].

speech generator training. Second, it could simultaneously learn mappings across multiple domains using a single generator network (Figure 2.16) and thus fully exploit available training data collected from multiple domains to capture latent features that are common to all the domains. Third, it was able to generate converted speech signals fast enough to allow real-time implementations.

A more detailed description of the StarGAN-VC architecture and training is provided in Section 3.1.1 and 3.3.

2.1.4 **Top-K GAN Approach**

As it have been shown until now, GANs have been a very successful tool for image and audio synthesis. However, despite the quality of their outputs GANs, in general, are difficult to train and in the last couple of years a lot of effort has been focused on how to modify their training procedure with the objective of reducing this difficulty. Very recently, at the end of 2020, Sinha et al. published a very innovative study [11] in this sense. This is the Top-K methodology for training GANs.

Their proposal has captivated the generative deep learning field due to its simplicity and efficiency. As shown in 2.17, it is based on the principle that for updating the weights of the generator in a GAN, not all the fake samples
Figure 2.15: Overview of StarGAN, consisting of two modules, a discriminator $D$ and a generator $G$. (a) $D$ learns to distinguish between real and fake images and classify the real images to its corresponding domain. (b) $G$ takes in as input both the image and target domain label and generates a fake image. The target domain label is spatially replicated and concatenated with the input image. (c) $G$ tries to reconstruct the original image from the fake image given the original domain label. (d) $G$ tries to generate images indistinguishable from real images and classifiable as target domain by $D$ [9].

from the generator should be used. Instead, only the best $K$ images or samples that were able to fool the discriminator will be used for training and updating the generator. This way, only the most realistic fake samples will have an effect on the learning of the generator, which are the ones that have produced a lower adversarial loss. This way, as the title of their work suggest, they do not consider the bad samples, and only the good samples produced from the generator are kept for training. The authors showed that by applying this change during the training there is a substantial increase on the quality of the results without increasing the computational cost.

More in-depth, when the parameters of the generator are updated during training a mini-batch of generated samples, they simply zero out the gradients from the elements of the mini-batch corresponding to the lowest critic outputs. Intuitively, as training progresses, the discriminator $D$, can be seen as a scoring function for the generated samples: a generated sample that is close to the target distribution will receive a higher score, and a sample that is far from the target distribution will receive a lower score. Therefore, by performing the Top-K operation on the critic predictions, the generator is only updated on the best generated samples in a given batch, as scored by the discriminator $D$ [11].

An issue that could rise is that if the Top-K methodology starts very early
Figure 2.16: Illustration of A-StarGAN training from [10]. The A network is designed to produce 2K probabilities, where the first and second K probabilities correspond to real and fake classes, and simultaneously play the roles of the real/fake discriminator and domain classifier.

during the training process, the discriminator might not be a reliable scoring function for the samples of the generator, as itself is not trained enough to distinguish real from fake samples. In this case, it would not be helpful to throw away gradients from samples poorly scored by the discriminator at the beginning of the training as we would just be throwing away random samples, which could be equivalent as just reducing the batch size. For this reason, they set $K = B$, where $B$ is the full batch size, at the start of the training and then, gradually, the value of $K$ is reduced by a decay factor $\gamma$, with a lower limit of $k = v$. The $v$ term is introduced so that the value of $K$ is not reduced to one element.

However, in the current deep learning literature there has not been a study that aims at finding the optimal moment for starting to apply the decay of the $K$ value, nor a study that suggests which should be the value of this $\gamma$ (i.e. if it should be a progressive decay or a quick decay in the value of $K$). That is the objective of this thesis: firstly, to apply the Top-K methodology to a state-of-the-art voice conversion model such as StarGAN-VC [20][10], secondly, to carry out a study on the optimal training moment for starting to apply the Top-K methodology and thirdly, to perform a study on the optimal value of the decay of $K$, $\gamma$. 
Figure 2.17: Diagram of top-k training of a GAN. The generator generates a batch of samples, which are scored by the critic. Only the $K$ samples with the highest scores are used to update the generator [11].
Chapter 3

Methodology

In this chapter, first, we introduce the dataset and the development environment employed for carrying out the different experiments. Next, we present in detail to the reader the StarGAN-VC architecture which was chosen as the baseline for this work, our proposed methodology which is the Top-K training of StarGAN-VC, along with an explanation and reasoning of the performed experiments in this dissertation.
3.1 Material and Development Environment

In this section we describe the dataset that was used for training and evaluating our model, the preprocessing and audio feature extraction from the voices belonging to the dataset and a description of the development environment (libraries, IDE and hardware) that was used for this thesis.

3.1.1 Dataset

In order to train and test the proposed training of Top-K StarGAN-VC, we used the CSTR VCTK (Voice Cloning Toolkit) English Multi-speaker Corpus [30], which is the same one that was used to train CycleGAN-VC and StarGAN-VC. This dataset includes speech data uttered by 109 English speakers with various accents. Each speaker reads out about 400 sentences, which were selected from a newspaper, the rainbow passage and an elicitation paragraph used for the speech accent archive. All speech data was recorded using an identical recording setup: an omni-directional microphone (DPA 4035), 96kHz sampling frequency at 24 bits and in a hemi-anechoic chamber of the University of Edinburgh. All recordings were converted into 16 bits, were downsampled to 48 kHz based on STPK, and were manually end-pointed.

For training both the baseline model StarGAN-VC and our models to do the experiments with the different parameters of the Top-K approach, we chose a subset of ten speakers from [30]. Particularly, we selected five pairs of male and female speakers corresponding to five different accents (Edinburgh, South England, Northern Irish Belfast, American New Jersey and India). This decision was made for making the model more robust against different domains such as gender and accent. As the data from each of the speakers consists in them reading the same sentences, we could have actually constructed a parallel scenario using this dataset. However, as our purpose is to simulate a non-parallel training scenario, it was decided not to take advantage of it.

Audio Data Preprocessing

The first step in the preprocessing of the voice data from the ten speakers was to downsample the original audiofiles to 16kHz. Then, for each utterance, we extracted the harmonic spectral envelope \( sp \), the fundamental frequency \( F_0 \), the aperiodic spectral envelope (relative to the harmonic spectral envelope) \( ap \), and 36 MFFCs every 5 ms using the WORLD vocoder [31] using the Python implementation PyWORLD [32]. Instead of the raw audio data from
the speakers, these MFCCs were used as the audio features for training our model implementation.

### 3.1.2 Development Environment

In this subsection we will describe the environment used for developing this thesis.

The project was developed using Python, which is an interpreted high-level general-purpose programming language. The Python implementation of our baseline model StarGAN-VC was developed in PyTorch and can be found in [17]. PyTorch [33] is an open source machine learning library used for applications such as computer vision and natural language processing. In this case, the chosen integrated development environment (IDE) was PyCharm [34].

The main libraries needed for running the StarGAN-VC code [17] were the following:

- **sox**: Sound eXchange is a cross-platform audio editing software [35]. It was used for downsampling the original voice data to 16kHz during the preprocessing.

- **librosa**: Librosa is a Python package for music and audio analysis [36]. It was used for loading the original voice data saving the converted samples.

- **pyworld**: PyWORLD is a fast and high-quality vocoder which parameterises speech data [32]. It was used for extracting the audio features from the original voice data and for resynthesizing the converted features again to audio data.

- **torch**: PyTorch was used for creating the StarGAN-VC architecture.

- **tensorboard**: TensorBoard provides the visualisation and tooling needed for machine learning experimentation [37]. It was used for visualizing the training data and curves on a web server.

- **ray**: Ray Tune is a hyperparameter optimisation framework for long-running tasks such as RL and deep learning training [38]. Ray Tune makes it easy to go from running one or more experiments on a single machine to running on a large cluster with efficient search algorithms. It was used for optimizing the parameters of the model, and trying different values for finding the most optimal ones.
All the experiments in this thesis were ran using a PC with an Intel(R) Core(TM) i7-10700F processor with 32GB of RAM, and a NVIDIA Quadro RTX 4000 GPU due to the high computational cost of the audio preprocessing and model training.

3.2 Background: StarGAN Voice Conversion (StarGAN-VC) as a Baseline Model

StarGAN-VC [20] by Kameoka et al., which was generally introduced in Section 2.1 is the baseline voice conversion system that was chosen with the objective to introduce the Top-K training methodology, and then to compare the obtained results. This decision was made because their state-of-the-art implementation of the StarGAN architecture [9] applied to the voice conversion task obtained the best quantitative and qualitative results compared to other non-parallel methods such as VAEGAN [19] and their previously proposed method CycleGAN-VC [8]. This is also the reason why in our results section, we only compare the results obtained with the different Top-K variants with the ones from StarGAN-VC, as it has already been proven to outperform previous non-parallel voice conversion approaches.

Therefore, in this section we will present to the reader an in-depth explanation of the StarGAN-VC pipeline, architecture and the code implementation in Python, with the objective of later understanding the Top-K training of this model.

3.2.1 Model Pipeline

As a reminder to the reader, the StarGAN-VC only has one pair of generator and discriminator that are capable of doing the conversions between different many-to-many domains, differently than CycleGAN, which had the limitation that it only learns one-to-one domains mappings for each pair of generators and discriminator.

Training Objectives

Let $G$ be a generator that takes a sequence of acoustic features $x \in \mathbb{R}^{Q \times N}$, where $Q \times N$ is the size of $x$, with an arbitrary attribute and a target attribute label $c$. This acoustic feature sequence is the input of the generator that generates $\hat{y} = G(x, c)$, which corresponds to the original sequence $x$ in the
target domain $c$. Therefore, $c$ is represented as a concatenation of one-hot vectors, each of which is filled with 1 at the index of a class in a certain category and with 0 everywhere else. For example, if we consider speaker identities as the only attribute category, $c$ will be represented as a single one-hot vector, where each element is associated with a different speaker.

One of the main goals of StarGAN-VC is to make $\hat{y} = G(x, c)$ sound as realistic as real speech features and belong to attribute $c$. For this reason, similarly than in CycleGAN-VC there is the introduction of the real/fake discriminator $D$, which in this version of StarGAN-VC [10] it is unified with the domain classifier as well, whose role is to also predict to which class or speaker a given input belongs (See Figure 3.1). Note that in [10], $D$ is referred as $A$ as it is considered an "augmented" classifier. However, in this work the real/fake discriminator + domain classifier will be referred to as $D$ for simplicity.

Figure 3.1: Illustration of A-StarGAN training. The $A$ network is designed to produce $2K$ probabilities, where the first and second $K$ probabilities correspond to real and fake classes, and simultaneously play the roles of the real/fake discriminator and domain classifier [10].

In this case, the output of $D$ are $2K$ probabilities $p_{D}(k|x)$ $k = 1, ..., 2K$, where $k = 1, ..., K$ and $k = K + 1, ..., 2K$ correspond to the real domain classes and the fake classes respectively. Then, by using this multi-class classifier, the adversarial loss for the generator $G$ and generator $D$ is defined as $\mathcal{L}_{adv}^G$ and $\mathcal{L}_{adv}^D$, respectively (see Equation 3.1 and 3.2).
\[ L_{adv}^G = -E_{k \sim p(k), x \sim p_d(x)} \{ \log p_D(k | G(x, k)) \} + E_{k \sim p(k), x \sim p_d(x)} \{ \log p_D(K + k | G(x, k)) \} \] (3.1)

\[ L_{adv}^D = -E_{k \sim p(k), y \sim p_d(y|k)} \{ \log p_D(k | y) \} - E_{k \sim p(k), x \sim p_d(x)} \{ \log p_D(K + k | G(x, k)) \} \] (3.2)

Following the adversarial idea, \( L_{adv}^D(D) \) becomes small when \( D \) correctly classifies \( y \sim p_d(y|k) \) as real speech in domain \( k \) and \( \hat{y} = G(x, k) \) as fake samples in domain \( k \). On the other hand, \( L_{adv}^G \) becomes small when \( G \) is able to fool \( D \) so that \( \hat{y} = G(x, k) \) is misclassified by \( D \) as real speech samples in domain \( k \) and not classified as fake samples.

Nonetheless, as stated in Section 2.1 training \( G \) and \( D \) using only the losses presented above does not guarantee that \( G \) will preserve the linguistic content of input speech. Therefore, similarly than with CycleGAN-VC, it is needed to include a cycle consistency loss to be minimised.

### 3.2.2 Baseline Model Training

In the Vanilla version of StarGAN-VC the generator weights are updated every five iterations of the discriminator updates according to Equation 3.3.

\[ G_{loss} = G_{real/fake} + \lambda_{rec} G_{rec} + \lambda_{clf} G_{clf} \] (3.3)

The vanilla training of \( G \) consists in the following steps:

1. A sequence of real MFCCs belonging to the original speaker \( o \), with a one-hot encoded target speaker domain \( k \) is passed through the generator \( G \). Therefore, \( G \) generates a fake sequence of MFCCs which aim to resemble the ones from the given target speaker domain \( k \).

2. This fake sequence of MFCCs resembling \( k \) is passed to the discriminator \( D \) which is in charge of determining whether this sequence is real or not, and to which speaker domain does it belong.

3. The first output of \( D \) is then used to compute the term \( G_{real/fake} \), which corresponds to the loss of the fake generated samples by \( G \) after being passed to \( D \) which determines which of them are more likely to be real.

4. The second output of \( D \) is used to compute the speaker classification loss \( G_{clf} \), which will be close to zero if \( D \) determines that the given
sequence of MFCCs corresponds to the target speaker domain $k$.

5. Finally, the $G_{rec}$ term corresponds to the reconstruction loss. Here, the fake MFCCs sequence previously generated by $G$, is passed back to $G$ with the label of the original speaker $o$ as the target speaker domain label. Here it is performed a target-to-original domain, to reconstruct the target fake sequence back to the original domain.

### 3.3 Proposed Model Architecture: Top-K StarGAN-VC

In the previous section we have explained in detail the architecture and training of StarGAN-VC which we have chosen as our baseline model. Therefore, in this section we will explain how the Top-K methodology introduced in Section 2.1 will be introduced to StarGAN-VC with the objectives of generating more realistic sounding voices, an earlier convergence of the model and more stability during the training process.

The Top-K methodology introduced by Sinha et al. in [11], consisted in using for the training only those images that were able to fool the discriminator more effectively. Therefore, the weights of the generator were only updated by using the loss generated by those images.

So far in literature, the Top-K methodology has only been applied to image-related tasks and to the regular GAN architecture, particularly to the SAGAN architecture [39]. Therefore, the key contribution of this thesis relies on combining such a novel and promising training scheme with a state-of-the-art voice conversion system such as StarGAN-VC. This corresponds to a challenging task as not only we are introducing it to a complex GAN architecture but also we are dealing with audio features (MFCCs) instead of image features.

According to the training process of Generator $G$ in the Vanilla version of StarGAN-VC described in the previous Section based on Equation 3.3, this scenario provides us with three different options to introduce the Top-K methodology. First, to choose only the Top-K fake MFCCs generated by $G$ which $D$ discriminated as real ones for the $G_{real/fake}$ term. Secondly, to select only the Top-K MFCCs generated by $G$ that were correctly classified as belonging to the target speaker $k$, to affect the $G_{clf}$ term in the Equation 3.3. Thirdly, combine both approaches, and choose the intersection of the Top-K generated MFCCs samples that were considered real and belonging to
the correct target speaker domain by $D$.

However, the problem for the second and third approaches is that the speaker domain classifier of $D$ is trained very fast, and the generator $G$ is able to quickly generate samples that resemble the correct target speaker $k$ (see Figure 3.2c). Therefore, the speaker domain classification loss $G_{clf}$ becomes very close to zero, and therefore selecting the Top-K best classified samples would not add much value. For that reason the main experiments performed in this thesis correspond to the first considered approach, which is to select the $K$ best samples that better fooled $D$ in resembling to real MFCCs sequences, despite being generated by $G$.

### 3.3.1 The Critical Point of GAN StarGAN-VC Training

In the original Top-K paper by Sinha et al. [11], it is not specified when is the best moment to start decaying the value of $K$ from the full batch size until $v$. They only mention that it should not be from the very beginning as then, the $G$ is not trained enough to start selecting the best generated samples, as all of them are of a poor quality. Therefore, in this part of the study we aim to determine when is the most optimal moment for starting to apply the Top-K methodology.

In a regular training of StarGAN-VC the training curves are represented in Figure 3.2. Here, we can observe that the $D_{loss}$ (Figure 3.2b) instead of smoothly going down to values close to zero, presents a considerable raise before the first half of the training and then starts lowering its values again. The same phenomena but in the opposite sense happens to the generator loss (Figure 3.2a) where once it is very close to zero, it goes down again and starts to not be able to fool $D$ with its fake samples. That corresponds to what we call the Critical Point of the StarGAN-VC training. Please note that the $G_{loss}$ curve in Figure 3.2a is inverted due to the negative sign of the adversarial nature of the StarGAN training.

At the Critical Point, discriminator $D$ is close to start being constantly fooled by the fake samples from $G$. This phenomenon can be seen from step 20,000 to 30,000 in Figure 3.2b. This would be the ideal situation for the training as it would mean that the fake samples generated by $G$ start looking very realistic to the eyes of $D$. However, the $G_{loss}$ instead of converging to zero at that point, goes down again meaning that it is starting to generate samples of not such a good quality, that are not able to fool $D$. Exactly at this point is where we hypothesise that it will be more effective to start applying the Top-K methodology. This is because not all the samples generated by $G$ are equally
bad. In fact, there are some that are better than the rest, but it is still a long way until the full batch size start looking real for $D$. Therefore, considering that at that point is safe to assume that some of them are starting to look real to the eyes of $D$, we consider it a good moment for starting to select only the $K$ better samples of $G$. Hence, these top $K$ best samples will be the ones affecting the $G_{\text{real/fake}}$, while the $D_{\text{loss}}$ will not be directly affected as that result does not affect its value, but by selecting the best samples from $G$ it will make it start raising again as it will start being fooled more often.

For all these reasons, in this thesis we perform a study for determining when is the best moment for starting to apply the Top-K training, as according to [11] it should not be too early, nor too late as probably, if we wait too much, $D$ will be already very well trained for detecting which samples are fake. In our particular case, the whole training of Vanilla StarGAN-VC are 120.000 steps. Therefore, we present a study by testing different moments through this training where to start applying Top-K. The idea is to test whether it is more optimal to introduce Top-K early, in the middle or at the end of the StarGAN-VC training. For this reason, the early version corresponds to start during the Critical Point, which the exact step is not decisive as the value of $K$ at the beginning does not reduce drastically. Therefore, the chosen moments for the experiments are steps 25.000 (early), 50.000 (middle) and 75.000 (end) shown in the second column of Table 3.1.
Table 3.1: Experimental Parameters for the Top-K Variants

<table>
<thead>
<tr>
<th>Top-K Variant</th>
<th>Start Top-K From Step</th>
<th>Gamma (γ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-K A</td>
<td>25.000</td>
<td>0.9999</td>
</tr>
<tr>
<td>Top-K B</td>
<td>25.000</td>
<td>0.99999</td>
</tr>
<tr>
<td>Top-K C</td>
<td>50.000</td>
<td>0.9999</td>
</tr>
<tr>
<td>Top-K D</td>
<td>50.000</td>
<td>0.99999</td>
</tr>
<tr>
<td>Top-K E</td>
<td>75.000</td>
<td>0.9999</td>
</tr>
<tr>
<td>Top-K F</td>
<td>75.000</td>
<td>0.99999</td>
</tr>
</tbody>
</table>

3.3.2 The Optimal Value of γ, the Decay of K

As explained in Section 2.1, the term γ in the Top-K methodology corresponds to the factor that every training step is in charge of reducing the value of K from the full batch size to the value of ν. This is the decay of K.

In the original Top-K article by Sinha et al. [11], the moment of when to start applying this methodology by reducing the value of K is not explicitly studied. Also, the way that K should be decreasing, this is the decay γ, is not provided. However, the value of γ is of paramount importance, as depending on it, the reduction of K will be slow and progressive, or quick and abrupt until reaching the value of ν.

In this thesis we also aim to study and determine if the decay of K should happen in a progressive or more abrupt way. For this reason, together with the study of the optimal moment from which starting to apply Top-K, explained in the previous subsection, we will also perform a study for helping to determine how the value of K should be reduced. For this reason, we will experiment with a slower decay value γ of 0.99999 and a faster, or more abrupt, decay of 0.9999, shown in the third column of Table 3.1.

By carrying out these experiments we aim to perform a general study of the GANs training in general and StarGAN-VC in particular, about the optimal moment of tarting to reduce the samples from the batch considered for the $G_{real/fake}$, and how this reduction should be performed.
Chapter 4

Results Analysis and Evaluation

Once the StarGAN-VC model has been described, it has been trained using two methods. The first one is the basic or *vanilla* version, which corresponds to the one described by Kameoka et al. in [10], that consists in updating the weights of the generator using all the fake generated samples present in the batch size. The second one is based on the Top-K methodology introduced by Sinha et al. in [11] where the generator is only updated with the $K$ best fake samples created by the generator, that were the ones able to *fool* the discriminator. This latter one also includes the experimental study for finding the most optimal moment to start applying the Top-K methodology, and the most optimal decay parameter of $K$, $\gamma$.

In this chapter we perform both a quantitative and qualitative evaluation of the described proposed methodologies as well as a comparison with the basic training of the StarGAN-VC model.
4.1 Quantitative or Objective Evaluation

In this section, we describe the quantitative evaluation for our proposed methodology using the common metrics in the field and we show and analyse the obtained results for the different proposed variants of the Top-K methodology for finding the most optimal moment for starting to apply it during training and the most optimal value of $gamma (\gamma)$ for reducing the value of $K$.

4.1.1 Mel-Cepstral Distance (MCD)

Evaluating voice conversions systems in an objective way is a challenge to the speech synthesis field. Due to the subjectivity of perception and the difficulty of finding an optimal audio feature representation there is no golden standard. For this reason, there is a need of performing both an quantitative and qualitative evaluation of the converted voices for assessing how good is the capacity of a voice conversion system.

The objective performance metric used for quantitatively evaluating the quality of the conversions is the average of the mel-cepstral distortions or mel-cepstral distance (MCD) [40] taken along the dynamic time warping (DTW) [41] path between the converted and the target audio feature sequences. The chosen dataset [30] consists of speech samples of each speaker reading the same sentences. Therefore, if we extract the MFCCs from the original samples of all the sentences of every speaker from the dataset it will be possible to use them later to compare them with the converted ones. This is because when we convert a given sentence from one source to a target speaker, we have the features of that same sentence uttered by the original target speaker, so we know how the true features should look like compared to those fake ones generated by our model. Therefore, using the MCD we can compare the performance of our models by calculating how similar the fake generated features from our model are from the true features uttered by the original target speaker.

4.1.2 Most Optimal Top-K Parameters Study

Losses of the Generator and Discriminator

Table 4.1 shows the performance in the training of StarGAN-VC of the different variants of Top-K described in Table 3.1. Here, we show the
Table 4.1: Ray Tune Results for the Different Top-K Variants compared to the Vanilla version.

<table>
<thead>
<tr>
<th>Training Type</th>
<th>Best Step</th>
<th>Loss G</th>
<th>Loss D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>106.000</td>
<td>0.0165601</td>
<td>10.192</td>
</tr>
<tr>
<td>Top-K A</td>
<td>59.000</td>
<td>0.0000546</td>
<td>20.974</td>
</tr>
<tr>
<td>Top-K B</td>
<td>74.000</td>
<td>0.0001500</td>
<td>17.108</td>
</tr>
<tr>
<td>Top-K C</td>
<td>92.000</td>
<td>0.0002488</td>
<td>8.048</td>
</tr>
<tr>
<td>Top-K D</td>
<td>114.000</td>
<td>0.0001400</td>
<td>11.223</td>
</tr>
<tr>
<td>Top-K E</td>
<td>104.000</td>
<td>0.0000249</td>
<td>9.612</td>
</tr>
<tr>
<td>Top-K F</td>
<td>115.000</td>
<td>0.0012760</td>
<td>12.748</td>
</tr>
</tbody>
</table>

minimum loss achieved by the generator $G$, the step when the model achieves the lowest loss, and the loss of the discriminator $D$ at that moment.

From these results, it is possible to appreciate that the Top-K variant that achieves the best performance is the A version. This is because it achieves the lowest value for the generator’s loss while maximizing the discriminator’s loss, in an earlier training step than the rest of the variants. The Top-K A version corresponds to starting the Top-K training at the step 25.000 with an abrupt decay in the value of $K$ with $\gamma$ being 0.9999 (Table 3.1).

MCDs of the Top-K Variants Compared with the Baseline

In Table 4.2 we show a detailed comparison of the MCDs of the different proposed variants of the Top-K training and our baseline, the vanilla version of StarGAN-VC. The first two columns show the original and target speakers that correspond to the conversions. For this study we selected two random female speakers (SF1 and SF2) and two male speakers (SM1 and SM2), in order to test our proposed model training against all the possible conversions combinations. This way we are able to determine if our proposed variants perform better than the baseline in close and distant domain conversions. Being the first ones the ones happening between speakers of the same gender, and the latter ones, the conversions between different genders. On the last row of the table its possible to see the average of the conversions between all the speakers of each model.

We observe from Table 4.2 that all the Top-K variants obtain lower MCDs results compared to the baseline. This means that the conversions of the Top-K variants are closer to the real speech of the target speaker than the ones performed with the vanilla training of StarGAN-VC. Particularly, the best performing model on the most of speaker conversions and on average is the Top-K A variant, which corresponds to an early start of the Top-K training
Table 4.2: MCD Comparison of the Top-K Variants with the Baseline on the VCTK Corpus Dataset.

<table>
<thead>
<tr>
<th>Speakers</th>
<th>Baseline</th>
<th>Top-K A</th>
<th>Top-K B</th>
<th>Top-K C</th>
<th>Top-K D</th>
<th>Top-K E</th>
<th>Top-K F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Target</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF1</td>
<td>SM1</td>
<td>4.05 ± 0.16</td>
<td>4.01 ± 0.21</td>
<td>4.03 ± 0.19</td>
<td>4.04 ± 0.17</td>
<td>4.02 ± 0.19</td>
<td>4.03 ± 0.18</td>
</tr>
<tr>
<td>SF2</td>
<td>SM1</td>
<td>3.97 ± 0.16</td>
<td>3.95 ± 0.17</td>
<td>3.95 ± 0.18</td>
<td>3.96 ± 0.17</td>
<td>3.95 ± 0.17</td>
<td>3.95 ± 0.19</td>
</tr>
<tr>
<td>SF1</td>
<td>SM2</td>
<td>4.0 ± 0.16</td>
<td>3.98 ± 0.18</td>
<td>3.98 ± 0.15</td>
<td>4.0 ± 0.14</td>
<td>3.98 ± 0.16</td>
<td>4.0 ± 0.15</td>
</tr>
<tr>
<td>SF2</td>
<td>SM2</td>
<td>3.99 ± 0.11</td>
<td>3.98 ± 0.12</td>
<td>3.98 ± 0.1</td>
<td>3.99 ± 0.1</td>
<td>3.99 ± 0.11</td>
<td>3.99 ± 0.1</td>
</tr>
<tr>
<td>SM1</td>
<td>SF1</td>
<td>3.88 ± 0.15</td>
<td>3.85 ± 0.16</td>
<td>3.86 ± 0.17</td>
<td>3.87 ± 0.17</td>
<td>3.86 ± 0.15</td>
<td>3.85 ± 0.16</td>
</tr>
<tr>
<td>SM2</td>
<td>SF1</td>
<td>3.96 ± 0.14</td>
<td>3.92 ± 0.13</td>
<td>3.92 ± 0.13</td>
<td>3.95 ± 0.15</td>
<td>3.93 ± 0.11</td>
<td>3.94 ± 0.11</td>
</tr>
<tr>
<td>SF2</td>
<td>SF1</td>
<td>3.91 ± 0.13</td>
<td>3.91 ± 0.13</td>
<td>3.9 ± 0.16</td>
<td>3.9 ± 0.13</td>
<td>3.91 ± 0.14</td>
<td>3.89 ± 0.16</td>
</tr>
<tr>
<td>SM1</td>
<td>SM1</td>
<td>3.96 ± 0.17</td>
<td>3.94 ± 0.19</td>
<td>3.95 ± 0.19</td>
<td>3.95 ± 0.17</td>
<td>3.95 ± 0.16</td>
<td>3.94 ± 0.19</td>
</tr>
<tr>
<td>SM2</td>
<td>SM1</td>
<td>3.92 ± 0.18</td>
<td>3.9 ± 0.19</td>
<td>3.9 ± 0.19</td>
<td>3.93 ± 0.19</td>
<td>3.92 ± 0.17</td>
<td>3.91 ± 0.21</td>
</tr>
<tr>
<td>SF2</td>
<td>SM1</td>
<td>4.02 ± 0.22</td>
<td>4.02 ± 0.24</td>
<td>4.01 ± 0.24</td>
<td>4.04 ± 0.23</td>
<td>4.03 ± 0.24</td>
<td>4.03 ± 0.23</td>
</tr>
<tr>
<td>SM2</td>
<td>SF2</td>
<td>3.88 ± 0.18</td>
<td>3.85 ± 0.21</td>
<td>3.86 ± 0.18</td>
<td>3.88 ± 0.17</td>
<td>3.86 ± 0.19</td>
<td>3.88 ± 0.19</td>
</tr>
<tr>
<td>All pairs</td>
<td></td>
<td>3.954 ± 0.158</td>
<td>3.934 ± 0.177</td>
<td>3.935 ± 0.17</td>
<td>3.951 ± 0.162</td>
<td>3.941 ± 0.162</td>
<td>3.941 ± 1.865</td>
</tr>
</tbody>
</table>

(Step 25,000) with an abrupt decay in the value of $K$ with $\gamma$ being 0.9999.

4.1.3 Convergence Study of the Top-K Methodology

One of the objectives of applying the Top-K methodology to the training of StarGAN-VC was not only that the conversions achieved a higher level of realness, which has already been quantitatively demonstrated on Table 4.2, but also that the model converged earlier, and that kept the training stable, solving the instability limitation of the GANs training.

In Table 4.1 we have shown how the Top-K A variant is capable of minimizing the generator’s loss and maximizing the discriminator’s loss by the step 59,000, showing a much quicker convergence than the vanilla version. However, we wanted to go one step further and that was the reason of performing the convergence study. Once we have determined that the Top-K variant was the overall best performing combination of Top-K parameters based on the MCDs shown in Table 4.2, on Table 4.3 we show a comparison of the MCDs of the conversions obtained using the vanilla StarGAN-VC and the Top-K approach, computed in different steps of the training.

In Table 4.3, if we take a look on the closer domains columns (SF1-SF2) and (SM1-SM2), we observe that at the beginning of the training (steps 10,000 and 25,000) the MCDs of the baseline and Top-K A are practically the same. This is coherent, as at this point we have not started applying the Top-K methodology yet. However, it can be seen that from the step 40,000 onwards, while the Vanilla version resists to converge, our proposed approach is already performing considerably better, as we started applying the Top-K training from...
Table 4.3: Convergence Study Results Comparing the MCDs at Critical Steps Between Vanilla and the Best Top-K Approach.

<table>
<thead>
<tr>
<th>Step</th>
<th>SF1-SF2</th>
<th>SF1-SM1</th>
<th>SM1-SM2</th>
<th>SM1-SF1</th>
<th>SF1-SF2</th>
<th>SF1-SM1</th>
<th>SM1-SM2</th>
<th>SM1-SF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.000</td>
<td>3.98 ± 0.16</td>
<td>4.04 ± 0.18</td>
<td>3.96 ± 0.13</td>
<td>3.99 ± 0.08</td>
<td>3.98 ± 0.15</td>
<td>4.05 ± 0.18</td>
<td>3.95 ± 0.14</td>
<td>4.0 ± 0.11</td>
</tr>
<tr>
<td>25.000</td>
<td>3.97 ± 0.18</td>
<td>4.04 ± 0.19</td>
<td>3.95 ± 0.16</td>
<td>4.0 ± 0.11</td>
<td>3.96 ± 0.18</td>
<td>4.03 ± 0.19</td>
<td>3.94 ± 0.16</td>
<td>3.99 ± 0.1</td>
</tr>
<tr>
<td>40.000</td>
<td>3.97 ± 0.16</td>
<td>4.05 ± 0.16</td>
<td>3.97 ± 0.15</td>
<td>3.99 ± 0.11</td>
<td>3.95 ± 0.19</td>
<td>4.03 ± 0.19</td>
<td>3.95 ± 0.15</td>
<td>3.99 ± 0.1</td>
</tr>
<tr>
<td>55.000</td>
<td>3.96 ± 0.18</td>
<td>4.03 ± 0.17</td>
<td>3.96 ± 0.12</td>
<td>3.99 ± 0.11</td>
<td>3.94 ± 0.17</td>
<td>4.01 ± 0.21</td>
<td>3.95 ± 0.15</td>
<td>3.98 ± 0.11</td>
</tr>
<tr>
<td>70.000</td>
<td>3.95 ± 0.19</td>
<td>4.03 ± 0.18</td>
<td>3.97 ± 0.13</td>
<td>4.0 ± 0.09</td>
<td>3.94 ± 0.17</td>
<td>4.02 ± 0.18</td>
<td>3.96 ± 0.14</td>
<td>3.98 ± 0.1</td>
</tr>
<tr>
<td>85.000</td>
<td>3.96 ± 0.19</td>
<td>4.04 ± 0.18</td>
<td>3.97 ± 0.14</td>
<td>4.0 ± 0.1</td>
<td>3.95 ± 0.18</td>
<td>4.03 ± 0.17</td>
<td>3.96 ± 0.14</td>
<td>3.99 ± 0.09</td>
</tr>
<tr>
<td>100.000</td>
<td>3.97 ± 0.16</td>
<td>4.05 ± 0.16</td>
<td>3.96 ± 0.14</td>
<td>3.99 ± 0.11</td>
<td>3.94 ± 0.18</td>
<td>4.03 ± 0.17</td>
<td>3.95 ± 0.14</td>
<td>3.98 ± 0.1</td>
</tr>
</tbody>
</table>

the step 25.000. In fact, in the step 40.000 of the SF1-SF2 conversion, the Top-K A variant already reaches a level of quality that the vanilla version does not obtain until the step 70.000. Also, in step 55.000 Top-K A already achieves the lowest MCD (3.94) among the whole comparison, that is a level of quality never reached by the vanilla version. This is because, thanks to the fact of keeping reducing the value of $K$, we are selecting only a reduced number with the best samples from the generator.

Moreover, if we observe the conversions between different genders (SF1-SM1 and SM1-SF1) which correspond to statistically more distant domains, we see that the same pattern repeats. Until the step 25.000 the MCD values are practically the same from the baseline and with our proposed training. However, from the step 40.000 onwards it can be seen that the MCD values of the Top-K version, are always lower and reach their lowest value considerably earlier than with the vanilla training of StarGAN-VC.

This phenomenon can also be observed in the loss function of the generator using the Top-K A variant (Figure 4.1a), where we started applying Top-K from step 25.000. We observe that from the step 25.000 of the generator onwards, the loss quickly reduces to values close to zero and remains stable without signs of overfitting like in the vanilla version (Figure 3.2a). This clearly confirms our hypothesis that by applying the Top-K methodology we obtain a quicker convergence of the model and a more stable training.

4.2 Qualitative or Subjective Evaluation

In this section, we describe the performed subjective evaluation with the objective of answering the qualitative research questions introduced in Section 1.2. This study with participants was performed in order to support the promising results from the quantitative evaluation, because of the limitations
of the MCD metrics and due to the inherent subjective perception of speech quality.

### 4.2.1 Subjective Listeners’ Evaluation of the Obtained Voice Conversions

We conducted subjective listening tests to compare the speech quality and naturalness of speech of the converted speech samples obtained with the proposed Top-K A methodology and with our baseline method StarGAN-VC. In addition we also performed the listening version of the convergence study explained in the previous section where we aim to confirm that not only the model converged earlier in terms of the MCD and the loss of the generator by applying the Top-K A methodology, but that this earlier and improved convergence is also significant perceptually for the participants.

With these objectives we created a web-based online survey using Google Forms, where participants were asked to listen to several audios containing voice utterances and to answer some questions related to their quality, naturalness and realness. In this sense, participants were requested to use earphones or headphones and to perform the tests in a quiet space. Participants were also presented during the study with definitions of some particular concepts regarding voice conversion such as the meaning of *utterance* or the definition of *speech naturalness* in order to familiarise them with the tasks that they were going to evaluate. The utterances that were used for the voice conversions presented in this qualitative evaluation were part of the CSTR VCTK English Multi-speaker Corpus [30].

In total, 25 listeners participated in the qualitative study study (14 male and 11 female), from which 19 had no previous experience with digital audio or speech synthesis; 3 were semi-professionals in this field with several years
of practice and skills confirmed; and 3 were experts or full-professionals in the field. All of them signed the informed consent in the beginning of the Google Form, understanding and agreeing with the conditions to participate in this study, allowing their answers to be used in the present and in future research related to this task.

Realness Study

The qualitative evaluation consisted in two parts. In the first one, participants were presented in each question with an audio of a cloned voice generated either using the baseline or with the best model from our proposed methodology. After listening to it they had to rate using a Likert scale [42] from one to five to how confident they were that the audio was produced by a human speaker, with one being unconfident and five being confident. This way we evaluated the mean opinion score (MOS) for each speech sample. This first part comprised four types of conversions (SF1-SF2, SM1-SM2, SM1-SF1, SF1-SM1) per model including intra-gender and cross-gender, for also being able to evaluate the performance of our models with conversions between the domain-distant conversions. The order in which the utterances were presented, was randomised in order to remove any kind of bias from the evaluation from the participants.

The obtained scores of this part are shown in Table 4.4. As the results show, in all the conversion types the average MOS is higher for the proposed model confirming the results obtained in the quantitative study with the MCD. This shows that on average, the listeners are more inclined to perceive the voice conversions more human or more real, when they are generated using the model trained with the Top-K A methodology than with the vanilla version of StarGAN-VC. This phenomena is especially interesting in the domain-distant conversions (SM1-SF1 and SF1-SF2) when the generated samples from our proposed model also outperform those from the baseline. Additionally, it is remarkable that overall there was not a major difference between the ratings from the experts and the non expert participants. Particularly, it is interesting to note that all of the expert participants consistently rated our model’s conversion as more real, and even one of them rated with the top realness score one of the conversions from our model.

Convergence Study

The second part of the subjective evaluation, consisted in an sort of A/B test were participants were presented with two cloned utterances, one using
Table 4.4: Results of the realness study performed in the qualitative study, comparing the baseline and the other using our proposed methodology, generated in different critical steps of the training, corresponding to the same as in the quantitative convergence study shown in Table 4.3. After listening to both of them, participants would have to determine which of the two samples sounded more natural to them, or if none of them sounded natural. This third option, was added due to the conversions from early steps where most likely none of the models that we are comparing are able to generate intelligible and natural sounding cloned voices. In this case only one conversion (F1-F2) was chosen in order not to unnecessarily lengthen the duration of the evaluation, and the same utterance was presented converted in different steps with both models, in order for the participants to have a reference for the evaluation.

The obtained scores of this part are shown in Table 4.5. We can observe than from the step 40.000, the vast majority of the participants perceive the voices generated with the Top-K A methodology as sounding more natural and therefore with a higher quality than those generated using the baseline model. Similarly than in the first part of the study, there were no major differences between the experts and the non experts when selecting the most natural model. These results also confirm the ones from the quantitative convergence study (Table 4.3), showing that our model is capable of obtaining better quality results earlier than the baseline. In this sense, the results obtained in step 25.000 are normal, as at this point the Top-K training did not start (the Top-K A variant starts being applied from this step onwards) and the generator is not trained enough to consistently generate intelligible and natural sounding cloned voices, however, once we start selecting only the best samples for the training (from the step 25.000, see Table 3.1) then perceived naturalness is always higher with our model. Moreover, the samples from the step 100.000

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Conversion Type</th>
<th>Model</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SF1-SF2</td>
<td>Vanilla</td>
<td>1.92 ± 1.15</td>
</tr>
<tr>
<td>1</td>
<td>SF1-SF2</td>
<td>Top-K A</td>
<td>2.6 ± 1.11</td>
</tr>
<tr>
<td>2</td>
<td>SM1-SM2</td>
<td>Vanilla</td>
<td>2.84 ± 1.41</td>
</tr>
<tr>
<td>2</td>
<td>SM1-SM2</td>
<td>Top-K A</td>
<td>2.92 ± 1.25</td>
</tr>
<tr>
<td>3</td>
<td>SM1-SF1</td>
<td>Vanilla</td>
<td>3.44 ± 1</td>
</tr>
<tr>
<td>3</td>
<td>SM1-SF1</td>
<td>Top-K A</td>
<td>3.76 ± 0.77</td>
</tr>
<tr>
<td>4</td>
<td>SF1-SM1</td>
<td>Vanilla</td>
<td>2.72 ± 1.17</td>
</tr>
<tr>
<td>4</td>
<td>SF1-SM1</td>
<td>Top-K A</td>
<td>2.8 ± 1.11</td>
</tr>
</tbody>
</table>
Table 4.5: Results of the Convergence Study Comparing the Perceived Speech Naturalness at Critical Steps Between Vanilla and the Best Top-K Approach by the Participants in the Qualitative study.

<table>
<thead>
<tr>
<th>Step</th>
<th>Vanilla</th>
<th>Top-K A</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.000</td>
<td>32%</td>
<td>24%</td>
<td>44%</td>
</tr>
<tr>
<td>40.000</td>
<td>12%</td>
<td>76%</td>
<td>12%</td>
</tr>
<tr>
<td>55.000</td>
<td>24%</td>
<td>60%</td>
<td>16%</td>
</tr>
<tr>
<td>70.000</td>
<td>12%</td>
<td>72%</td>
<td>16%</td>
</tr>
<tr>
<td>85.000</td>
<td>12%</td>
<td>76%</td>
<td>12%</td>
</tr>
<tr>
<td>100.000</td>
<td>36%</td>
<td>52%</td>
<td>12%</td>
</tr>
</tbody>
</table>

correspond to the best model of the vanilla version, which means that at this point the baseline model starts generating more natural sounding voices, however, even in this case, participants perceived higher quality in the ones generated with the proposed Top-K approach.
Chapter 5

Discussion and Conclusions

In this chapter, first, we will summarise how our research questions were answered, together with a discussion and reflection on the obtained results from the experiments. Secondly, we will review the ethical considerations of speech synthesis and voice conversion technologies. Finally, the future work that could be done in relation to the research line followed in this dissertation will be presented.

5.1 Conclusions from the Research Questions

The main objective of the present work consisted in developing a more efficient way of generating high-quality voice conversions using deep learning. Mainly, we were interested in determining whether the application of a novel training technique for GANs, Top-K, would significantly improve the results and training stability of a state-of-the-art deep learning voice conversion system such as StarGAN-VC. For trying to answer the main research question from this study we carefully designed and implemented a quantitative and qualitative study. This decision was made due to the lack of a golden standard metric in the field of voice conversion for evaluating a voice conversion system quantitatively, and also due to the inherent subjectivity component of a person’s interpretation and perception of speech.

With the objective of answering the first three research subquestions in the quantitative study which are related to the moment of when to start applying the Top-K methodology, and how the value of $K$ should be reduced, we performed a study where we tried different values of $\gamma$, the decay of $K'$, and different steps for starting to reduced the value of $K$, as shown in Table 3.1. From the results obtained regarding the loss functions of the generator and
discriminator (Table 4.1), we can determine that by starting to apply the Top-K methodology earlier and with an abrupt reduction of the $K$’ samples, we achieve to minimise the generator’s loss and maximise the discriminator’s loss earlier than in the vanilla version. Also, we prove that by applying any kind of the proposed variants of Top-K we achieve significantly better results than by doing a regular StarGAN-VC training.

Regarding the fourth and fifth research subquestions belonging to the quantitative study, we performed the convergence evaluation of the Top-K methodology. In Table 4.3, we show that by using Top-K not only we obtain a quicker convergence of the StarGAN-VC model, but also, we obtain a better conversion quality that those obtained with a regular or vanilla training. Additionally, we demonstrate that even in one of the most challenging tasks for a voice conversion system which is the conversion between distant domains, the Top-K training methodology provides a quicker convergence to lower values of MCD and losses from the generator, and also a lower value of these.

Additionally, due to the limitations of the MCD metric and to the inherent subjective perception component of speech quality, with the objective of confirming the results obtained in the quantitative study, we performed a subjective study with human listeners.

With the objective of answering the first qualitative research question, about whether it was possible for a human listener to mistake a voice generated from an AI with the voice of a real human, in the first part of the qualitative evaluation we performed the Realness Study. From the results shown in Table 4.4, we show that especially when the samples corresponded to the ones generated from our proposed model Top-K A, participants were more inclined to believe that these were generated by a human speaker. Particularly, in the domain-distant conversion SM1-SF2 we observe that in general the most of them thought that audio corresponded to a human speaker. This way we show, that even for the challenging domain-distant conversions our proposed model outperforms the chosen baseline, even subjectively for the listeners. This also, answers our second research question, as on average participants were more fooled with the samples generated from our model than with the baseline.

For confirming the quantitative results of the Convergence Study, we performed an almost analogue version in the second part of the qualitative study. Here we presented the participants from samples generated at different times of the training of the generator with both the Top-K A variant and the vanilla version of StarGAN-VC. In Table 4.5, it is possible to observe that since the moment when we start applying the Top-K methodology (from step 25.000 onwards in the Top-K A variant) in the different critical steps of the training,
on average, participants perceived the samples generated by our proposed method as more natural than those from the baseline. This second part of the qualitative study shows that the samples generated using Top-K present higher quality for the listeners.

Overall, in this dissertation we have achieved to prove quantitatively and qualitatively that the Top-K methodology not only achieves a better performance earlier regarding the training objectives but also perceptually for human listeners. We have also successfully studied when and how the decay of this $K$ should be reduced, showing that an early and abrupt reduction of $K$ will converge earlier and will increase the training stability in a state-of-the-art voice conversion system as StarGAN-VC. Nevertheless, it is important to highlight that when we aim to introduce the Top-K methodology to any particular GAN architecture, as each model and dataset are different from each other, it is crucial to, first, run a training with the unchanged (Vanilla) parameters. This way, we will be able to observe the training curves for determining when is the Critical Point taking place, and therefore, when is the most optimal moment for applying the Top-K methodology during the training of a GAN.

5.2 Ethical Considerations and Implications of Speech Synthesis and Voice Conversion

The ethical use of voice technologies, such as speech and voice recognition or voice cloning and conversion, is becoming more important every day. Devices such as smart speakers, smartphones or smartwatches collect massive amounts of data from users thanks to the wide range of activities they allow (e.g. asking questions, setting reminders, checking bank accounts, accessing calendars, etc.). This data is often personal or private by nature. Therefore companies offering services through these gadgets and researchers developing them now have to assure not only a legal processing of user’s data but also an ethical one.

"The voice is a deep reflection of character... voice is the fingerprint of the soul." These were the words uttered by Daniel Day-Lewis when describing the magnitude of the challenge of preparing to play Abraham Lincoln in Steven Spielberg’s movie in 2012. Our voices are arguably the most precious tools we have to assert agency in our lives. There’s a sacredness with which humans have treated their voices such that any improper attribution of an utterance concerns us. A good example could be that we never want to "put words in others’ mouths". In this sense, our historical appreciation
of voice as our ultimate way of expression, can explain the shock that we suffer when technology appears to disrupt the "natural order". This is why when we see deep fakes, partly because of the uncanny valley effect, but also because we have an intellectual appreciation for the harm that synthetic media could unleash, that we feel an instinctive and automatic concern towards them. However, the use of voice conversion technology, similar to every AI application, carries a considerable potential risk and ethical implications but also holds a significant societal, personal, and economic potential for good.

One of the most important speech synthesis application is providing a voice to those who cannot speak or suffer from a severe vocal impairment. Those born with cerebral paralysis or who foster a condition like ALS or Parkinson’s disease can benefit enormously from a synthetic voice. Simplifying this cycle yet with high quality is the focal point of initiatives like the Voice Preservation Clinic from VocaliD [43]. However, in this case accuracy is sometimes not the ultimate goal, as many users want these voices to be a bit different from the ones that they had previously (younger, older, or accentuated on one dimension or another).

Due to the rise in volume of written information generated in the internet, there is an increasing demand for voice overs. This is aimed for people who are unable to read visually, but also for sighted people on the go, driving or listening to headphones. Producing high-quality voice-overs of all that content in a timely fashion is a tall task that synthetic voice can help address.

A population group that could be presumably at risk from the emergence of speech synthesis technologies are voice actors. One could think that with high-quality enough voice conversions and synthesisers systems their jobs would be replaced. However, these technologies, more specifically voice conversion could be used to scale and complement their voice acting and dubbing work. For example, partnering with companies that would use their voice as part of the training of these systems and sharing profits with them could be an attractive passive income stream for this group. Also, in some particular voice acting jobs such as gaming may require loud screams which can be harmful for their vocal cords, then using voice conversion for these kinds of situations might be beneficial also for them, in order to help them preserve a long-lasting career.

The downside to the human voice option is always about flexibility. If we change or add content, we need the talent to record it, which takes time and money. This adds up quickly for games or films. For studios and actors alike, synthetic voices are a useful fallback option for edits, rather than a whole-scale substitute for voice acting work. However, the creative potential of purely
synthetic voices is almost none, and that is where voice conversion has its biggest potential. If voice actors could engage their linguistic utterances in terms of tone or emphasis with voices from historical characters, for instance, this could represent a huge opportunity for more quality media content and a collaboration between humans and AI. This can be achieved through voice conversion technologies, which differently from pure speech synthesis from text, it still needs the voices from the original speakers.

However, the threats and risks that speech synthesis and voice conversion technologies pose to societies must also be considered. For example, in 2018 and 2019, in the press and news, deepfakes were analysed as only considering video or image deepfakes. However, deepfakes can be created not only for videos or images, but this possibility also exists for audio recordings too. The combination of both can potentially create a complete imitation of a person, which in the case of politicians, could potentially lead to a very dangerous manipulation of the society. In addition, with the ability to imitate voices, convincing phone phishing attacks can be carried out against companies. Furthermore, the basic existence of the technology gives people the ability to reject video or audio evidence as fake, whether that evidence is true or false [44].

As a society, in the past we have not been adequately anticipating the potential harmful consequences of technological innovation. Now we have the responsibility of not allowing the evolution of technology to outpace our anticipation to its challenge, specially with the advances in generative deep learning. Two good recent initiatives in this line are the AITHOS Coalition [45], which has introduced a guide to ethics in synthetic media [46], that hopes to imbue the industry with “mindful technology,” and the Open Voice Network [47] which is advocating for ethical guidelines for voice synthesis as part of its broader agenda.

In this sense, technology companies will also need to find reliable methods of flagging, disclaiming, or removing content that is synthesised, and whether it has been created and posted within rights. Tech companies could also monetise this technology through official, “verified voices,” a digital watermark of synthetic veracity [48].

Particularly in our research and due to the inherent architecture of the used model, StarGAN-VC, the fact that we train a generator of fake speech features against a discriminator of real or fake speech samples, gives us the opportunity to, at the same time that we are training the generator to generate higher quality samples, we are simultaneously training the discriminator to recognise whether a given fake sample corresponds to a real voice uttered by
a human or to a fake generated voice generated by the AI.

5.3 Future Work

There is still work that could be done following the Top-K research line. If we focus on the deep learning side of this thesis, one could perform a study trying to determine which is the most optimal percentage of the batch size \( v \) for being the lower limit of the reduction in the value of \( K \), as in [11] Sinha et al. just mention that it should not end up being one. For the presented study, the half of the batch size was chosen as the lowest value to which \( K \) could be reduced to.

The development of generative deep learning models keeps growing fastly as we write these lines, and in the month of July 2021 the voice conversion version of StarGANv2 [22], StarGANv2-VC [49] was published by Li et al. with auspicious results. Therefore it could be relevant to study if the Top-K methodology could also improve the obtained results in this recent study.

Additionally, it would be interesting to test how our developed model with the Top-K A variant would behave when trained and tested in a cross-lingual dataset with the objective of obtaining higher quality voice conversions between speakers from different languages.
References


[38] “Ray tune,” https://docs.ray.io/en/ray-0.4.0/tune.html.


