



Adversarial Text-to-Speech for low-resource languages

Ashraf Elneima¹, Mikołaj Bińkowski^{2*}

¹Mater's Student, AMMI

²Research Scientist, DeepMind

*Project Supervisor

Objectives

- Train a fast and efficient TTS system for the Arabic language using a publicly available speech dataset.
- Propose an extension to MelGAN model which makes it more amenable to knowledge transfer between languages.
- Propose a quantitative metric for Arabic speech generation based on Fréchet distance.

Introduction

- GAN TTS are highly parallelizable and more suitable to run efficiently on modern hardware.
- For many languages (e.g. Arabic) we lack resources to create sufficiently large labelled datasets.
- Arabic language has a large global population, it is a complex language to model.
- The high-frequency similarities between languages can be exploited to learn better speech synthesis models for low-resource languages.
- We explored raw waveform generation of Arabic using auxiliary data, and taking MelGAN as our baseline model.

Methodology

0.1 Model Architecture

- We used the MelGAN architecture with an amended down-sampling schedule that we found to perform better in our early experiments.
- We introduced auxiliary data to the model through an additional discriminator, designed to operate on short segments of speech to capture high-frequency similarities.
- The additional discriminator accepted a batch of small segments 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.

0.2 Datasets

- Our main dataset was the Arabic Speech Corpus, which is a 2 hours of standard Arabic dataset.
- The auxiliary datasets includes: The LJSpeech English dataset. The Tunisian_MSA Tunisian dialect Arabic dataset. The AMMI_Speech standard low-quality Arabic dataset.

0.3 Evaluation Metrics

- Mean Opinion Score.

- Conditional Fréchet Wav2Vec Distance: Inspired by the DeepSpeech Distances. A pre-trained Wav2Vec2ForCTC Arabic speech recognition model was used in place of the DeepSpeech model to extract high-level features from raw Arabic audio.

Main Experiment

We trained the proposed architecture with the main and the auxiliary datasets. The additional discriminator batches are derived by sub-sampling small segments from the 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.

Results

- MelGAN + Extra Dic trained only with the main data outperforms the MelGAN model. Adding auxiliary dataset increases the performance even further.
- Different language datasets (English) with high quality produce better results than the same language or dialects (Standard or Tunisian Arabic) datasets with low or average quality
- The mixing ratio can extremely affect the results.

Model	Ratio	FWD	MOS	95%CI
+Extra Dic	1 : 0	22.94	3.29	±0.057
+Extra Dic	2 : 0	12.15	3.40	±0.056
+Extra Dic	1 : 1	16.95	3.55	±0.058
+Extra Dic	1 : 2	11.16	3.63	±0.056
MelGAN	—	18.01	3.10	±0.063

Table 1: MOS and average of the last five Conditional Fréchet Wav2Vec Distance scores.

Model	Ratio	FWD	Auxiliary Dataset
+Extra Dic	1 : 1	18.64	AMMI_Speech (Arb)
+Extra Dic	1 : 1	18.56	Tunisian_MSA (Arb)
+Extra Dic	1 : 1	16.95	LJSpeech (Eng)

Table 2: Average of the last 5 Conditional Fréchet Wav2Vec Distance scores for different auxiliary datasets.

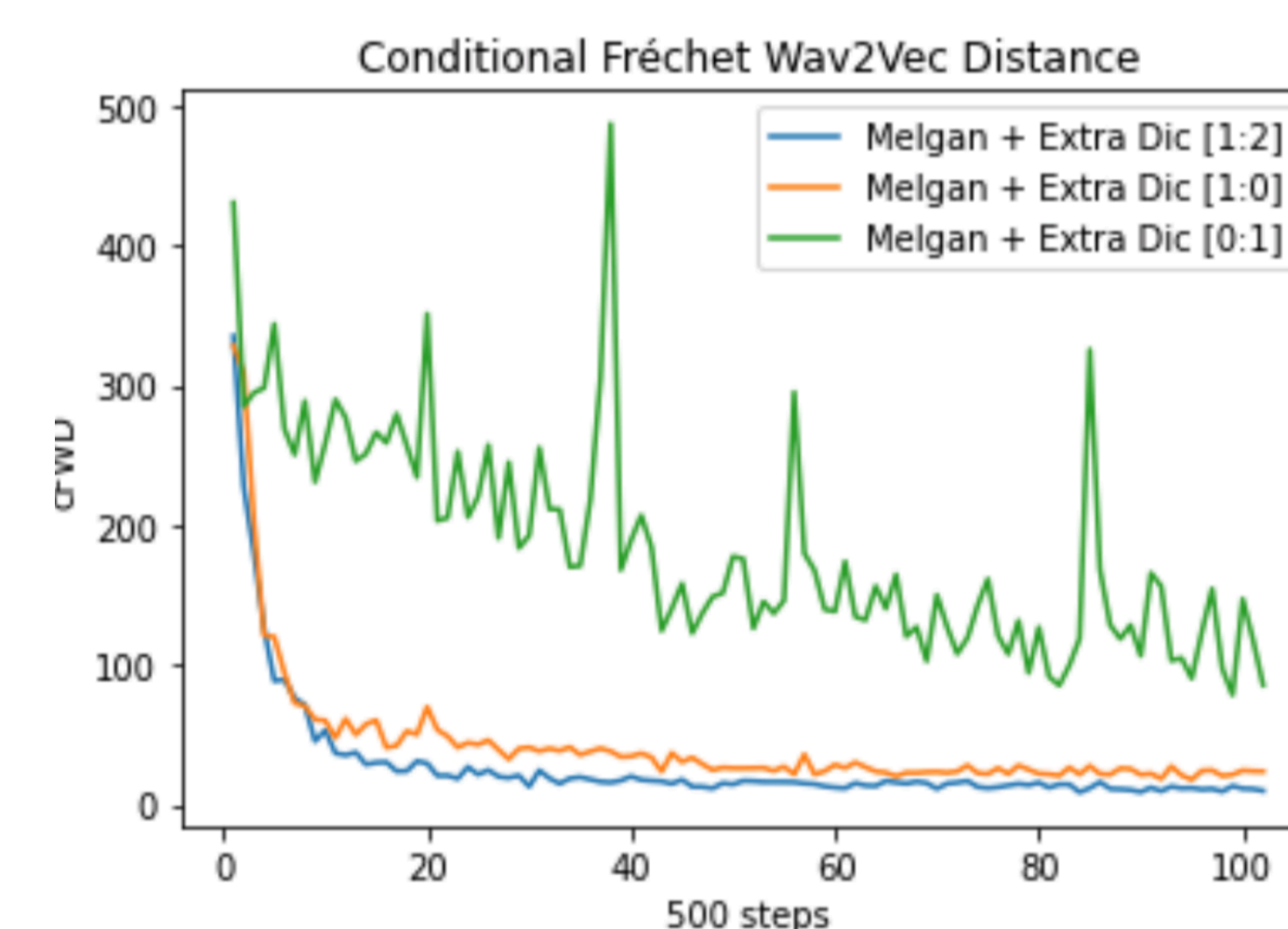


Figure 1: The importance of the mixing ratio between the Arabic and English segments. Distance reported every 500 steps during training