Using multi-speaker models for single speaker Spanish synthesis in the Blizzard 2021

Christian Saam, João P. Cabral

Trinity College Dublin, Ireland

saamc@tcd.ie, cabralj@tcd.ie

Abstract

Our entry to the Blizzard 2021 builds on the training of Spanish and mixed Spanish/English models using a Conformer-based FastSpeech 2 system. The mixed Spanish/English models were used for the task of synthesising Spanish texts containing a small number of English words. We used external public multi-speaker Spanish and English datasets to build our models. The main reasons were to avoid problems of data scarcity using the relatively small (around 5 hours long) single speaker Spanish dataset provided by the organisers of the Blizzard, and to be able to build a mixed Spanish/English synthesis model. Our assumption was that we could model well the target Spanish speaker voice in the multi-speaker model and obtain an high-quality Spanish voice. Unfortunately our models produced synthetic speech with quality lower than expected, both in speaker similarity and naturalness. The synthetic voices sounded particularly monotonic, which could have a strong effect on the results. The complexity of consecutively fine-tuning models obtained using multiple datasets to achieve the final voice models could have contributed to inaccurate modelling of the prosody and eventually the speech spectrum characteristics. Another possible reason is that the multi-speaker Spanish dataset we used only included 4 speakers, which may not have been sufficient. We are conducting further analysis to better understand the reasons for the output quality of the system.

Index Terms: multi-speaker speech synthesis, FastSpeech, code-switching

1. Introduction

In the Blizzard Challenge 2021, participants had to build a synthetic voice using the shared training dataset. This data consisted of around 5 hours of European Spanish speech from one native female speaker. The speaking style is read speech and the text transcriptions were included in the data.

The challenge was divided into two tasks:

- Hub task (SH1): To build a voice from the provided European Spanish data to synthesise texts containing only Spanish words
- Spoke task (SS1): To build a voice from the provided European Spanish data to synthesise Spanish texts containing a small number of English words in each sentence

Participants were allowed to use external data in addition to the provided audio files. In our entry, we used external data to obtain a pre-trained model for Spanish, which is described later.

For our team, this was the first time we built a synthetic voice for Spanish. The Blizzard challenge was an opportunity for us to test building a synthetic voice in a new language with minimal human work hours, given our time constraints. Another motivation for our participation was the novel task to synthesise Spanish texts with English words. We have not done experiments with building synthetic voices to address this challenge of code-switching before and this task was a good opportunity for us to try to tackle this problem.

The Text-to-Speech Synthesis (TTS) method we used for the Blizzard is based on deep neural networks. More specifically, we chose to use a non-autoregressive sequence-to-sequence model with multi-head attention, a Conformer-based [1] FastSpeech 2 [2] architecture with pitch prediction akin to Fastpitch [3]. We chose this model because it showed to be robust and provided good quality in past experiments we did with other languages. From our past experiments, auto-regressive models can provide more natural sounding speech than non auto-regressive ones but they showed to be less reliable due to problems in learning the attention and consequently producing poorer quality when it fails to align properly. In this work, we also used ancillary Transformer TTS [4] models as a source of alignment information for the duration prediction of the non-autoregressive models.

Our approach was focused on leveraging previous work wherever possible and avoiding to train new models from scratch. Thus we tried to benefit from knowledge transfer in the form of transfer learning and multi-speaker modelling. To synthesise bilingually, we chose to apply a speaker embedding learned on the Spanish subset of the data on the English inputs. For modelling pronunciation we used the phoneme based tokenisation provided by the VCTK recipe of ESPnet [5] for the English data but used character based tokenisation for Spanish. Character tokenisation worked well in earlier experiments on Italian which has similarly straightforward letter-to-sound rules. Our final model for task SS1 was trained on multi-speaker Spanish and English data with mixed input token inventories.

Another choice we made in building the synthesis models in this work was to use external data, for both the SH1 and SS1 tasks. The reason was to tackle the problem of building a synthetic voice with a relatively small shared dataset. In both tasks we used publicly available datasets of audiobooks with multiple speakers. For task SS1, using external English data was particularly important because the shared dataset only included Spanish data. Another approach could have been to go on an extensive search for an English speaker with similar voice characteristics and use single-speaker English data that is maximally compatible with the Spanish recordings. We chose to avoid this overhead. An important effect of using multi-speaker data is that we needed to take particular care of modelling speaker embedding well, in order to be able to synthesise a voice as similar as possible to the target female Spanish speaker.

The evaluation experiment conducted by the organisers was similar to the format of recent Blizzard Challenge events. It included sections to evaluate speech naturalness and speaker similarity using Mean Opinion Scores (MOS), as well as sections to evaluate intelligibility using the Word Error Rate (WER), in which participants are asked to play the stimuli and have to type the words that they heard. A difference in the evaluation experiment compared to previous years was the part to evaluate the Acceptability (MOS scores) of listeners to synthesis of Spanish text with English words (relative to task SS1).

The paper is organised as follows. Section 2 describes the external dataset used to build the pre-trained model and the data made available by organisers to build the Spanish voice. Our systems for tasks SH1 and SS1 are explained in Section 3. The results of the evaluation are presented in Section 4 and discussed in Section 5. Finally, the conclusions are given in Section 6.

2. Databases

We used external data, because the data provided for the Blizzard is small comparatively with the size of datasets commonly used to produce high-quality voices with neural TTS models. For example, a typical dataset used in the literature to report results of TTS evaluation on single speaker dataset is LJSpeech (US English), which is around 24 hours long. However, we have not experimented to build a synthetic voice using the shared Spanish data only to verify the quality of a speaker dependent model.

We used an external dataset to build a pre-trained Spanish model for task SH1, the Spanish M-AILABS Speech Dataset [1]. It is divided into three parts: female, male, and mixed. We used the female and male subsets that consist of 1 female speaker (10h 37m), and 2 male speakers (55h 5m and 17h 19m). In total, the duration of the external speech data used is approximately 83h.

We also used an external English dataset to build a mixed English-Spanish model for task SS1, which is the CSTR VCTK Corpus [6]. VCTK is an English Multi-speaker Corpus that includes 109 English speakers with various accents, totaling around 44 hours of speech data and respective text transcripts.

Please note that we also took advantage of our own pre-trained models obtained in past experiments, in addition to the external datasets indicated above. We used our pre-trained Italian model to warm-start training the Spanish model in task SH1 and our pre-trained English model to warm-start the training of the multi-speaker mixed language model for task SS1. The first was trained on a selected subset of Italian M-AILABS consisting of around 20 hours of speech and the second on LibriTTS [7].

3. System

Our systems are based on ESPnet2 [5]. They are derivations of its Transformer [4] and FastSpeech 2 [2, 3] implementations which use internal speaker and gender embeddings instead of relying on externally supplied x-vectors. The Transformer models were only trained to supply alignment information for training the FastSpeech 2 models.

All models of the same type have the same hyper-parameters, except for the number of speaker and input embeddings. An overview of the most important hyper-parameters is given in Table 1. The FastSpeech 2 models also employ sub-networks for predicting and embedding duration, pitch and energy, which follow standard recipe hyper-parameterisation. We used ESPnet's variable batch size that is roughly constant in the number of batch bins (summed input/output sequence states) targeted at 18E+6 bins. The network was regularised with dropout of 0.2 except for decoder pre- and post-nets which have a dropout of 0.5.

All models used as input 80-dimensional log-mel-filterbank features extracted from 24 kHz speech signals. The recorded speech of the Blizzard Spanish dataset was downsampled from 48 kHz for compatibility with training voices using our pre-trained models that were built from speech sampled at 24 kHz. The acoustic feature extraction was done following standard ESPNet TTS recipes. A short-time Fourier transform of 2048 points was computed from the speech signal by using a Hanning window with 50 ms duration (corresponding to a vector of 1200 samples padded with zeros), with shifting of 12.5 ms (300 samples). The resulting magnitude spectra were warped to an 80 band mel-scale limited to 80-7600Hz, and then converted to natural logarithmic scale. All features were normalised with global mean and variance normalisation. For FastSpeech 2 models, from the same short-term spectra, energy was extracted as RMS and the fundamental frequency (F0) via the DIO [8] plus Stonemask algorithm. Then, their mean values were taken over the per-token frames.

Since we only used the Transformer models to generate alignments, not to synthesise speech, we only describe the specifics of the training process of the FastSpeech 2 models in the next sections. The training criterion of our FastSpeech 2 models is the sum of spectral L1 loss with pitch, energy and duration losses.

The system uses an implementation of the Parallel WaveGAN vocoder [9] to generate the speech waveform from the spectra generated by the acoustic model during synthesis[2]. We used a pre-trained model of the Parallel WaveGAN vocoder made available by the author of this implementation, which was trained on the VCTK corpus.

Table 1: *System configurations. In this table, e/d stands for encoder/decoder.*

	Transformer	**FastSpeech 2**
transformer layers e/d	6/6	4/4
layer dimensions e/d	1024/1024	1536/1536
conv layer kernel size e/d	n/a	7/31
conv layer filters	n/a	384
self-attention heads	8	2
self-attention dimensions	512	384
src-attention heads	8	n/a
src-attention dimensions	512	n/a
pre-net layers	2	n/a
pre-net dimensions	256	n/a
post-net layers	5	5
post-net conv kernel size	5	5
post-net conv filters	256	256
embedding dimensions	512	512
optimizer	RAdam	RAdam
learning rate	0.001	0.001
weight decay	0.01	0.01

3.1. Hub Task

For the hub-task, we built a character based, speaker-independent FastSpeech 2 system with a view to using it as the

[1] Available at https://www.caito.de/2019/01/the-M-AILABS-speech-dataset/

[2] This implementation is available at https://github.com/kan-bayashi/ParallelWaveGAN

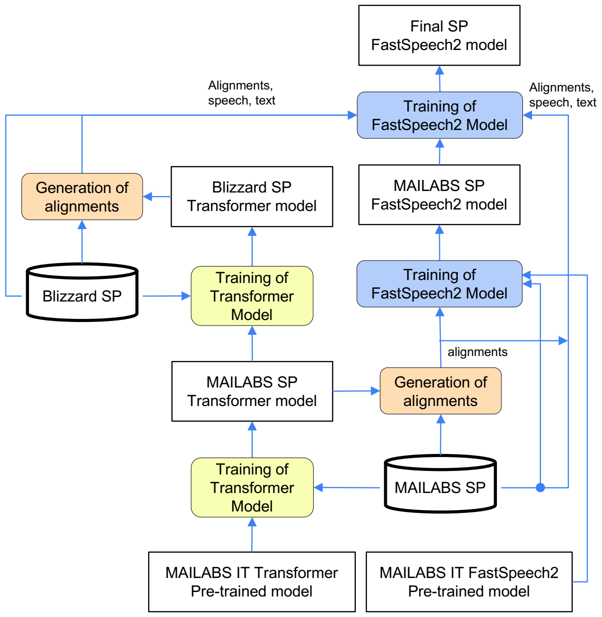
Final SP
FastSpeech2 model
Alignments, speech, text
Alignments, speech, text
Training of FastSpeech2 Model
Generation of alignments
Blizzard SP Transformer model
MAILABS SP FastSpeech2 model
Blizzard SP
Training of Transformer Model
Training of FastSpeech2 Model
alignments
MAILABS SP Transformer model
Generation of alignments
Training of Transformer Model
MAILABS SP
MAILABS IT Transformer Pre-trained model
MAILABS IT FastSpeech2 Pre-trained model

Figure 1: *Block diagram representing the voice building stages for obtaining the final Spanish voice model for task SH1.*

seed system for warm-starting the SS1 system. Figure1 shows the different stages to build the final Spanish synthetic voice from the pre-trained models and speech datasets.

In order to generate alignments for Spanish speech data, we built a character based, speaker-independent Transformer model on the M-AILABS Spanish database that was warm-started from an existing Italian model trained on the M-AILABS Italian database. The resulting model was further adapted on the Blizzard dataset. Finally, these Transformer models built from the M-AILABS Spanish and Blizzard data were used to generate alignments for these datasets respectively.

On the aligned M-AILABS Spanish dataset we trained an initial character based speaker-independent FastSpeech 2 model, which was warm-started from a model pre-trained on M-AILABS Italian. First, its speaker and token embeddings were adapted for 50 epochs (of 500 iterations each) while the rest of the network remained frozen. Then, the full model was adapted for further 870 epochs. The best loss values on the validation set were achieved at epoch 960, with l1_loss=0.715, duration_loss=0.257, pitch_loss=0.360, and energy_loss=0.566. Next, the five best validated models were averaged and used as the seed for the model to be trained on an extended dataset consisting of the aligned M-AILABS and Blizzard datasets. Here, we adapted the embeddings for 20 epochs and trained the model up to a total of 1110 epochs. The best validation loss was achieved at epoch 1100 with l1_loss=0.582, duration_loss=0.163, pitch_loss=0.248, and energy_loss=0.398. Finally, the five best models with respect to validation loss were averaged to yield our final model that was used to synthesise speech for this task.

Please note that the word epoch in the text above and in the following sections refers to pseudo-epochs, which are not full iterations over the dataset but are arbitrarily limited to 500 or 1000 batches for the purpose of intermittent validation.

The token inventory over the Spanish databases consisted of 36 (partly accented) characters, 4 punctuation marks and 4 special tokens.

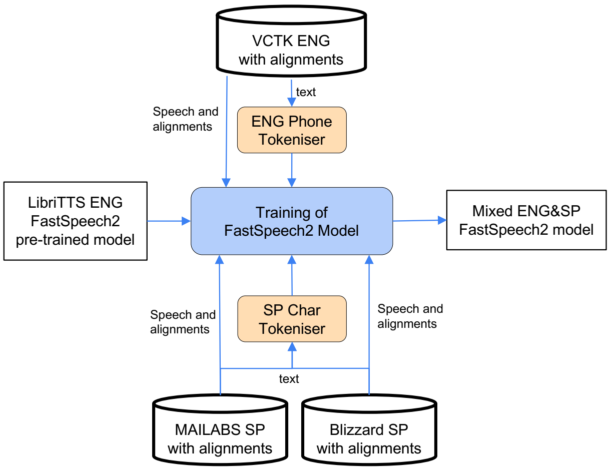
VCTK ENG
with alignments
text
Speech and alignments
ENG Phone Tokeniser
LibriTTS ENG FastSpeech2 pre-trained model
Training of FastSpeech2 Model
Mixed ENG&SP FastSpeech2 model
Speech and alignments
SP Char Tokeniser
Speech and alignments
text
MAILABS SP with alignments
Blizzard SP with alignments

Figure 2: *Block diagram showing the voice building process for obtaining the mixed Spanish/English voice model for task SS1.*

3.2. Spoke Task

For the spoke task, we built a speaker-independent FastSpeech 2 system with mixed character and phone token inputs. Figure 2 shows the block diagram for training the synthesis voice using this system. We re-used the aligned Spanish databases built for the hub task and added an existing pre-aligned VCTK corpus with phone tokens from the CMU pronouncing dictionary. Our character tokenisation preserved word boundaries by inserting space tokens, in contrast to the phone tokenisation. The token inventory of the VCTK database consisted of 69 (partly stress marked) phones, 5 punctuation marks and three special symbols. The combined token inventories consisted of 75 character and phone tokens, 8 punctuation marks and 4 special symbols.

The training began with 20 epochs (500 iteration each) of embedding adaptation followed by full-network training up to 1600 epochs. The model achieved its best validation score at epoch 1575 with l1_loss=0.737, duration_loss=0.236, pitch_loss=0.283, and energy_loss=0.337. The final model was again averaged over the five best validated epoch models.

4. Evaluation Results

4.1. Hub Task

In terms of intelligibility results, our system obtained a Word Error Rate (WER) for Sharvard test of 7.4%, which is above average compared with the other systems. However, the Pairwise Wilcoxon signed rank tests showed that the differences between our system and the others are not statistically significance at 1% level. The results of intelligibility obtained for SUS (semantically unpredictable sentences) are worse than for Sharvard test, as expected. Also, in the SUS test our system is significantly worse than four other systems.

The results of speech quality showed that our system obtained unsatisfactory Mean Opinion Scores (MOS) and similarity to the original speaker, being placed in the bottom part of the ranking. We believe that one of the main factors to explain the poor results is the flat prosody and inefficient adaptation of our pre-trained multi-speaker model to the target speaker using the shared data. The possible factors that explain these results are further discussed later.

4.2. Spoke Task

Our system also ranked in the bottom part with lowest Acceptability scores in the Spoke task. This result indicates that the quality of synthesis of the English words was not good enough. Our approach to be able to synthesise English words was simple and we had limited time to make improvements on some problems we later detected in our implementation. Another expected contribution to this result is that the quality of synthesis voice was not natural enough, because it was similar to the quality of the voice built for the hub task.

5. Discussion

The results obtained by our system in both tasks are worse than we expected. Our goal was to train the synthetic voices for these tasks using a state-of-the-art system, the FastSpeech 2 system, minimising the amount of human work hours needed. In our own past experiments, we built high-quality voices on other languages so we assumed it could work well for building the Spanish voice using the high-quality Blizzard dataset. A difference to our previous experiments was that we built voices either from a single multi-speaker dataset or a single speaker dataset, whereas here we consecutively fine-tuned models built on different multi-speaker datasets towards a final multi-speaker model that includes the target speaker. We assume that this process of fine-tuning the models was not optimal and ended up in a lower quality voice compared with other voices we built before with FastSpeech 2.

The choice of a speaker-independent system for a speaker-dependent task, such as the hub-task, may not seem obvious. However, during initial planning, we aimed to re-use this system as the seed model for the spoke task. Since our approach for the spoke task was to transfer speaker characteristics via the speaker embedding, it made sense to warm-start the final model from an already speaker-independent one. Ironically, due to resource constraints, we did not actually explore seeding the spoke model from this one, as we had prioritised to explore the adaptation of a speaker-independent model trained on a much larger variety of speakers from the LibriTTS database. On the whole, the process for building the hub-task system appears overly complex. Yet, to achieve good speaker generalisation one needs a large variety of speakers. Unfortunately, the M-AILABS Spanish dataset, while representing more than three speakers, only has identity labels for three of them. We started with this labeled subset and trained a speaker-independent Transformer model intended as a source for alignments. This system, however, did not generate satisfactory alignments. Our next attempt was to seed a Transformer system from a Italian model trained on 51 speakers (shown to produce good alignments in past experiments), which succeeded. Possibly the failure to generate good alignments from the initial Transformer model trained on Spanish data is due to the low speaker coverage of the three speaker corpus. Lack of speaker coverage may also contribute to the reduced naturalness of the multi-speaker FastSpeech 2 hub system despite the large size (86 hours) of the combined M-AILABS and Blizzard dataset on which this system was trained. We did not have enough time to increase the speaker space coverage by identifying the speakers in the unlabeled subset of M-AILABS Spanish to extend the dataset used in the experiments.

In terms of tunning the system, we did not do any hyperparameter optimisation to find settings appropriate to the specific conditions. Instead, we relied on settings that worked in similar experiments on other data. Thus, all our models represent first attempts in each stage of model building, which may not be the best models that can be achieved with the system. Another possible explanation for the non-optimal training of the synthesis models is that the models may have been over-trained in an effort to optimize the speaker embedding representation.

One aspect we neglected to model was the phonetic overlap of the two languages in the SS1 task. This was because our combined character and phone tokenisation did not account for the overlap in both inventories. The one-letter phone symbols map to character tokens but the information of what language these tokens represent is lost. This may create ambiguity that is hard to overcome and cause mixed language pronunciations within words. There may be benefit in tying the input token embedding representations of the two languages when phonemes overlap and using a separate embedding layer to signal which language's phonic realisation to use. We are going to test this hypothesis in future work.

One aspect that we find puzzling is that in terms of validation losses our models are around 25-50% better than those of the reference models for ESPnet's FastSpeech 2 implementation trained on VCTK that are provided by ESPnet's authors. For this comparison, the reference model losses were gleaned from the training loss curves distributed with the model. Since we used the same vocoder as in their implementation, there may be a mismatch between our synthesiser outputs and the inputs required by the vocoder. Perhaps a purpose trained vocoder would have yielded better results.

From listening to the synthetic speech, one of the clear problem with our Spanish voice is that it produced speech with flat prosody. There are also occasional artifacts but speech distortion does not seem to be a major reason to explain the low scores, given that the system is not significantly worse than any other system in terms of intelligibility. We need to conduct further detailed analysis of the intermediary stages of voice building, the vocoder and trained models to better understand the reasons for the result of quality lower than expected.

There have been significant advances in speech synthesis in recent years and the latest results in the Blizzard Challenge show that the best synthesis systems produce speech quality almost indistinguishable from recorded speech. It may be the case that in evaluations experiments like the Blizzard that compares different systems, listeners in this experiment tolerated less synthetic speech with flat prosody and penalised more the systems that are significantly worse than the best quality systems. In other words, for listeners the reference of synthetic speech quality is higher compared with past Blizzard experiments, in which there was a larger difference in quality between the synthesis quality of the systems and human speech.

6. Conclusions

We embraced the challenge to participate in the Blizzard this year, motivated to build a synthetic voice in a new language, Spanish, and the task of code-switching between the Spanish and English languages. We took a standard approach of using a robust and high-quality DNN-based system, the FastSpeech 2 system, in this attempt. We also chose the approach of building multi-speaker models instead of single speaker ones. One reason was to take advantage of a much larger publicly multi-speaker Spanish dataset compared with the shared single speaker data. The other reason was to be able to synthesise the voice of the target Spanish speaker for the model trained from a mix of English and Spanish speakers.

One of the main challenges encountered was in our initial attempt to build a Spanish voice using a large external multi-speaker Spanish dataset. It did not seem to produce a good result so we got around this problem by using one of our pre-trained Italian models to warm-start the training of the Spanish model. Another main challenge was to build a model using both Spanish and English datasets for the code-switching task. Our approach was simple but effective by using a different tokeniser for the Spanish and English text processing, phone and character based tokenisers respectively. The evaluation results of our entry were lower than our expectations. Unfortunately, we had no time to further investigate the causes of the low synthesis quality and perform significant improvements. Our participation was valuable for us in terms of what we learned in building the Spanish synthetic voices. Also, the results are quite interesting and intriguing as a case study in future work for increasing our knowledge of what produces unexpected results in training neural TTS models.

7. Acknowledgements

The authors are part of the VoiceTune project that is co-funded by the European Regional Development Fund (ERDF) under Ireland's European Structural and Investment Funds Programme 2014-2020. This research is also supported by TCHPC (Research IT, Trinity College Dublin) and by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie), at Trinity College Dublin. The opinions, findings and conclusions, recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

8. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2021.

[3] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 6706–6713, Jul. 2019.

[5] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.

[6] C. Veaux, J. Yamagishi, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[7] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *CoRR*, vol. abs/1904.02882, 2019. [Online]. Available: http://arxiv.org/abs/1904.02882

[8] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," *Journal of the Audio Engineering Society*, February 2009.

[9] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," 2020.