# CS-E4891 Deep Generative Models
# Lecture 7: Text-conditional generation for images and audio

Lauri Juvela

Department of Information and Communications Engineering (DICE)
Aalto University

April 12, 2025

# Outline

Part 1: Overview of text-to-image and text-to-audio models

- Intended learning outcome: **recognize the common patterns** in text-conditional image and audio generation systems
- Encoder-Decoder models with diffusion and GANs
- Latent space generative models
- Text conditioning with CLIP and CLAP
- Discrete representation learning and language models

Part 2: More detailed case-study of a a text-to-speech system (for home exercise)

- Tacotron
- HiFi-GAN

Reading: reference list at the end (mostly nice to know, only Tacotron and HiFi-GAN appear in the exercise)

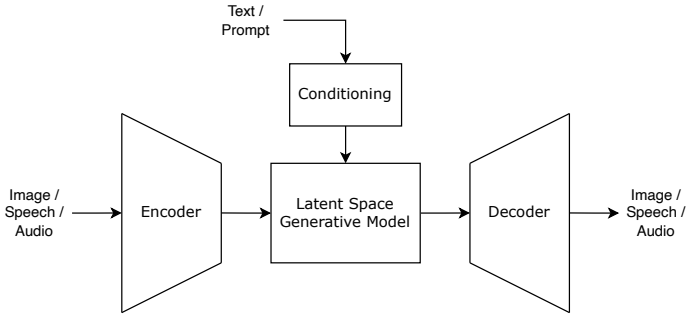
# Latent-space generative models for images, speech and audio

Problems fundamental problems

1. Images and audio are high dimensional data, it's difficult and costly to build generative models directly in data space
2. Text conditioning is usually available, but it's not clear where exactly text and audio or images. Text can be for example speech transcripts, or image alt-text captions.

Solutions usually include

- Encoder-Decoder model for learning more compact **latent space representation**
- Text conditioning model with some **alignment** mechanism
- Generative model in latent space (autoregressive, diffusion, GAN)



# Latent diffusion for image synthesis (basis for Stable Diffusion 1.0)

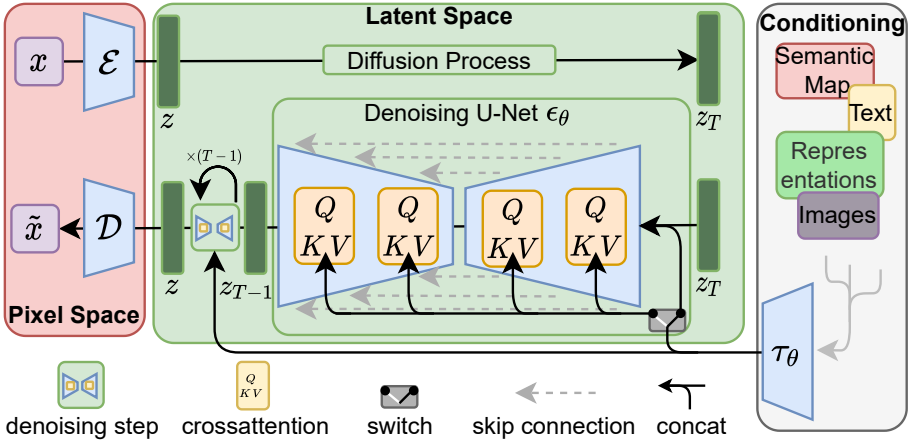



Figure: Latent diffusion model [Rombach et al., 2021]

# Contrastive Language-Image Pre-training (CLIP)

- Compute cosine distance between all pairs of text and image encoder output vectors
- Contrastive training: make true pairs (on diagonal) similar and all the other pairs dissimilar
- How does this compare to GAN training?

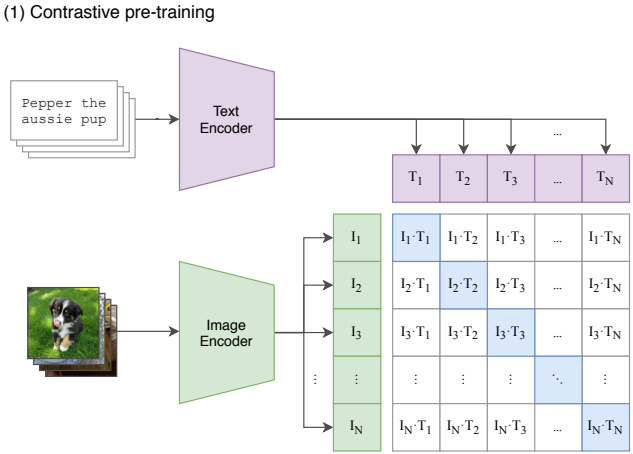



Figure: CLIP pre-training [Radford et al., 2021]

# Zero-shot classification with CLIP

- Insert class labels to a template sentence and compute CLIP text embeddings for all classes
- Compute image embedding for test image and compare with the text embeddings
- Choose the closest text embedding as the classification result

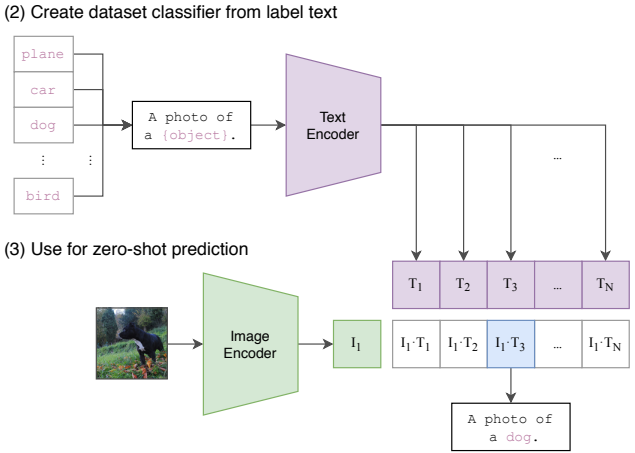



Figure: [Radford et al., 2021]

# Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL-E 2)

- First train a CLIP embedding model to get a joint latent representation for text and images (above dashed line)
- Use an autoregressive or diffusion prior to produce an image embedding
- Diffusion decoder model generates images from embeddings

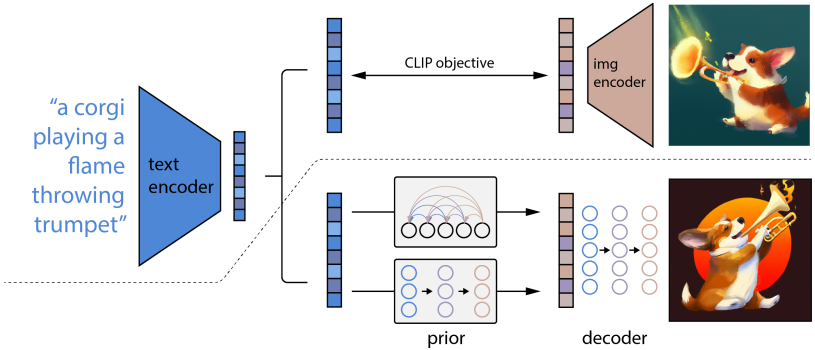



Figure: [Ramesh et al., 2022]

# Generated images from DALL-E 2


Figure: From [Ramesh et al., 2022]

# Diffusion Transformers (DiT)

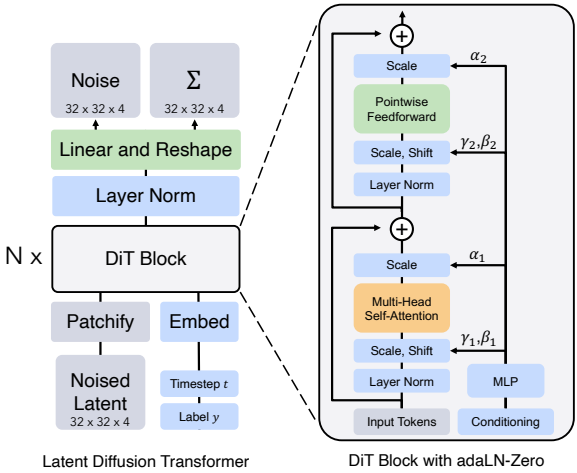
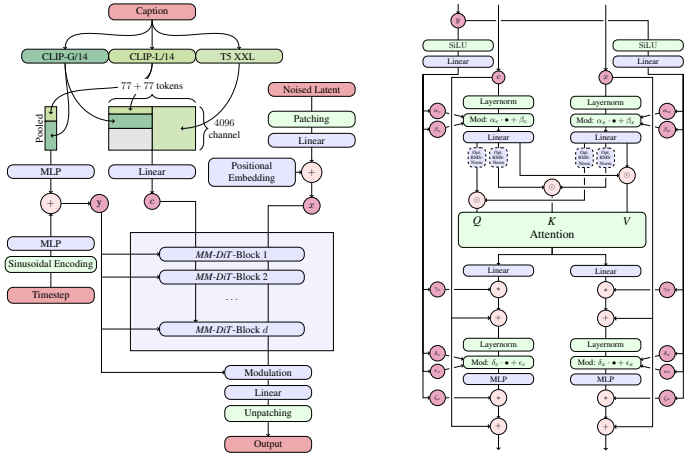



Figure: From [Peebles and Xie, 2022]

# Stable Diffusion 3 – (getting increasingly complicated)

- Rectified flows with diffusion Transformers, conditioned on CLIP and T5 embeddings.




(a) Overview of all components.

(b) One *MM-DiT* block

Figure: From [Esser et al., 2024]

# Stable Diffusion 3 generated images

a space elevator, cinematic scifi art

A cheeseburger with juicy beef patties and melted cheese sits on top of a toilet that looks like a throne and stands in the middle of the royal chamber.

a hole in the floor of my bathroom with small gremlins living in it

a small office made out of car parts

This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest.

human life depicted entirely out of fractals

an origami pig on fire in the middle of a dark room with a pentagram on the floor

Figure: Generated images, from [Esser et al., 2024]

# CLAP – CLIP for audio

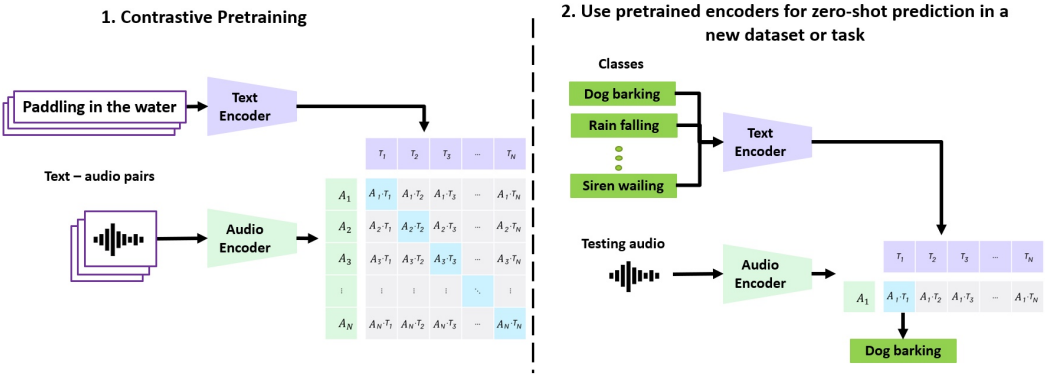



Figure: From [Elizalde et al., 2023]

# Latent diffusion for music generation (Stable Audio)

- Variational autoencoder with adversarial training
- Apply a Diffusion Transformer (DiT) model on the latents

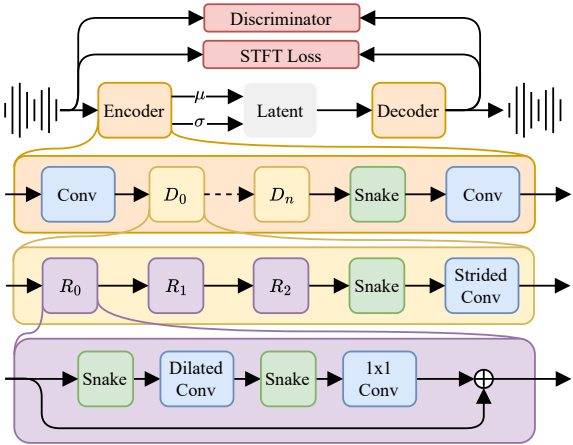



Figure: Autoencoder model, from [Evans et al., 2024]

# Latent diffusion for music generation (Stable Audio)

- Variational autoencoder with adversarial training
- Apply a Diffusion Transformer (DiT) model on the latents
- Audio samples `https://stableaudio.com`

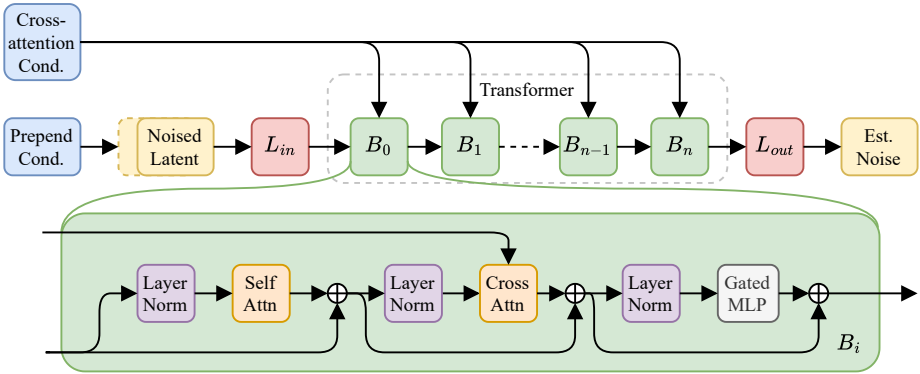



Figure: Diffusion Transformer, from [Evans et al., 2024]

# Vector-quantized variational autoencoder (VQ-VAE)

- Learn a codebook of vectors in the embedding space
- Encoder outputs are quantized to the nearest codebook vector
- Quantitation is not differentiable – on backward pass, copy gradients coming from the decoder through the quantization operation

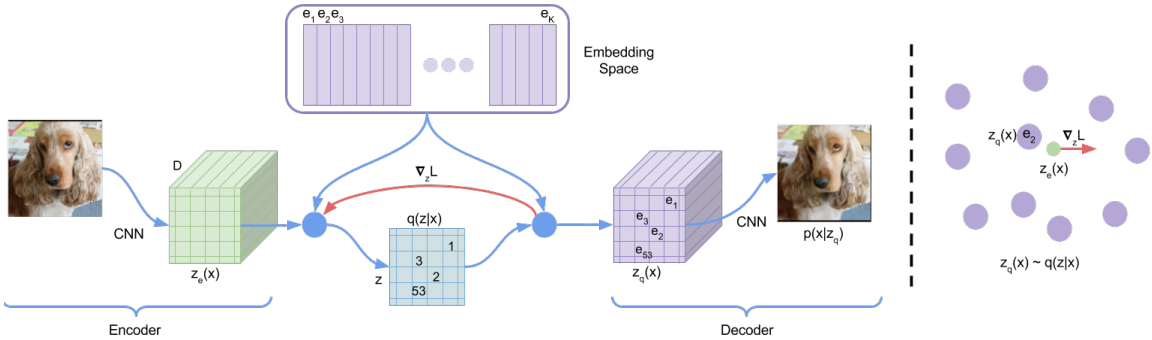



Figure: From [van den Oord et al., 2017]

# Jukebox – VQ-VAE for music generation

- Encoder-Decoder model with a convolution net backbone
- For generation, uses an autoregressive Transformer model on the quantized codes

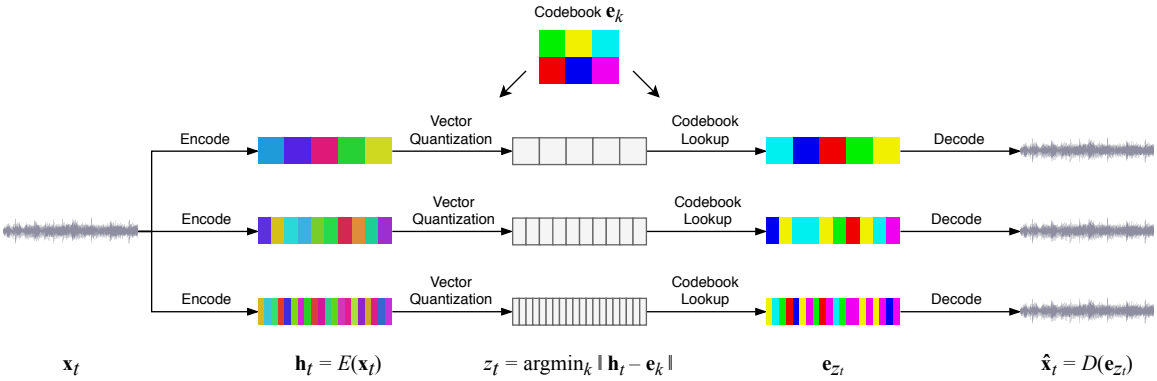



Figure: From [Dhariwal et al., 2020]

# EnCodec neural audio codec

- VQ-VAE model for speech and audio
- Uses residual vector quantisation
- Adversarial training with Discriminators + regression losses in waveform and spectral domains

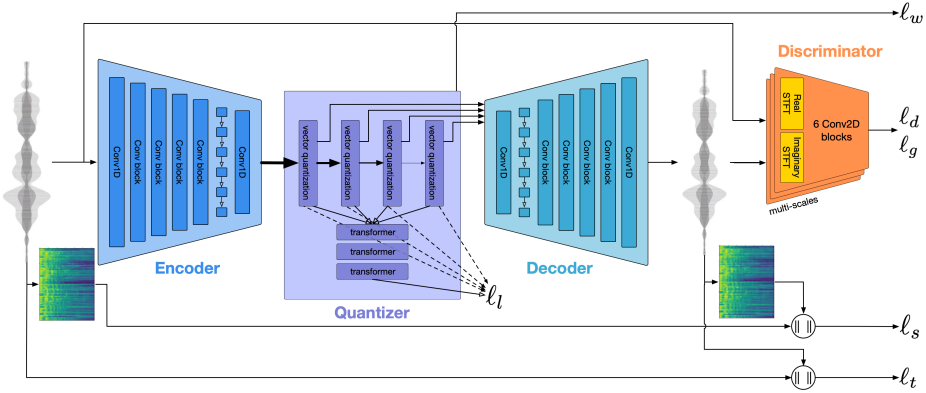



Figure: From [D'efossez et al., 2022]

# Residual vector quantization

- Use multiple codebooks to quantize the residual from previous codebook
- More parameter efficient: 8 times $2^{10} = 1024$ codes is better than one codebook with $2^{8 \times 1024}$ codes

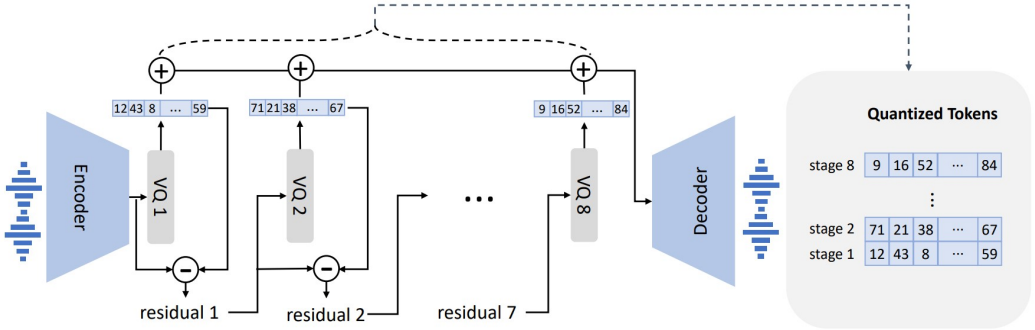



Figure: From [Wang et al., 2023]

# VALL-E – Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

- Language model on neural audio codec tokens and text-based tokens
- Acoustic prompts enable instant voice cloning

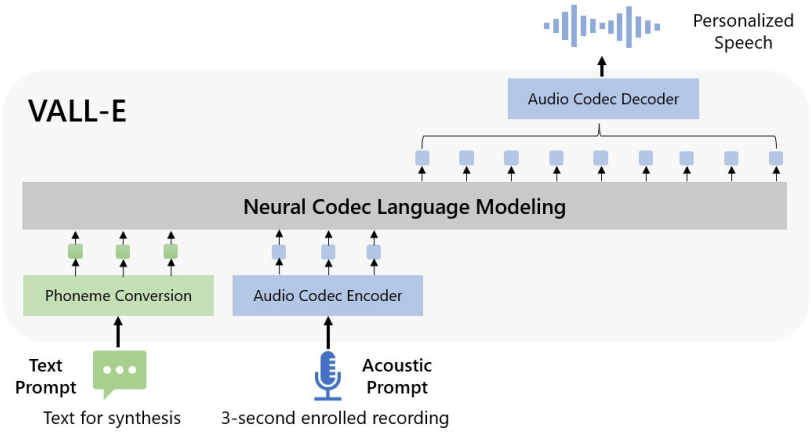



Figure: From [Wang et al., 2023]

# MELLE – Autoregressive Speech Synthesis without Vector Quantization

- Continuous latents (mel-spectrogram), autoregressive model
- VAE-style latent sampling module to enhance diversity
- Starting to look like good old Tacotron (next)
- Demo page `https://www.microsoft.com/en-us/research/project/vall-e-x/melle/`

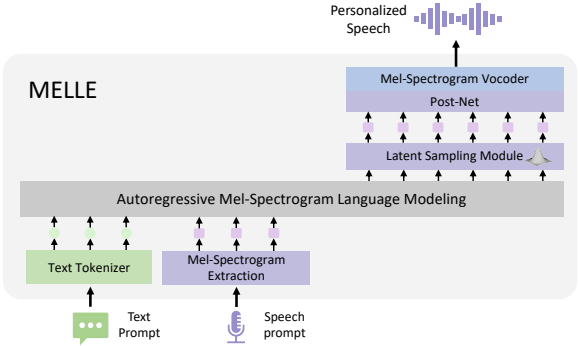



Figure: From [Meng et al., 2024]

# MELLE – Building blocks

- VAE-style latent sampling module to enhance diversity
- Predict a stopping probability for each generated frame
- Starting to look like good old Tacotron (next)

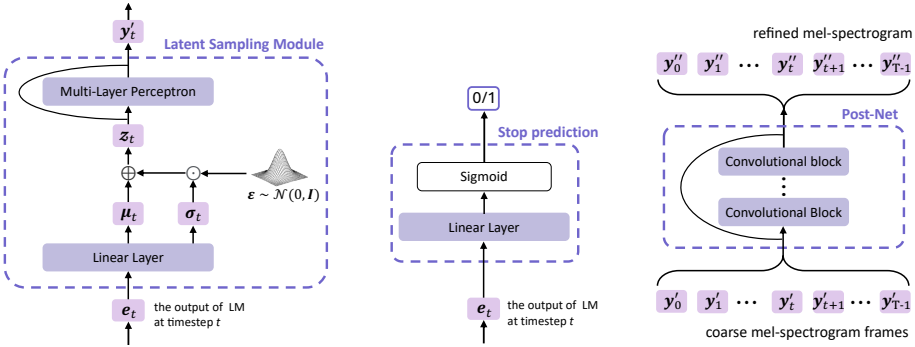
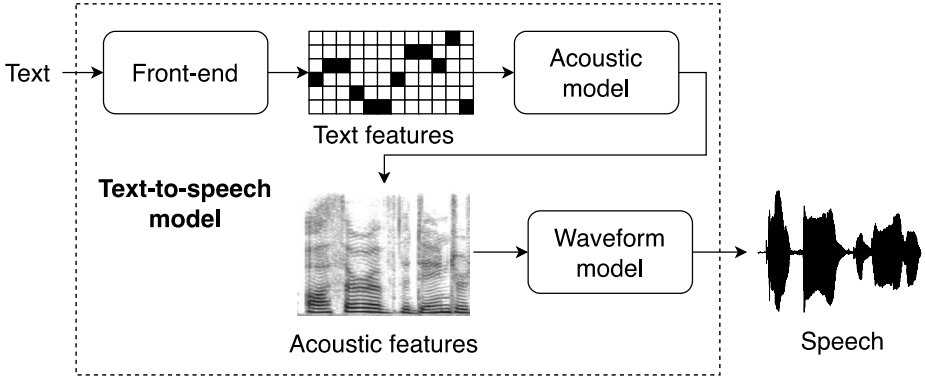



Figure: From [Meng et al., 2024]

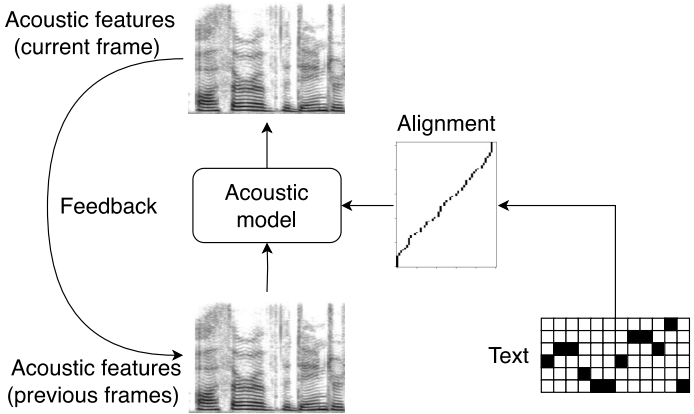
# Text-to-speech (TTS) systems

- Typical text-to-speech system
- Text processing front-end maps characters to symbolic text features (phonemes etc.)
- Acoustic model (generative model) maps text features to acoustic features (e.g., mel-spectrograms), these can be latent or observable
- Waveform model (decoder) maps acoustic features to a speech waveform)



# Autoregressive acoustic models

- Acoustic model predicts next acoustic feature frame from previous frames and text context
- Some alignment mechanism is needed, usually based on cross-attention, but classically used HMM-based ASR alignment and separate phoneme duration models



# Tacotron 1

- Autoregresssive model on mel-spectrograms
- Single cross-attention head does the all alignment between text encoder and acoustic decoder

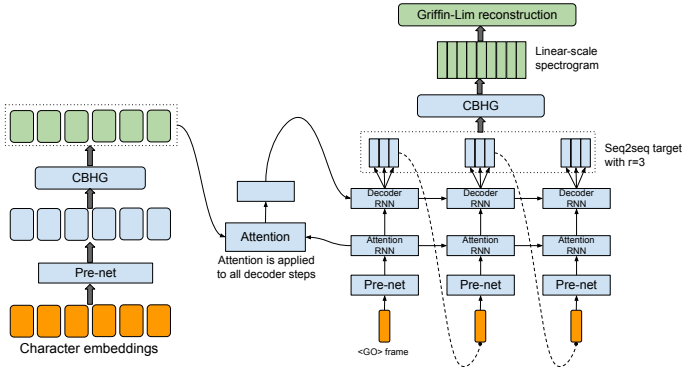



Figure: From [Wang et al., 2017]

# Tacotron 1

- CBHG module; 1-D convolution bank + highway network + bidirectional GRU
- Average activations over channels in attention $QK^T$ map can be used as an alignment plot

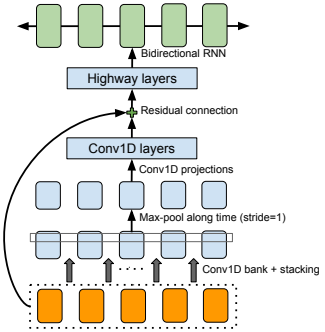



Figure: [Wang et al., 2017]

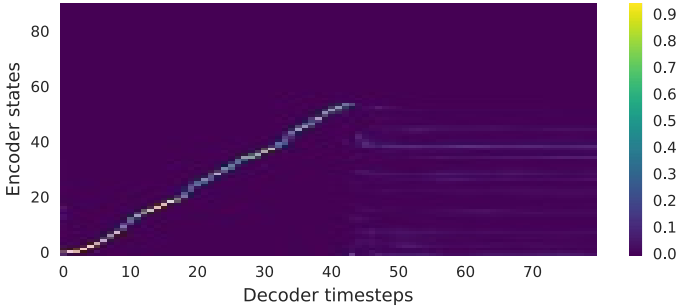



Figure: [Wang et al., 2017]

# Tacotron 2

- Simplified architecture over Tacotron
- Autoregressive model on mel-spectrograms, WaveNet for waveform decoding
- Single cross-attention head does the all alignment between text encoder and acoustic decoder

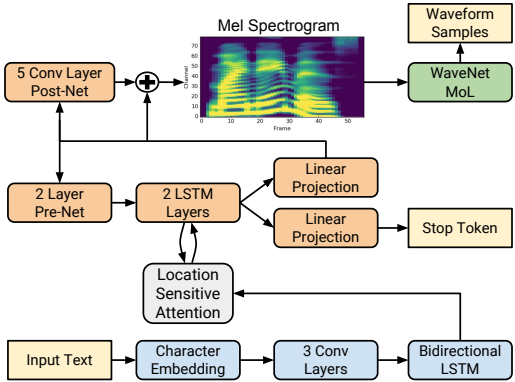
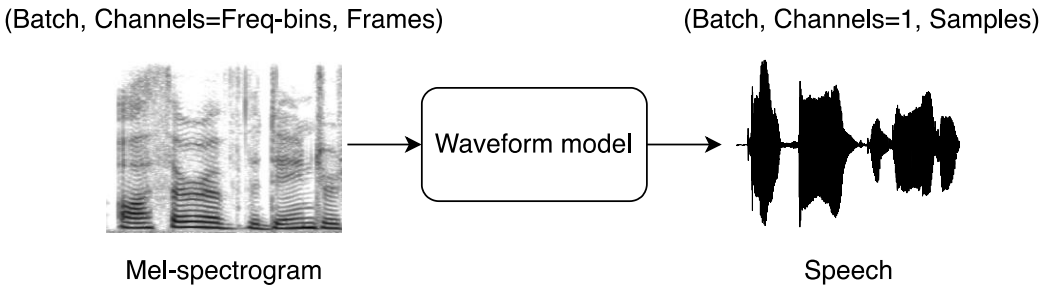



Figure: From [Shen et al., 2017]

# Waveform Model

- Waveform model (Decoder) upsamples the the latent representation (or acoustic features) and generates the missing detail
- GANs are a popular solution for this, diffusion models work as well, but are usually slower
- Upsampling factor is determined by the acoustic feature hop size (i.e., stride), typically around 256 samples at 16 kHz sample rate




Samples = Hop-size (stride) $\times$ Frames

# HiFi-GAN generator

- Generator upsamples mel-spectrograms up to $|k_u|$ times to match the temporal resolution of waveforms
- MRF module adds features from $|k_r|$ parallel residual blocks of different kernel sizes and dilation rates
- Residual blocks consist of dilated 1D convolution layers and Leaky ReLU activations

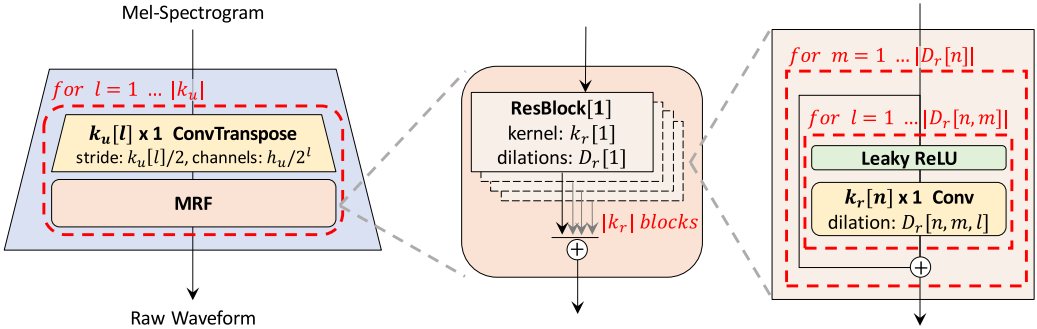



Figure: From [Kong et al., 2020]

# HiFi-GAN discriminators

- Multi-scale discriminator (MSD) apply various downsampling operations (using average pooling) and process the signal with 1D convolutions
- Multi-period discriminator (MPD) applies a framing operation to the signal and uses 2D convolutions to process the signal
- Note: later work like EnCodec use complex valued STFT inputs for Discriminators, this is more DSP informed and works equally well

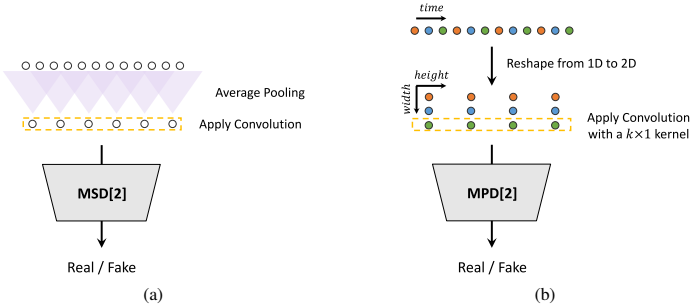



Figure: From [Kong et al., 2020]

# HiFi-GAN loss functions

- $x$ is signal waveform, $s$ is generator input mel spectrogram
- Least-squares GAN adversarial losses

$$\mathcal{L}_{\text{Adv}}(D;G) = \mathbb{E}_{x,s}\left[(D(x)-1)^2 + (D(G(s)))^2\right]$$
$$\mathcal{L}_{\text{Adv}}(G;D) = \mathbb{E}_{s}\left[(D(G(s))-1)^2\right]$$

- Mel-matching regression L1 loss, $\phi$ is mel-spectrogram transform

$$\mathcal{L}_{\text{Mel}}(G) = \mathbb{E}_{x,s} = [\|\phi(x) - \phi(G(s))\|_1]$$

- Feature matching loss ($T$ is the number of layers in discriminator, $D^i$ are hidden activations at layer $i$

$$\mathcal{L}_{\text{FM}}(G;D) = \mathbb{E}_{x,s}\left[\sum_{i=1}^{T}\frac{1}{N_i}\|D^i(x) - D^i(G(s))\|_1\right]$$

# Summary

Part 1: Overview of text-to-image and text-to-audio models

- Intended learning outcome: **recognize the common patterns** in text-conditional image and audio generation systems
- Encoder-Decoder models with diffusion and GANs
- Latent space generative models
- Text conditioning with CLIP and CLAP
- Discrete representation learning and language models

Part 2: More detailed case-study of a a text-to-speech system (for home exercise)

- Tacotron
- HiFi-GAN

# References I


D'efossez, A., Copet, J., Synnaeve, G., and Adi, Y. (2022).
High fidelity neural audio compression.
*ArXiv*, abs/2210.13438.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020).
Jukebox: A generative model for music.
*ArXiv*, abs/2005.00341.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. (2023).
Clap learning audio concepts from natural language supervision.
*ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Muller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. (2024).
Scaling rectified flow transformers for high-resolution image synthesis.
*ArXiv*, abs/2403.03206.

# References II


Evans, Z., Parker, J., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. (2024).
Long-form music generation with latent diffusion.
In *International Society for Music Information Retrieval Conference*.

Kong, J., Kim, J., and Bae, J. (2020).
Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.
*ArXiv*, abs/2010.05646.

Meng, L., Zhou, L., Liu, S., Chen, S., Han, B., Hu, S., Liu, Y., Li, J., Zhao, S., Wu, X., Meng, H., and Wei, F. (2024).
Autoregressive speech synthesis without vector quantization.
*ArXiv*, abs/2407.08551.

Peebles, W. S. and Xie, S. (2022).
Scalable diffusion models with transformers.
*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182.

# References III



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021).
Learning transferable visual models from natural language supervision.
In *International Conference on Machine Learning*.


Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022).
Hierarchical text-conditional image generation with clip latents.
*ArXiv*, abs/2204.06125.


Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021).
High-resolution image synthesis with latent diffusion models.
*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.


Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2017).
Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.
*2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

# References IV



van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017).
Neural discrete representation learning.
In *Neural Information Processing Systems*.


Wang, C., Chen, S., Wu, Y., Zhang, Z.-H., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. (2023).
Neural codec language models are zero-shot text to speech synthesizers.
*IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.


Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R. A. J., and Saurous, R. A. (2017).
Tacotron: Towards end-to-end speech synthesis.
In *Interspeech*.