The Blizzard Challenge 2021
23 October 2019, Vienna, Austria

The AHOLAB Text-to-Speech system for Blizzard Challenge 2021

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

In this paper we present the Text-to-Speech synthesis system proposed for the 2021 Blizzard Challenge by Aholab Signal Processing Group. The goal of this challenge is to build a synthetic voice from a provided speech corpus recorded in European Spanish. The challenge comprises two tasks: synthesising text containing only Spanish words and synthesising Spanish texts containing a small number of English words. Our system uses Tacotron-2 to compute mel-spectrograms from the input sequence, followed by WaveGlow as neural vocoder to obtain the audio signals from the spectrograms. A Spanish linguistic front-end module was used to transform grapheme sequences into phoneme sequences. In order to improve the robustness of the system and make the learning of the alignments in the acoustic model easier, a prior knowledge based loss was added to it. Evaluation shows that our systems had a good performance on both tasks.

Index Terms: DNN based Speech Synthesis, Text to speech, Tacotron-2

1. Introduction

Blizzard Challenge 2021 is the seventeenth edition of an annually hold challenge [1] intended to compare speech synthesis models and techniques applied over a common corpus. Participants have to extract the released corpus, build synthetic voices and synthesise a determined set of test sentences. In this year's challenge, almost 10 hours of speech data from a female native speaker of European Spanish are provided and two tasks are proposed:

- Hub task 2021-SH1: The hub task goal is to build a voice from the provided European Spanish data to synthesise texts containing only Spanish words.
- Spoke task 2021-SS1: The spoke task consists in building a voice from the provided European Spanish data to synthesise Spanish texts containing a small number of English words in each sentence.

The output from each task undergoes subjective evaluation through listening tests covering intelligibility, naturalness, similarity to the original speaker and, in the Spoke task, acceptability of the English words.

The goal of Text-to-Speech (TTS) systems is to achieve human-like synthetic speech from input written language. Traditionally, unit selection (US) based concatenative synthesis [2, 3] and statistical parametric (SP) speech synthesis [4, 5] have been used to develop TTS systems.

Nowadays, deep neural networks (DNN) achieve state of the art performance in the development of speech synthesis systems [6]. Neural networks have benefited TTS systems by largely improving the quality and naturalness of the synthetic speech with respect to traditional methods. Furthermore DNNs allow to train and design the systems in an end-to-end (E2E) fashion [7, 8], reducing traditional multi-stage pipelines complexity at the expense of an increased data dependency.

E2E systems usually contain two components; a feature prediction network that extracts intermediate feature representations of the acoustic signals, and a vocoder that synthesises speech from the generated intermediate representations. In the Spanish TTS system that we propose for this challenge we make use of three main components:

- A Spanish linguistic front-end that cleans and converts the input text into a phoneme sequence, using the SAMPA alphabet [9].
- A Tacotron-2 [10] based feature prediction network, with an added loss term that contains aligning information.
- A pretrained WaveGlow [11] neural vocoder, fine-tuned with the provided Spanish corpus.

This paper is organised as follows. Section 2 introduces our proposed system for both tasks, with a detailed description of the architecture and the data preparation. Section 3 covers the results obtained with our system and conclusions are drawn in Section 4.

2. Methods

In this section we will describe the framework and methodologies used in the systems we proposed for 2021-SH1 and 2021-SS1 tasks. Figure 1 shows the base architecture of the proposed system. An adaptation of the linguistic Front-End was applied for the 2021-SS1 task. A description of the data and the modifications applied to it will be covered in the next subsection, followed by an explanation of each module of the system.

2.1. Data preparation

The data provided by the Blizzard challenge committee consists of 9.58 hours of recordings by a single Spanish female speaker, including 4920 sentences and their corresponding orthographic transcriptions. In order to train the system, both audio and text were processed. First, all audio signals were down-sampled from 48kHz to 22.05kHz. To prevent out-of-memory errors, audio signals longer than 15 seconds were left out (i.e. a 0.5% of audios from the provided corpus). Phoneme sequences were obtained from the text using a Spanish linguistic Front-End developed by our team. The alignment between phonemes and audio is an additional information required for training the acoustic model. To obtain the alignment we made use of Montreal Forced Aligner (MFA)[12], a speech recognition Kaldi [13] based model that returns the timestamps of each phoneme. MFA uses a pronunciation dictionary to look up which phonemes correspond to each word. The pronunciation dictionary required for this model was built-up from the transcriptions obtained from our front-end.

When obtaining the time-aligned phoneme sequences, we observed some inconsistencies between the pauses in the tran-

10.21437/Blizzard.2021-11

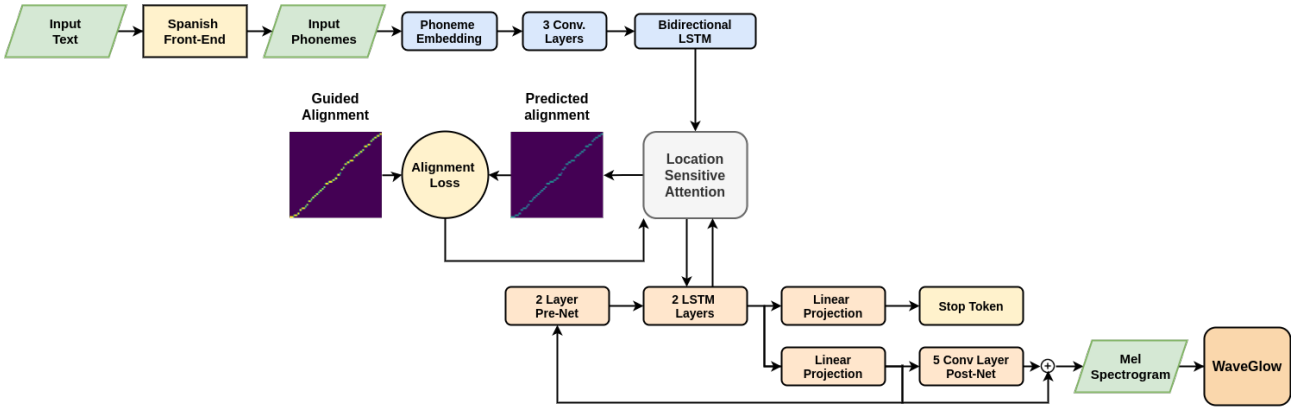

Figure 1: *Architecture of the system.*

scriptions and the actual pauses in the audio signals. This could potentially harm the learning of the alignments in the acoustic model, so we re-positioned the pauses in the phoneme sequences to match the silences in the audio signals.

2.2. Front-End

For the processing of the provided text we used the AhoTTS [14, 15] Spanish linguistic front-end developed by our team. This front-end has two main modules: a initial text processor and a linguistic processor. The text processor expands the numbers and the acronyms into directly readable words. The output of this module gets fed into the linguistic processor, a rule based module that returns the corresponding SAMPA phone sequence along with the stress level of each phone.

The approach we took for the 2021-SS1 task is fully built in this component of the system. The first step was developing an automatic detection tool that could identify all English words located in the sentences of the test set. For this purpose we opted for a dictionary based strategy, considering as English words all those present in the CMU dictionary[1], but excluding the ones in common with a publicly available Spanish dictionary [2]. In addition to this list of Spanish words, some conjugated verbs were included in it.

Once a word is identified as an English one, the aim of our proposal is to replicate the actual pronunciation of the target word with the available phones in the Spanish phone set. For this purpose the CMU dictionary was used again. CMU dictionary contains a list of tuples, including each word in the English language and its corresponding phones represented in IPA alphabet [16]. As our linguistic front-end extracts phones in SAMPA alphabet, we adapted the phones from the CMU dictionary to the SAMPA notation. This step was done via a simple rule based system, which transforms an IPA symbol (or a set of IPA symbols) to the corresponding SAMPA representation.

2.3. Acoustic model

The aim of the acoustic model is to represent the relation between the input phonemes and the corresponding acoustic features of the audio signal. The acoustic model that we used in our proposal is based on Tacotron-2 [10]. Tacotron-2 is a sequence-to-sequence model that originally maps character embeddings to mel-scale spectrograms. This model involves an encoder-decoder architecture with attention mechanism.

The encoder converts a character sequence into a hidden feature representation which is later used by the decoder. In contrast to the original model, the input to the encoder in our proposal is no longer a character sequence but a phoneme sequence obtained from the Spanish front-end. Therefore, the input character embedding is substituted by a phoneme embedding which includes the symbols of the Spanish SAMPA alphabet [3], along with the corresponding accent marks. The output of the phoneme embedding is then fed into a stack of 3 convolutional layers that model long-term relations between the phonemes. Finally, a Bi-directional LSTM generates the encoder hidden outputs from the output of the last convolutional layer.

The output of the encoder is consumed by the attention mechanism to produce a context vector. The attention mechanism provides the decoder with the information required to refer to the correct parts of the encoding sequence at each decoding step. Tacotron-2 uses a custom location-sensitive attention mechanism [17] that employs cumulative attention weights from previous decoding steps as an additional feature.

The decoder is an auto-regressive recurrent neural network that predicts one frame at each decoding step. It consists on a 2 layered pre-net, a 2 Layer LSTM network and a convolutional post-net. The prediction from the previous decoding step is passed through the pre-net and the output is concatenated to the context vector. This concatenation is then fed into the 2 layer LSTM. The output of this stage and the context vector are again concatenated and then passed through two different projection layers: one that predicts the stop token, and another one that predicts the target spectrogram frame. A final convolutional post-net predicts a residual that combined with the whole spectrogram gives as result the final mel spectrogram.

Although this model achieves state-of-the-art results in terms of speech quality, it may run into alignment issues derived from the attention module that cause major speech degradation [18, 19, 20]. To improve the model stability we opted for the approach taken in [18], including an additional loss computed from the predicted alignment and the ground truth alignment of the sentence.

The training of this model was performed on a single NVIDIA TITAN RTX GPU. Batch size was set to 64 and learn-

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] https://github.com/javierarce/palabras/blob/master/listado-general.txt

[3] https://www.phon.ucl.ac.uk/home/sampa/spanish.htm

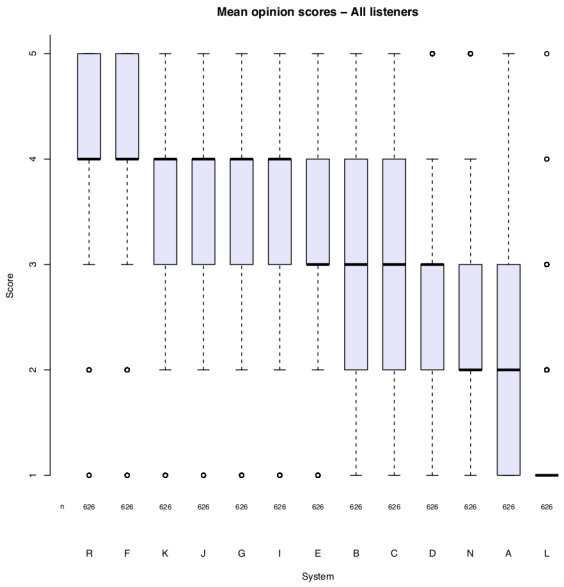

Figure 2: *Task 2021-SH1 Mean Opinion Score on naturalness*

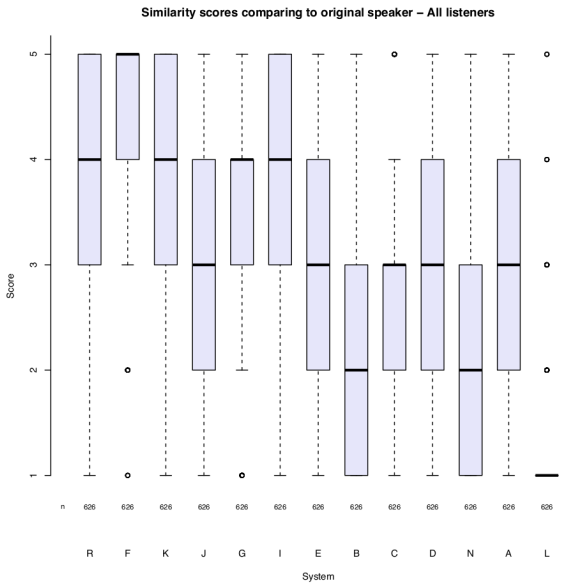

Figure 3: *Task 2021-SH1 Mean Opinion Score on similarity comparing to original speaker*

ing rate remained constant at 1e-3. In order to prevent overfitting, decoder and attention dropouts were set to 0.4. The training was early stopped at 26000 training iterations.

2.4. Neural vocoder

The neural vocoder reconstructs the waveform audio from the mel spectrogram obtained in the acoustic model. For this purpose WaveGlow generative network [11] was adopted, as it provides audio quality close to WaveNet but with faster inference times. The architecture of this vocoder combines insights of Glow [21] and WaveNet [22].

WaveGlow generates audio by sampling from a simple distribution. The distribution needs to have the same number of dimensions as the target output. For reconstructing the waveform, this simple distribution goes through a series of layers trained to perform invertible transformations to it until reaching the target and a more complex distribution.

During training, the network is trained in the opposite way: The complex distribution is subjected to a series of invertible transformations until becoming the simple distribution, in this case a zero mean spherical Gaussian. This sequence of transformations is called ”normalising flow” [23].

In order to improve the quality of the synthetic speech generated with this vocoder, we used a pretrained model provided by [24]. This model was trained on LJSpeech corpus [25], so we fine-tuned it with the provided audios.

The fine-tuning of the model was performed for 230000 iterations, with a batch size of 3 and a constant learning rate of 1e-4. The rest of the parameters remained unchanged.

Due to deadline issues, the fine-tuning of the model was stopped before reaching the optimal error. Despite this, informal listening tests confirmed that this fine-tuning improved the quality with respect to the unmodified pretrained model.

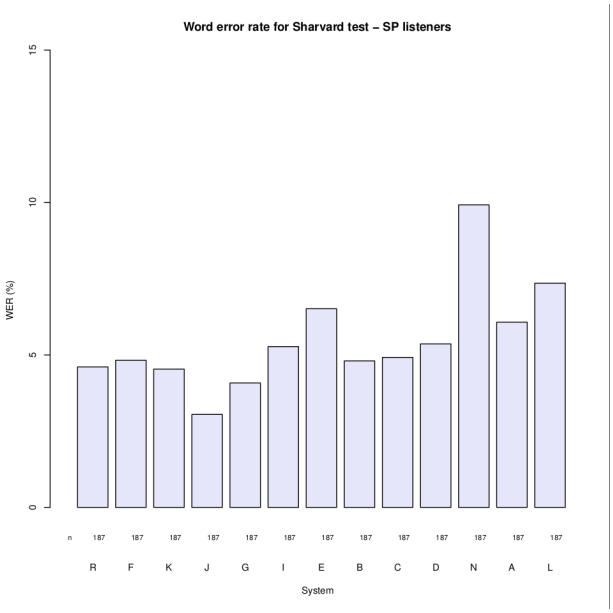

Figure 4: *Task 2021-SH1 Word error rate obtained from Sharvard test*

3. Results

In this section we will introduce the official evaluation results of our proposals. This year the challenge presented two different tasks, and we submitted our proposals to both of them: Task 2021-SH1 and Task 2021-SS1. In the first task (2021-SH1), a total of 12 systems plus a reference natural voice were evaluated. The participating systems are labelled as A/ B/ C/ D/ E/ F/ G/ I/ J/ K/ L/ N, being ”E” the label assigned to our proposed system. In most evaluations natural speech was also considered

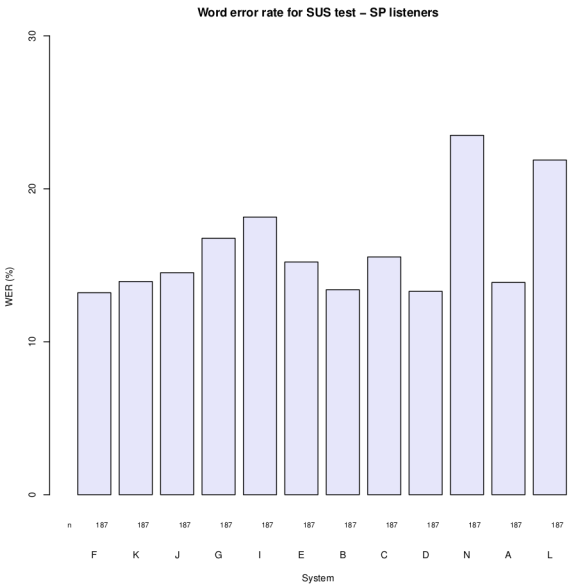

Figure 5: *Task 2021-SH1 Word error rate obtained from SUS test*

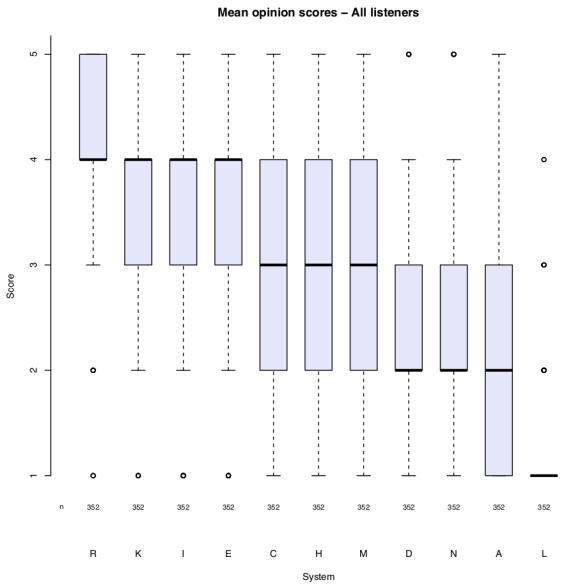

Figure 6: *Task 2021-SS1 Mean Opinion Score on naturalness*

and it was labelled as "R". In the second task (2021-SS1), a total of 10 systems plus a reference natural voice were evaluated. In this case the systems are labelled as A/ C/ D/ E/ H/ I/ K/ L/ M/ N, being "E" our proposed system and again "R" the reference natural voice. The following subsections will comment on the results for each task.

3.1. Task 2021-SH1

Synthetic speech submitted to this task undergoes an online evaluation with three different sections that attend to different characteristics of the audios:

- Naturalness: A measurement of how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].
- Similarity to the original speaker: A measurement of how similar the synthetic voice sounded to the original voice in comparison to some reference samples on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person].
- Intelligibility: A measurement of the comprehensibility level of the synthetic utterances. Reviewers listened to the synthetic speech and typed what they heard, then word error rates are obtained for each system.

Figure 2 shows the scores achieved by all teams in the Hub naturalness evaluation. Figure 3 shows the similarity scores achieved in the Hub task. It can be seen that our model has an average performance among all teams in both dimensions.

Regarding the intelligibility of the synthetic speech provided by all systems, two different test datasets were used for conducting this evaluation: Sharvard corpus containing phonetically balanced sentences, and SUS corpus containing semantically unpredictable sentences. Figures 4 and 5 show the respective word error rates from all systems in each dataset. Attending to both figures it can be seen that our system has not one of the lowest WER but there are no statistically significant differences with respect to those performing best.

3.2. Task 2021-SS1

In this task the online evaluation of the systems had three different sections:

- Naturalness: A measurement of how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].
- Similarity to the original speaker: A measurement of how similar the synthetic voice sounded to the original voice in comparison to some reference samples on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person].
- Acceptability: In this test participants evaluated how acceptable or unacceptable the English words contained in the Spanish sentences sounded, from 1 [Not Intelligible] to 5 [Perfect]. The Spanish words in the sentence were not evaluated.

Figure 6 presents the overall MOS obtained by each system in the Spoke task. It can be seen that our proposal achieved a good performance. Regarding similarity to the original speaker, Figure 7 presents the scores obtained by each system. In this evaluation we obtained a slightly worse result but still achieved an average performance with respect to other systems.

Finally, Figure 8 presents the scores of the part of the test where the acceptability of the English words was evaluated. In this section our system achieved a good score.

4. Conclusions

In this paper we present the submission of the AHOLAB Text-to-Speech system based on Tacotron-2 to the 2021 Blizzard Challenge. We submitted a proposal to both tasks: Hub task 2021-SH1 and Spoke task 2021-SS1. In both tasks our system

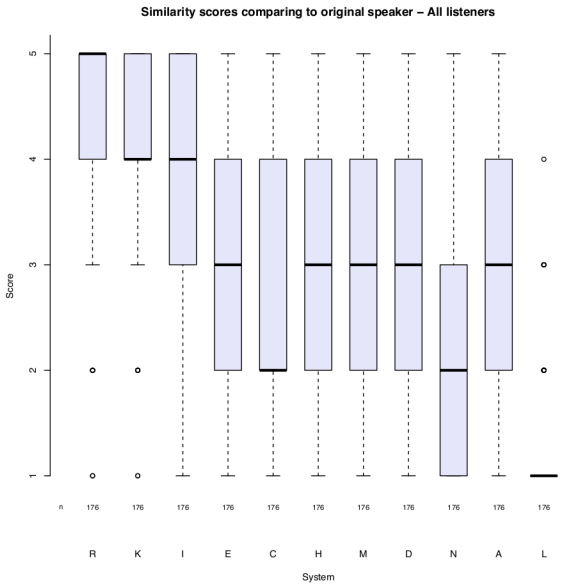

Figure 7: *Task 2021-SS1 Mean Opinion Score on similarity comparing to original speaker*

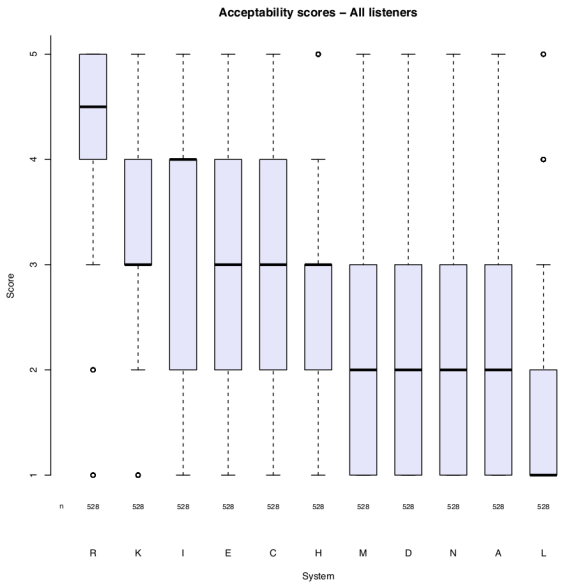

Figure 8: *Task 2021-SS1 Acceptability of English words*

presented an average performance among all participating systems. Despite having a good overall performance, we believe that our system could have achieved better results with a longer adaptation of the vocoder to the provided voice.

In future work we will experiment with different vocoders and training techniques. The aim is to reduce the number of artifacts and improve the overall quality of the synthetic speech in terms of naturalness of the signal and similarity to the original speaker.

5. Acknowledgements

This work has been funded by the Basque Government (Project refs. PIBA 2018-035, IT-1355-19) and Agencia Estatal de Investigacion ref. PID2019-108040RBC21/AEI/10.13039/501100011033.

6. References

[1] "The blizzard challenge website," https://www.synsig.org/index.php/Blizzard_Challenge.

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

[3] 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.

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

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

[6] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.

[7] 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.

[8] 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.

[9] J. C. Wells *et al.*, "Sampa computer readable phonetic alphabet," *Handbook of standards and resources for spoken language systems*, vol. 4, pp. 684–732, 1997.

[10] 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.

[11] 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.

[12] 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.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[14] I. Hernaez, E. Navas, J. L. Murugarren, and B. Etxebarria, "Description of the ahotts system for the basque language," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[15] 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," *Proc. FALA*, pp. 67–70, 2010.

[16] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[17] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[18] 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.

[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] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[21] 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.

[22] 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.

[23] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] NVIDIA, "Waveglow repository and pretrained model," https://github.com/NVIDIA/waveglow, 2018, online; accessed 10 September 2021.

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