Exploring the limits of neural voice cloning: A case study on two well-known personalities

Ander González-Docasal 12, Aitor Álvarez 1, Haritz Arzelus 1

1 Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia–San Sebastián, Spain
2 University of Zaragoza, Department of Electronics, Engineering and Communications, Pedro Cerbuna 12, 50009 Zaragoza, Spain

{agonzalezd,aalvarez,harzelus}@vicomtech.org

Abstract

This work describes one successful and one failed Voice Cloning processes of two famous personalities in order to be broadcast in a high-impact podcast and in a Spanish public television program. Whilst a good quality synthesised voice could be generated for the first public figure, the second one was not adequate enough for its broadcast on television given its low speech quality. In this study, we explore the limits of the neural voice cloning considering the different conditions of the training material employed in each case and, based on several objective measures (amount of training data, phoneme coverage, SNR, MCD and PESQ), we analysed the main features to be considered for a high-quality synthetic voice generation. In addition, a webpage is provided in which samples of the resulting audios are available for each cloning model.

Index Terms: Voice Cloning, Neural Speech Synthesis, Tacotron-2, WaveGlow, Corpus Evaluation

1. Introduction

Voice cloning aims to generate synthetic utterances that are very similar to those from a given speaker, i.e. the cloned speaker. This practice has many interesting applications, particularly in the media domain, such as long-form reading (emails, webpages...), audiobooks, voiceovers or dubbing, among others [1].

Given the improvement in the quality of these systems thanks to the application of Deep Learning techniques, voice cloning is currently in high demand, mainly in applications related to deepfaking [2, 3]. As an example, the Salvador Dali Museum recently released an interactive piece of art featuring a deepfake of the artist [4]. Regarding only audio, the project AhoMyTTS [5] aims to generate a bank of synthetic voices to help both orally disabled people or those who lost their own voice. Similarly, the Speech-to-Speech Parrotor model [6] normalises atypical speech converting it to the voice of a canonical speaker with no speech disorders, enhancing its intelligibility. Another example of a successful voice cloning was the recovery of the voice of an ex-American football player diagnosed with ALS with the help of a database with his recordings [7], within the framework of the project Euphonia [8]. Given the risks of this technology at ethical and legal levels, the scientific community has developed counter-measures [9], released audio [10] and video [11] datasets, and organises more and more challenges [12, 13] to improve deepfake detection techniques.

Two methods are mainly used in the context of voice cloning: (1) Voice Conversion, which aims to convert input voice to match the voice of the target speaker and (2) Text-to-Speech (TTS), where audios are generated from an input text.

In this work, we describe the voice cloning processes of two well-known personalities to be broadcast in the media following the TTS approach, and explore the limits and possibilities of the technology depending on the quality of the training material. The first voice corresponded to the Spanish dictator Francisco Franco, chief of state from 1939 until his death in 1975. The development of his voice was contracted to Vicomtech by the Story Lab1 producer with the aim of being integrated in the XRey podcast. Both a letter and an interview of the dictator which were never spoken or recorded were synthesised. The podcast won The Ondas awards in 2020, and it is available in Spotify with an added special track2 in which the generation of the cloned voice is explained by the authors.

The second voice to clone corresponded to a famous Spanish cyclist, winner of multiple Grand Tours with international significance. Their cloned voice was contracted to Vicomtech as well, and it was required alongside a face-swap deepfake for a television prank3 in the show ‘Gala Inocente, Inocente 2021’. Although the training data were more recent than in the XRey project, the high variability and scarcity of the audios made this process much more challenging, so the cloned voice was not present in the program.

In addition, in this work we also analyse the issues that made the quality of the first model much higher, mainly focusing on the quality of the training material. To this end, we computed objective metrics to measure the quality of the data and the synthetic audios generated, following the conclusions of previous works [14, 15, 16] that establish some quality criteria related to acoustics (signal-to-noise ratio, SNR), high correlation between the audio and text pairs, a steady prosody along the multiple sentences and a good coverage of phones. Finally, we provide a webpage4 where audio samples of each cloning voice are provided for listening.

The remainder of this paper is organised as follows: Section 2 explains the compilation of each corpus, whilst in Section 3 the process of building the cloning models is described. Section 4 analyses and discusses about the quality of the training data and results, and Section 5 draws the main conclusions.

2. Compilation of the acoustic corpora

In order to clone the voice of any human being following the TTS approach, a dataset of their own voice has to be gathered.

1https://www.storylab.com/
2https://open.spotify.com/episode/0Vkoa3yyS998P2XkKNah9m2
3These particular use cases will be regulated by the EU AI Act: https://artificialintelligenceact.eu/
4https://vicomtech.github.io/voice-cloning/
Ideally, this dataset has to meet some quality conditions to ensure a good modelling of this data. First, these audios must be of the highest possible acoustic quality, given that the final TTS system will replicate the training audio material, including any noise, reverb or any other acoustic artefacts. Similarly, good audio coding formats are usually demanded, with low audio compression and optimal features of sample rate and bit depth. Finally, a clear and ordered diction is also expected from the input audios, with spontaneous speech being a more complicated type of data to transcribe and use to correctly train TTS systems.

Once the audio corpora are compiled, the next step includes the filtering and transcription of these audios. The filtering aims to remove the audio events that are unsuitable for training the TTS system. The transcriptions are necessary since the TTS systems are trained with pairs of audio and text, and this process is usually aided by an Automatic Speech Recognition (ASR) engine. In this work, both filtering and automatic transcriptions were performed by an ASR system based on the \texttt{nnet3} DNN setup of the Kaldi toolkit [17]. Particularly, we used a chain acoustic model based on a Convolutional Neural Networks (CNN) followed by a factorised time-delay network (TDNN-F), and a 3-gram language model for decoding. These models were trained with generic and media data following the CNN-TDNN-F setup explained in [18]. The raw transcriptions were then manually corrected and forced-aligned using the same acoustic model to generate the final pairs of audio and literal texts at sentence level. This alignment was performed in two steps: (1) first using a beam-size of $10^9$ and a retry-beam of $10^8$ to assure that the long audios were aligned with their corresponding text, and (2) using a smaller beam-size of 1 and a retry of 2 at sentence-level to discard non-literal transcriptions.

The corpora for the generation of the XRey and Inocente cloned voices were prepared following the above described process composed of 5 main steps: gathering, filtering, transcribing, post-editing and alignment. In the next subsections, a more detailed information is presented for each case.

2.1. XRey

The gathering of the dictator’s source data was performed from the Internet, mainly from contents published on Youtube and on the A la Carta web portal (nowadays, RTVE play\footnote{https://www.rtve.es/play/}) of the Spanish National Television (RTVE), summing up a total of 17 hours and 55 minutes. This source data were mainly composed by Christmas discourses and public politic speeches, ranging from 1936 to 1974. While the discourses included audios recorded indoor in front of a microphone of the time, the public speeches presented more challenging artefacts such as background noise, reverberation, emotional speech, low energy and poor presence of the principal voice.

During filtering, given the huge acoustic variability of the source data and considering the specific historical period and the formal speech style required for the Podcast, some of the Christmas discourses from 1955 to 1969, divided in three acoustically similar groups, were finally selected. After the transcription, post-editing and alignment processes, the final training material used to build the TTS model summed up a total of 3 hours and 12 minutes, as it is shown in Table 1.

### 2.2. Gala Inocente

The case of Inocente was completely different. Although the appearances of this figure are more recent than Franco’s, the nature of their public activities meant the available data covered a high variety of audio sources, such as television interviews, documentaries, phone-calls or voice messages, among others. These sources featured a high variety of audio formats and codecs, background noises or music, overlapping, spontaneous speech and voice qualities, which makes this content unsuitable for generating a high quality voice. Initially, a total of 3.77 hours of raw data were gathered.

The filtering process included discarding many parts of low acoustic quality, in addition to selecting manually only the segments with the desired speaker. After transcribing, post-editing and aligning these contents, the final corpus considered as suitable for voice cloning contained 29 minutes and 52 seconds from interviews and documentaries, as it is presented in Table 1. This final dataset was less than one sixth in duration compared to XRey.

<table>
<thead>
<tr>
<th>Content</th>
<th>t (s)</th>
<th>t (h:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X) 1955-1956</td>
<td>3604.62</td>
<td>1:00:05</td>
</tr>
<tr>
<td>(X) 1960-1961</td>
<td>3769.98</td>
<td>1:02:50</td>
</tr>
<tr>
<td>(X) 1968-1969</td>
<td>4202.26</td>
<td>1:10:02</td>
</tr>
<tr>
<td>(X) Total</td>
<td>11576.86</td>
<td>3:12:57</td>
</tr>
<tr>
<td>(I) Interview</td>
<td>622.95</td>
<td>0:10:23</td>
</tr>
<tr>
<td>(I) Documentary</td>
<td>1169.10</td>
<td>0:19:29</td>
</tr>
<tr>
<td>(I) Total</td>
<td>1792.05</td>
<td>0:29:52</td>
</tr>
</tbody>
</table>

**Table 1:** Final adaptation data for XRey (X) and Inocente (I)

### 3. Voice Cloning System

The voice cloning task was addressed using a TTS system based on the setup of Tacotron-2 [19]. This model consists of a sequence-to-sequence model, which includes an encoder, a decoder with attention and a post-processing final convolutional neural network (CNN). The encoder is fed with the embedding representation of the input characters, generated by a 1D-CNN+RNN prenet and trained simultaneously with the whole TTS system. Concerning the Vocoder, since the systems for each cloned personality were generated at different times, both Griffin-Lim [20] algorithm and WaveGlow neural network [21] were employed for the voices of XRey and Inocente respectively. For the same reason, the TTS models were built on the top of different toolkits\footnote{https://github.com/Rayhane-mamah/Tacotron-2 (Xrey) and https://github.com/NVIDIA/tacotron2 (Inocente)} and implementations.

#### 3.1. XRey

Given the in-domain data scarcity to generate a new Tacotron-2 based model from scratch (3 hours and 12 minutes only), an initial model was trained on a proprietary database designed and recorded in studio for TTS purposes, which contained 20 hours of high quality audios from a Spanish female speaker. This base model was trained with audios sampled at 16kHz, using 80 mel-spectrogram channels, 50 ms frame length, 25 ms frame shift and 1024-point Fourier transform. We employed Adam optimiser [22] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of $10^{-3}$ exponentially decaying to $10^{-5}$ starting after 50k iter-
iations. We also applied $L_2$ regularisation with a weight of $10^{-7}$ for a total of 150k training time steps with a batch-size of 64.

Afterwards, this initial model was fine-tuned with the in-domain data composed of the Christmas discourses described in Table 1 for 20k new time steps, using the same hyperparameters described above but with a batch-size of 12. However, mixing data from different decades (1955-1969) with notable perceptual differences in the voice and speaking style of an increasingly older Franco (died in 1975) resulted on an unstable intermediate model generating synthetic voices which resembled the heterogeneous conditions of the new training data. Therefore, this intermediate model was refined with the audios of the years that best fit the period of the Podcast (1960-1961), resulting in the final Tacotron-2 based voice cloning model.

Regarding the Vocoder, in this case we obtained better results with the Griffin-Lim algorithm than fine-tuning the Wavenet [23] neural model we previously trained on the 20 hours of the Spanish female speaker, probably due to the low quality and the small amount of audios of the in-domain data.

### 3.2. Gala Inocente

The challenge of generating a model capable of cloning the voice of this personality was much more difficult, due to the conditions of the training material (29 minutes and 52 seconds of low quality and heterogeneous data). Similarly to XRey, we used a base model generated from the same high quality acoustic database of the Spanish female voice. This base model was trained with audios sampled at 22050Hz, using 80 mel-spectrogram channels, 1024 samples of frame length, 256 samples of frame shift and 1024-point Fourier transform. We also employed Adam optimiser with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and an initial learning rate of $10^{-3}$. We applied $L_2$ regularisation with a weight of $10^{-6}$ for a total of 50k training time steps (due to the use of a different toolkit) with a batch-size of 64.

Different approaches were tested to generate a good enough fine-tuned model, evaluating many intermediate models and using different combination of data. The first approach with better results was focused on fine-tuning the base model using all the in-domain data composed of interviews and documentaries for 20k new steps and a batch-size of 32. With the aim of improving the results, two new models were trained on the two separate datasets, by fine-tuning the previous model for 10k steps and batch-size of 32. However, these last two models reached worse results than the previous one.

With regard to the Vocoder, Inocente’s cloned spectrograms were converted into the final audios through the flow-based WaveGlow vocoder. Given the poor quality of the in-domain data, we used a universal model available online.

### 4. Results and discussion

Although the linguistic characteristics that an optimal corpus for TTS should meet are difficult to measure quantitatively and are usually considered once the corpus is built [15, 24], the following subsections try to build a path to answer the question of what is needed to successfully clone the voice of a given personality by computing some objective metrics: quantity of the training data, phoneme coverage, SNR, Mel Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ), on both presented projects.

#### 4.1. Quantity of the training data

The quantity of training data is a key role on any machine learning task in general, and on TTS in particular. Even though high fidelity TTS system can be obtained with less than half an hour of audio per speaker using neural architectures [25], these models tend to use a multi-speaker approach, with the aid of corpora consisting of even higher amounts of audios than mono-speaker datasets. Although there are this type of resources available for languages such as English (e.g. LibriTTS [25] or VCTK [26]), none are known for Spanish. Furthermore, although transfer learning approaches have been successfully applied to fine-tune a pretrained Tacotron-2 model in one language with few data (2.5 hours) in other language [27], the final quality of the model strongly depends on the quality and homogeneity of these data.

In our case, the quantity and quality of the in-domain data could be definitive for the final resulting models within each project. While the more than 3 hours of the XRey project were enough to fine-tune a pretrained Spanish model, the high variability and scarcity (less than 30 minutes) of the Inocente’s data were insufficient to create a good final synthesis model.

#### 4.2. Phoneme coverage

As it was concluded in [28], if the training data include all the common graphemes, the adaptation data requirements can be significantly lowered. In our case, this aspect was specially relevant for Inocente’s voice, since few in-domain data were available to build de model. At phoneme and diphone levels, supposing a total of 26 Spanish phonemes, while every phoneme was correctly represented in our both datasets, 428 and 384 diphonemes were covered in the XRey and Inocente’s corpora respectively. Of course, not all of the $26^2 = 676$ diphone combinations were possible due to Spanish phonotactics. For comparison, the coverage of diphonemes on the Spanish Common Voice and the well-known English LJ Speech [29] datasets was of 593 in the former and 1217 diphonemes (supposing 37 English phonemes) in the latter. Moreover, although corpora in different languages are not fully comparable, every corpus displays a relatively similar distribution at diphone frequencies, as it is shown in Figure 1.

![Figure 1: Absolute frequencies of diphonemes in Spanish Common Voice (blue), XRey (red), Inocente (used data in yellow, all in green) and LJ Speech (burgundy) in logarithmic scale, with the x axis normalised to the maximum of diphonemes](https://commonvoice.mozilla.org/)

In addition to the amount of adaptation data, this distri-
bution of diphonemes served to correctly train XRey’s model. In contrast, it was not enough to build a good model for Inocente. In general, this failed model generated audios with high prosodic and acoustic variability, and low intelligibility in some specific correlations of input graphemes.

4.3. Signal-to-Noise Ratio

In general, the quantity and phonetic richness of the training data is not enough if the acoustic quality is not appropriate. We analysed this point by computing the SNR of the training audios following the method proposed in [30], assuming that the background noise is Gaussian, and that the amplitude distribution of clean speech can be approximated by the Gamma distribution with a shaping parameter of 0.4. Based on these assumptions, we estimated the SNR by examining the amplitude distribution of the noisy speech. In Figure 2, the distributions of the SNR values for both corpora can be found.

![Figure 2: Histograms and statistics of SNR for the audios of XRey, years 1955-1969 (left) and Inocente, gathered audios (right)](image)

As it can be observed in the histograms in Figure 2, the training audios of XRey maintained a relative stable distribution, whilst the ones gathered for Inocente displayed more variability in terms of SNR values. The corresponding table presents similar mean values of SNR in both corpora, but the standard deviation (Std) denotes the high variability of the Inocente’s audios, which implied a more exhaustive pruning that definitely affected the stability of the synthesis model.

4.4. MCD and PESQ

Once the training data were analysed, MCD and PESQ were also computed to quantitatively measure the quality of the synthesised audios. To this end, ground-truth-aligned spectrograms of the final real audios for each adaptation data were firstly generated by the feature prediction network following the approach shown in [19]. It allowed us to obtain spectrograms that were both predicted by the neural network and aligned with its corresponding original input audio. These spectrograms were then converted to waveforms using the vocoders corresponding to each TTS system in order to compute the MCD and PESQ metrics on the audio signals.

The computation of the MCD metric was performed over the Mel-Frequency Cepstral Coefficients (MFCCs) calculated with librosa\(^7\) and normalised on mean and variance per utterance. The PESQ metric was computed on the wide-band mode with the toolkit available at [31], which encompasses an audio quality assessment model standardised as the ITU-T recommendation P.862 [32]. The results of each metric can be seen in Figure 3.

![Figure 3: Histograms of MCD (left) and PESQ (right) values for XRey (blue) and Inocente (orange). Both distributions are normalised for comparison and the y axis lacks statistical meaning](image)

In this case, the MCD metric shows quite similar distributions, which suggests that both Tacotron-2 models have learned to generate phonetic information in similar ways, at least in terms of MFCCs. In contrast, when comparing the PESQ values obtained for each project, the results show that, as previously stated, the audios generated for XRey were of higher quality compared to those for Inocente. More specifically, the XRey audios achieved PESQ scores between 2 and 3, that correspond to acceptable values on a common Mean Opinion Score (MOS) scale from 1 (bad) to 5 (excellent). It is worth noting that the MOS metric would be the most accurate method to measure the quality of the synthesised audios. Nevertheless, since human evaluators are needed for its computation and considering the short period of time (few weeks) given by the producers to gather and process the training material in order to build the models and generate the final audios, this subjective evaluation finally corresponded to our clients who ultimately decided on their definitive broadcast in each case.

5. Conclusions

Voice cloning of historical or famous personalities often has to deal with the high difficulty of having low quality and scarce data to build a robust model. Given the nature of the current modelling techniques, this issue can make a project successful or unsuccessful. Since in this type of cloning it is often necessary to search for audio material from different sources with diverse audio qualities, this work aimed to be an example of the aspects that must be considered in order to achieve a good quality synthesis model. This study adds to previous works in the literature in establishing in 2-3 hours the amount of data that may be necessary to correctly fine-tune a Tacotron-2 based pre-trained model. These data should also cover all the phonemes of the language and the vast majority of diphonemes, and be of a homogeneous and appropriate acoustic quality with low variation in SNR. In addition, objective metrics such as PESQ provide added value to the process to check the quality of the synthetic audios of the intermediate models. In our case, PESQ scores of 2 or higher already indicated audio of good enough quality given what our clients expected considering the low quality of the adaptation data.

As future work, based on more demands we have about the generation of new voices of personalities, we will continue to establish specific metrics and criteria to measure quality in the synthesis process. Moreover, we will also explore other cloning techniques based on technologies such as voice conversion.

---

\(^7\)https://librosa.org
6. References


