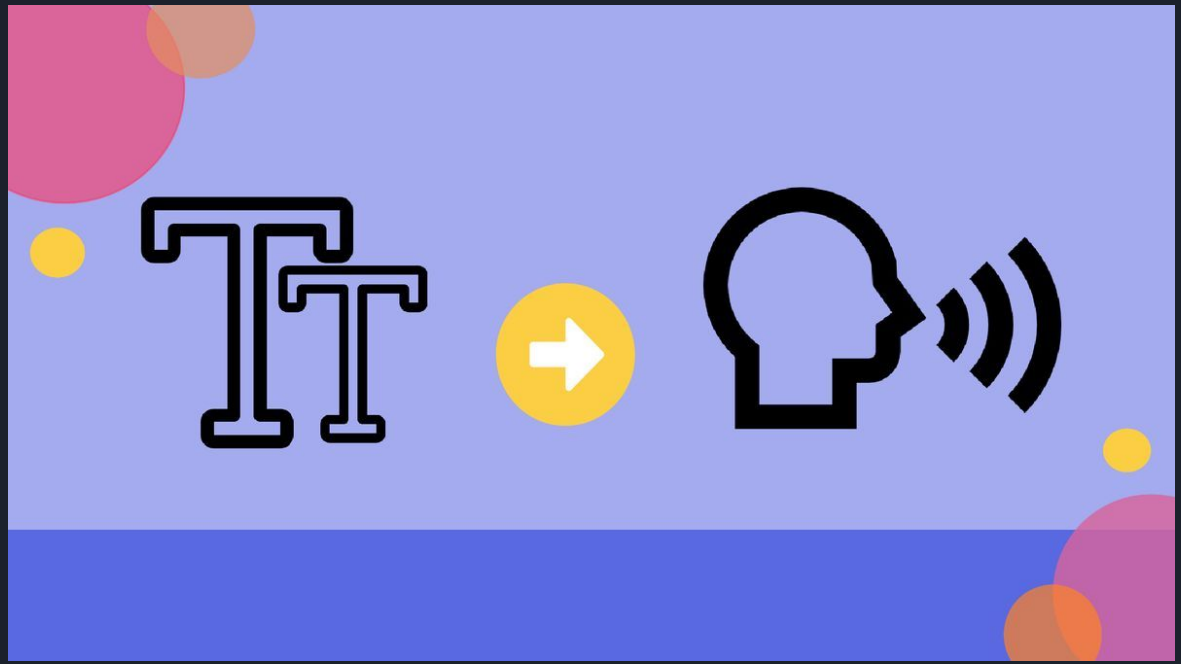
Text To Speech Synthesis

Miller Hickman & Cash Close

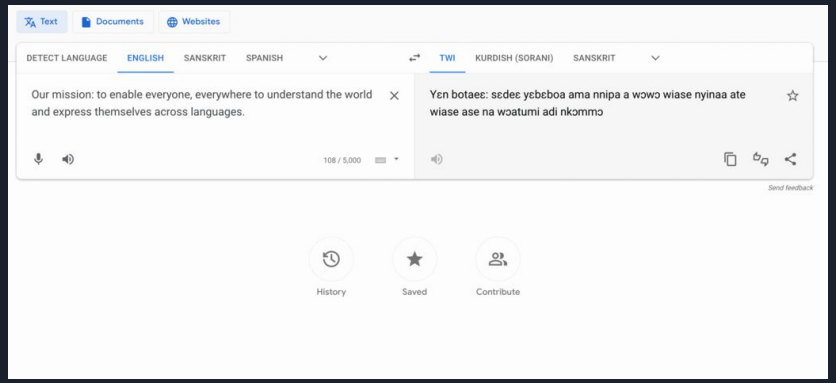
Text To Speech Synthesis



TTS Example

Why is TTS important?

- Visually impaired
- Reading comprehension
- Learning Disabilities (Young)
- Language translation
- Voice assistants







Key Stages

- Text Analysis - Generate linguistic features from text
- Acoustic Models - Generate acoustic features from linguistic features
- Vocoders - Generate waveform from acoustic features

Text Analysis

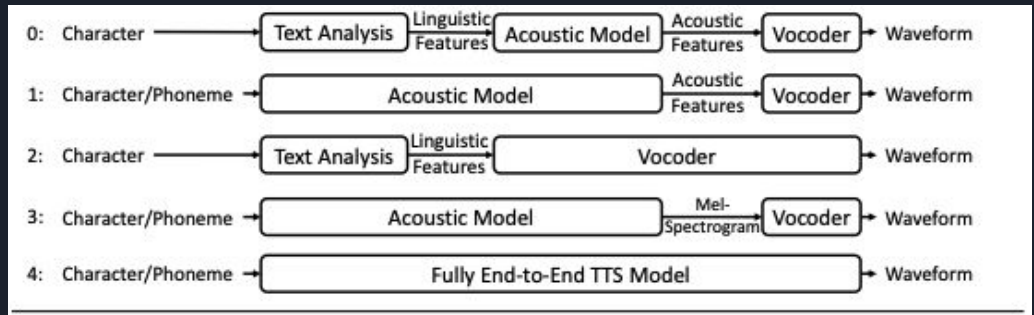
- Word Segmentation
  - Necessary for languages with character-based words (e.g. Chinese)
- Part of Speech (POS)
  - Tagging each word's POS allows for better prosody prediction and G2P accuracy
- Prosody Prediction
  - Prosody encapsulates intonation and other performative aspects of speech
  - Allows for more realistic speech synthesis
- Grapheme to phoneme Conversion (G2P)
  - Graphemes are the different ways a phoneme can be represented in text
    - Ex: /k/ can be written as "c", "k", "ck", "qu", "ch"
  - Generate phonemes from text graphemes
  - Typically used as a safety net for words that are not in a models vocabulary

Acoustic Features

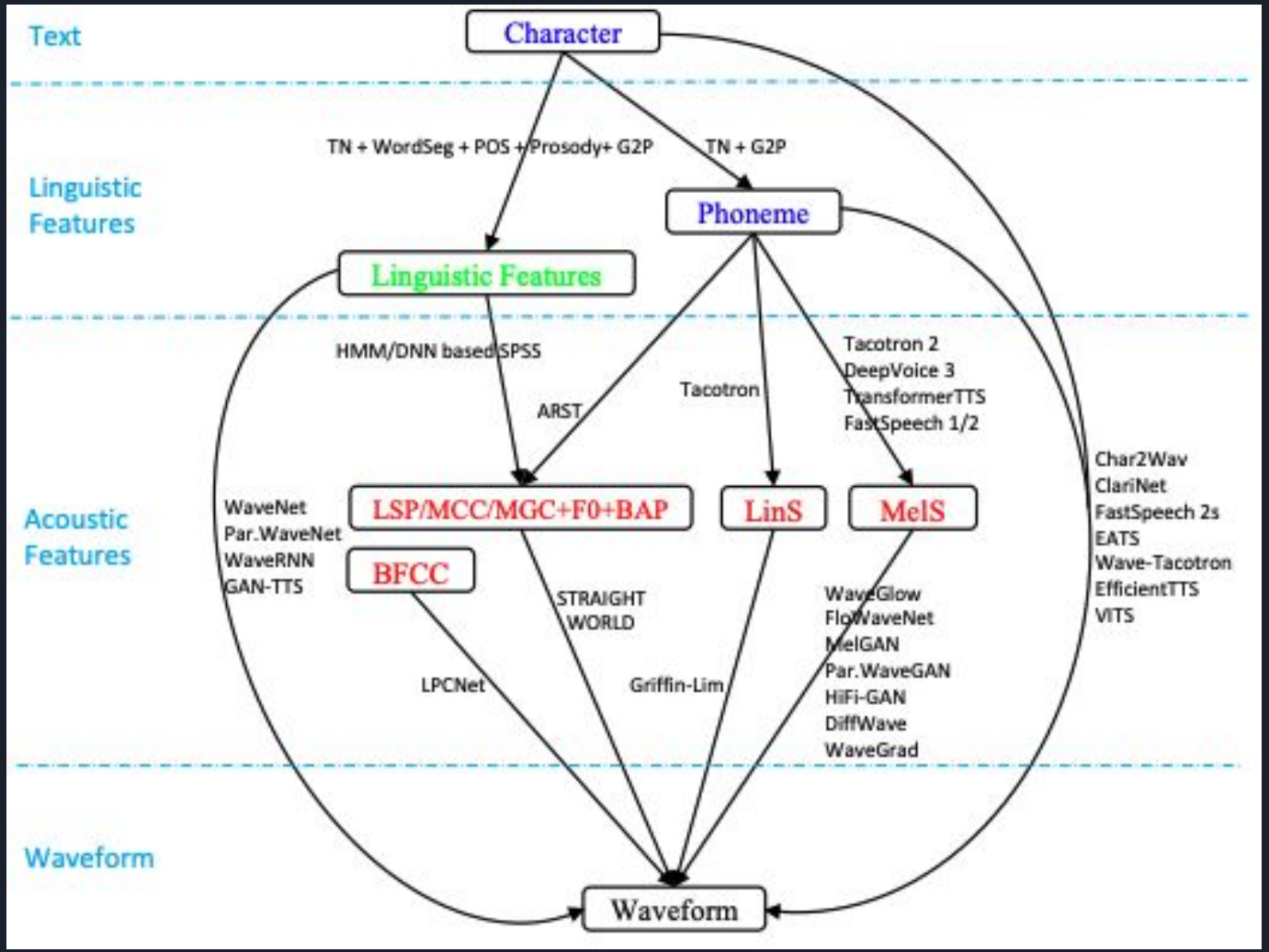
- Mel Cepstral Coefficients (MCC)
- Mel Spectrograms
- Fundamental Frequency (F0)
- Voiced/Unvoiced
- Bark-Frequency Cepstral Coefficients

Vocoders

- Analysis/Synthesis
  - Uses acoustic features, such as MCCs, band aperiodicity, and fundamental frequency (F0)
  - Creates waveform from acoustic features
- Autoregressive Vocoders
- Flow-based Vocoders
- GAN-based Vocoders
- Diffusion-based Vocoders



Stage	Models
0	SPSS [416, 356, 415, 425, 357]
1	ARST [375]
2	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
3	DeepVoice 3 [270], Tacotron 2 [303], FastSpeech 1/2 [290, 292], WaveGlow [279], FloWaveNet [163]
4	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

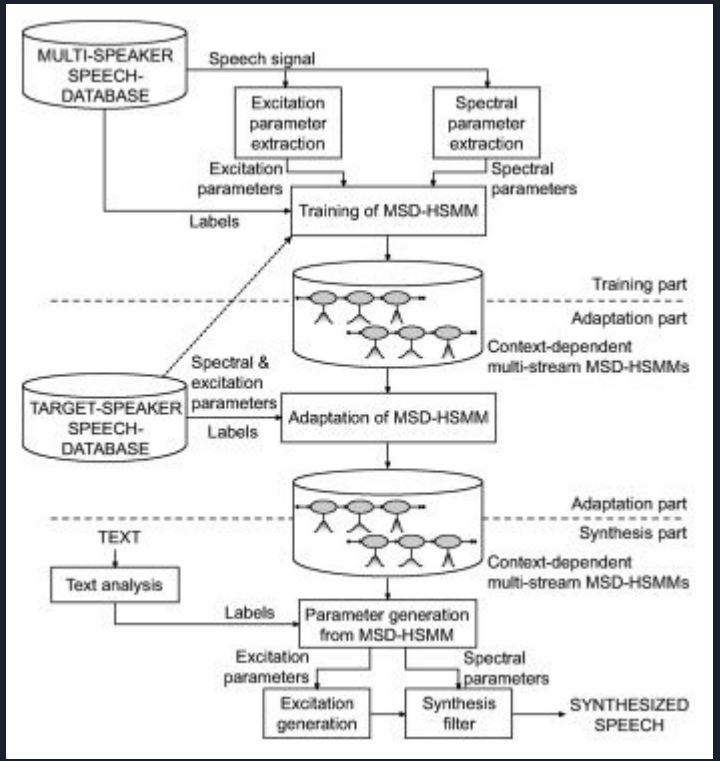

Text
Character
Linguistic Features
TN + WordSeg + POS + Prosody+ G2P
TN + G2P
Phoneme
Linguistic Features
HMM/DNN based SPSS
ARST
Tacotron
Tacotron 2
DeepVoice 3
TransformerTTS
FastSpeech 1/2
Acoustic Features
WaveNet
Par.WaveNet
WaveRNN
GAN-TTS
LSP/MCC/MGC+F0+BAP
LinS
MelS
BFCC
STRAIGHT
WORLD
LPCNet
Griffin-Lim
WaveGlow
FloWaveNet
MelGAN
Par.WaveGAN
HiFi-GAN
DiffWave
WaveGrad
Char2Wav
ClariNet
FastSpeech 2s
EATS
Wave-Tacotron
EfficientTTS
VITS
Waveform
Waveform

Types of Models

- Older signal-processing methods
  - Articulatory
  - Formant
  - Concatenative
- Statistical Parametric Speech Synthesis (SPSS)
  - HMM
- Neural Methods
  - RNN-based
  - CNN-based
  - GANs
- Newer Probabilistic Methods
  - Flow models
  - Diffusion models

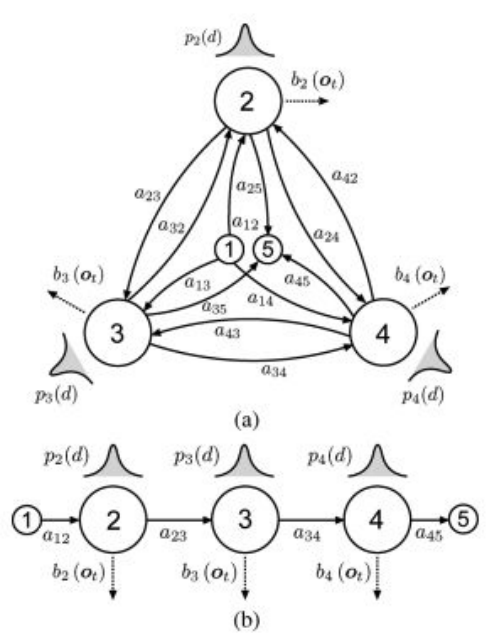
SPSS - HMM Approach (Yamagishi et al.)

- Multispace probability distribution hidden markov model (MSD-HMM)
- Speech Analysis - Extract MCCs, F0, and aperiodicity measures
  - Aperiodicity measures - ratio between smoothed lower and upper spectral envelopes, averaged over different frequency ranges
- Acoustic Models and Labels
  - Features at phoneme, segment, syllable, word, and utterance levels
  - Labels also include speaker gender
- Mixed-Gender Modeling
  - Gender of speaker is used as clustering context during training
  - One acoustic model encapsulates gender information

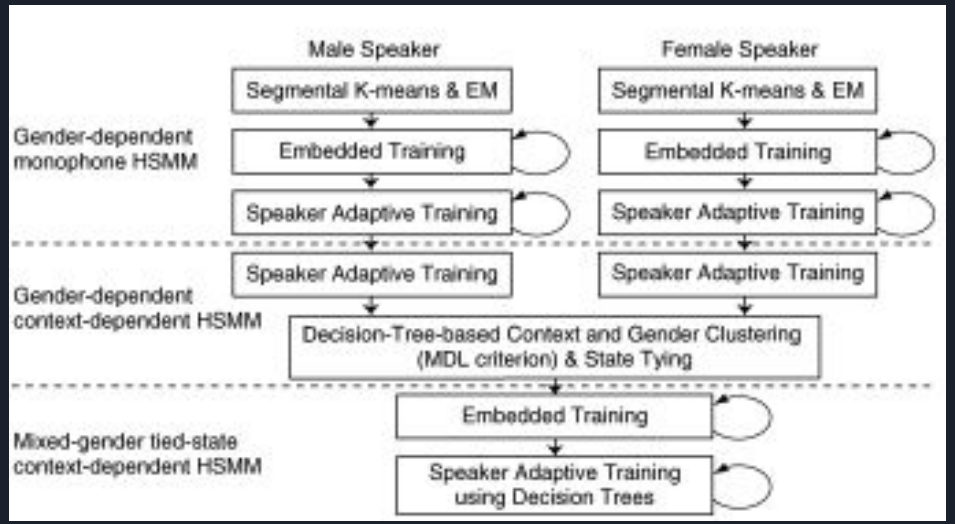


SPSS - HMM

Each state has an output distribution $b_i(o_t)$ as well as duration distribution $p_i(d)$

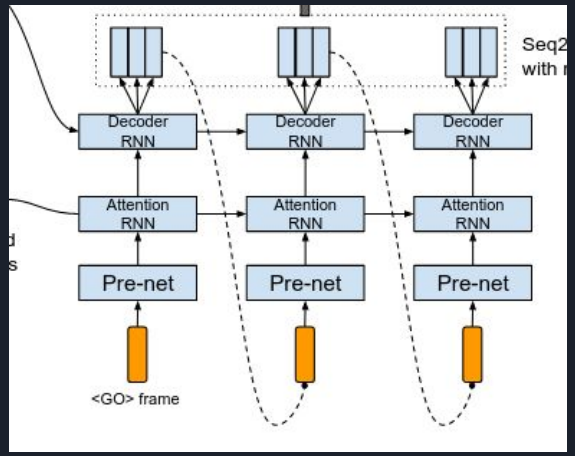


Overview of mixed-gender modeling:



RNNs

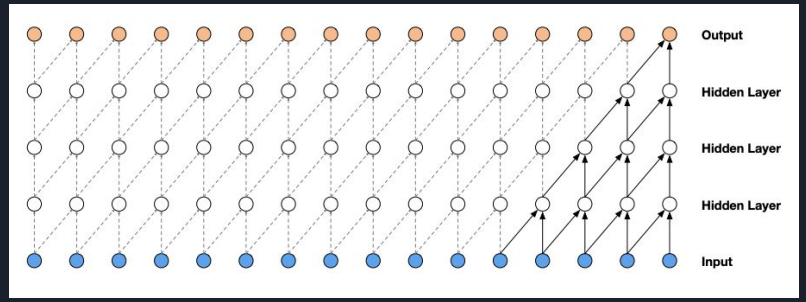
- RNNs:
  - Include temporal information by using model states in the next input
  - Autoregressive



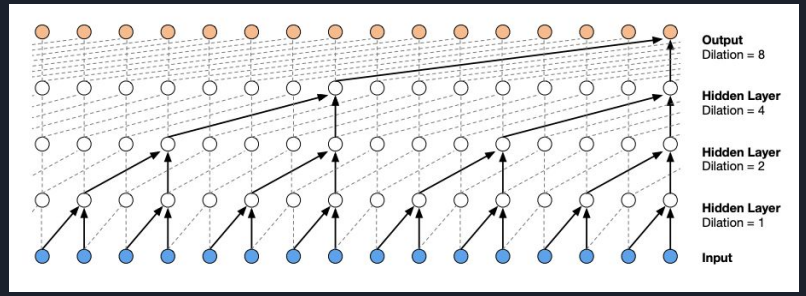
Tacotron (Wang et al., 2017)

Neural - Google WaveNet (Oord et al.)

- Causal Convolutional Layers



- Dilated Causal CNN



- Autoregressive Vocoder
- CNN based method
- Works directly on waveform or acoustic feature
- Each sample x_t is influenced by previous samples
- Faster to train (no recurrent connections)

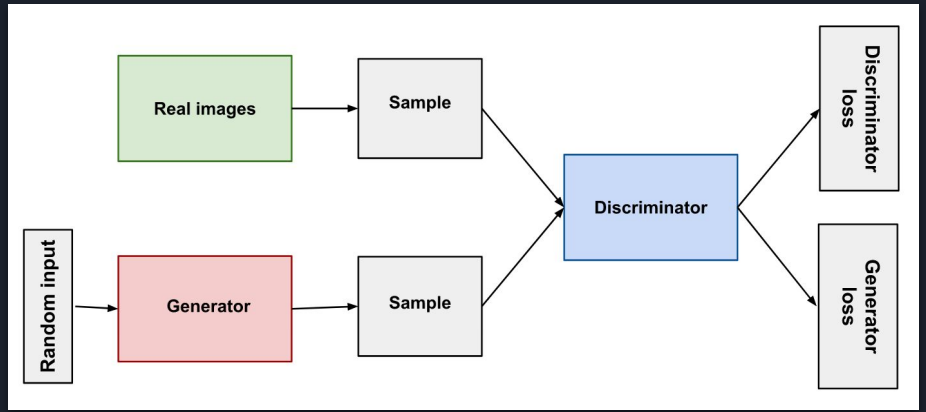
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

Speechify

GAN - HiFi-GAN (Kong et al.)

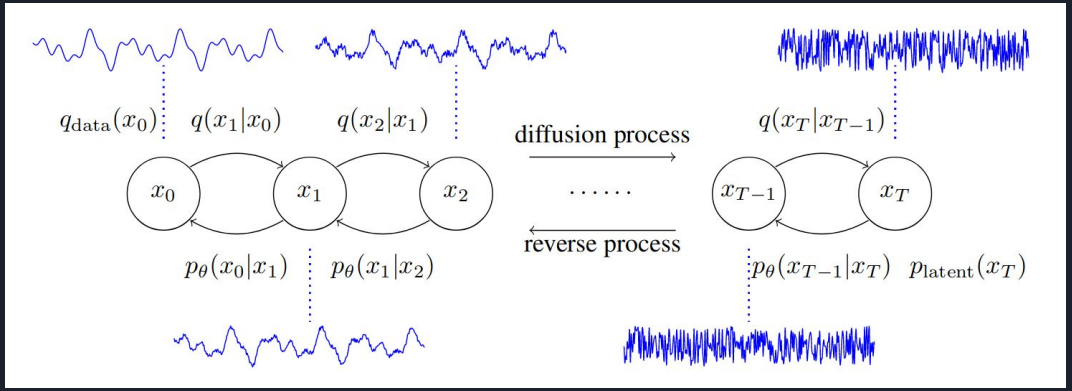
- Generator creates the speech; Discriminator tries to distinguish it from real speech
  - Encourages fine details
- Multiple discriminator losses apply to different levels of structure
  - Multi-scale: time domain
  - Other methods are multi-scale in spectrogram
- Generator losses
  - L1 Mel-Spectrogram
  - Feature Matching Loss in the Discriminators
- Training
  - Doesn't specify procedure
  - But many GANs train generator alone first
- Examples



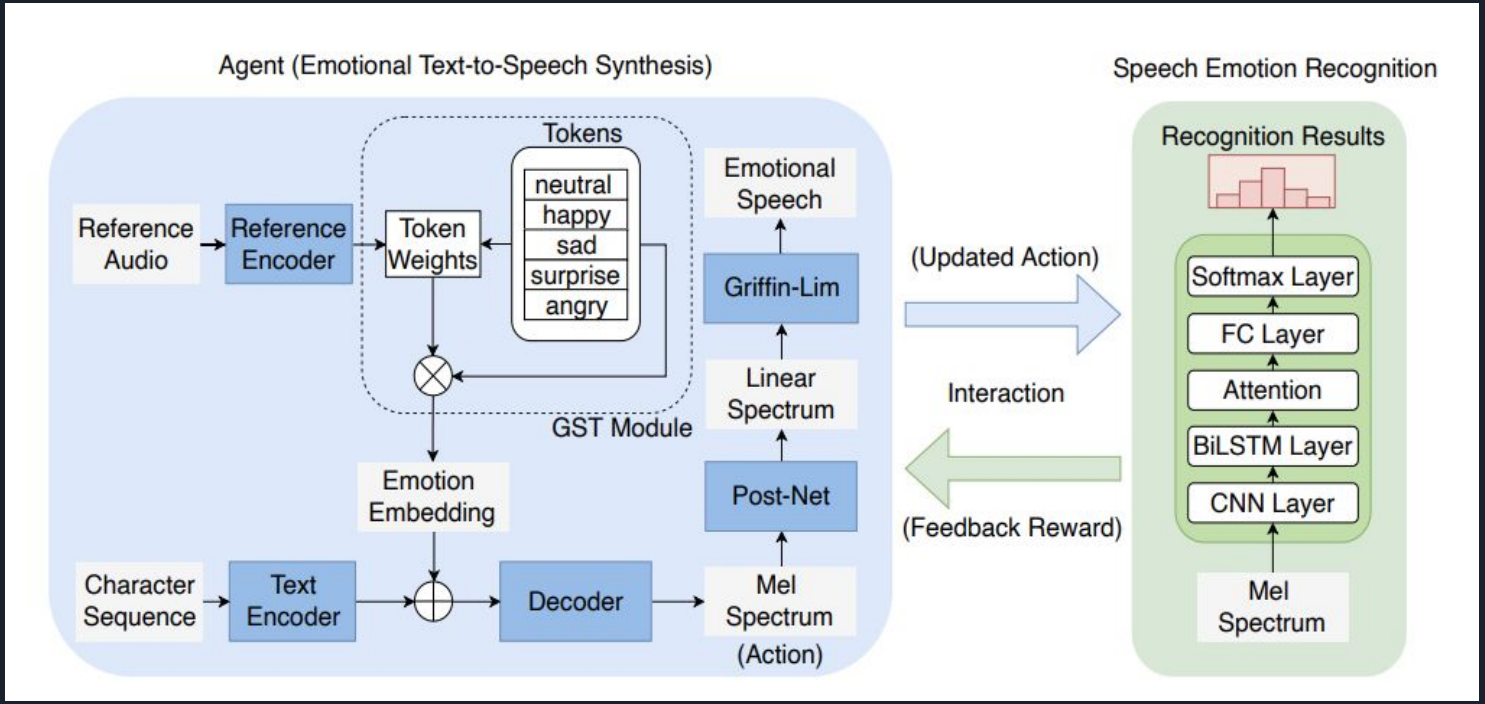
https://developers.google.com/machine-learning/gan/gan_structure

Diffusion Models - DiffWave (Kong et al.)

- Model input as a probability distribution
- Diffusion process:
  - Iteratively add noise to the input
- Reverse process:
  - Start with noise, reverse. Neural network estimates the reverse process probability distributions. After generation (using the reverse distributions from the NN), you can compare the output with the ground truth.
  - For conditional generation, the NN takes conditional input as a bias term at each layer
- High quality, but very slow (~100,000s of iterations)
- Examples



Emotional Text to Speech (ETTS) (Liu et al.)



[Examples]

Limitations/Future Research Directions

- More realistic, expressive speech generation
- Improved ETTS
- Improve generalization of models to handle cases not encountered in training sets
- Data efficiency in training - TTS is limited to languages with large amounts of recorded speech *with* annotated text
- Parameter efficiency - Neural models have extremely large parameter sets, limiting compatibility with devices that have less memory and power capabilities
- Generation speed - improvement of non-autoregressive models will allow for parallel inference

References

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

J. Yamagishi *et al*., "Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208-1230, Aug. 2009, doi: 10.1109/TASL.2009.2016394.

Young, Mary Cece, "The Effects of Text-to-Speech on Reading Comprehension of Students with Learning Disabilities" (2017). *Theses and Dissertations*. 718.

Liu, R., Sisman, B., and Li, H., “Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability”, 2021.

Tan, Xu et al. “A Survey on Neural Speech Synthesis.” *ArXiv* abs/2106.15561 (2021): n. Pag.

M. Macchi, "Issues in text-to-speech synthesis," *Proceedings. IEEE International Joint Symposia on Intelligence and Systems (Cat. No.98EX174)*, 1998, pp. 318-325, doi: 10.1109/IJSIS.1998.685467.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33, 2020.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In ICLR, 2021.



Questions?