TMiT

MŰEGYETEM 1782

BME FACULTY OF ELECTRICAL ENGINEERING AND INFORMATICS

Emotional TTS

Challenges & Applications

Rami Kammoun
Rami.kammoun@tmit.bme.hu

SmartLab
Intelligent Interactions

Table of contents

01 Problem Statement

Listen closely, does it sound human?

Listen closely, does it sound human?

In terms of Naturalness, TTS has evolved thanks to new deep learning emerging models,

Is it only for English?

For Arabic Languages

Written Text

Diacritized text has become more available

Audios

Audio recordings are more accessible and open source

Published Work

For MSA we have 3 published datasets and 1 for Tunisian

What’s the next step?

Emotionality

02 Emotional TTS

Why is it important?

Healthcare

Accessibility

Technology Development

Challenges in Emotional TTS

Emotional Challenges

Capturing Emotions

Speech Nuances

intonation, pitch, timing, and rhythm

Context Dependant

The words don't necessarily convey the true meaning

Limited Datasets

Low-resource languages still suffer

Challenges in Emotional TTS

Existing Models

Generalization issue

TTS models were not designed for emotional tasks. Finetuning them on multiple emotions proved hard since they are generally pre-trained on one emotion

Quality v.s. speed

FastSpeech2 and Tacotron2 trade-off.

Limited Range of Emotions?

03 Experiments

Challenges with the Tunisian?

Emotional Audios?

For open source data, it is rare to find clips of emotional Tunisian audio recordings for one speaker over lengthy hours.

Annotation?

Since there are no conventions of writing in Tunisian, annotation is basically non-existent.

Diacritized Annotation?

Diacritization is another problem for arabic-based languages, is that it gives them sound which is more expensive to be done.

An example of a diacritized Arabic text

Chosen Datasets

TunArTTS

- Has 3 hours of spoken speech.
- Has 2 main emotions: Neutral & Anger.
- Is manually annotated.

KazEmoTTS

- Has 6 main emotions (Neutral, Angry, Happy, Sad, Scared, and Surprised).
- Has total duration of 74,85 hours.
- Has mainly 3 speakers.

The inspiration behind the Tunisian Experiment

ArTST: Arabic Text and Speech Transformer

Hawau Olamide Toyin,* Amirbek Djanibekov,* Ajinkya Kulkarni, Hanan Aldarmaki
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
{hawau.toyin;amirbek.djanibekov;ajinkya.kulkarni;hanan.aldarmaki}@mbzuai.ac.ae

Abstract

We present ArTST, a pre-trained Arabic text and speech transformer for supporting open-source speech technologies for the Arabic language. The model architecture follows the unified-modal framework, SpeechT5, that was recently released for English, and is focused on

2020), which enable the utilization of large unlabeled datasets for multiple potential downstream tasks. Pre-trained self-supervised models like Wav2Vec2.0 (Baevski et al., 2020), and its multilingual variant (Babu et al., 2022), have mostly replaced traditional acoustic features like MFCCs and filter banks in the speech domain. These pre

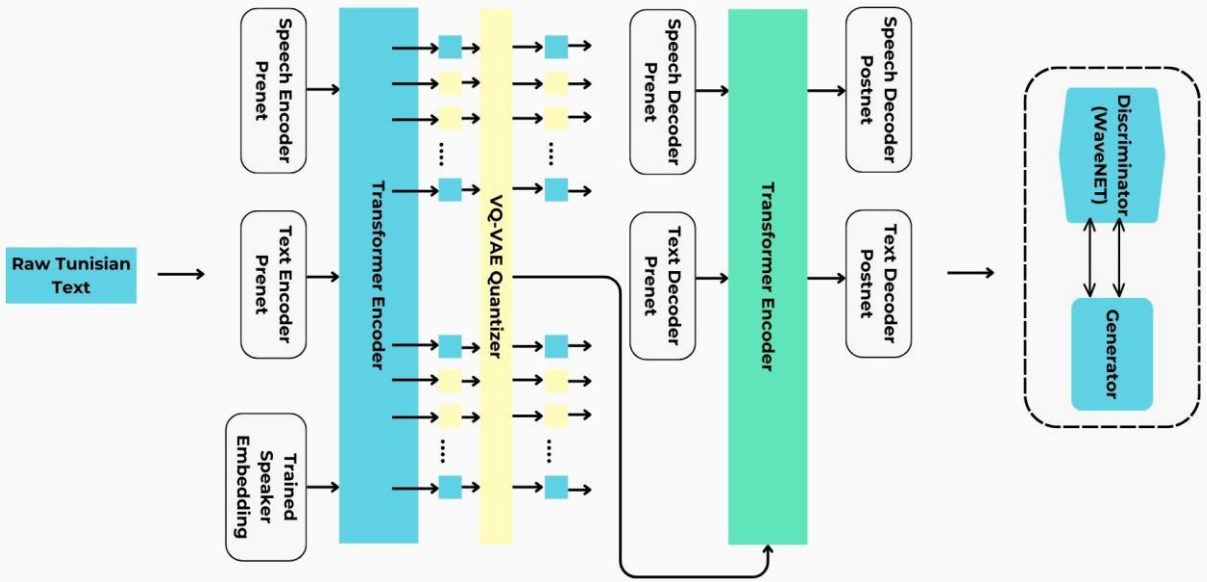
Pre-treatment

1

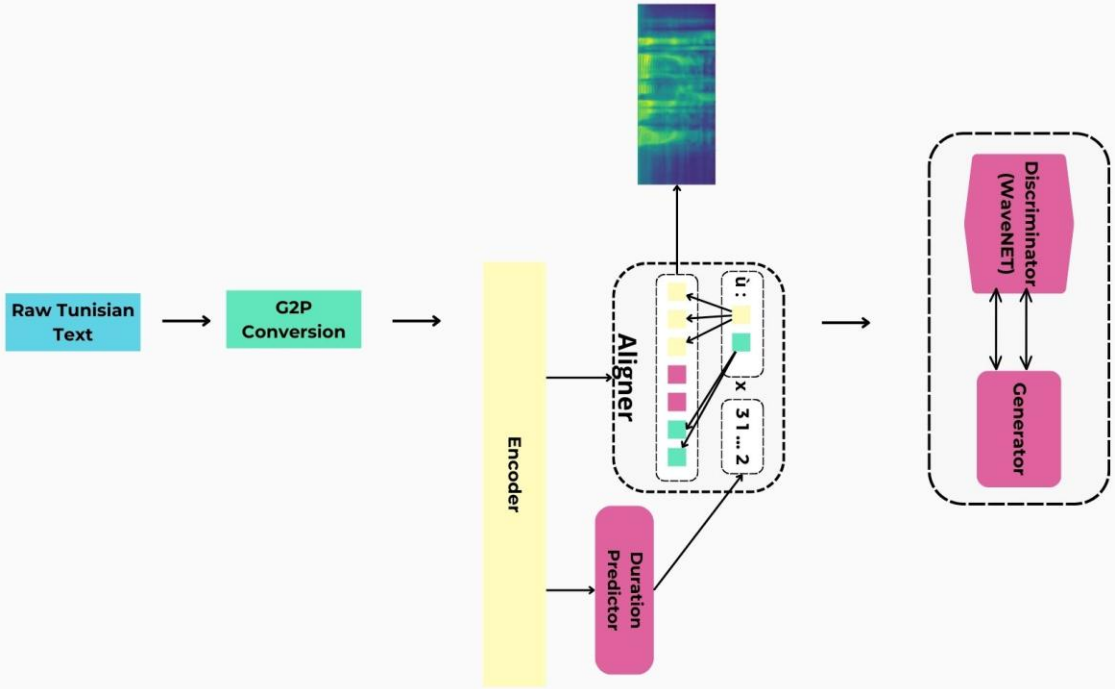
2

3

Newly Proposed System Architecture



Newly Proposed System Architecture

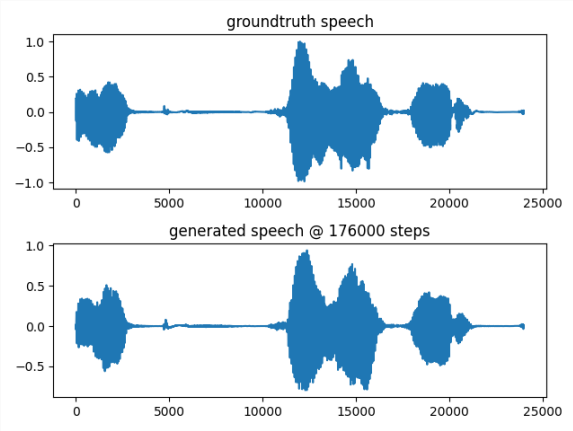


Preliminary Results

TunArTTS

- Model still being Trained and finetuned.
- Could be enhanced to train a speaker embedding model on the Tunisian dialect dataset.
- Retraining a parallel WaveGAN on the new 16KHz audio recordings,

KazEmoTTS

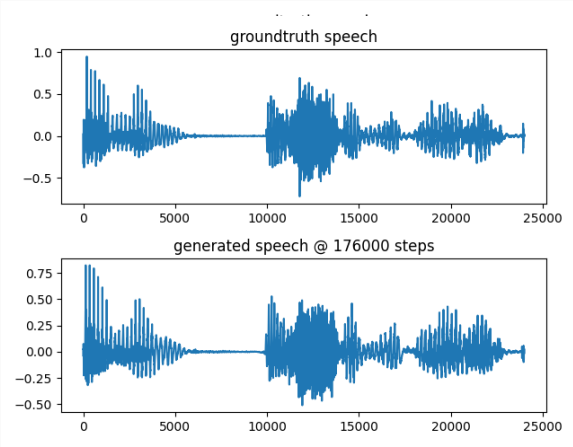


Preliminary Results

TunArTTS

- Model still being Trained and finetuned.
- Could be enhanced to train a speaker embedding model on the Tunisian dialect dataset.
- Retraining a parallel WaveGAN on the new 16KHz audio recordings,

KazEmoTTS



Preliminary Results

TunArTTS

- Model still being Trained and finetuned.
- Could be enhanced to train a speaker embedding model on the Tunisian dialect dataset.
- Retraining a parallel WaveGAN on the new 16KHz audio recordings.

KazEmoTTS

Gen

Ref

Preliminary Results

TunArTTS

- Model still being Trained and finetuned.
- Could be enhanced to train a speaker embedding model on the Tunisian dialect dataset.
- Retraining a parallel WaveGAN on the new 16KHz audio recordings.

KazEmoTTS

- For the **HiFiGAN**, CER is 4,54 for the neutral voice M1.
- For **Parallel WaveGAN**, CER is 4,22 for the neutral voice M1 thus far (196000 steps of training).

What about the “VITS” model?

04 Cocnlusion

Conclusions

SpeechT5

SpeechT5 could be the next road to emotionality for less needed data.

Parallel WaveGAN

Parallel WaveGAN proves to be better than other vocoders for emotional data

Self-Supervision

Self-Supervised models require a huge amount of data that’s unavailable for low-resource languages

Perspectives

End-to-end

The next model needs to be an end-to-end model not a cascade as presented

Multi-lingual

French is the next to be added and other languages will follow suit

Thanks!

Do you have any questions?

Rami.kammoun@tmit.bme.hu
+34 70 40 38 154