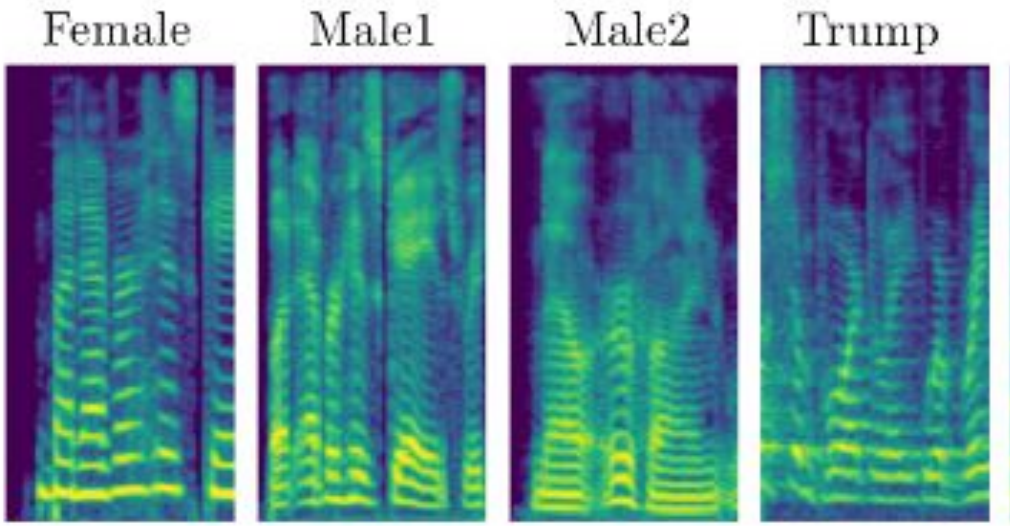
Voice embeddings

tts, vc, vc

https://arxiv.org/pdf/2004.11284.pdf



voice features

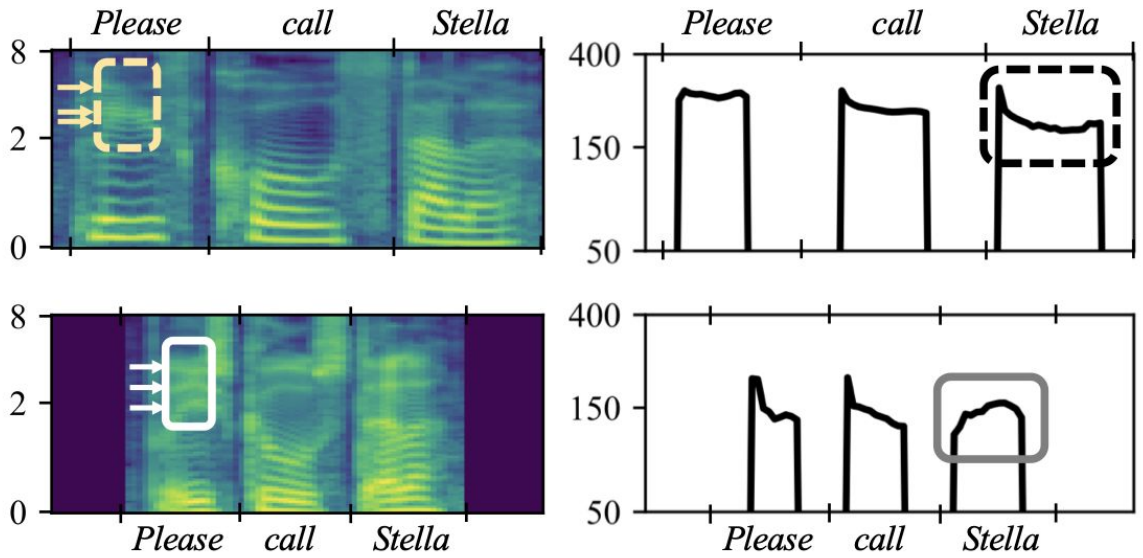
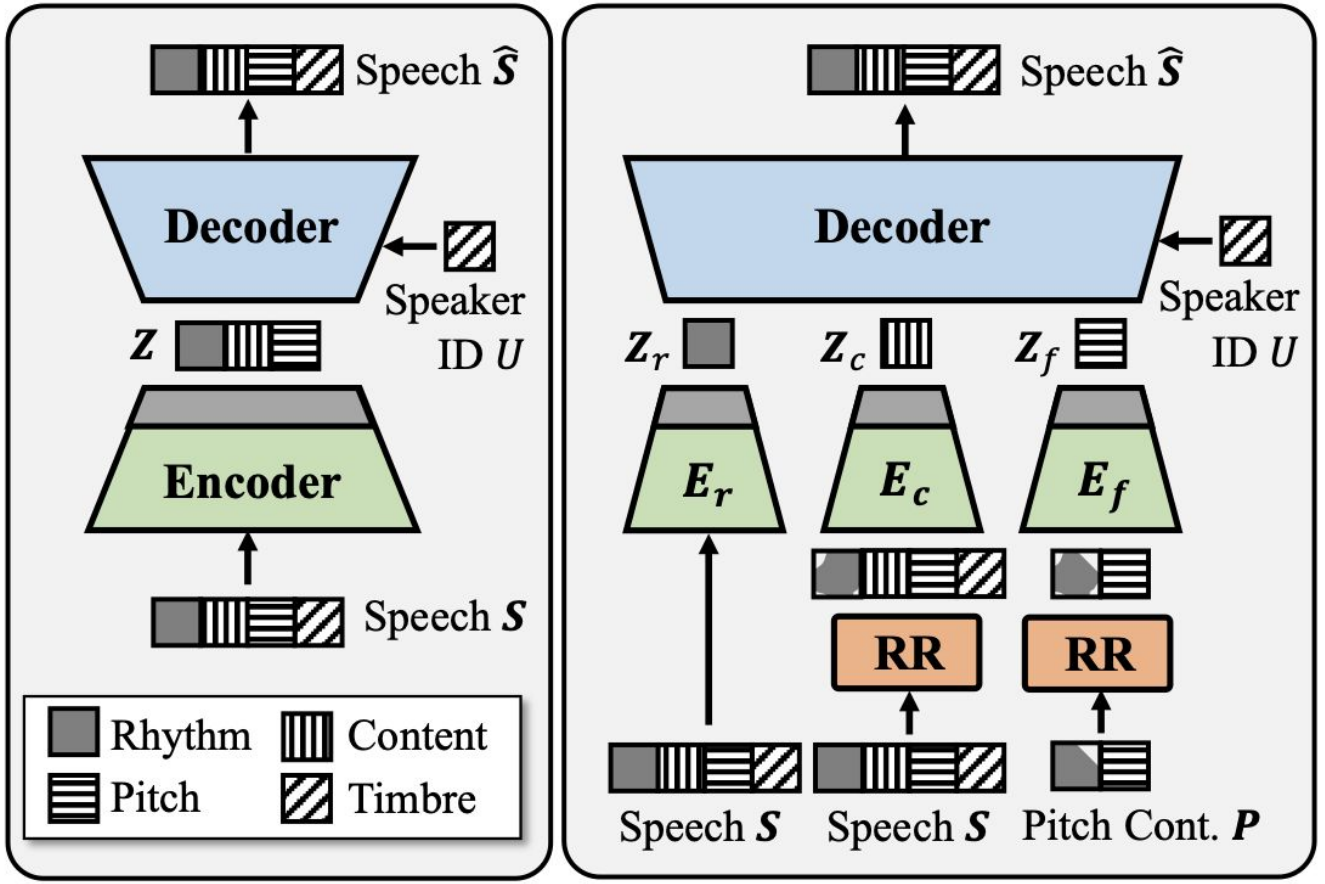


Figure 1. Spectrograms (left) and pitch contours (right) of two utterances of the same sentence *'Please call Stella'*. The left rectangle marks highlight the formant structures of the phone *'ea'*. The arrows mark the frequencies of the second, third and fourth formants. The right rectangle marks highlight the pitch tones of the word *'Stella'*.

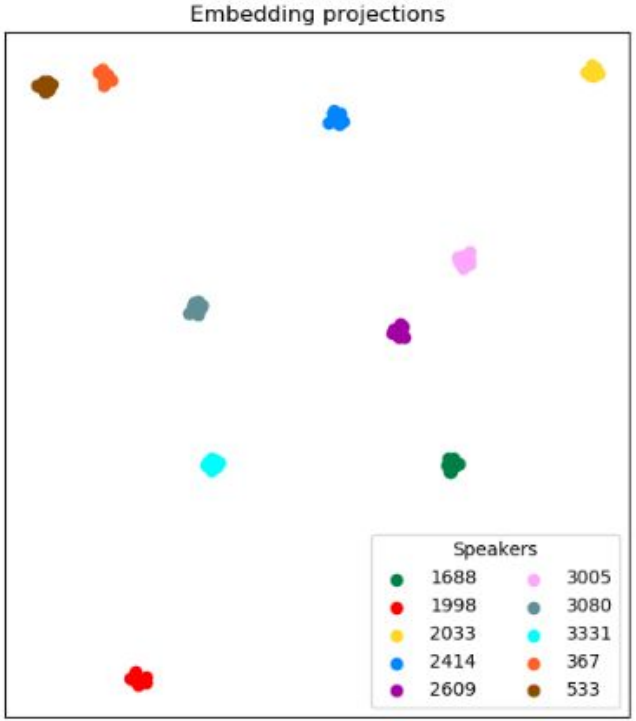
https://arxiv.org/pdf/2004.11284.pdf

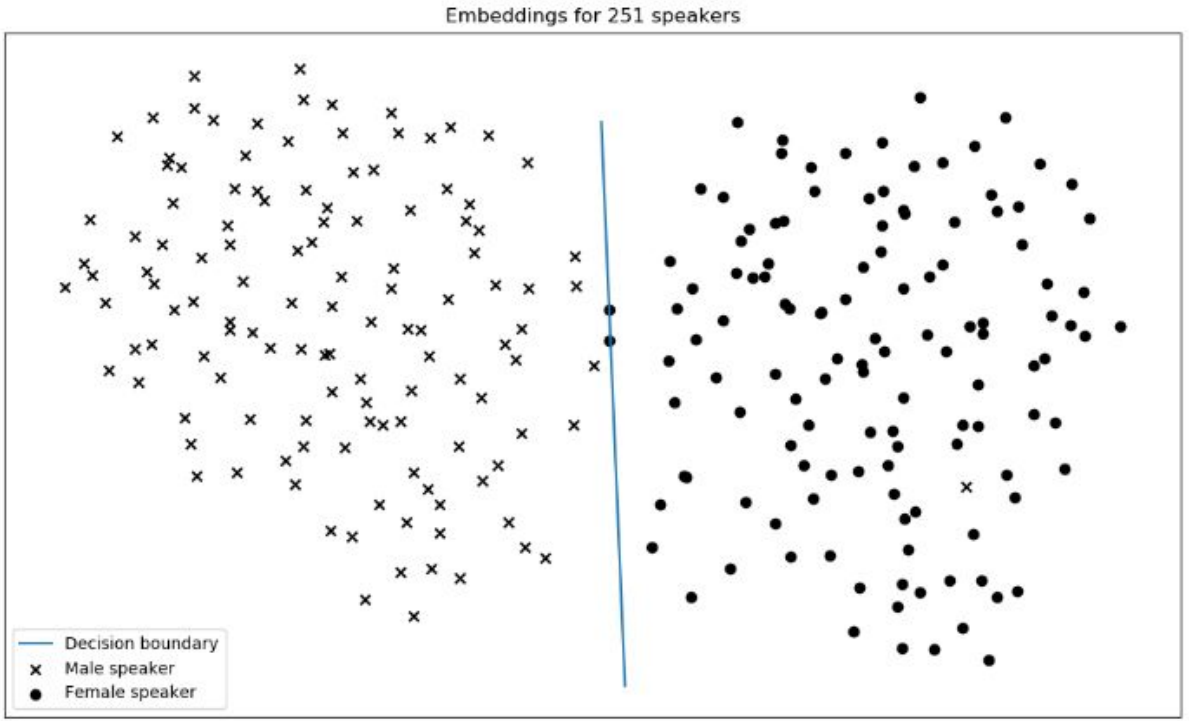


(a) AUTOVC (b) SPEECHSPLIT

voice embeddings

https://github.com/resemble-ai/Resemblyzer





text to speech


text
text2spectro
spectro2wave


vocoder
text
text2spectro
spectro2wave

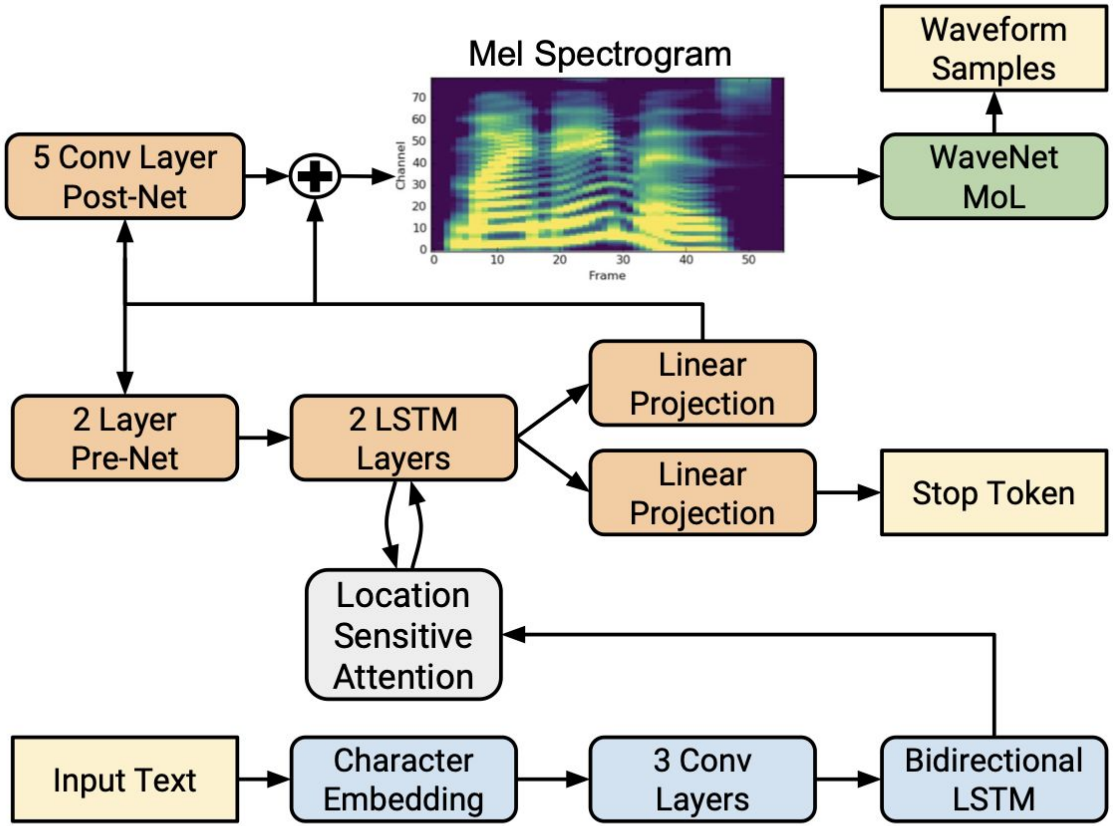
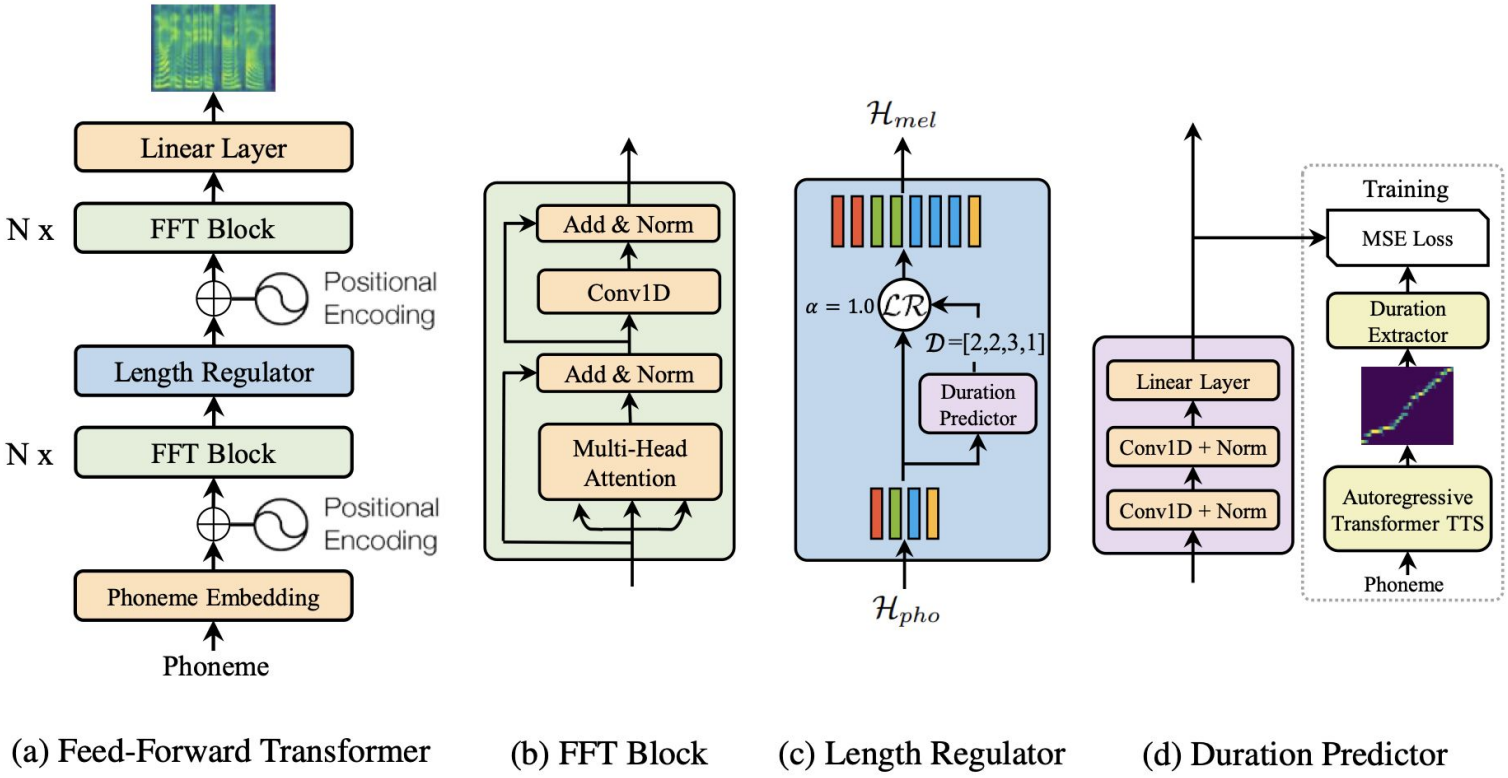


Fig. 1. Block diagram of the Tacotron 2 system architecture.

https://github.com/NVIDIA/tacotron2

https://arxiv.org/pdf/1905.09263.pdf



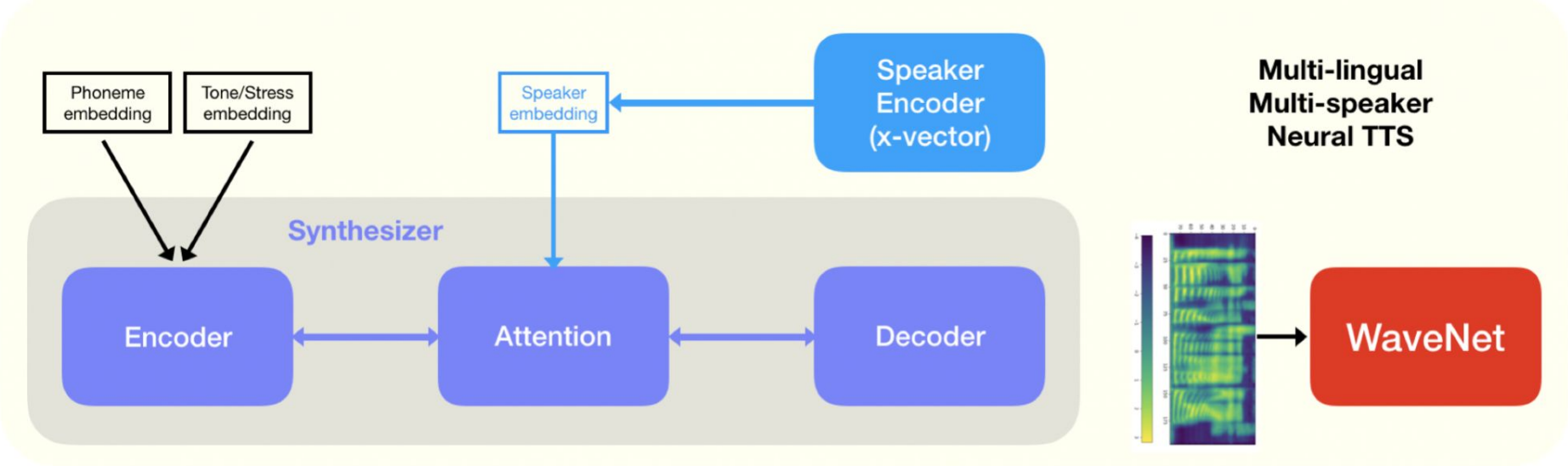
(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Figure 1: The overall architecture for FastSpeech. (a). The feed-forward Transformer. (b). The feed-forward Transformer block. (c). The length regulator. (d). The duration predictor. MSE loss denotes the loss between predicted and extracted duration, which only exists in the training process.

https://github.com/espnet/espnet

voice cloning

voice embeddings + tts = voice cloning

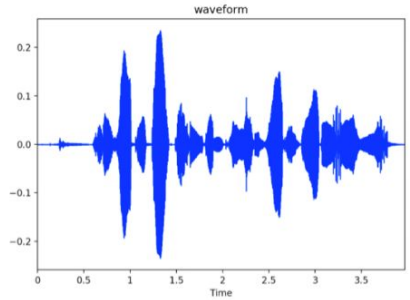


https://github.com/espnet/espnet

https://github.com/CorentinJ/Real-Time-Voice-Cloning

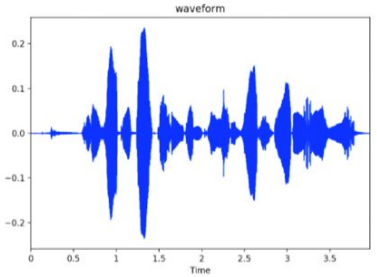
voice conversion

https://github.com/andabi/deep-voice-conversion



Net1: phoneme classifier

Net2: speech synthesizer



A's Waveforms **Speech Recognition** **Speech Synthesis** **B's Waveforms**

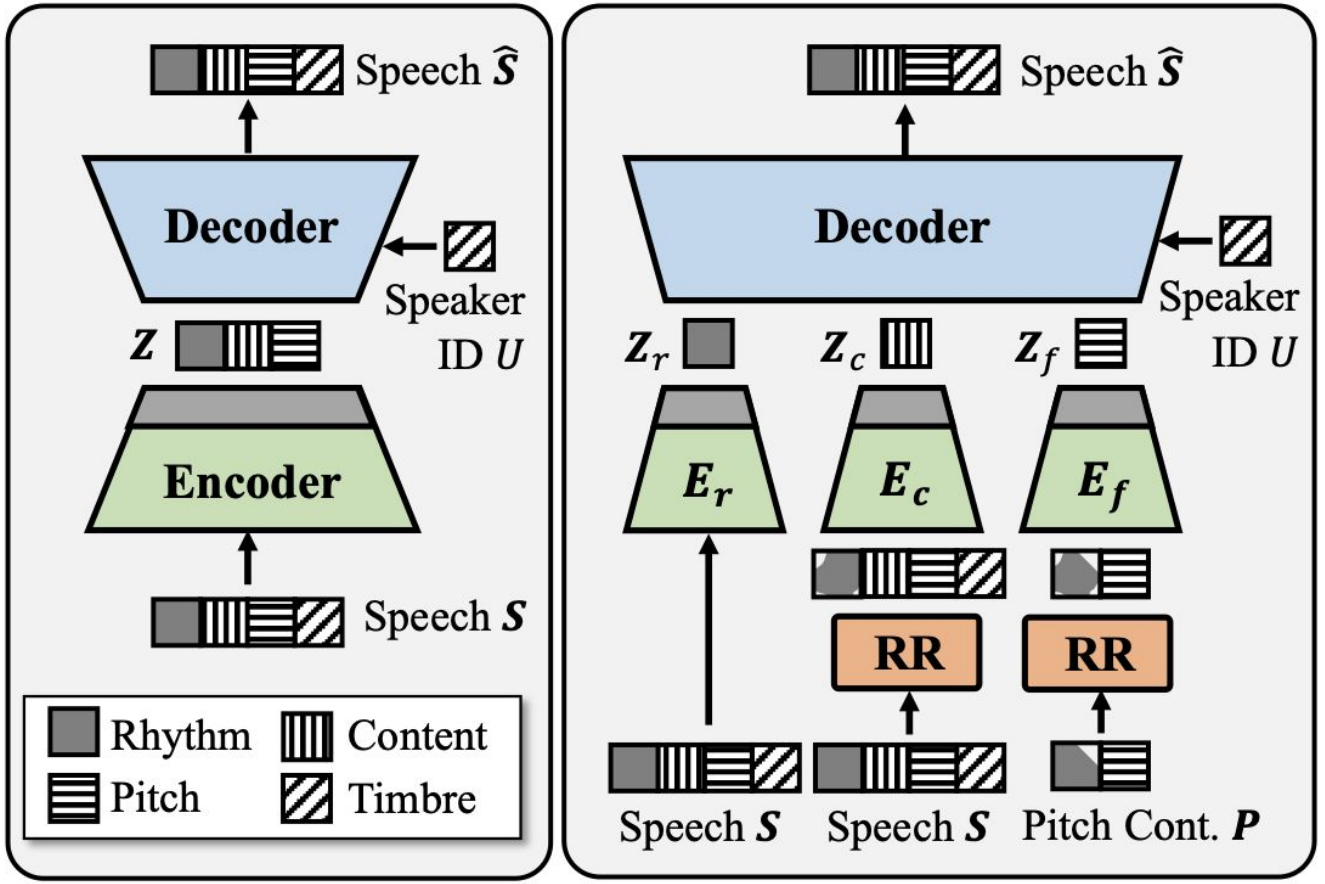
Train1 \w small parallel dataset

Train2 \w large non-parallel dataset

"My name is Avin!"

"My name is Avin!"

https://arxiv.org/pdf/2004.11284.pdf



(a) AUTOVC (b) SPEECHSPLIT

Генерация спектрограмм

1. CycleVAE — хитро тренируемый VAE. Уже старый подход. есть код
2. VQVAEExtended — сжимает фичи “текста речи” в конечное множество + использует unsupervised CPCLoss. есть код
3. AgainVC — использует InstanceNorm, чтобы очищать фичи звука от свойств голоса говорящего. есть код
4. NoiseVC — по-факту объединяет (2) и (3), нет кода
5. AttentionEmbedding — attention на вектор “голоса”. Можно встроить в (4)
6. SpeechSplit — новый подход: делим речь на 4 составляющие, а потом восстанавливаем, есть код
7. GAZEV — GAN со speaker-embedding’ами из коробки. Крутое демо, кода нет

1. PhoneticPosteriograms — end-to-end, PWGAN внутри, использует фичи из ASR и adversarial loss’ы, есть код

Спектрограммы в звук

1. MelGAN — старый, есть код
2. PWGAN — новый, все последние статьи его используют, есть код