Presentation 4

Wavenet: A Generative Model for Raw Audio

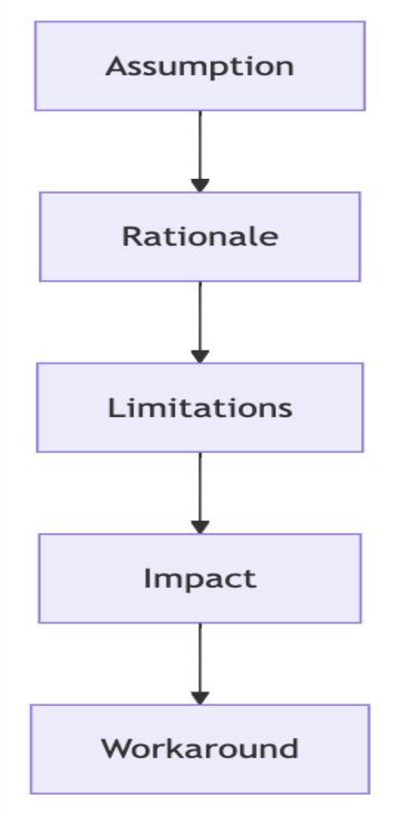
Areas for Improvements

Li Cao (licao), Furi Xiang (furix)

CMU 18-789 Deep Generative Modelling

Assumptions & Limitations

Assumptions	Limitations
1. AR Model	Slow Inference Speed
2. Quantization	Quality Degradation
3. Linguistic Features	Error Propagation
4. Stationarity of Audio	Lack of Control
5. Dilated Convolutions	Limited Receptive Field
6. Data Availability	Multi-speaker Scalability



Autoregressive Generation – Slow Inference Speed

- Assumption: Audio samples' probability distribution depends only on previous samples.
- Rationale: Mimics human speech generation
- Limitation: Slow inference due to sequential sampling.
- Impact: Impractical for real-time applications.
- Workaround: Parallel or non-autoregressive architectures (e.g., Parallel WaveNet) for faster inference.

Workaround - Parallel Wavenet

- Masked Autoregressive Flow (MAF)

  Fast likelihood evaluation, slow sampling -> parallel training based on MLE.
- Invertible Autoregressive Flow (IAF)

  Fast sampling, slow likelihood evaluation -> paralell real-time generation.

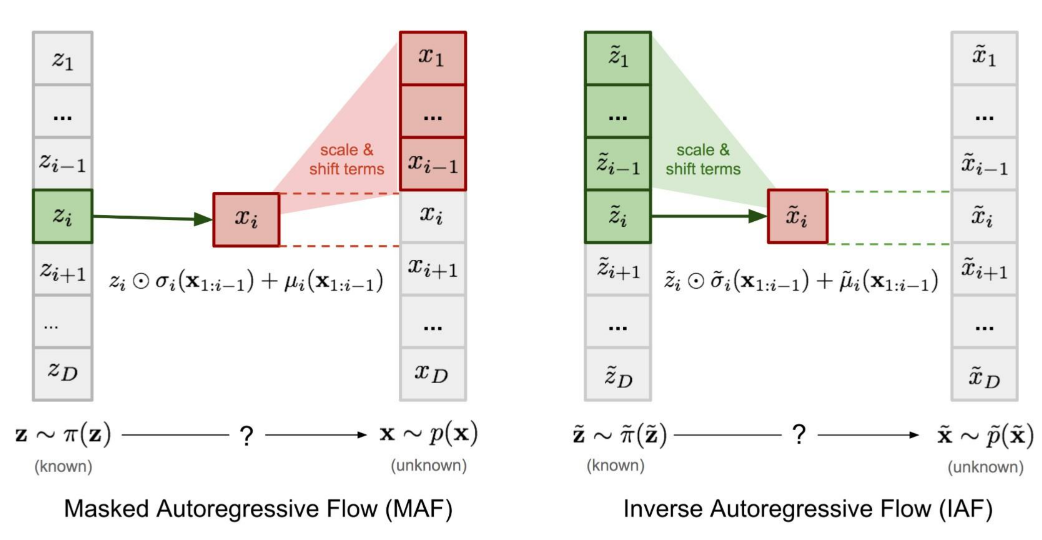


Figure 1: Comparison of MAF and IAF. The variable with known density is in green while the unknown one is in red.

Workaround - Parallel Wavenet

- Two part training with a teacher model (MAF) and student model (IAF).
- Once Teacher is trained in parallel via MLE, initialize a student model parameterized by IAF.

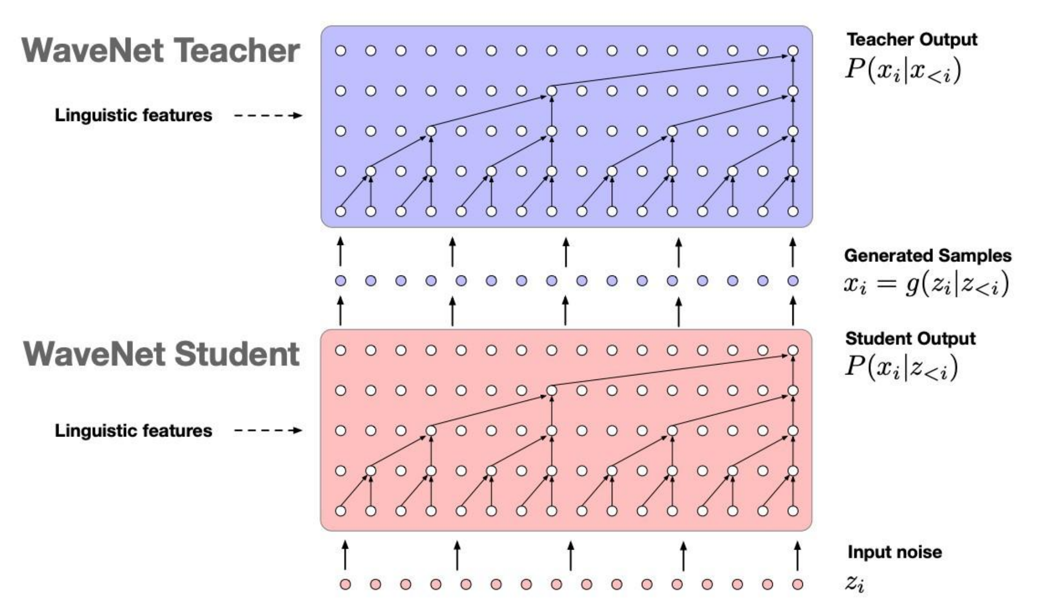


Figure 2: Overview of Probability Density Distillation

Workaround - Parallel Wavenet

- **Probability density distillation**:

Student distribution is trained to minimize the KL divergence between student (s) and teacher (t)

$$D_{KL}(S, t) = \mathrm{E}_{x \sim s}[\log s(x) - \log t(x)]$$

- Evaluation and optimization of the objective only requires efficient operations

- At training time:
  1. Train teacher model via MLE.
  2. Train student model via minimizing D_{KL} with teacher model.
- At Test-time:

  Use student model for inference / generation.

Workaround - Parallel Wavenet

- Improves inference speed by 1000x compared to the original wavenet.
- Successfully deployed in Google Assistant in 2017.

RESEARCH

WaveNet launches in the Google Assistant

4 OCTOBER 2017

Aäron van den Oord, Tom Walters

Share

Just over a year ago we presented WaveNet, a new deep neural network for generating raw audio waveforms that is capable of producing better and more realistic-sounding speech than existing techniques. At that time, the model was a research prototype and was too computationally intensive to work in consumer products.

But over the last 12 months we have worked hard to significantly improve both the speed and quality of our model and today we are proud to announce that an updated version of WaveNet is being used to generate the Google Assistant voices for US English and Japanese across all platforms.

Using the new WaveNet model results in a range of more natural sounding voices for the Assistant.

Figure 3: WaveNet launches in the Google Assistant

Quantization – Audio Quality Degradation

- Assumption: Audio generation framed as a classification problem, discretizing raw 16-bit audio into 8-bits (256 values).
- Rationale: Reduces softmax output dimensions from 65,536 to 256, making training computationally tractable.
- Limitation: Quantization introduces approximation errors, creating high frequency noise and limiting the dynamic range.
- Impact: Quantization artifacts degrade audio fidelity.
- Workaround: Continuous waveform modeling. Parallel WaveNet: Replaced softmax with a mixture of logistics to model continuous audio signal.

Workaround - Continuous Waveform Modelling

- **Parallel Wavenet** The PDF of a Mixture of Logistics distribution defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \frac{1}{s_k} \cdot \sigma\left(\frac{x-\mu_k}{s_k}\right) \cdot \left(1 - \sigma\left(\frac{x-\mu_k}{s_k}\right)\right)$$

where:

K is the number of logistic components

π_k is the weight of the k-th component

μ_k is the mean of the k-th logistic component

s_k is the scale of the k-th component

$\sigma(z)$ is the sigmoid function

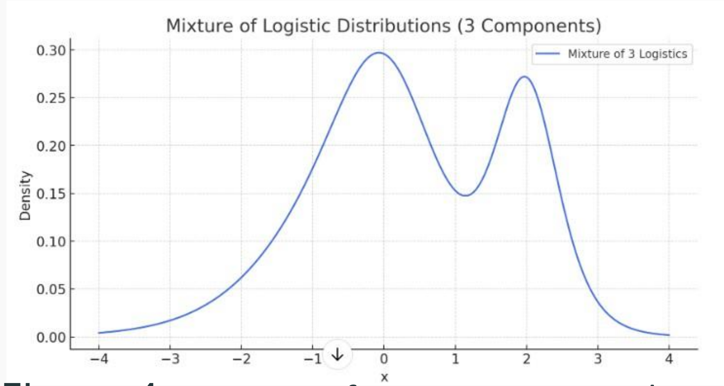


Figure 4: Mixture of 3 Logistics example PDF

Precomputed Linguistic Features – Error Propagation

- Assumption: Relies on precomputed linguistic features (phonemes), only handles features to audio generation.
- Rationale: To leverage well-established linguistic feature extraction NLP tools developed over decades in traditional TTS System.
- Limitation: Handcrafted and inflexible features, mistakes in feature extraction (misaligned phonemes) directly degraded output quality.
- Impact: Limited adaptability, error propagation.
- Workaround: Models like VITS integrate text-to-spectrogram and spectrogram-to-waveform steps into a single end-to-end neural network.

Workaround - Single Step End-to-End Modeling

- Wavenet only handles Spectrogram to Audio Waveform Synthesis.
- Need an end-to-end model for complete TTS system (e.g., VITS).

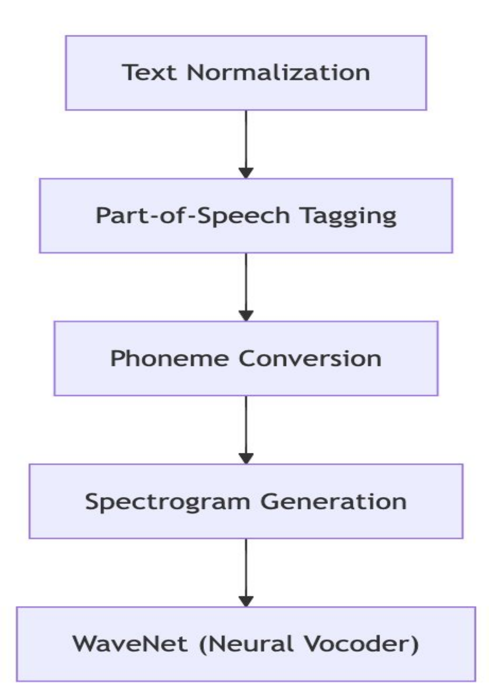


Figure 5: Traditional TTS System Modules

Stationarity of Audio - Lack of Fined-Grained Control

- Assumption: Statistical properties (e.g., mean, variance) of the audio signal remain consistent over time.
- Rationale: Simplifies modeling by treating audio as a stationary process.
- Limitation: Lacks fine-grained control over speed, prosody, pitch, or tone, unless explicitly conditioned.
- Impact: Monotonic or unnatural-sounding speech.
- Workaround: Latent representation that captures both linguistic content and prosodic features (e.g., VITS).

Workaround - Variational Autoencoder

Component	Contribution to Speech Control
Variational Autoencoder	Learn a low-dimensional latent space that captures speech attributes in a structured way
Adversarial Training	Ensures that generated waveforms are indistinguishable from real speech
Pitch / Duration Predictor	Allow for conditioning and control of intonation and rhythm.

Table 1: VITS Components Enabling Speech Style Control

Dilated Convolutions - Limited Receptive Field

- Assumption: Dilated convolutions alone suffice to model both long-term and short-term dependencies.
- Rationale: Dilations can effectively expand the receptive field.
- Limitation: May under-represent local patterns or shorter-term interactions critical for naturalness.
- Impact: Loss of coherence in synthesized speech over extended durations.
- Workaround: Use attention mechanism to model both long range and short range dependencies (e.g., VITS).

Workaround - Attention Module

- Capture long and short range dependencies with self-attention module within prior encoder
- Feed the context-aware text to VAE

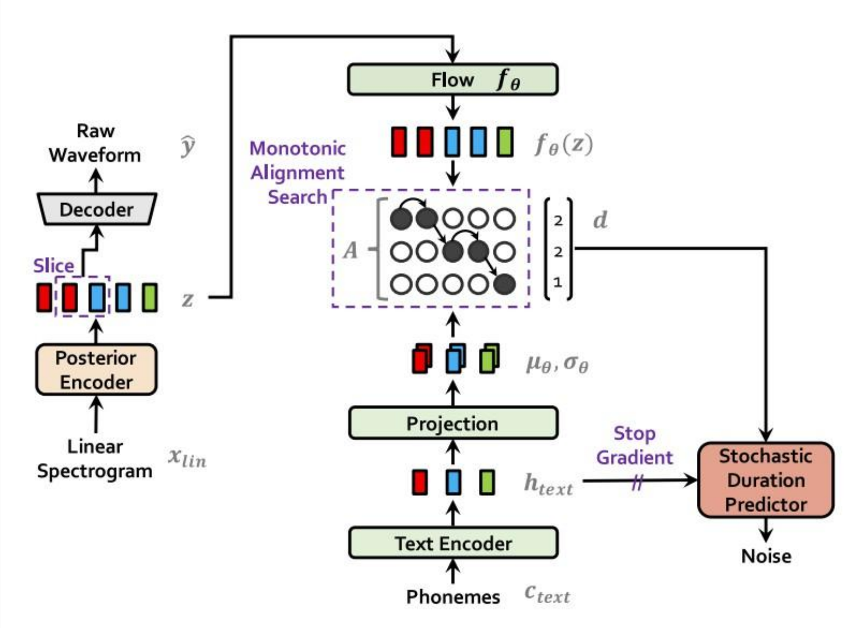


Figure 6: Overview of the VITS model.

Data Availability - Multi-speaker Scalability

- Assumption: Large and diverse dataset is required to model different speaker voices
- Rationale: To support multiple speakers, WaveNet uses speaker embeddings as conditional inputs. Each embedding must encode unique vocal traits for each speaker.
- Limitation: Needs hours of data per speaker to generate voices that doesn't sound generic.
- Impact: The model struggles to synthesize speakers or accents with little data.
- Workaround: Zero-shot / Few-shot multi-speaker generation reduces reliance on data (e.g., VITS).

Workaround - Zero-shot training

Component	Contribution to Multi-Speaker Capability
Speaker Embeddings	Inject speaker identity into encoder, duration predictor, and decoder to condition generation on target speaker.
Speaker Encoder	Zero-shot / Few-shot synthesis by extracting speaker embeddings from reference audio.
Variational Inference	Separates speaker identity from content for voice diversity.

Table 2: VITS Components Enabling Multi-Speaker Synthesis

Summary

Wavenet Limitations	**Solutions**	**Key Subsequent Models**
Slow inference Speed	Parallel sampling /non-AR Model	Parallel WaveNet, DiffWave
Quantization	Continuous Waveform Modelling	Parallel WaveNet, WaveGlow
Pre-computed features	End-to-end training	VITS
Limited receptive field	Attention mechanism	VITS, GST-Tacotron
Limited controllability	Latent Space Representation	VITS, GST-Tacotron
Multi-speaker scalability	Few-shot/zero-shot adaptation	VITS

Table 3: WaveNet Limitations, Solutions and Key Subsequent Models

References i

DeepMind.
Wavenet launches in the Google assistant, November 2017.

A. Gibiansky.
Wavenet and Tacotron aren't TTS systems, 2023.
Personal blog post.

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu.
Efficient neural audio synthesis, 2018.

J. Kim, J. Kong, and J. Son.
Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021.

R. Prenger, R. Valle, and B. Catanzaro.
Waveglow: A flow-based generative network for speech synthesis, 2018.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926. PMLR, July 2018.

L. Weng. Flow-based deep generative models. *lilianweng.github.io*, 2018.