Commentator Voice Synthesizer

Fabienne Nicolaij, Hubert Matuszewski, Veronica Valente, Massa Baali

Mentor: Cenk Demiroglu

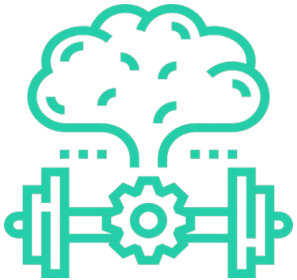
MOTIVATION

- High growth in the gaming industry’s revenue

Reference:
https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide

Problem

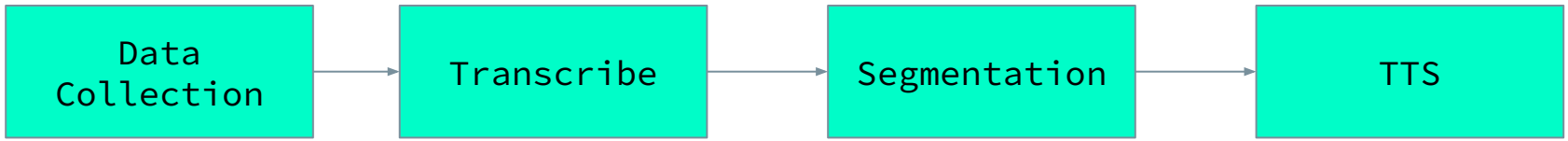
- 1k of voice variations are being recorded for video games (e.g. Fifa)



SOLUTION

- Creating a speech generator for a football commentator in order to make the video game more engaging.

PROJECT PIPELINE



DATA COLLECTION

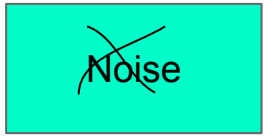
- Youtube (1.5 hours) using youtube_dl
- Issam Shouali (Tunisian Commentator)

TRANSCRIBING

- Large Pre-trained Arabic ASR model
- Vowelization using Farasa API

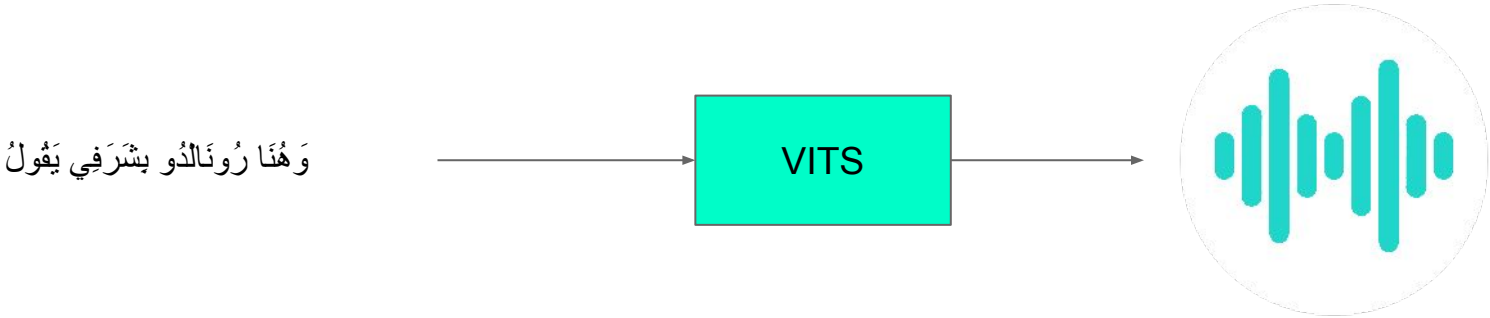
SEGMENTATION

- Voice Analysis Detection (noise, music) labels
- Manual segmentation according to three different classes:{Very Excited, Excited, Neutral}
- CTC Segmentation to align utterances with the audio

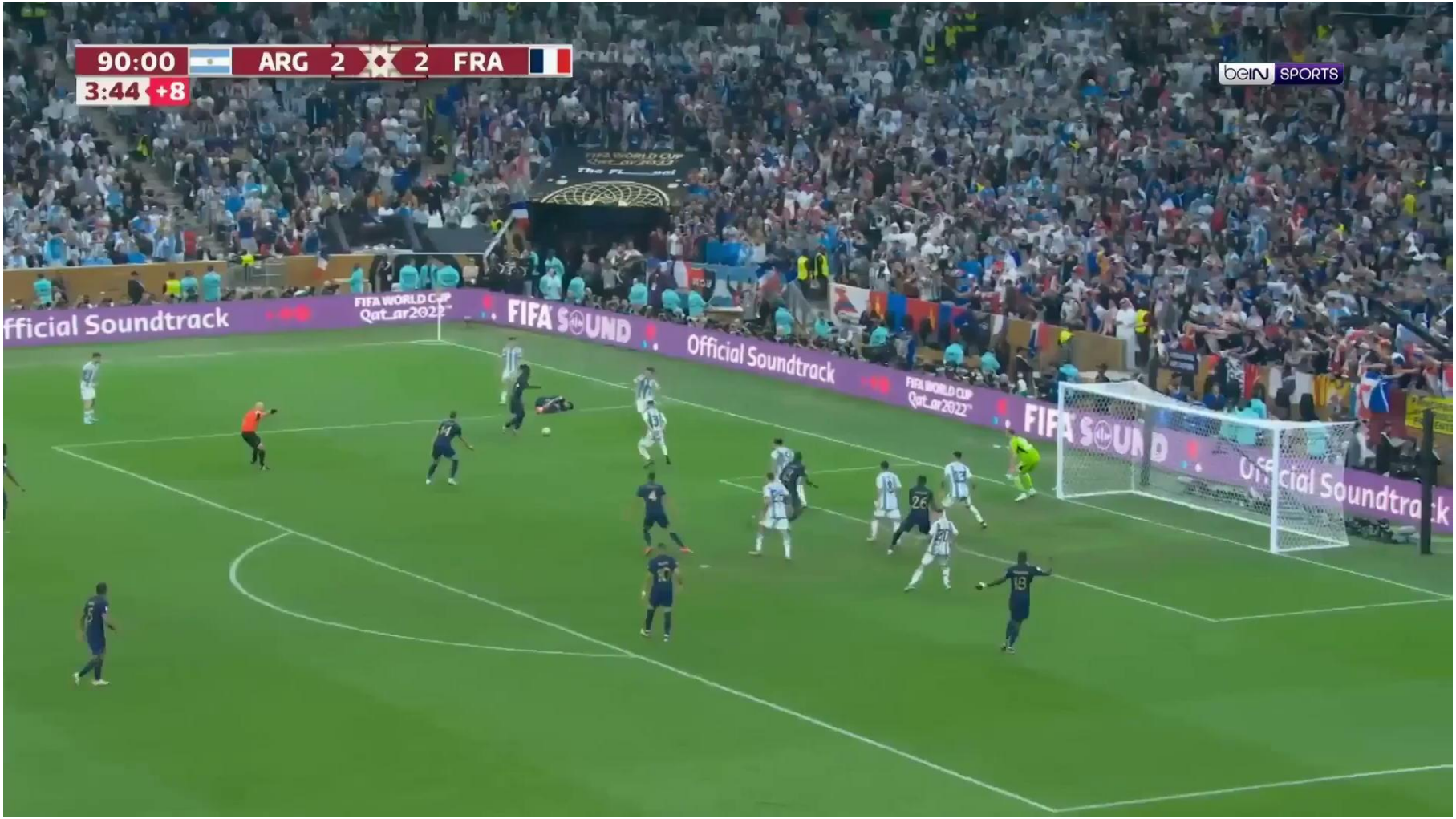


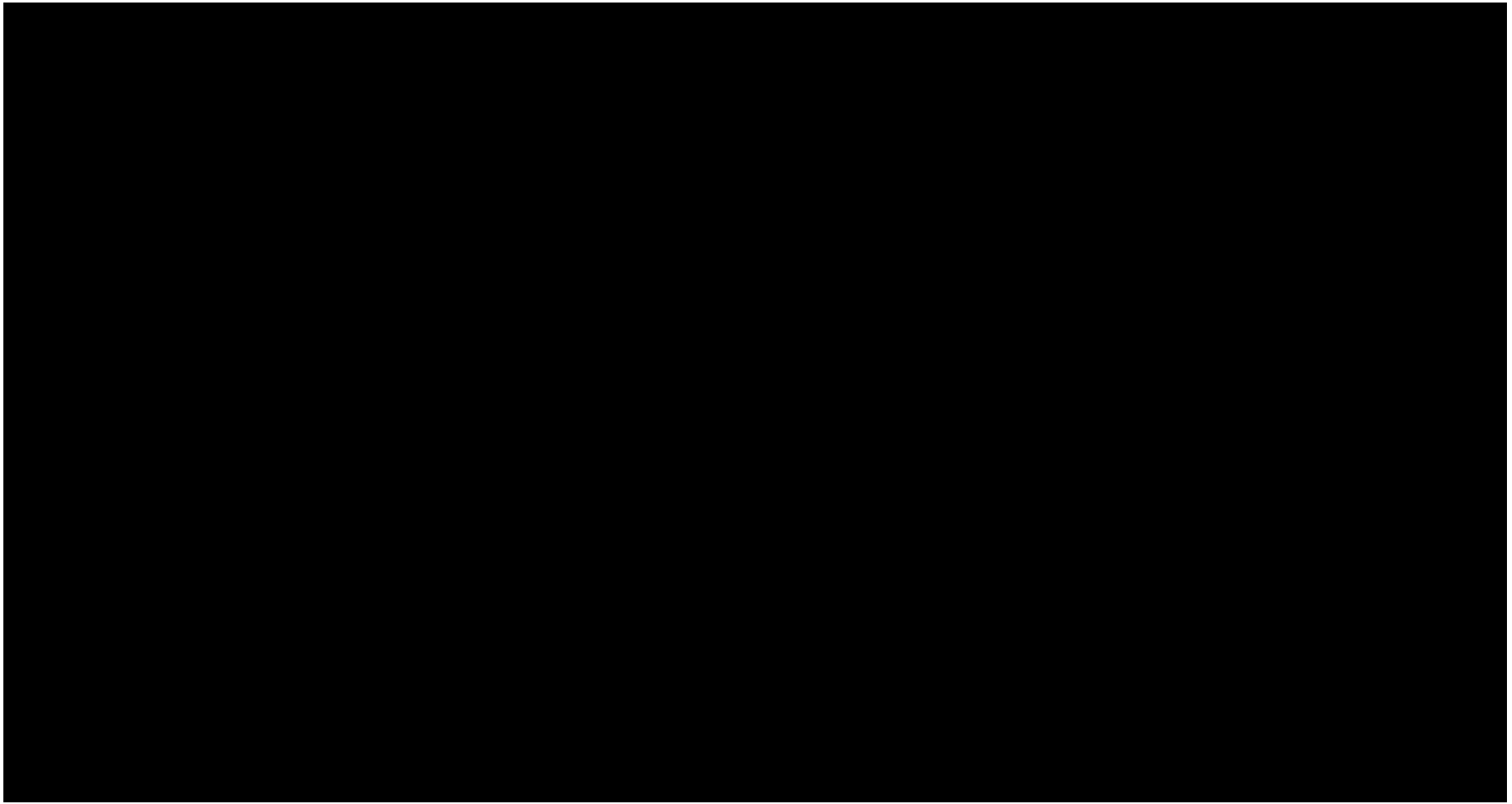
Text to speech

- VITS: which is a text to waveform model that adopts a conditional variational autoencoder with normalizing flows and GAN-based optimizations.
- Pre-trained Arabic VITS model trained on one hour of data
- Finetune it on 15 mins of the commentator speech



DEMO


90:00
ARG 2
2 FRA
3:44 +8
beIN SPORTS
Official Soundtrack
FIFA WORLD CUP Qatar2022
FIFA SOUND



Challenges & Limitations

- Inconsistency in the pitch of the player’s names (Messiii, Messi, Meeeeeessi)
- Noisy data
- Data annotation
- Transcription is inconsistent e.g. (Tunisian dialect, and French words)

Thank you!