01

HOW TEXT TO SPEECH WORKS?

@learn.machinelearning

Text

Speech

WHAT IS TEXT TO SPEECH(TTS)?

- Speech synthesis is simply a form of output where a computer or other machine reads words to you out loud in a real or simulated voice played through a loudspeaker; the technology is often called text-to-speech (TTS).
- Nowadays the goal of TTS is not to simply have machines talk, but to make them sound like humans of different ages and gender.

HOW TO MEASURE QUALITY??

Intelligibility

The quality of the audio generated, or the degree of each word being produced in a sentence.

Naturalness

The quality of the speech generated in terms of its timing structure, pronunciation and rendering emotions.

Preference

preference and naturalness are influenced by TTS system, signal quality and voice, in isolation and in combination.

Comprehensibility

The degree of received messages being understood.

DIFFERENT TEXT-TO-SPEECH SYSTEMS

Concatenative TTS

This technique relies on high-quality audio clips(or units) recordings, which are then combined together to form the speech.

Parametric TTS

This method generates speech by combining parameters like fundamental frequency, magnitude spectrum etc. and processing them to generate speech.

Deep Learning

DNN maps the input texts to the output audio with some approximation functions. It should not use any hand engineered features, and rather learn new high dimensional features to represent what makes speech, human.

05

learn.machinelearning

Text input

Voice input

Text encoder

Speaker encoder

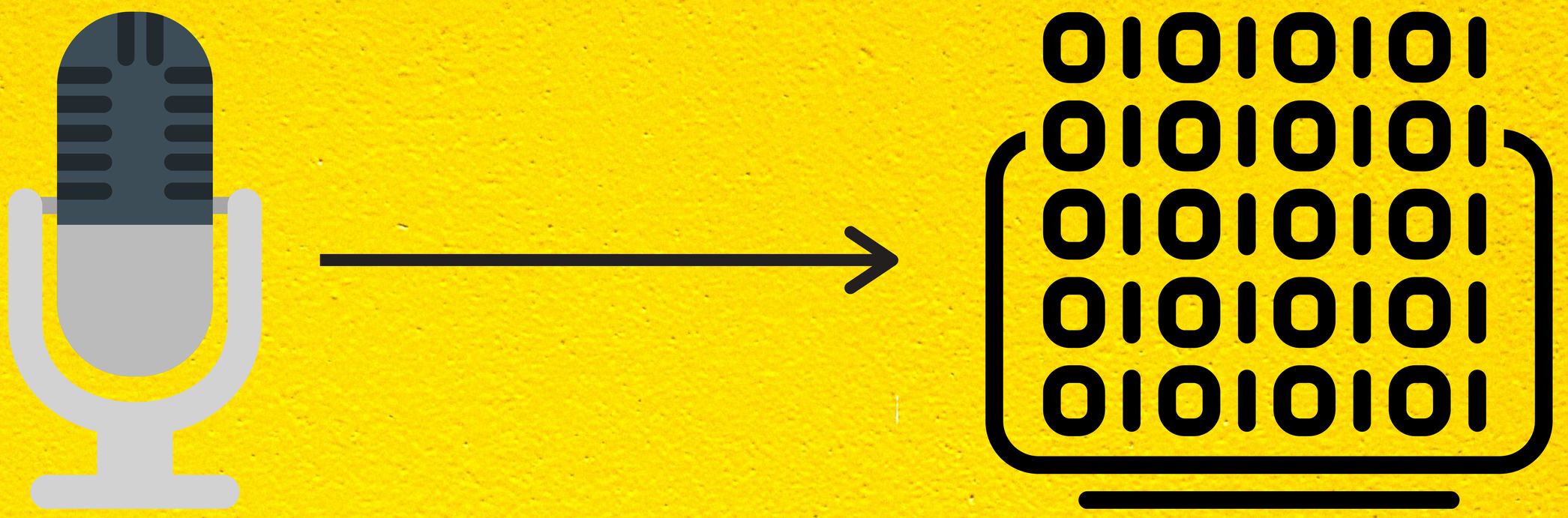
combine

Decoder

Vocoder

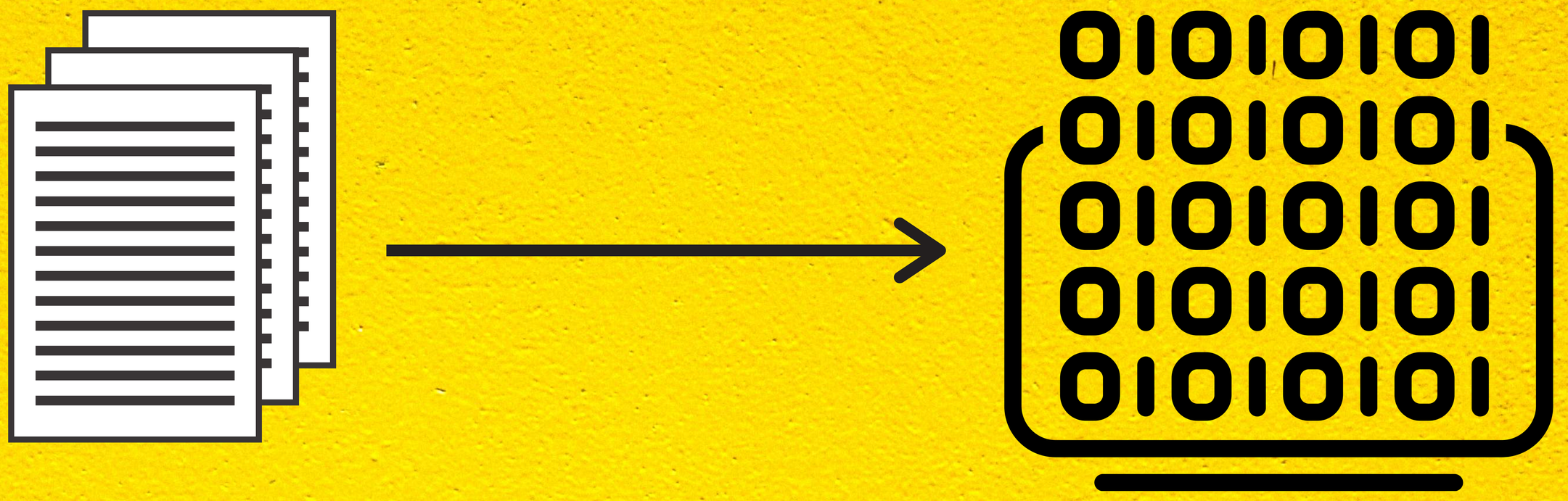
Audio output

PROJECT FLOW (VOICE CLONE)



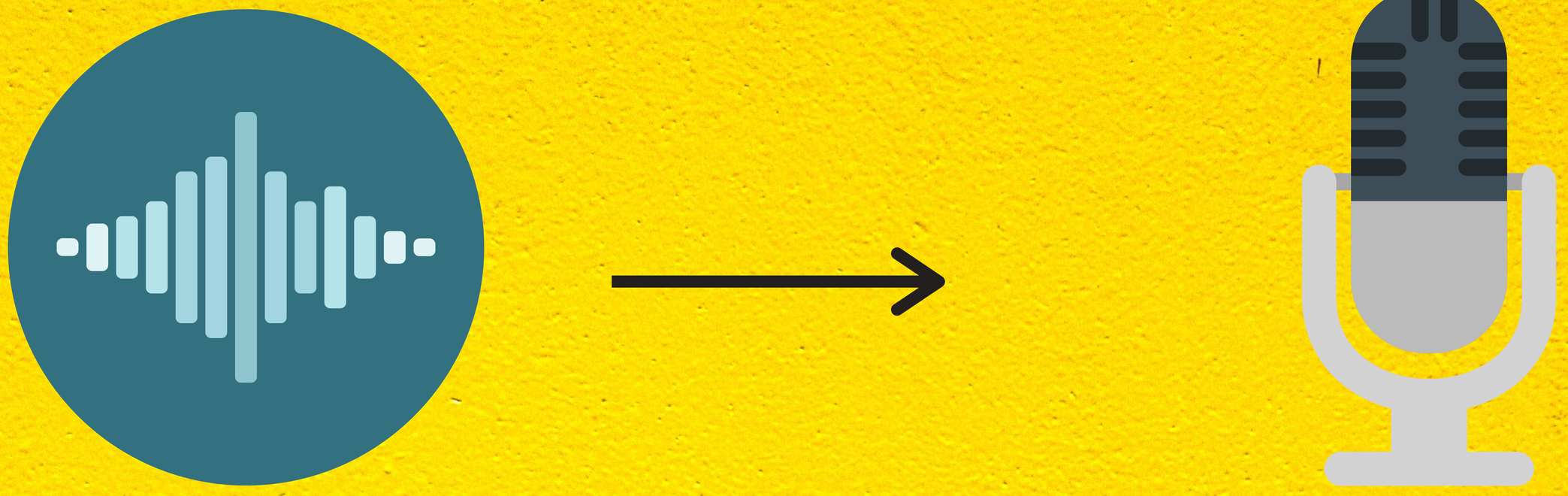
SPEAKER ENCODER

- It take's some input audio of a given speaker, and output an embedding that captures how the speaker sounds.
- The speaker encoder concentrates only on the voice of the speaker, e.g., high/low pitched voice, accent, tone, etc. It doesn't care about the words or background noice.
- Best speaker encoder method (GE2E)



SYNTHESIZER (TEXT ENCODER + COMBINE + DECODER

- The goal of the Synthesizer is to convert the input text into mel spectrograms.
- Goal of text encoder is Given a piece of text, it encode's the text into a vector representation.
- Later it combines both vector representations (text and speaker) and pass that to the decoder (normal NN or attention models).
- Finally this decoder will generate the mel spectrograms as output.
- The mel spectrogram is compared to the original target to create a loss, which is then optimized
- Best method is Tacotron2



VOCODER

- The goal of vocoder is to generate raw audio waveforms from the mel spectrograms.
- Best Attention Methods are - Guided Attention, Forward Backward Decoding, Graves Attention, Double Decoder Consistency.
- Best vocoder methods are - MelGAN:, MultiBandMelGAN, GAN-TTS discriminators.

Most used TTS models are WaveNet, Tacotron, Deep Voice, Parallel WaveNet, VoiceLoop, SV2TTS, etc.....

09

learn.machinelearning

APPLICATIONS OF TEXT-TO-SPEECH

- Text to speech for learning a new language
- Text to speech for people with visual impairment
- In Gaming — Talking Avatars
- Notification Systems
- Reading Applications
- voice to IoT
- E-Learning
- MANY MORE....

RESOURCES

CLICK THE LINKS TO GET RESOURCES

@learn.machinelearning

- Find all resources here