<https://scholar.google.com/scholar?as_ylo=2018&q=text+to+speech+voice+clonin>

<https://arxiv.org/pdf/2104.01818.pdf> -------1

THE MULTI-SPEAKER MULTI-STYLE VOICE CLONING CHALLENGE 2021

* The Multi-speaker Multi-style Voice Cloning Challenge (M2VoC) aims to provide a common sizable dataset as well as a fair testbed for the benchmarking of the popular voice cloning task.
* Specifically, we formulate the challenge to adapt an average TTS model to the stylistic target voice with limited data from target speaker, evaluated by speaker identity and style similarity

SUBMITTED SYSTEMS AND MODELS

1. . Acoustic model
2. Vocoder
3. Speaker and style modelling

DATA AND TRACKS

four audio/text data sets to the participants at different stages of the challenge. All audio data is mono, 44.1KHz sampling rate, 16 bits, equipped with transcripts.

1. Multi-speaker training set (MST)
2. Target speaker validation set (TSV)
3. Target speaker test set (TST)
4. Test text set (TT)

DRAW BACKS:

. The challenge demonstrated the performance of current voice cloning technologies – with the advances of deep learning.

. few-shot voice cloning has achieved reasonably good performance but one-shot voice cloning is still an unsolved problem.

. In real-world voice cloning applications, low-quality (noisy) audio and time/cost constrains of training/adaptation/inference are also important factors that cannot be ignored.

[https://www.isca-speech.org/archive\_v0/Interspeech\_2020/pdfs/1464.pdf -------2](https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/1464.pdf%20-------2)

Multi-Lingual Multi-Speaker Text-to-Speech Synthesis for Voice Cloning with Online Speaker Enrolment

* Recent studies in multi-lingual and multi-speaker text-to-speech synthesis proposed approaches that use proprietary corpora of performing artists and require fine-tuning to enrol new voices.
* we investigate a novel approach for generating high-quality speeches in multiple languages of speakers enrolled in their native language.
* text-to-speech (TTS) synthesis methods……

MODELS PROPOSED:

* e grapheme-to-phoneme mode
* phoneme duration model
* segmentation model
* fundamental frequency estimation model and synthesis model.
* Deep Voice [2] presents a neural TTS system which replaces each separate component with a neural net-based model.
* uses an attention mechanism to convert a sequence of text directly to its corresponding sequence of vocoder features, from which speech audios may be generated using a vocoder.
* 1Actually “end-to-end” here only means that both Char2Wav and Tacotron generate vocoder features, not speech audios, from some representation of input texts.
* The multi-task learning approach and duplicates the output layer for each of its training speakers so that each speaker is trained with its own speaker-dependent output layer while sharing other hidden layers in the model.

MODEL PROPOSED:-

1. Inputs: Phoneme, Tone and Stress Embeddings
2. Speaker Encoder
3. Mel-spectrogram Synthesizer
4. WaveNet
5. Synthesis of Native and Accented Speech

DRAW BACKS:

* In this paper, we investigate a multi-lingual and multispeaker TTS approach to synthesize high-quality speech in three languages and speakers enrol in their own native speech
* . We further find that the WaveNet could be trained in any of the supported languages in this paper and then used to synthesize speech in the other languages well.

<https://arxiv.org/pdf/2005.08484.pdf--------3>

Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding

* On account of growing demands for personalization, the need for a so-called few-shot TTS system that clones speakers with only a few data is emerging.
* To address this issue, we propose Attentron, a few-shot TTS model that clones voices of speakers unseen during training.

SPECIAL ENCODERS:

* A fine-grained encoder extracts variable-length style information via an attention mechanism.
* coarse-grained encoder greatly stabilizes the speech synthesis, circumventing unintelligible gibberish even for synthesizing speech of unseen speakers.

INDEX TERMS:

* few-shot
* text-to-speech (TTS)
* neural TTS
* multi-speaker modelling
* speaker embedding.
* We propose an attention mechanism that finds only the relevant positions among the audio frames of the multiple references. It allows the model to take any number of the reference samples, and the quality improves with more reference samples.
* We compare the proposed model with state-of-the-art methods for multi-speaker TTS that can clone unseen speakers

ABLATION STUDY:

* y. We evaluate key components of the proposed model by tweaking Attentron(8-8). First, we further analyze the impact of multiple reference inputs.
* We address the impact of the coarse-grained encoder. Without coarse-grained encoder (denoted as w/o CE), it loses its stability to generate an intelligible speech accompanying 35 and 143 attention collapse counts for seen speakers and unseen speakers, respectively.