applied sciences

MDPI

Article

Evaluation of Tacotron Based Synthesizers for Spanish and Basque

Víctor García *, Inma Hernáez and Eva Navas

HiTZ Basque Center for Language Technologies—Aholab, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain; Inma.hernaez@ehu.eus (I.H.); eva.navas@ehu.eus (E.N.)
* Correspondence: victor.garcia@ehu.eus

Abstract: In this paper, we describe the implementation and evaluation of Text to Speech synthesizers based on neural networks for Spanish and Basque. Several voices were built, all of them using a limited number of data. The system applies Tacotron 2 to compute mel-spectrograms from the input sequence, followed by WaveGlow as neural vocoder to obtain the audio signals from the spectrograms. The limited number of data used for training the models leads to synthesis errors in some sentences. To automatically detect those errors, we developed a new method that is able to find the sentences that have lost the alignment during the inference process. To mitigate the problem, we implemented a guided attention providing the system with the explicit duration of the phonemes. The resulting system was evaluated to assess its robustness, quality and naturalness both with objective and subjective measures. The results reveal the capacity of the system to produce good quality and natural audios.

Keywords: speech synthesis; robustness; text to speech; Spanish; Basque

Citation: García, V.; Hernáez, I.; Navas, E. Evaluation of Tacotron Based Synthesizers for Spanish and Basque. *Appl. Sci.* **2022**, *12*, 1686. https://doi.org/10.3390/app12031686

Academic Editors: Francesc Alías, Valentín Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira

Received: 27 December 2021
Accepted: 2 February 2022
Published: 7 February 2022

1. Introduction

The aim of Text to Speech (TTS) systems is to create synthetic speech from input written language. In recent years, very rapid progress has been made in this technology, improving the quality of the synthetic voices and increasing its use in consumer applications such as virtual agents, automated call assistance and many others. Traditionally, two approaches have been used to produce synthetic voices: unit selection (US) based concatenative synthesis [1,2] and statistical parametric speech synthesis (SPSS) based on Hidden Markov Models (HMM) [3,4]. These approaches are based on complex multi-stage pipelines and require from large domain expertise, impeding wider access to this technology.

Nowadays, Deep Neural Network (DNN)-based systems are state-of-the-art in speech synthesis [5,6]. DNNs provide TTS systems with the capacity to capture in a more efficient way the nonlinear complex relations between the voice acoustic parameters and the symbolic representation of speech. This improvement results in the generation of higher quality and more natural speech than the one obtained by traditional methods. Some examples of architectures based on DNNs are the Feed-Forward Networks (FF) [7], Recurrent Neural Network (RNN) [6] and the WaveNets [8].

Over the last few years, the implementation of neural networks has been used as an alternative approach to complex traditional multi-stage pipelines. Current systems are predominantly trained and designed in an end-to-end (E2E) fashion, meaning that audio is directly extracted from the input text without using separated models. E2E systems introduce three main advantages: (i) They prevent the propagation of errors through different components; (ii) they lower the need for feature engineering; and (iii) they reduce the requirement of manual annotations, minimizing costs.

Different neural network based E2E architectures have been proposed in the deployment of TTS synthesizers. Deep Voice [9] was the first proposed E2E system. The

quality obtained by this model was not as high as the quality obtained by previous models, therefore several improvements were developed in the subsequent versions Deep Voice 2 [10] and Deep Voice 3 [11]. Another example of a system built in an E2E fashion is Char2Wav [12], which was able to produce audio directly from text but revealed some artefacts in the speech signal.

All the TTS synthesizers implemented in this work were developed using Tacotron 2 [13] as a basis. Tacotron 2 is a system that addresses the sequence-to-sequence learning problem by means of an architecture inspired on the proposal by Sutskever et al. [14]. In Tacotron 2, a sequence of input character embeddings is mapped to a sequence of mel-scale spectrogram frames. To transform the spectrograms into waveforms, we used a neural vocoder named WaveGlow [15]. WaveGlow is a network that combines insights of Glow [16] and WaveNet [8] to produce high-quality audio at a fast speed thanks to its parallel processing capabilities.

Although E2E TTS systems show excellent results in terms of quality and naturalness of the generated synthetic speech, there are two main issues that must be addressed. Firstly, these systems require a sizeable set of audio and text pairs to be properly trained [17]. The strong dependence these system have on large amounts of data make them less accessible than traditional synthesis techniques, especially for low-resourced languages. Secondly, these models suffer from alignment issues, specially in long utterances. Alignment failures between the input text and the generated speech lead to unintelligible or incomplete signals.

The aim of this work is to evaluate the performance of Tacotron 2 based systems when trained on limited amounts of speech data. With this purpose in mind, we developed different TTS systems for Spanish and Basque. The main contribution of this research is the analysis of the performance of these systems on robustness, quality and naturalness. We evaluated the robustness of these systems against alignment errors by synthesizing sentences with a very different structure from the ones used in training. To alleviate the robustness issues, we used the guided attention mechanism proposed by Zhu et al. [18]. To automatically detect unintelligible synthetic speech, we propose a novel method that inspects the alignments searching for errors. This method was used to measure the impact of the changes we applied with the purpose of improving the robustness. The quality and naturalness of the synthetic voices were assessed by means of MOS evaluations and an objective measure proposed by Mittag and Möller [19].

The paper is organized as follows. Section 2 describes the data and methods we applied to build our TTS systems. Section 3 includes the description of the models used to create the synthetic voices. This section also details the method proposed to assess the robustness against alignment errors. The evaluation of the different aspects regarding the synthetic voices is presented in Section 4. Finally, some conclusions are drawn in Section 5.

2. Materials and Methods

In this section, we provide a description of the materials and frameworks used in the development of the work. For the purpose of training the models, datasets containing speech signals and their corresponding orthographic transcriptions are required. For the evaluation of the models, only the text is required. The TTS systems we built comprise two different parts: an acoustic model based on Tacotron 2 and a neural vocoder named WaveGlow. A description of both frameworks is provided at the end of the section.

2.1. Training Dataset

Every neural network based system completely depends on the data used to train it. For the TTS task it is necessary to have a speech corpus with its corresponding orthographic transcriptions. On the one hand, we used two phonetically balanced datasets of around 4100 utterances each, one in Basque and one in Spanish. These corpora had been recorded in the premises of our laboratory by one female and one male bilingual speakers. The datasets are composed by utterances with an average of 10.10 ± 2.81 and 12.84 ± 3.93 words per utterance in Basque and Spanish, respectively. The duration of the speech signals in each corpus amounts to approximately 4 h 30 min.

On the other hand, and as an extension to the work presented in [20], we included an additional Spanish dataset. This addition is intended to evaluate the performance of the system when trained on a larger corpus with different sentence length distributions. The corpus was provided by the Blizzard Challenge Organisation and corresponds to the dataset supplied to the participants of the Blizzard Challenge 2021 (https://www.synsig.org/index.php/Blizzard_Challenge_2021 (accessed on 20 December 2021)). The data consists of 5.25 h of high quality audio recordings by a single female speaker, including the corresponding 4920 orthographic transcriptions.

Table 1 shows a summary of the most relevant details regarding the structure of the training datasets used in this work. It can be seen that despite having less lexical variety, the utterances in the Spanish Blizzard corpus have a wider range of lengths than in the previous Spanish corpus. This detail can also be observed in Figure 1, where the distribution of the sentence lengths across all datasets is shown.

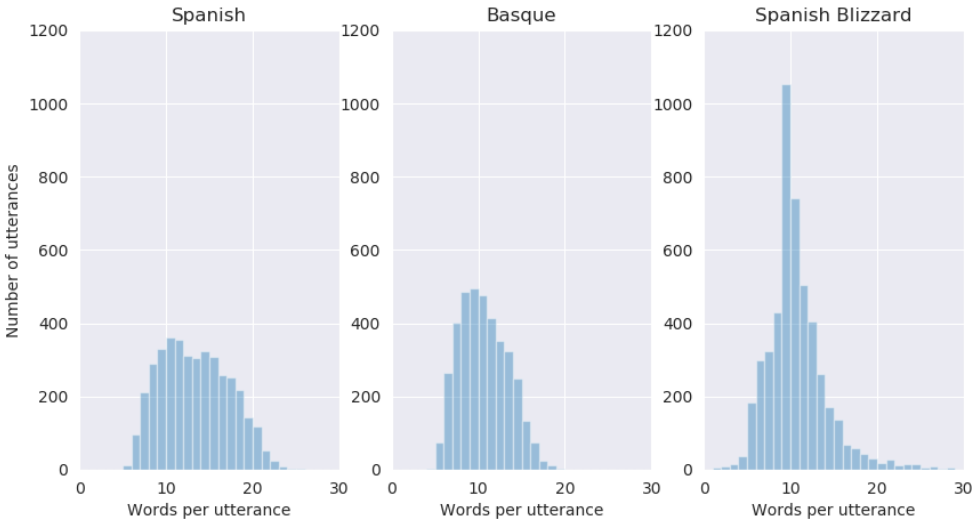


Figure 1. Distribution of words per utterance in the training dataset.

Table 1. Description of the training datasets.

	Spanish	Basque	Spanish Blizzard
Number of utterances	3993	3795	4920
Number of words	51,301	38,339	50,732
Unique words	10,687	12,352	8661
Words in shortest utterance	4	4	1
Words in longest utterance	26	19	50
Avg. words per utterance	12.84 ± 3.93	10.10 ± 2.81	10.30 ± 3.88

Before training the models, both the text and the speech signals were processed. The processing of the transcriptions was performed by means of a Spanish and Basque linguistic Front-End [21]. This Front-End is composed by two main modules: a text processor and a linguistic processor. The text processor normalizes the input text by expanding acronyms and converting numbers into directly readable words. The output of this module is directly fed into the linguistic processor, which returns the SAMPA [22] phone sequence along with the stress level of each phone.

Regarding the audio, tail silences were reduced to 150 ms and silences at the beginning of the utterances were completely taken out, as this process eases the learning of the alignment between text and audio. We decimated the original sampling frequency from 48 to 22.05 KHz to correspond the sample rate of a pre-trained model that would be later adapted using this corpus.

An additional processing step was performed to the Spanish Blizzard corpus. Inspection of the audio files revealed some misalignment issues between the pauses indicated by punctuation signs in the transcriptions and the actual pauses of the speech signals. This misalignment issue harms the training of the models, so we automatically re-positioned the pauses in the phoneme sequences using the actual silences in the audio signals as reference.

2.2. Testing Datasets

The deployed TTS systems were evaluated using two different written text datasets. The naturalness and quality evaluation of the synthetic voices was performed using in-domain sentences taken from novels and tales, similar to the sentences used in the training corpora. To assess the robustness of the systems, we used a dataset that contains sentences extracted from the Basque Parliament.

(a) Texts from tales and novels: This dataset contains utterances extracted from different books. It is a phonetically balanced dataset with 450 sentences both in Basque and Spanish. This is the corpus used to assess the naturalness and quality of the synthetic speech. The right column of Table 2 presents a summary of the text length related information in this corpus. As shown in the table, the text length in this dataset follows a similar distribution to the one used in the training corpus. This choice was intentionally made to prevent failures due to alignment issues in long utterances.

Table 2. Description of the testing datasets.

	Parliamentary Texts		**Texts from Novels and Tales**	
	Spanish	**Basque**	**Spanish**	**Basque**
Number of utterances	20,000	20,000	450	450
Number of words	425,467	389,252	3748	3064
Unique words	10,704	18,563	1378	1703
Words in shortest utterance	1	1	3	2
Words in longest utterance	268	133	15	14
Avg. words per utterance	19.46 ± 16.81	21.27 ± 16.55	8.33 ± 2.44	6.81 ± 8.75

(b) Parliamentary texts: This dataset was supplied by MintzAI project [23] (http://www.mintzai.eus/indice.html (accessed on 20 December 2021)). The sentences in the dataset are transcriptions from the Basque Parliament speeches, in both Basque and Spanish. This dataset is used to assess the robustness of the deployed models. From the entire dataset, 20,000 utterances were randomly selected. The left column of Table 2 shows the most relevant information of the size and lengths of this dataset. It can be seen that the lengths of the utterances contained in this corpus, with an average of 21.27 ± 16.55 and 19.46 ± 16.81 words per sentence in Basque and Spanish, greatly differ from those of the training dataset. In particular, the presence of very long utterances is an especially difficult scenario for Tacotron 2 model [24].

Figure 2 reveals the distribution of the number of words per utterance in each testing dataset. Compared to the distributions shown in Figure 1, a high resemblance can be found between the sentence length distribution of the training dataset and the corpora extracted from novels and tales. On the contrary, the length distribution of the sentences extracted from the parliamentary texts greatly differ from that of the training dataset. The motivation behind this choice is to assess the robustness of the systems in a more challenging environment.

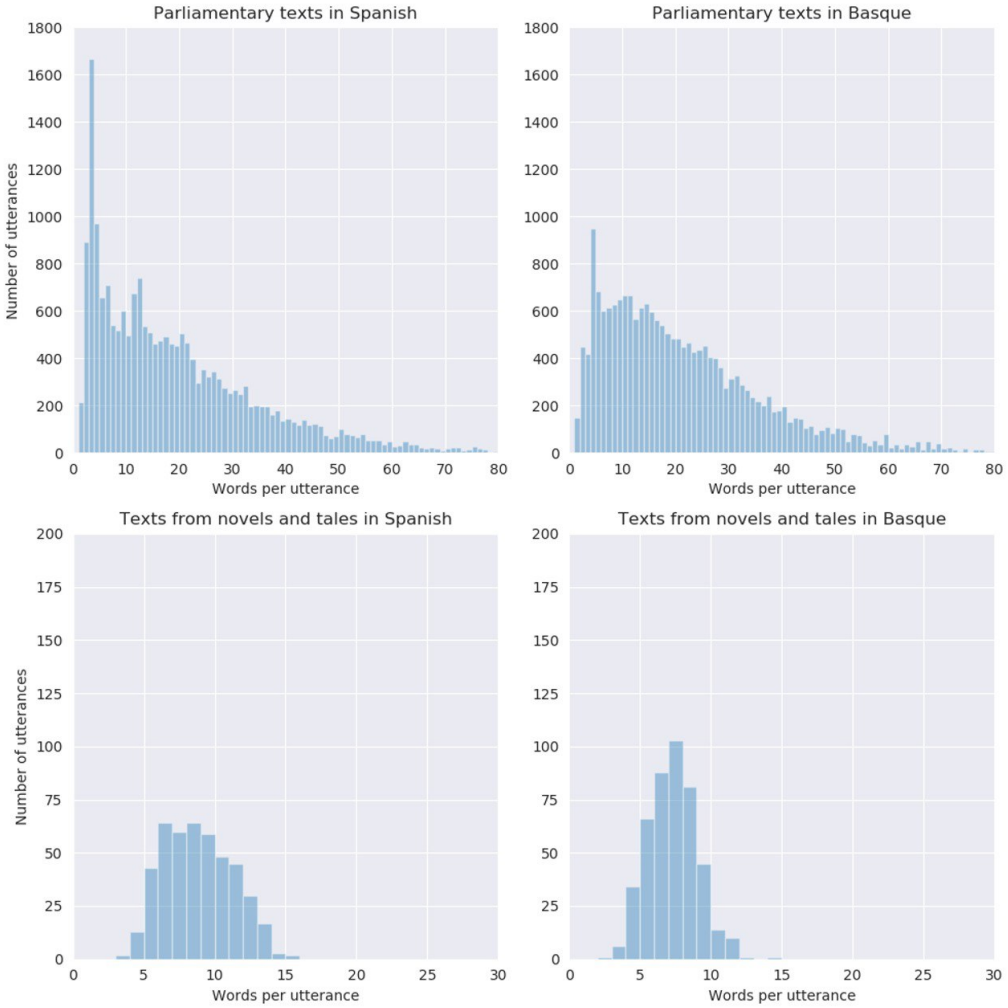


Figure 2. Distribution of words per utterance in the testing dataset.

These datasets were processed according to the same processing applied to the transcriptions of the training dataset.

2.3. Tacotron 2

To represent the relation between the audio signal and the input text, an acoustic model is required. In this work, we used Tacotron 2 as acoustic model. Tacotron 2 is a model that uses a sequence of characters as input and transforms it into a mel-scale spectrogram. Different open-source implementations of this model are publicly available. In this work, we used the open-source implementation provided by NVIDIA (https://github.com/NVIDIA/tacotron2 (accessed on 20 December 2021)) as the basis for our system. The architecture of this model is comprised of an encoder, an attention mechanism and a final decoder. To allow a better understanding on how this model works, a brief description of each module is provided:

- The encoder handles and transforms the input of the acoustic model. In the encoder the input sentence is converted into a hidden feature representation (that is later consumed by the decoder). First, each character of the input sequence is represented by a 512-dimensional vector using character embedding. The sequence of vectors is then fed into a three-convolutional-layer stack that models the long-term relationships between the input characters. Finally, using the output obtained from the last convolutional layer, a bidirectional Long short-term memory (LSTM) network generates the encoder hidden output (a vector of 512 $\times$ T, being T the length of the input sequence).

- The attention mechanism consumes the output of the encoder to produce a context vector at each decoding step. This attention mechanism is one of the most important parts of the model. It is at this point where the alignment between the input text and the frame-level acoustic features is learnt. The context vector provides the decoder with the necessary information to refer to the corresponding part of the encoder sequence at each decoding step. Tacotron 2 uses a custom location-sensitive attention mechanism [25], with an additional feature that is computed from the cumulative attention weights of the previous decoding steps.
- The decoder of Tacotron 2 is a recurrent neural network that predicts one frame at each decoding step in an auto-regressive fashion. During training, the decoder makes use of the context vector and the previous ground truth frame to compute the current step frame. This way of training the network is referred to as "teacher-forcing", and it is employed to ensure that each predicted frame is correctly aligned with the features of the target audio. During inference, as the ground truth frames are not available, the decoder uses the frame computed in the previous decoding step. The architecture of the decoder consists on a 2 layered pre-net, a 2 layer LSTM network and a convolutional post-net. The prediction from each decoding step is passed through the pre-net and then concatenated to the context vector. The resulting concatenation is fed into the two-layer LSTM. This output is again concatenated to the context vector and then passed through two different projection layers: one that predicts the stop token, and another one that predicts the target spectrogram frame. The final mel spectrogram is a combination of the whole spectrogram and a residual obtained in an ending convolutional post-net.

2.4. Waveglow

To reconstruct the audio waveforms from the mel spectrograms obtained in the acoustic model, we made use of a neural vocoder. Conventionally, WaveNet [8] has been used for this purpose as it produces almost human-like speech. Nevertheless, the auto-regressive nature of WaveNet makes the inference process extremely slow. For this reason, the WaveGlow generative network [15] was adopted in this work, as it provides audio quality close to WaveNet but with faster inference times. The publicly available implementation provided by NVIDIA was used in this work (https://github.com/NVIDIA/waveglow (accessed on 20 December 2021)).

The architecture used in the WaveGlow vocoder is mostly similar to the architecture proposed in the Glow generative model, implementing a series of "steps of flow". Each step of flow is comprised of an invertible 1×1 convolution followed by an affine coupling layer. During the training, the input samples are subjected to a series of invertible transformations until obtaining samples that follow a simple distribution, in this case a zero mean spherical Gaussian. This sequence of transformations is called "normalizing flow" [26].

In inference, WaveGlow generates audio by sampling from a zero mean spherical Gaussian distribution. This distribution needs to have the same number of dimensions as the target output. To reconstruct the waveform, the samples obtained from the zero mean spherical Gaussian distribution go through the steps of flow. These steps of flow perform the inverse transformations learnt during training, so that finally the target audio waveform with a complex distribution is achieved.

3. Methodology

In this section, we detail the implementation of the synthesizers and the strategies adopted to provide robustness to the systems. We also describe a new error detection post-processing stage used for evaluation of the system robustness. Finally, we report on the distinct neural vocoder models used in this research.

The final architecture of the system is shown in Figure 3.

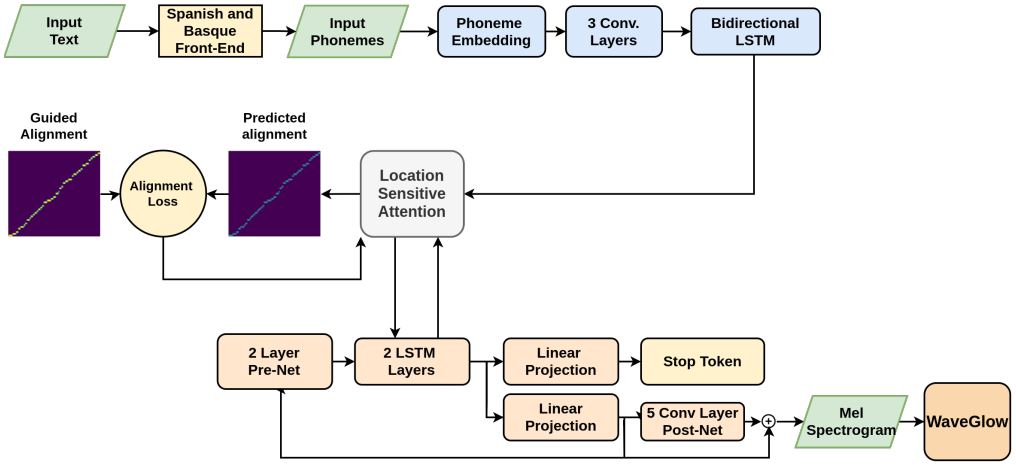


Figure 3. Architecture of the system trained using phonemes and pre-alignment guided attention.

3.1. Baseline

The first step in the development of this work was building a baseline model that could serve as reference for further evaluation. For this purpose, we trained the Tacotron model described in Section 2.3 using the data described in Section 2.1. This model remained unaltered with respect to the original open-source implementation, except for the inclusion of the specific characters existing in both Spanish and Basque (i.e., the stressed vowels and the characters 'Ñ' and 'Ü').

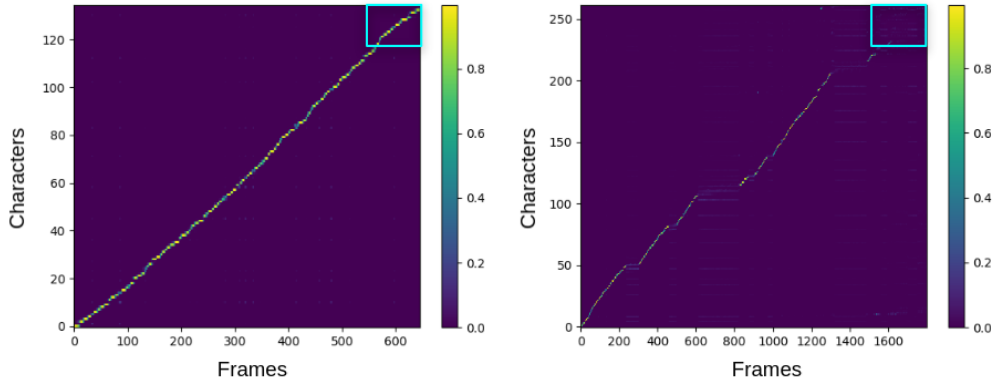
A very restrictive issue regarding Tacotron's training process is the requirement of large amounts of paired data. As specified by Chung et al. [17], the best speech quality is achieved when using between 10 and 40 h of recordings. Training the models with less than 10 h of data is feasible, but the signals obtained with this number of data present higher Mean Cepstral Distortion (MCD). As we lacked such amount of high quality paired data, we decided to apply a transfer learning approach over a pre-trained model that is publicly available. We used the model provided by NVIDIA (https://github.com/NVIDIA/tacotron2 (accessed on 20 December 2021)), which has been trained on LJSpeech corpus [27], an English dataset that contains approximately 24 h of high quality data recorded from a single female speaker. To fine-tune the model for a target language different from the source (i.e., English), the language dependant layer weights must be ignored. In this case, the pre-trained model character embedding weights were not loaded into the training. In total, five different models were trained, each one corresponding to each available training dataset. The training process of the five models was performed using a NVIDIA TITAN RTX GPU with a batch size of 64. To prevent over-fitting, the dropout of the attention module was set to 0.4 and dropout of the decoder was set to 0.1. We used a reduction factor of 1, denoting that a single frame is predicted at each decoding step. Learning was not changed through the whole process and stayed constant at a value of 0.001.

3.2. Error Detection Strategy

An informal listening test of the audio signals obtained with the baseline models revealed that at certain decoding steps the attention module was occasionally failing to address to the proper input character. Indeed, as stated in [18,24], auto-regressive attention-based systems are susceptible to alignment instability. Three different kind of errors can be produced as a consequence of this instability: a word in the sentence being repeated; a word being skipped; or in the worst cases the alignment being completely lost at some point in the sentence and producing no signal or babbling. This latter case is the most concerning one, because the generated speech becomes completely unintelligible.

Detecting these errors manually in a large number of synthetic files is a very time consuming task. To automatically detect them, we opted to design and append a post-processing stage at the end of the acoustic model. The function of this post-processing stage

is to scan the last decoding steps of the inference process in the alignment, ensuring that the attention weights assigned to the last characters of the sentence are higher than a threshold. Figure 4 displays the alignment matrix of two different sentences, with the vertical axis representing the input characters and the horizontal axis representing the output frames. A correct alignment is shown in the left image whereas the right image shows that the alignment has been lost midway. The rectangle area highlighted in the figure corresponds to the final decoding steps of the inference process. The algorithm proposed in this post-processing stage scans row-wisely the assigned weights in the selected area, looking for values exceeding a threshold. If no values exceeding the threshold are found, the algorithm concludes that the alignment has not reached the end of the sentence and therefore it is classified as a wrongly aligned sentence. In this work, the threshold was set to 0.3.



(**a**) (**b**)

Figure 4. Alignment of two different sentences and inspection area of the error detection algorithm: (**a**) correctly aligned sentence; (**b**) incorrectly aligned sentence.

3.3. Robustness Improvement

The attention instability during inference makes the models unreliable when synthesizing a large amount of utterances without any supervision. This robustness issue of Tacotron 2 has already been observed and investigated. Some works propose modifications over the current attention mechanism with the intention of reducing the alignment errors, as related in [28,29]. Other techniques propose the injection of prior knowledge into the system to improve the training process of the current attention mechanism [18,30]. In this work, we applied the latter, specifically we implemented a pre-alignment guided attention following the work described in [18]. This method improves the model stability while also enhances the training efficiency.

The principle of pre-alignment guided attention is introducing an explicit target in order to compute an additional loss term described by Equation (1).

$$Loss_{align}(A, \alpha) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T'} (A_{ij} - \alpha_{ij})^2 \quad (1)$$

where A_{ij} denotes the ground truth time aligned phoneme sequence and α_{ij} is the predicted alignment by the attention module.

This new loss eases the learning of the attention during the training process. The explicit target provided to the model is the ground truth time-aligned phoneme sequence. To create it, we used the Montreal Forced Aligner [31] and the phoneme sequence obtained with the linguistic front-end for Spanish and Basque described in Section 2.1.

Figure 3 shows the final architecture of the system. The input of the baseline system is now the phoneme sequence instead of the original character sequence. This architecture includes the additional loss term computed using the ground truth alignment of the phoneme sequence and the predicted alignment obtained in the attention model.

The training of the system was performed according to the same hyper-parameters used in the training of the baseline. In this case, no pre-trained models were used and the five voices were trained from scratch. These models are hereinafter referred to as Taco-PAG.

3.4. Neural Vocoder

As mentioned in Section 2.4, we used a WaveGlow neural vocoder to convert the mel-spectrograms into audio waveforms. We used an unaltered architecture of the network during the development of this research, but different models were fine-tuned and used.

With the baseline Tacotron described in Section 3.1, we used the pre-trained WaveGlow model accessible in [32]. This model was originally trained on the LJSpeech dataset [27], and we did not fine-tune it for its use along with the Tacotron baseline implementation.

The synthetic speech signals obtained with this baseline model present lower quality (due to the presence of a higher number of audible artefacts) for our male voices than for the female voices. Thus, to improve the quality of the male voices, we fine-tuned a second WaveGlow model using the previously mentioned pre-trained model with a total of 19 h of male voice. For this training, we used a batch size of 3 and a constant learning rate of 0.001. This WaveGlow model was used along with the Tacotron acoustic model described in Section 3.3, and it is referred to as WG-MA (i.e., WaveGlow-male adapted).

A third WaveGlow model was trained in the context of the Blizzard Challenge. To achieve the highest possible quality, we fine-tuned the model that can be accessed in [33] with the data provided by the Blizzard Challenge Organisation. Thus, this case uses a more advanced version of the public model provided by NVIDIA, adapted with 5.25 h of female voice. The same hyper-parameters as in the previous adaptation were used. This mode is referred to as WG-BA (i.e., WaveGlow-Blizzard adapted).

4. Evaluation

This section describes the evaluations carried out on the models. First, we performed an evaluation of the robustness of the implemented models. Then, a subjective evaluation of the naturalness and the quality of the synthetic speech signals was conducted. For this evaluation, we performed a Mean Opinion Score (MOS) test. In addition to the subjective evaluation, we also evaluated the naturalness of the synthetic signals with a deep learning based assessment algorithm using the models and the methodologies proposed in [19].

4.1. Robustness

To assess the robustness of the systems, we conducted an evaluation where 20,000 utterances were synthesized. These sentences were obtained from Corpus (b) described in Section 2.2. The systems used to synthesize the utterances were the baseline system described in Section 3.1 (denoted as Baseline) and the Tacotron with pre-alignment guided attention described in Section 3.3 (Taco-PAG). We used the error detection algorithm detailed in Section 3.2 in order to detect the speech signals with critical errors.

The relative improvement in terms of the number of generation errors from the baseline implementation to the Taco-PAG is shown in Table 3. As can be observed, there is a considerable improvement in all cases. A special result can be observed in the case of the Blizzard Spanish voice, where the initial number of errors is considerably lower than in the rest of the cases. We deem this good result to the fact that the Blizzard corpus contains longer sentences, more related in terms of length to those of the evaluation corpus. The additional processing in which the pauses of the Blizzard transcriptions were re-positioned could also have had a positive impact leading to this good result.

Considering that the baseline model is more prone to alignment failures, the following quality and naturalness evaluations were performed using the acoustic model with the pre-alignment guided attention (i.e., the Taco-PAG model).

Table 3. Number of utterances with errors and relative improvement in percentage.

	Baseline	Taco-PAG	Improvement
Female Spanish	1791	1103	38.41%
Female Basque	2596	1941	25.23%
Male Spanish	1077	95	91.18%
Male Basque	1206	274	71.31%
Blizzard Spanish	361	90	75.07%

4.2. Naturalness and Quality

In this section, the results of two independently developed MOS tests are presented. The first one was carried out for the voices of the two corpora recorded in the premises of our laboratory and was organized by our research group; the second one is part of the external evaluation campaign provided by the the Blizzard Challenge Organisation 2021 [34], where we participated with the synthesis system described in Section 3.3. In this latter case, we just present the results for reference.

In our test, listeners were asked to evaluate the naturalness and the quality for each utterance using a 1–5 point scale, being 1 the lowest score assigned to the signal and 5 the highest. For this evaluation, we used the testing dataset described in Section 2.2, Corpus (a). The DNN-based systems used to synthesize the sentences were composed of the Taco-PAG acoustic model along with the neural vocoder model WG-MA described in Section 3.4.

In addition to the assessment of the signals obtained with the DNN based systems, we also included signals from other two sources in the evaluation:

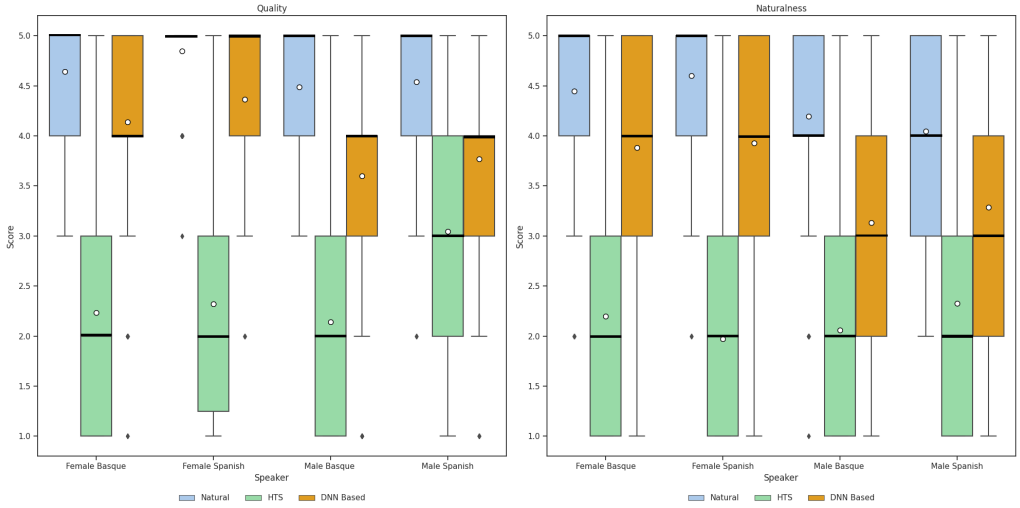
- Signals from the natural reference speech.
- Synthetic signals obtained using an HMM-based speech synthesis sytem (HTS) based TTS system for Spanish and Basque developed in our research group (https://sourceforge.net/projects/ahotts/ (accessed on 20 December 2021)) [21].

In total, 33 listeners took part in the evaluation. As the test includes sentences in both Basque and Spanish, all listeners were bilingual in both languages. From the 450 available utterances, each participant rated 6 random signals per language (2), speaker (2) and method (3), i.e., a total of 72 utterances.

Figure 5 shows the results obtained in this evaluation. The box-plot graph representing the assessment of the quality evaluation is shown in Figure 5a. The scores assigned to the signals obtained with the DNN-based approach are in all case higher than those assigned to the signals obtained with HTS-based synthesis. Nevertheless, the score remains lower than that of natural speech. A preference for the female voice is shown for the natural speech in both languages. This preference is also shown in the evaluation of the DNN based signals, where female speech obtained higher rates than male speech.

Figure 5b presents the box-plot graph representing the naturalness ratings. As in the quality evaluation, DNN-based systems were ranked above the HTS systems, but they received a lower score comparing with the natural speech signal. In the naturalness evaluation, participants also preferred female voices over male voices in natural speech signals, which is reflected in the scoring of the DNN based systems as well.

Different aspects of the speech signals were rated in the evaluation test performed by the Blizzard Challenge Organisation (i.e., intelligibility, similarity to to original speaker and naturalness). In this work, we only include the naturalness scores as a reference, but the complete evaluation can be accessed in [34]. In total, 12 systems participated in the challenge. The synthetic signals obtained with these systems were rated by 313 participants. Figure 6 shows the results of this evaluation. Our proposed system was labelled as letter 'E' and the reference natural speech was labelled as letter 'R'.



(**a**) (**b**)

Figure 5. Results of the quality and the naturalness assessments, white dots represent the mean: (**a**) quality assessment; (**b**) naturalness assessment.

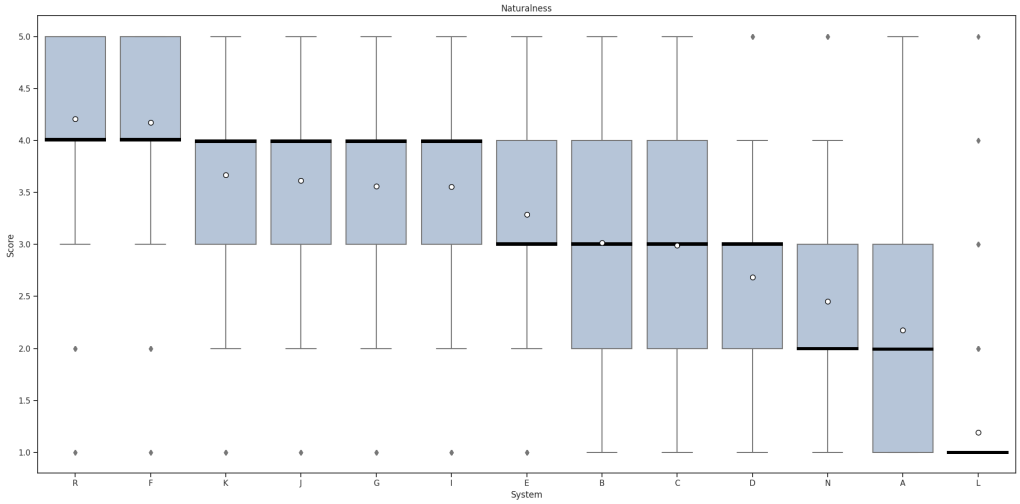


Figure 6. MOS evaluation of the naturalness performed by the Blizzard Challenge Organisation. White dots represent the mean.

4.3. NISQA

NISQA [19] model is a deep learning based estimator of speech naturalness. Results obtained with this model are in 1–5 scale, as in MOS test. According to the authors, NISQA model works language independently. In this evaluation, we used the sentences from the testing dataset described in Section 2.2, Corpus (a). These sentences were synthesized using the HTS-based system and the Taco-PAG acoustic model along with two different neural vocoder models, WG-MA and WG-BA. No HTS based model for the Blizzard Spanish voice was available by the time the evaluation was performed.

Table 4 shows the averaged scores achieved by the different systems using NISQA as evaluation model. In most cases, the DNN-based approach got better values of naturalness in comparison to the HTS based system. Even though all female voices obtained a higher score using the Taco-PAG + WG-BA, the male voices performed worse. Despite the large improvement achieved in the case of the female voices, the absence of male voices in the adaptation of the WG-BA vocoder model negatively affected the naturalness of the male synthetic speech.

Table 4. NISQA assessment scores with 95% confidence interval.

	HTS Based	Taco-PAG+WG-MA	Taco-PAG+WG-BA
Female Spanish	2.97 ± 0.05	3.34 ± 0.05	$\mathbf{4.20 \pm 0.04}$
Female Basque	3.07 ± 0.04	3.46 ± 0.07	$\mathbf{3.76 \pm 0.01}$
Male Spanish	3.62 ± 0.04	$\mathbf{3.64 \pm 0.08}$	3.58 ± 0.01
Male Basque	2.95 ± 0.04	$\mathbf{3.56 \pm 0.08}$	3.35 ± 0.01
Blizzard Spanish	-	3.63 ± 0.01	$\mathbf{4.38 \pm 0.03}$

5. Conclusions

Neural networks allow the generation of synthetic voices that show high resemblance to natural ones, but require large amounts of data to train them. In languages with difficult access to large amounts of data, it is important to get the most out of the available data. One of the limitations of neural networks trained with few data is the appearance of synthesis errors. We propose a novel algorithm to automatically assess these errors based on the alignment of the last decoding steps of the acoustic model. This algorithm allows evaluating the results of the baseline Tacotron 2 model and showing the improvement achieved when the guided attention is introduced. The results show that the implementation of the guided attention decreases the number of errors for all the developed voices.

In the MOS evaluation of quality and naturalness, female voices obtained better results than male voices for both languages and dimensions. This is the case even for the natural recordings, which implies that using recordings with high MOS values is important to get good subjective results in the synthetic voices. Therefore, it would be interesting to obtain the MOS values of several speakers to select the most suitable one prior to recording the database.

We also applied an objective evaluation measure that assesses the naturalness of synthetic signals, NISQA. The best results were obtained for all voices by the guided attention Tacotron 2 system combined with a WaveGlow vocoder adapted to the gender of the speaker.

Further work includes the evaluation of the effect that data augmentation techniques may have on the quality and naturalness of the generated synthetic voices. Another way of training with more data would be combining data from both languages and/or several speakers in the same model. Taking into account the phonetic resemblance between Spanish and Basque, mixing sentences from both languages may contribute to a more robust training of the common phonemes.

Author Contributions: Conceptualization, V.G., I.H. and E.N.; Data curation, V.G.; Formal analysis, V.G., I.H. and E.N.; Funding acquisition, I.H. and E.N.; Investigation, V.G., I.H. and E.N.; Methodology, V.G., I.H. and E.N.; Project administration, I.H. and E.N.; Software, V.G.; Supervision, I.H. and E.N.; Validation, I.H.; Visualization, V.G. and E.N.; Writing—original draft, V.G., I.H. and E.N.; Writing—review and editing, V.G., I.H. and E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Basque Government (Project refs. PIBA 2018-035, IT-1355-19). This work is part of the project Grant PID 2019-108040RB-C21 funded by MCIN/AEI/10.13039/501100011033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Basque and Spanish databases are accessible via the following link http://metashare.ilsp.gr:8080/repository/search/?q=ahosyn (accessed on 20 December 2021). For the Spanish Blizzard database, please contact the Blizzard Challenge Organisation.

Acknowledgments: We would like to thank the organizers of the Blizzard Challenge 2021 for providing the data and carrying out the evaluations. We would also like to express our gratitude to all the participants of the evaluation we performed.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Black, A.W.; Taylor, P. Automatically Clustering Similar Units for Unit Selection Speech Synthesis. In Proceedings of the EUROSPEECH 1997, Rhodes, Greece, 22–25 September 1997; pp. 601–604.
2. Campbell, N.; Black, A.W. Prosody and the Selection of Source Units for Concatenative Synthesis. In *Progress in Speech Synthesis*; Springer: New York, NY, USA, 1997; pp. 279–292.
3. Wu, Y.J.; Wang, R.H. Minimum generation error training for HMM-based speech synthesis. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP), Toulouse, France, 14–19 May 2006; Volume 1.
4. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [CrossRef]
5. Ze, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
6. Fan, Y.; Qian, Y.; Xie, F.L.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proceedings of the INTERSPEECH 2014, Singapore, 14–18 September 2014; pp. 1964–1968. [CrossRef]
7. Qian, Y.; Fan, Y.; Hu, W.; Soong, F.K. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3829–3833. [CrossRef]
8. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
9. Arık, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. Deep voice: Real-time neural text-to-speech. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 195–204.
10. Gibiansky, A.; Arik, S.Ö.; Diamos, G.F.; Miller, J.; Peng, K.; Ping, W.; Raiman, J.; Zhou, Y. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. In Proceedings of the Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.
11. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.O.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep voice 3: 2000-speaker neural text-to-speech. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018; pp. 214–217.
12. Sotelo, J.; Mehri, S.; Kumar, K.; Santos, J.F.; Kastner, K.; Courville, A.; Bengio, Y. Char2wav: End-to-end speech synthesis. In Proceedings of 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–6.
13. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
14. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
15. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3617–3621.
16. Kingma, D.P.; Dhariwal, P. Glow: Generative flow with invertible 1×1 convolutions. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 10215–10224.
17. Chung, Y.A.; Wang, Y.; Hsu, W.N.; Zhang, Y.; Skerry-Ryan, R.J. Semi-supervised Training for Improving Data Efficiency in End-to-end Speech Synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6940–6944.
18. Zhu, X.; Zhang, Y.; Yang, S.; Xue, L.; Xie, L. Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis. *IEEE Access* **2019**, *7*, 65955–65964. [CrossRef]
19. Mittag, G.; Möller, S. Deep Learning Based Assessment of Synthetic Speech Naturalness. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020; pp. 1748–1752.
20. Garcia, V.; Hernaez, I.; Navas, E. Implementation of neural network based synthesizers for Spanish and Basque. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 225–229. [CrossRef]
21. Erro, D.; Sainz, I.; Luengo, I.; Odriozola, I.; Sánchez, J.; Saratxaga, I.; Navas, E.; Hernáez, I. HMM-based speech synthesis in Basque language using HTS. In Proceedings of the FALA. RTTH, Vigo, Spain, 10–12 November 2010; pp. 67–70.
22. Wells, J.; Barry, W.; Grice, M.; Fourcin, A.; Gibbon, D. Standard Computer-Compatible Transcription. Esprit Project 2589 (SAM), Doc. no. SAM-UCL-037, 1992; Volume 37. Available online: https://www.phon.ucl.ac.uk/home/sampa/ (accessed on 20 December 2020).
23. Etchegoyhen, T.; Arzelus, H.; Gete, H.; Alvarez, A.; Hernaez, I.; Navas, E.; González-Docasal, A.; Osácar, J.; Benites, E.; Ellakuria, I.; et al. MINTZAI: Sistemas de Aprendizaje Profundo E2E para Traducción Automática del Habla MINTZAI: End-to-end Deep Learning for Speech Translation. *Soc. Española Para Proces. Del Leng. Nat.* **2020**, *65*, 97–100.
24. Ren, Y.; Qin, T.; Ruan, Y.; Zhao, S.; Liu, T.Y.; Tan, X.; Zhao, Z. FastSpeech: Fast, robust and controllable text to speech. *arXiv* **2019**, arXiv:cs.CL/1905.09263.

25. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. In Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2015; pp. 577–585.
26. Kobyzev, I.; Prince, S.; Brubaker, M. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3964–3979 [CrossRef] [PubMed]
27. Ito, K.; Johnson, L. The LJ Speech Dataset, v1.1. 2017. Available online: https://keithito.com/LJ-Speech-Dataset/ (accessed on 20 December 2020).
28. He, M.; Deng, Y.; He, L. Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS. In Proceedings of the INTERSPEECH 2019, ISCA, Graz, Austria, 15–19 September 2019; Volume 2019, pp. 1293–1297.
29. Battenberg, E.; Skerry-Ryan, R.J.; Mariooryad, S.; Stanton, D.; Kao, D.; Shannon, M.; Bagby, T. Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6194–6198.
30. Liu, R.; Sisman, B.; Li, J.; Bao, F.; Gao, G.; Li, H. Teacher-student training for robust tacotron-based tts. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6274–6278.
31. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In Proceedings of the INTERSPEECH 2017, ISCA, Stockholm, Sweden, 20–24 August 2017; pp. 498–502.
32. NVIDIA. 2020. Available online: https://ngc.nvidia.com/catalog/models/nvidia:waveglow_ljs_256channels (accessed on 10 March 2020).
33. NVIDIA. 2020. Available online: https://drive.google.com/file/d/1rpK8CzAAirq9sWZhe9nlfvxMF1dRgFbF/view (accessed on 12 January 2021).
34. Ling, Z.H.; Zhou, X.; King, S. The Blizzard Challenge 2021. In Proceedings of the Blizzard Challenge Workshop 2021, Online, 23 October 2021.