IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Implementation of neural network based synthesizers for Spanish and Basque

Víctor García Romillo, Inma Hernáez Rioja, Eva Navas

HiTZ Center - Aholab, University of the Basque Country (UPV/EHU)

victor.garcia@ehu.eus, inma.hernaez@ehu.eus, eva.navas@ehu.eus

Abstract

This paper describes the implementation of neural-network based Text-to-Speech (TTS) synthesizers for Spanish and Basque. In order to develop this research, the voices of one male and one female speakers, both bilinguals, are used in a data set of around 4 and a half hours for each voice and language. The system uses Tacotron to compute mel-spectrograms from the input text sequence and Waveglow to obtain the resulting audios.

Training the mentioned models with a limited amount of data leads to synthesis errors in some utterances, affecting the naturalness of the audios and even producing unintelligible speech. In this paper, we describe the method followed to automatically detect erroneously synthesized audios and the strategy followed to address the causes of the errors. The designed method has been validated by testing the TTSs using a large set of out-of-domain sentences. In the end a fully operational system is developed, with capacity to generate good quality and natural audios, as showcased by the evaluation conducted.

Index Terms: speech synthesis, robustness, text to speech, Basque, Spanish

1. Introduction

Text-to-Speech (TTS) systems transform input written language into synthetic speech. Traditionally, two main approaches have been used: unit selection (US) based concatenative synthesis and statistical parametric (SP) speech synthesis. The former uses big databases segmented into sub-word units, and attempts to select the best sequence of them to match the target sentence [1, 2]. The latter approach generates mathematical models that attempt to relate input text features with acoustic features [3, 4]. Hybrid approaches have been also proposed, with the intention of combining the segmental naturalness of US and the consistency offered by SP [5, 6].

Nowadays, deep neural network (DNN) based systems are state-of-the-art in speech synthesis [7, 8]. Neural networks have benefited the speech synthesis field by improving the quality and naturalness of the synthetic voices with respect to the traditional systems. Another contribution made by neural networks is the possibility of training and designing the systems in an end-to-end (E2E) fashion. While traditional multi-stage pipelines are complex and require from large domain expertise, E2E systems reduce the complexity by extracting the audio directly from the input text without needing separated models.

There are different neural network based E2E architectures for TTS [9, 10]. The TTS systems built in this work are based on Tacotron 2 [11]. Tacotron 2 is a sequence-to-sequence [12] architecture that maps character embeddings to mel-scale spectrograms. To transform the output spectrograms into waveforms, we use WaveGlow [13]. WaveGlow is a neural vocoder that combines insights of WaveNet [14] and Glow [15] to produce high-quality audio using parallel capabilities of GPUs.

Although end-to-end TTS systems have shown excellent results in terms of audio quality and naturalness, there are some issues to be faced. On the one hand, these systems usually suffer from low training efficiency, requiring a sizable set of text and audio pairs to train properly. On the other hand, synthesized speech is usually not robust, due to alignment failures between input text and speech during the generation.

In this paper we describe the implementation of four TTS systems, two for Spanish and two for Basque based on the previously mentioned architectures. To evaluate them, several out-of-context utterances were synthesized. We propose a method to automatically detect sentences where a poor alignment leads to unintelligible synthetic speech. We also describe the strategy followed to address the causes of the issues found during the evaluation of the initial implementation. Finally, we conduct an evaluation to check if the changes improve the robustness of the models while maintaining good quality and naturalness over the synthetic voices.

The paper is organized as follows. Section 2 describes the data used to train and evaluate all models. Section 3 contains a description of the construction of the systems, along with an analysis of their issues and the approaches taken to address them. Subjective and objective evaluations of the final implementation are shown in section 4. Finally, some conclusions are drawn.

2. Materials

In order to train the models, datasets containing speech signals and their corresponding transcriptions are needed. For evaluation purposes we only required the text. The following sections describe all the data used in this work and the processing applied to it.

2.1. Training dataset

The neural-network approach taken in the construction of the TTS system demands having available a speech corpus with its corresponding transcriptions. In this work we have used two phonetically balanced corpora of about 4000 sentences, one for Basque and one for Spanish, which have been recorded by one male and one female bilingual speakers. Figure 1 shows a distribution of the amount of words per utterance in the Spanish and Basque corpora. The datasets have an average of 12.84 ± 3.93 and 10.10 ± 2.81 words per sentence in Spanish and Basque. Each recorded corpus has a duration of approximately 4 hours and 30 minutes.

In order to obtain faster convergence times along with higher synthesis quality, speech signals and their corresponding transcriptions needed to be processed. Regarding the audio, silences at the beginning of the sentences were removed and silences at the end were trimmed to 150ms, as this process eases the learning of the alignment between text and audio. The original sampling frequency of 48000 Hz was decimated to

10.21437/IberSPEECH.2021-48

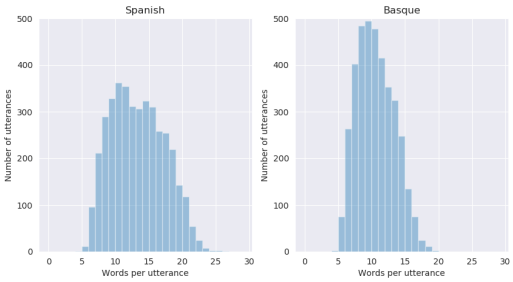

Figure 1: *Distribution of words per utterance in the Spanish and Basque corpora used for training*

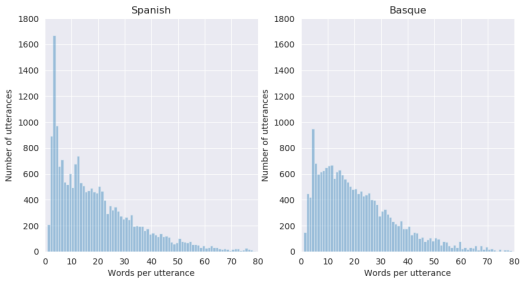

Figure 2: *Distribution of words per utterance in the Spanish and Basque corpus used for robustness evaluation*

22050 Hz to match the sample rate used in a pre-trained model that would be later adapted with the Basque and Spanish voices. Finally, recordings longer than 9 seconds were removed to avoid convergence issues due to memory restrictions from the available GPUs.

In relation to the transcriptions, the first step was standardizing the encoding of all texts to UTF-8. A linguistic Front-End in Spanish and Basque [16] was used to normalize and clean the utterances. The same Front-End was used to extract the phonetic transcriptions of all the utterances expressed in SAMPA alphabet [17].

2.2. Testing datasets

For the evaluation of the deployed TTS systems two different written text datasets were used. To evaluate the robustness of the systems a dataset containing out-of-domain sentences was used. The evaluation of the quality and naturalness of the synthetic voices was conducted using in-domain sentences.

a) Parliamentary texts: This corpus was provided by the MintzAI project [18] [1]. The utterances in the corpus were obtained from transcriptions of the Basque Parliament sessions, both in Basque and Spanish. Being transcriptions of spoken parliamentary speeches, the complexity of the sentences differ greatly from that of the sentences in the training dataset. Also, the utterances have varying lengths from a few words to very long sentences, a particularly difficult scenario for the Tacotron 2 model [19]. From the complete dataset we randomly selected 20000 sentences. Figure 2 shows the distribution of words per utterance in this dataset. As it can be seen, the distribution in this dataset differs from the one used in training. Testing datasets have an average of 19.46 ± 16.81 and 21.27 ± 16.55 words per sentence in Spanish and Basque.

b) Texts from novels and tales: This corpus contains sentences extracted from different tales and novels. It is a phonetically balanced corpus with 450 sentences in Spanish and Basque. As this corpus is used to evaluate the quality and naturalness of the synthetic signals, we used sentences with a length distribution similar to the one shown in Figure 1. The motivation behind this choice was preventing synthesis failures due to alignment issues in long sentences.

The processing applied to both corpora was the same as the one applied to the transcriptions of the training corpora.

[1] http://www.mintzai.eus/indice.html

3. Methodology

In this section we describe the procedure followed in the development of the final neural TTS systems. As will be described in subsection 3.1, we started with an initial implementation of the systems using slightly modified versions of the reference architectures described in the literature. However, the system suffered from some issues that will be described in subsection 3.2. Subsection 3.3 describes the steps taken towards a final implementation that aims to solve the previously mentioned issues.

3.1. Initial implementation

The TTS architecture used in this work consists of two components: a feature prediction network that predicts mel-spectrogam frames from the input text, and a neural vocoder able to generate speech from mel-spectograms. The initial implementation of the developed TTS system uses a Tacotron 2 [11] based model as feature prediction network and Waveglow [13] as neural vocoder. Other neural vocoders like Wavenet [14] and MelGAN [20] were tested, but Waveglow was chosen as it offers a high quality voice with easy training and fast synthesis times.

A restrictive issue when it comes to training Tacotron is the high data volume demand. According to [21], the best audio quality is obtained when using between 10 and 40 hours of data, whereas using less data still produces good quality albeit with certain degradation. As we do not have such big amount of high quality transcribed data available, we opted to apply transfer learning over a publicly available pre-trained model provided by NVIDIA [22]. This model was trained with LJSpeech dataset [23], an English corpus with approximately 24 hours from a single female speaker. As the main objective of this work was developing a TTS system with male and female voices in Spanish and Basque, 4 different models were trained using the training database described in section 2.1. The training of the 4 models converged after 15k steps, lasting approximately one day for each model with a NVIDIA TITAN RTX graphics card and batch size of 64. To prevent over-fitting, attention layer dropout was set to 0.4 and decoder dropout rate was set to 0.1. Learning rate remained constant through the whole training at 0.001.

Regarding the neural vocoder, a pre-trained model was also used due to the computational cost of training a new model from scratch. The obtained models produced good quality voices, but a few issues were identified when synthesizing speech with them. The following section covers all of them.

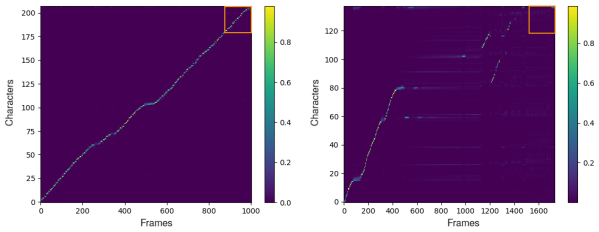

Figure 3: *Correctly aligned (left) and badly aligned (right) sentence alignments*

3.2. Issues found on the initial implementation

To evaluate the trained models, a set of unseen sentences were synthesized. These sentences are obtained from dataset a) described in section 2.2. The synthesized signals displayed two main issues: poorer naturalness for male voices and lack of robustness.

3.2.1. Poorer naturalness for male voices

A first informal listening test of the synthesized speech signals allowed to detect that, regardless of the input text, the naturalness for the two male voices (Basque and Spanish) was poorer than that of the female voices, showing unpleasant noises and buzziness. Visual inspection of the mel-spectrograms generated for female and male voices showcased no major degradation in the latter. We deemed this error to the fact that the Waveglow's pre-trained model corresponds to a female voice. This effect has also been reported in [24].

3.2.2. Lack of robustness

Closer inspection over the output attention graphs of the synthesized utterances displayed that, occasionally, the model was failing to identify the generation stop token. This results in noisy fragments at the end of the generated audio signals. In addition to this, some attention graphs also showed that the model was failing to attend to the correct input character at certain decoding steps, resulting in speech generation instability. As stated in [25, 19], autoregressive attention-based systems are prone to alignment instability, causing word skippings, repetitions and in the worst cases unintelligible speech.

In order to measure the number of audio files affected by either the noise at the end or the lost of alignment during synthesis, a post processing stage was added to the model. The function of this stage is to check the last decoding steps of the inference process, verifying that the attention matrix weights given to the last characters are higher than a threshold. Figure 3 shows the alignment matrix of 2 different synthesized sentences, with the input characters at the vertical axis and the output frames at the horizontal axis. The left image shows a correct alignment and the right image shows that the attention has been lost midway. The figure shows in a rectangle the area that corresponds to the final decoding steps. The algorithm checks row-wise the assigned weights in the selected area, searching for values over a threshold. The values for the area size (10x50) and the threshold (0.3) for the weights have been chosen empirically.

3.3. Final implementation

Addressing the aforementioned issues was crucial to develop natural voices in Basque and Spanish in a robust manner. The quality difference between the female and male voices was large and the attention instability during synthesis made the model unreliable for the task of generating a large amount of synthetic sentences without supervision.

The amount of available data did not suffice to apply transfer learning over the pre-trained neural vocoder model. Nonetheless, as stated in [26], Waveglow is able to generalize to unseen speakers. This feature enables the use of additional corpora containing male voices to improve the quality of the synthesized waveforms. In this case, a single speaker male voice extracted from an audiobook in Spanish was used. The final length in Spanish and Basque resulted in approximately 19 hours of male voices. The available memory in the GPU used for training forced to set the batch size to 3. The learning rate remained constant at a value of 0.001 and the training converged after 18 days using a NVIDIA TITAN RTX graphics card.

Regarding the lack of robustness observed in Tacotron 2 different approaches can be applied. Some studies propose modifying the existing attention mechanism aiming to improve the model instability, as related in [27, 28]. Other technique is based on injecting prior knowledge into the model to improve the training of the existing attention mechanism. We opted for the latter, specifically by implementing pre-alignment guided attention [25] as it improves the model stability when synthesizing long utterances, while also enhancing the training efficiency.

The basic idea of pre-alignment guided attention is introducing an explicit target to guide the model to learn the attention during the training process. These targets are time-aligned phoneme sequences. In order to obtain them, the linguistic Front-End for Spanish and Basque was used along with a forced alignment tool (Montreal Forced Aligner [29]).

Once all the necessary input files were obtained and the model was modified to include the new attention loss metric, the training of the 4 models was done using the same hyperparameters as in the initial implementation. All the models converged after 12k iterations and the training lasted for approximately 12 hours.

4. Evaluation

In this section the robustness of the implemented models is evaluated, along with a Mean Opinion Score (MOS) evaluation of the quality and naturalness of the final implementation. Furthermore, we also evaluate the naturalness of the synthetic signals through a deep learning based assessment proposed by [30] called "Non-Intrusive Speech Quality Assessment" (NISQA) for TTS.

4.1. Robustness

To evaluate the robustness of the models we conducted a test where 20000 utterances were synthesized. These utterances are obtained from corpus a) described in section 2.2. The error detection algorithm proposed in section 3.2.2 was used to detect the files with critical synthesis errors. The relative improvement from the initial to the final implementation in terms of number of generation errors is shown in Table 1. As it can be seen, there is an important improvement in all cases.

4.2. MOS

The quality and naturalness of the final implementation of the systems were evaluated in a Mean Opinion Score test where participants had to rate both aspects in a 5-point scale. Utter-

Table 1: *Number of sentences with errors and relative improvement in percentage*

	Initial	Final	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

ances from the dataset described in section 2.2 b) were used to generate the synthetic signals for the evaluation. As this dataset contains sentences in Spanish and Basque, bilingual subjects were required for the evaluation. Three different methods were used:

- Natural speech signals
- Synthetic speech signals generated with the DNN based TTS systems described in section 3.3
- Synthetic speech signals generated with the HTS based TTS systems previously developed in our research group [16][2]

Out of the 450 available sentences, each participant in the evaluation rated 6 randomly selected sentences per speaker (2), language (2) and method (3) (i.e. a total of 72 sentences). Overall, 33 subjects participated in the evaluation (only one among them was expert in speech technologies)[3].

Figure 4 shows the quality scores averaged for all models, languages and speakers together with 95% confidence intervals. In all cases the subjects conferred higher scores to the signals obtained with the Tacotron based systems than to those generated with the HTS systems. However the score is still lower than that of natural speech. On the other hand, female speech was rated higher than male speech in both languages for the DNN based systems. We deem this occurs because the neural vocoder training does not suffice to produce signals with the same quality for female and male voices. Furthermore, this preference is also shown for natural speech signals.

Figure 5 shows the averaged naturalness ratings of all models with 95% confidence intervals. As occurred in the quality evaluation, DNN based systems were also rated below natural speech but they scored higher than the HTS systems. Subjects showed preference for female voices over male voices in natural speech signals, and this also happened in the DNN based systems in both languages.

4.3. NISQA

NISQA-TTS [30] model is a speech naturalness estimator based on deep learning. According to the authors, it works language independently. Table 2 shows the scores provided by the NISQA-TTS model. As it can be seen, HTS based systems received more generous scores than the ones obtained in the MOS evaluation. Regarding the DNN based systems, NISQA-TTS model produces mixed results, being those more conservative in the case of female voices and more generous for male voices. In all cases DNN based systems score higher than HTS based ones, as happened in the MOS evaluation.

[2]https://sourceforge.net/projects/ahotts/

[3]Some examples can be found in http://aholab.ehu.eus/users/victor/IB2020.html

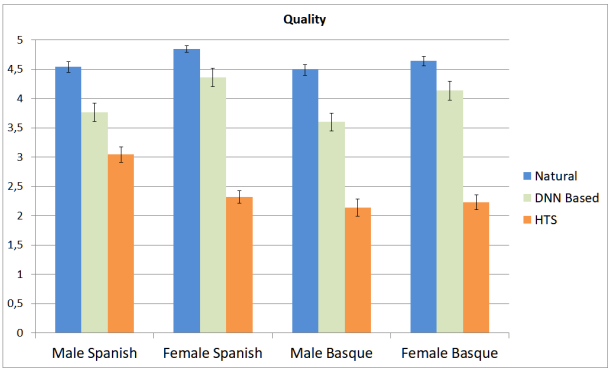

Figure 4: *Results of the quality assessment*

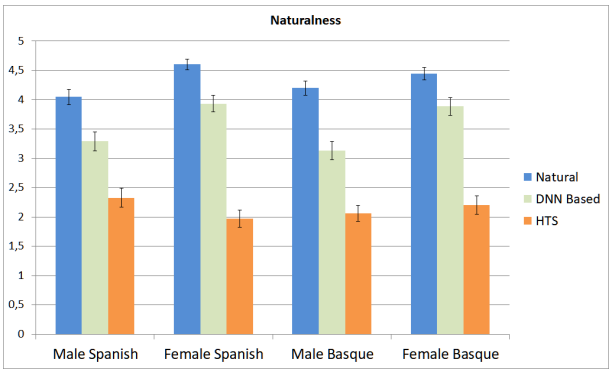

Figure 5: *Results of naturalness assessment*

Table 2: *NISQA evaluation scores with 95% confidence interval*

	HTS based	DNN based
Female Spanish	2.97 ± 0.05	3.34 ± 0.05
Female Basque	3.07 ± 0.04	3.46 ± 0.07
Male Spanish	3.62 ± 0.04	3.64 ± 0.08
Male Basque	2.95 ± 0.04	3.56 ± 0.08

5. Conclusions

This paper describes the implementation of several DNN based TTS systems for Spanish and Basque. Results from the conducted subjective and objective evaluation demonstrate that the robustness of the final systems improved and they are able to synthesize good quality and natural signals in both languages. The system is currently being used to generate training material to conduct speech to speech translation [18].

Future work includes improving the naturalness and robustness of the systems by increasing the amount of data used during the training. We also consider researching on different architectures for the feature prediction network to address the alignment issues without loosing quality.

6. Acknowledgements

This work has been funded by the Basque Government (Project refs. PIBA 2018-035, IT-1355-19 and MintzAI project KK-2019/00065).

7. References

[1] A. W. Black and P. Taylor, "Automatically Clustering Similar Units for Unit Selection Speech Synthesis," in *Proceedings of EUROSPEECH*. ISCA, 1997, pp. 601–604.

[2] N. Campbell and A. W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis," in *Progress in Speech Synthesis*. Springer New York, 1997, pp. 279–292.

[3] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proceedings of ICASSP*, vol. 1. IEEE, 2006.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, nov 2009.

[5] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1278–1288, 2010.

[6] I. Sainz, D. Erro, E. Navas, and I. Hernáez, "A hybrid tts approach for prosody and acoustic modules," in *Proceedings of INTERSPEECH*. ISCA, 2011, pp. 333–336.

[7] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of ICASSP*. IEEE, 2013, pp. 7962–7966.

[8] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Proceedings of INTERSPEECH*, 2014, pp. 1964–1968.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of INTERPSEECH*. ISCA, 2017, pp. 4006–4010.

[10] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proceedings of ICLR*, 2017.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proceedings of ICASSP*. IEEE, 2018, pp. 4779–4783.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, pp. 3104–3112, 2014.

[13] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proceedings of ICASSP*. IEEE, 2019, pp. 3617–3621.

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

[15] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in neural information processing systems*, 2018, pp. 10 215–10 224.

[16] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based speech synthesis in Basque language using HTS," in *Proceedings of FALA*. RTTH, 2010, pp. 67–70.

[17] J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon, "Standard computer-compatible transcription," *Esprit project 2589 (SAM), Doc. no. SAM-UCL*, vol. 37, 1992.

[18] T. Etchegoyhen, H. Arzelus, H. Gete, A. Alvarez, I. Hernaez, E. Navas, A. González-Docasal, J. Osácar, E. Benites, I. Ellakuria, E. Calonge, and M. Martin, "MINTZAI: Sistemas de Aprendizaje Profundo E2E para Traducción Automática del Habla MINTZAI: End-to-end Deep Learning for Speech Translation," *Sociedad Española para el Procesamiento del Lenguaje Natural*, vol. 65, pp. 97–100, 2020.

[19] Y. Ren, T. Qin, Y. Ruan, S. Zhao, T. Y. Liu, X. Tan, and Z. Zhao, "FastSpeech: Fast, robust and controllable text to speech," *arXiv*, may 2019.

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 910–14 921.

[21] Y. A. Chung, Y. Wang, W. N. Hsu, Y. Zhang, and R. J. Skerry-Ryan, "Semi-supervised Training for Improving Data Efficiency in End-to-end Speech Synthesis," in *Proceedings of ICASSP*, vol. 2019-May. IEEE, 2019, pp. 6940–6944.

[22] NVIDIA, https://github.com/NVIDIA/tacotron2, 2018, online; accessed 20 December 2020.

[23] K. Ito and L. Johnson, "The LJ Speech Dataset, v1.1," https://keithito.com/LJ-Speech-Dataset/, 2017, online; accessed 20 December 2020.

[24] B. Külebi, A. A. Alp¨oktem, A. Peiró-Lilja, S. Pascual, and M. Farrús, "CATOTRON-A Neural Text-to-Speech System in Catalan," in *Proceedings of INTERSPEECH*. ISCA, 2020, pp. 490–491.

[25] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65 955–65 964, 2019.

[26] S. Maiti and M. I. Mandel, "Speaker Independence of Neural Vocoders and Their Effect on Parametric Resynthesis Speech Enhancement," in *Proceedings of ICASSP*. IEEE, 2020, pp. 206–210.

[27] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proceedings of INTERSPEECH*, vol. 2019-September. ISCA, jun 2019, pp. 1293–1297.

[28] E. Battenberg, R. J. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis," in *Proceedings of ICASSP*. IEEE, 2020, pp. 6194–6198.

[29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proceedings of INTERSPEECH*. ISCA, 2017, pp. 498–502.

[30] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *Proceedings of INTERSPEECH*, pp. 1748–1752, 2020.