Adversarial Text-to-Speech for low-resource languages

Ashraf Elneima
African Institute for Mathematical Sciences
aelneima@aimsammi.org

Mikołaj Bińkowski
DeepMind
binek@deepmind.com

Abstract

In this paper we propose a new method for training adversarial text-to-speech (TTS) models for low-resource languages using auxiliary data. Specifically, we modify the MelGAN (Kumar et al., 2019) architecture to achieve better performance in Arabic speech generation, exploring multiple additional datasets and architectural choices, which involved extra discriminators designed to exploit high-frequency similarities between languages. In our evaluation, we used subjective human evaluation, MOS - Mean Opinion Score, and a novel quantitative metric, the Fréchet Wav2Vec Distance, which we found to be well correlated with MOS. Both subjectively and quantitatively, our method outperformed the standard MelGAN model.

1 Introduction

Text-to-speech (TTS) is the task of generating natural speech that corresponds to a given text. TTS systems play essential roles in a wide range of applications, ranging from human-computer interaction to assistance for people with vision or speech impairments.

In recent years the field of TTS has been dominated by the neural auto-regressive models for raw audio waveform such as WaveNet (Oord et al., 2016a), SampleRNN (Mehri et al., 2016) and WaveRNN (Kalchbrenner et al., 2018). However, inference with these models is inherently slow and inefficient given the high frequency of audio data; because of the auto-regressive behaviour and the sequential generation of the audio samples. Thus, auto-regressive models are usually impractical for real-time applications. Researchers put much effort into enabling parallelism of the TTS models, which resulted in a number of *non-auto-regressive* ones, such as Parallel WaveNet (Oord et al., 2018) which distils a trained auto-regressive decoder into a flow-based convolutional student model, WaveGlow (Prenger et al., 2019) which is a flow-based generative model based on Glow (Kingma and Dhariwal, 2018) as well as the Generative Adversarial Network (GAN (Yi et al., 2019))-based models such as MelGAN (Kumar et al., 2019) and GAN-TTS (Bińkowski et al., 2019). They are highly parallelizable and more suitable to run efficiently on modern hardware. However, those recent developments often came at the price of scale, and hence may be impractical for certain applications with limited compute or data budgets.

Deep neural networks have revolutionized the field of TTS achieving human-level performance on particular languages by leveraging massive collections of good-quality datasets, e.g. The LJ Speech Dataset[1]. However, these successes came at cost since creating these large datasets typically requires a great deal of human effort to manually record and label individual data samples. This cost can be particularly extreme when recording and labelling requires expert supervision (for example, recording high quality audio requires a professional studio and staff). For many languages we lack resources to create sufficiently large labelled datasets, which limits the widespread adoption of TTS techniques.

The lack of available resources makes it extremely valuable to study the relationship between the different languages. The high-frequency similarities between languages can be exploited to learn better speech synthesis models for low-resource languages. However, not much work has focused so far on exploring this direction. The notable exceptions include some multi-lingual TTS models (Do et al., 2021). In Lee et al. (2018) they pre-trained a speech synthesis network using datasets from both high-resource and low-resource languages, and fine-tuned the network using only low-resource data. The results showed that the learned phoneme embedding vectors are located closer if their pronunciations are similar across the languages.

[1] https://keithito.com/LJ-Speech-Dataset/

In this work, we explore raw waveform generation for low-resource languages using auxiliary data, taking Arabic as our case study and MelGAN (Kumar et al., 2019) as our baseline model. This study examines the Arabic language since it has a large global population, it is a complex language to model,[2] and there is a scarcity of Arabic TTS datasets, making it a low-resource language. Our main contributions are as follows:

- We train a fast and efficient TTS system for the Arabic language using a publicly available speech dataset[3].
- We propose an extension to MelGAN (Kumar et al., 2019) model which makes it more amenable to knowledge transfer between languages and evaluate its efficiency for low-resource speech datasets, focusing on co-training between vastly different languages/dialects and learning from low-quality samples.
- We propose a quantitative metric for Arabic speech generation based on Fréchet distance (Eiter and Mannila, 1994), the metric inspired by the DeepSpeechDistance for English language (Bińkowski et al., 2019), where we replace the DeepSpeech network with the Wav2Vec2ForCTC Arabic audio recognition network[4].

2 Background

The generative Adversarial Networks (GANs) Goodfellow et al. (2014) are a class of implicit generative models trained by adversarial means between two networks: the generator and the discriminator. Generators attempt to produce data that resemble reference distributions, while the discriminator tries to distinguish real data from generated data, providing a useful training signal.

Due to the high temporal resolution of raw waveform, the presence of structure at different time scales, and the short- and long-term interdependencies among these structures, audio synthesis is a challenging task. Most approaches simplify the problem by modelling a lower-resolution intermediate representation that can be efficiently computed from the raw temporal signal and preserves enough amount of information to allow a faithful inversion back to audio. It is therefore common to decompose text-to-speech (TTS) systems into two stages: the first stage maps text into the intermediate representation, while the second stage transforms it into audio waveform. Among the most commonly used intermediate representations are aligned linguistic features (Oord et al., 2016b) and Mel-spectrograms (Shen et al., 2018; Gibiansky et al., 2017). In this work, we use Mel-spectrogram as an intermediate representation and focus on the second stage. Considering the Mel-spectrogram inversion stage, the TTS systems can be categorized into three distinct families: the pure signal processing techniques, the auto-regressive models and the non-auto-regressive models. The auto-regressive models like the WaveNet (Oord et al., 2016a) produced the state-of-the-art results in text-to-speech synthesis (Sotelo et al., 2017; Shen et al., 2018) but inference with these models is inherently slow and inefficient due to the sequential generation of audio. The non-auto-regressive models hence are highly parallelizable and can exploit modern deep learning hardware like GPUs and TPUs. Well known examples are the WaveGlow (Prenger et al., 2019) which is a flow-based generative model based on Glow (Kingma and Dhariwal, 2018), and GAN-based TTS models like MelGAN (Kumar et al., 2019) and GAN-TTS (Bińkowski et al., 2019).

MelGAN generator is a fully convolutional feed-forward network which takes Mel-spectrogram as input and outputs a raw waveform. The generator is trained adversarially against a multi-scale architecture comprised of three discriminators that have identical network structures but operate on different audio scales. On the other, End-to-end architectures like the Tacotron (Wang et al., 2017), EATS (Donahue et al., 2020) and WaveGrad 2 (Chen et al., 2021) are introduced in the field of TTS to reduce the compound error of two-stage TTS systems. Tacotron is a generative text-to-speech model based on a seq-to-seq model with an attention mechanism (Sutskever et al., 2014), whereas Tacotron 2 (Shen et al., 2018) is a follow-up work that eliminates the non-neural network elements used in the original Tacotron.

Many works covered Arabic TTS synthesis to generate human-like speech, such as Abdel-Hamid

[2] Worldwide there are more than 420 million native Arabic speakers who speak over 25 dialects of the language, each of which has its own unique characteristics and dialectal words.

[3] http://en.arabicspeechcorpus.com/

[4] https://huggingface.co/docs/transformers/model_doc/wav2vec2#transformers.

et al. (2006), Rebai and BenAyed (2016) and Fahmy et al. (2020), but none of them adopted the GAN-based TTS models for the Arabic language. Fahmy et al. (2020) describes how to use a modified deep architecture from Tacotron 2 (Shen et al., 2018) to generate Mel-spectrograms from Arabic diacritic text as an intermediate feature representation followed by a WaveGlow (Prenger et al., 2019) architecture acting as a vocoder to produce a high-quality Arabic speech. The proposed model is trained using a published pre-trained Tacotron 2 English model using a dataset with a total of 2.41 hours of recorded speech [3]. To the best of our knowledge, this is the best Arabic TTS available.

3 Methodology

In this section, we present the details of the architectures of our models, the datasets, and the evaluation metrics we used. In MelGAN's official repository[5], generator weights are publicly available, but discriminator weights are not. We use various methods of knowledge transfer between languages, including fine-tuning and co-training.

3.1 Model Architecture

In our analysis, we used the MelGAN architecture (Kumar et al., 2019) with an amended downsampling schedule that we found to perform better in our early experiments. With the proposed schedule, we ensure that there is no common divisor between downsampling factors to encourage focus on different frequencies across discriminators. We used factors 3 and 5 to downsample audio before passing it to the second and third discriminators. The downsampling is done by a strided average pooling layer.

MelGAN's multi-discriminator architecture incorporates an inductive bias that aims to exploit different structures at various temporal resolutions. In addition, we are interested in investigating another inductive bias that aims to exploit the considerable overlap between the phonemes of different languages and dialects, which may be helpful to improve the performance of low-resource languages. In the proposed approach we introduce auxiliary data to the model through an additional discriminator, designed to operate on short segments of speech to capture high-frequency similarities. We found optimal segment length for this extra discriminator to be 512-time steps. We consider two ways of feeding the extra data to the model:

- As part of first setting, the additional discriminator is fed a batch of 512-time step segments of two types, one generated directly by passing a small window of the auxiliary dataset mel-spectrogram to the generator, and another produced by sub-sampling the audio generated with the main dataset conditioning to pass to the main discriminators.

- While in the second setting, the additional discriminator accepted a batch of 512-time step segments both are sub-sampled from the audio generated with the main dataset conditioning to pass to the main discriminators, but to introduce the auxiliary dataset, part of the ground truth segments are replaced by random segments of the auxiliary dataset.

The mixing ratio between the two types of segments in both settings is a hyper-parameter that we optimise experimentally.

Passing the auxiliary data to the generator in the first setting provides a more complicated task for the generator to learn, while in the second setting the generator's task remains unchanged; however the additional discriminator is provided with more ground truth samples and hence enriches the adversarial signal passed back to the generator. Finally, the additional discriminator uses half of the standard MelGAN discriminators' capacity[6], which we found to perform roughly on par with the full capacity variant.

3.2 Datasets

We used the Arabic Speech Corpus dataset[3] as our main dataset. The training set contains 1813 spoken utterances of a standard Arabic dialect recorded by a single speaker, covering a duration of 2 hours; additional 100 samples form a test set. The data is labelled with diacritic Arabic text (Sweet, 1877). In addition to the main dataset, we used three auxiliary datasets as described in the table 1. The auxiliary datasets include LJSpeech[1], Tunisian_MSA[7] and AMMI_Speech datasets[8]. The AMMI_Speech dataset is gathered by AMMI[9] student. The

[5] https://github.com/descriptinc/melgan-neurips

[6] half the number of convolution filters

[7] https://www.openslr.org/46/

[8] https://github.com/besacier/AMMIcourse/tree/master/STUDENTS-RETURN/Arabic4

[9] African Master of Machine Intelligence - https://aimsammi.org

Name	Language	Dialect	Speakers	Quality	Hrs
LJSpeech	English	-	1	high	24
Tunisian_MSA train	Arabic	Tunisian	118	low	11
Tunisian_MSA test	Arabic	Tunisian/Libyans	4	average	2
AMMI_Speech	Arabic	Standard	3	low	6
Arabic Speech Corpus	Arabic	Standard	1	high	2

Table 1: The datails of the auxiliary datasets used.

Tunisian_MSA train and test set are separated into two auxiliary datasets due to their varying quality.

3.3 Evaluation Metrics

For evaluation, two metrics are employed: the Mean Opinion Score (MOS) and a novel quantitative metric, the Conditional Fréchet Wav2Vec Distance (cFWD).

Mean Opinion Score In order to compare the performance of our models, we carried out Mean Opinion Score (MOS) tests. We gathered 100 samples generated by the different models using the same conditioning, along with 100 original samples. All the generated samples were not seen during training. MOS scores were computed on a population of 53 individual raters; each of them had to evaluate blindly a subset of 150 samples drawn randomly from the overall pool and assign a score from 1 to 5. Our tests were crowdsourced over multimedia platforms and testers were asked to wear headphones and be Arabic speakers. Additionally, we computed the 95% confidence intervals for the scores:

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} s_{i,k}$$

$$CI_i = \left[\hat{\mu}_i - 1.96\frac{\hat{\sigma}_i}{\sqrt{N_i}}, \hat{\mu}_i + 1.96\frac{\hat{\sigma}_i}{\sqrt{N_i}}\right]$$

Conditional Fréchet Wav2Vec Distance This metric is inspired by the DeepSpeech Distances (Bińkowski et al., 2019) and analogous to Fréchet Inception Distance (FID, Heusel et al., 2017) commonly used in generative modelling of images. In order to extract the high-level features from raw Arabic audio, the DeepSpeech2 model was replaced by the pre-trained Wav2Vec2ForCTC Arabic speech recognition model found in the HuggingFace Transformers library[4].

To obtain reasonable estimates of this metric it is preferred to use sufficiently large sets of samples. The original implementation used 50 thousand samples (Soloveitchik et al., 2021). However, as this would be too resource-intensive, we artificially expand the generated and real sets by randomly subsampling small windows from each audio.

The distribution for a set of waveforms is formed by sub-sampling thirty 2-second-long sub-samples from each audio; this way we construct fixed-length sub-samples from arbitrary-long ones, covering their whole length and putting equal weight to short and long samples. Finally, the features extraction is done by framing each sub-sample using a 40ms window of raw audio at 16kHz and stride of 20ms, passing the frames to the speech recognition model, and extracting the 512-dimensional output of the $feature\_projection$ layer, and then taking the average of the features along the temporal dimension. The Fréchet distance is calculated by comparing the distributions of such representations of real and generated samples from our test set, which has 100 samples, resulting in 3000 samples after subsampling. For representations $X \in \mathbb{R}^{m\times d}$ and $Y \in \mathbb{R}^{n\times d}$, where d is the representation dimension, and m is the number of samples, the (squared) Fréchet distance is obtained using the following estimator:

$$\widehat{\text{Fréchet}}^{\,2}(X,Y) =$$

$$\|X - \mu_Y\|_2^2 + Tr\left(\Sigma_X + \Sigma_Y - 2\left(\Sigma_X\Sigma_Y\right)^{1/2}\right)$$

An initial evaluation of the metric involved calculating the Fréchet distance between a reference sound and the same sound after adding multiple levels of Gaussian noise separately. The results are shown in figure 1.

4 Experiments

In this section we provide details on the experiments, including baselines and ablation study. We train our models using our main dataset, the Arabic Speech Corpus dataset[3], either with or without addition of the one of the auxiliary datasets described

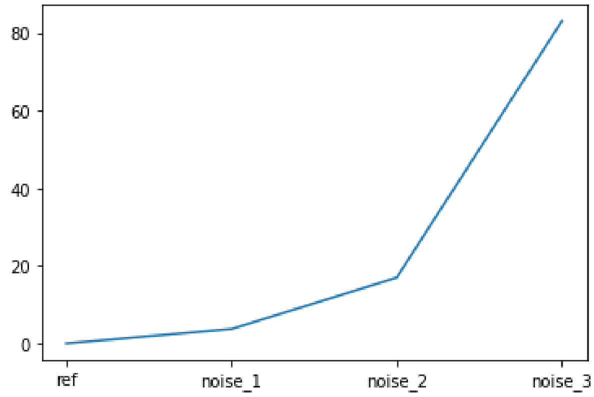

Figure 1: An initial evaluation for the Conditional Fréchet Wav2Vec Distance using different levels of Gaussian noise.

in table 1. In all experiments, unless stated otherwise, the English dataset[1] is used as an auxiliary dataset. The MelGAN model (Kumar et al., 2019) with an amended downsampling schedule was used in all experiments, and we added one additional discriminator when auxiliary datasets were analyzed. Currently, no clear strategies have been developed for GANs with auxiliary data; thus, fine-tuning and training from scratch using both the main and auxiliary datasets seems reasonable and we explored both here.

4.1 Baselines

We compare MelGAN model with a model described by Fahmy et al. (2020) to evaluate its effectiveness for Arabic language synthesis. Based on a modified deep architecture from Tacotron 2 (Shen et al., 2018), the model creates a mel-spectrogram of diacritical Arabic text as an intermediate feature representation, before using WaveGlow (Prenger et al., 2019) as a vocoder to synthesize high-quality Arabic speech. To develop the final model, Fahmy et al. (2020) started from English pre-trained model and fine-tuned using Arabic Speech Corpus dataset[3].

To examine the effectiveness of the additional discriminator (through which the auxiliary data is introduced), we compare the baseline MelGAN with the results obtained with different mixing ratios for the main and auxiliary segments that are passed to this additional discriminator.

4.2 Fine-tuning

In this experiment, we carry out transfer learning in its plain form, i.e. we start with a model pre-trained on an auxiliary dataset and then fine-tune using our main dataset. We use the standard MelGAN architecture (Kumar et al., 2019), with no additional discriminators. The initial pre-training is done on English data[1], followed by fine-tuning on 2 hours of Arabic data[3].

Transfer learning in our setting involves additional challenge that is specific to adversarial models: it seems crucially important to ensure that the min-max game between the generator and discriminator is balanced both during pre-training and fine-tuning. The latter becomes difficult e.g. in a situation when only one of the networks is avialable with pre-trained weights. This unfortunately happens to be the case with MelGAN, whose generator weights are publicly available from official repository[5], but discriminator weights are not shared. Of course pre-training both generator and discriminator from scratch using the English dataset is technically an option, however it is also computationally intensive, and was beyond capacity of our resources. In order to address this issue, we fine-tuned the discriminator alone with the main dataset for 2K steps while fixing the generator weights before fine-tuning the entire model. The discriminator was initially initialized either randomly or using the weights of a pre-trained Arabic discriminator.

4.3 Training GANs with auxiliary data

In this set of experiments we introduce an auxiliary dataset by developing a variant of MelGAN architecture with an additional discriminator. Original discriminators in MelGAN use longer segments than discriminators in GAN-TTS. In training the proposed architecture, we used both the main dataset and a range of auxiliary ones; including an English dataset[1], two Arabic dialect datasets[7], or a low-quality standard Arabic dataset[8]. According to how the auxiliary dataset is introduced to the model, the experiments can be divided into two parts as follows:

Generator with auxiliary segments In this setting, we send to the generator the mel-pectrogram of 512-time steps windows of the auxiliary dataset. The resulting segments are added to the discriminator along with 512-time steps segments subsampled from the audio generated given the main dataset conditioning. Mixing ratio refers to the ratio between these two types of segments.

Extra ground truths for discriminator In this setting, as illustrated in figure 2, we present a way

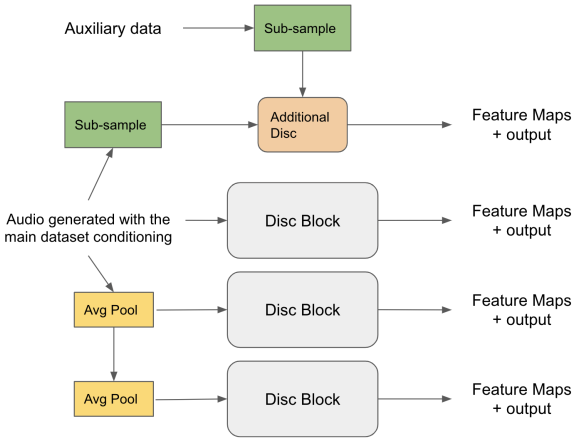

Figure 2: An illustration of the second part of training GANs with auxiliary data experiments, where we pass as extra ground truths for the discriminator.

to incorporate auxiliary data into the model without complicating the generator task. The additional discriminator batches are derived by subsampling 512-time steps segments from audios generated given the main dataset conditioning. The ground truths for part of this segment are replaced with random segments from the auxiliary datasets, but the rest remain fixed. Mixing ratio refers to the ratio between these two types of segments. Through this, we can improve the discriminator adversarial signal being fed back to the generator. A small window was used to concentrate on the high-frequency features. Different segments sizes were tested and 512 was found to perform the best.

4.4 Efficiency analysis of various speech datasets as auxiliary dataset

We present here a discussion of the effects of using various auxiliary datasets. For the comparison, each of the auxiliary datasets is introduced separately as additional ground truths for the extra discriminator with a mixing ratio of 1:1 between the main and the auxiliary datasets respectively.

4.5 Ablations

The proposed model combines several hyper-parameters and we have two approaches to introducing auxiliary datasets to the model; we hence conduct an ablation study to understand how different choices impact the model. In light of our limited resources, the ablation study was carried out using English as the auxiliary dataset, which provided the best results compared to other auxiliary datasets. Our experiments examined different ratios for mixing the Arabic and English segments passed to the extra discriminator. Further, we compared how well the auxiliary dataset worked either as additional ground truths or as a generator input. Finally, we evaluated the effect of smaller segment lengths and the full capacity of the extra discriminator.

4.6 Training Details

All the training is performed on the Arabic Speech Corpus train-set[3] and one of the three additional datasets. The training settings is the same as described in the MelGAN paper (Kumar et al., 2019). The experiments ran on Google Cloud Virtual Machine with a 4-Core CPU and Nvidia T4 GPU. Each model is trained for 500000 steps.

5 Results

This section summarizes all the results of the experiments described in the Experiments section 4. We evaluated the performance on the test set of the Arabic Speech Corpus dataset3 using the MOS and the average of the last five Conditional Fréchet Wav2Vec Distance scores. It is worth noting that the mean of the best and the mean of the last five scores produced almost the same ordering. Also, in all tables and figures, the mixing ratio represents the ratio between main and auxiliary segments respectively we feed to the additional discriminator.

Table 2 presents the quantitative results of the proposed model incorporating the English dataset[1] as additional ground truths for the extra discriminator, as well as the MelGAN (Kumar et al., 2019) model and WaveGlow model (Prenger et al., 2019). The table shows the models that have 4 or less additional signals compared to the MelGAN model. The addition of one segment of the Arabic dataset would result in adding two additional signals: one to the generator's adversarial loss and one to the discriminator's adversarial loss, while the addition of one segment of the English dataset would result in one signal added to the discriminator's adversarial loss. The results show that MelGAN is able to achieve a performance that is comparable to WavGlow in the synthesis of Arabic speech. Furthermore, the study shows that MelGAN + Extra Disc outperforms both MelGAN and WaveGlow models, and adding auxiliary dataset increases the performance even further. MelGAN + Extra Disc and mixing ratio of 1:2 between Arabic and English data sets respectively provided the best performance across all models. Figure 3 shows the importance of adding a mixture of Arabic and English segments compared to the extreme cases.

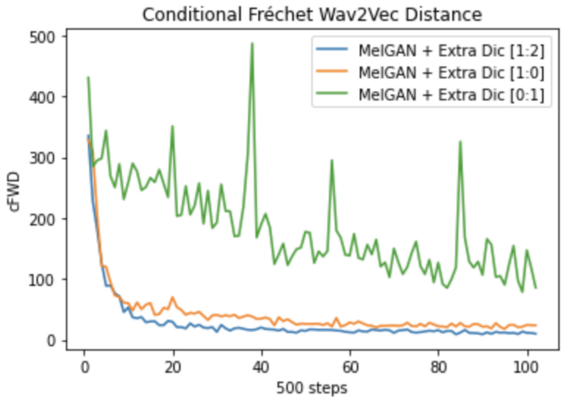

Figure 3: Conditional Fréchet Wav2Vec Distance reported every 500 steps during training of MelGAN + Extra Disc model with three different mixing ratios.

Table 3 represents the quantitative results of using different auxiliary datasets 1 as additional ground truths for the extra discriminator in the proposed model. The mixing ratio between the main and auxiliary datasets was 1:1. The results shows that different language auxiliary datasets (English[1]) with high quality produce better results than the same language or dialects (Standard[8], Tunisian[7] or Libyan Arabic[7]) auxiliary datasets with low or average quality.

FWD	Auxiliary Dataset
27.50	Tunisian_MSA trian
18.64	AMMI_Speech
18.56	Tunisian_MSA test
16.95	LJSpeech

Table 3: Average of the last five Conditional Fréchet Wav2Vec Distance for MelGAN + Extra Disc models trained with different auxiliary datasets fixed mixing ration if 1 : 1. The extra segments is added as an additional ground truths.

Tables 4, 5, 6 shows the results of the ablation study. According to the study, MelGAN + Extra Disc with 1:2 mixing ratio between Arabic[3] and English[1] data sets provided the best performance across all models. As well, adding auxiliary datasets as additional grounds truths in the extra discriminator is better than including the auxiliary dataset in the generator itself. Last but not least, by using full capacity extra discriminator and reducing segment lengths, we would achieve better results than with the current settings.

FWD	How Auxiliary Date Introduced
13.57	Generator with auxiliary segments
11.16	Extra ground truths for discriminator

Table 5: Average of the last five Conditional Fréchet Wav2Vec Distance for MelGAN + Extra Disc models with different ways of introducing the extra segments to the models and finxed mixing ratio of 1 : 2.

FWD	Capacity	Length
22.94	Half	512
18.85	Full	512
13.57	Full	256
10.46	Full	128

Table 6: Average of the last five Conditional Fréchet Wav2Vec Distance for MelGAN + Extra Disc models with different extra discriminator's capacity and segment length and mixing ratio of 1 : 1.

6 Ethical considerations

This paper aims to advance the field of text-to-speech and hence all considerations related to potential nefarious applications of such technology apply to this work. This includes the potential use of such systems to imitate voice of a certain individual in order to present a message that such person has never uttered. We also acknowledge that TTS systems carry a bias towards the dialect/accent of the population whose speech was used as a training data. However, we hypothesise our model might be suitable to counter such effects: as it has been designed for low-resource languages, it might well be used to improve TTS systems for underrepresented dialects or accents of otherwise well-modelled languages, in turn reducing geographical bias affecting certain populations.

Nevertheless, we believe that overall benefits of improved text-to-speech models outweight these and other ethical risks.

7 Conclusion

In this work, we have proposed an extension for MelGAN that utilizes information of auxiliary high-resource languages/dialects to help training of low resource language audio synthesis models. The proposed approach outperformed standard MelGAN

Model	Mixing Ratio	FWD	MOS	95%CI
WaveGlow	–	–	3.13	±0.061
MelGAN	–	18.01	3.10	±0.063
MelGAN + Extra Disc	1 : 0	22.94	3.29	±0.057
MelGAN + Extra Disc	2 : 0	12.15	3.40	±0.056
MelGAN + Extra Disc	1 : 1	16.95	3.55	±0.058
MelGAN + Extra Disc	1 : 2	**11.16**	**3.63**	±0.056
Original	–	–	3.88	±0.061

Table 2: Mean Opinion Score and average of the last five Conditional Fréchet Wav2Vec Distance scores for the MelGAN + Extra Disc models that have 4 or less additional signals compared to the MelGAN model. The extra segments is added as an additional ground truths. Note here, for MOS of WaveGlow model the samples are generated using the predicted mel-spectrogram not the ground truth mel-spectrogram.

English Arabic	0 segments	1 segments	2 segments	3 segments	4 segments
0 segments	18.01	105.51	–	–	–
1 segments	22.94	16.95	**11.16**	27.46	19.80
2 segments	12.15	11.68	17.30	18.27	17.37
3 segments	22.03	13.54	12.24	22.03	16.85
4 segments	13.07	18.59	16.84	18.73	15.41

Table 4: Average of the last five Conditional Fréchet Wav2Vec Distance for MelGAN + Extra Disc models with different mixing ratios. The extra segments is added as an additional ground truths.

model as well as the baseline WaveGlow in both the quantitative and subjective human evaluation. We demonstrated in an ablation study the importance of different components of the system to achieve good results. We hope to see how this approach can help training of the audio synthesis models in the future. Before that, we have trained the MelGAN model for conditional Arabic TTS using a publicly available dataset.

Furthermore, We have proposed a quantitative metric for generative models of Arabic speech that we called Conditional Fréchet Wav2Vec Distance, and demonstrated experimentally that it ranks models in line with Mean Opinion Scores obtained through human evaluation. The metric is based on the available Wav2Vec2ForCTC Arabic speech recognition model. Our quantitative results as well as subjective evaluation of the generated samples showcase the efficiency of our proposed approach for speech generation.

References

Ossama Abdel-Hamid, Sherif Mahdy Abdou, and Mohsen Rashwan. 2006. Improving arabic hmm based speech synthesis quality. In *Ninth International Conference on Spoken Language Processing*.

Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. 2019. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. 2021. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*.

Phat Do, Matt Coler, Jelske Dijkstra, and Esther Klabbers. 2021. A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages. *Proc. Interspeech 2021*, pages 16–20.

Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. 2020. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*.

Thomas Eiter and Heikki Mannila. 1994. Computing discrete fréchet distance.

Fady K Fahmy, Mahmoud I Khalil, and Hazem M Abbas. 2020. A transfer learning end-to-end arabic text-to-speech (tts) deep architecture. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 266–277. Springer.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Younggun Lee, Suwon Shon, and Taesu Kim. 2018. Learning pronunciation from a foreign language in speech synthesis networks. *arXiv preprint arXiv:1811.09364*.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016a. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

AVD Oord, S Dieleman, H Zen, K Simonyan, O Vinyals, A Graves, N Kalchbrenner, A Senior, and K Kavukcuoglu. 2016b. A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Ilyes Rebai and Yassine BenAyed. 2016. Arabic speech synthesis and diacritic recognition. *International Journal of Speech Technology*, 19(3):485–494.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Michael Soloveitchik, Tzvi Diskin, Efrat Morin, and Ami Wiesel. 2021. Conditional frechet inception distance. *arXiv preprint arXiv:2103.11521*.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Henry Sweet. 1877. *A handbook of phonetics*, volume 2. Clarendon Press.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*, 164.

Xin Yi, Ekta Walia, and Paul Babyn. 2019. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552.