Samsung AI Center – Moscow

Audio Synthesis and Bandwidth Extension

Audio Processing Problems

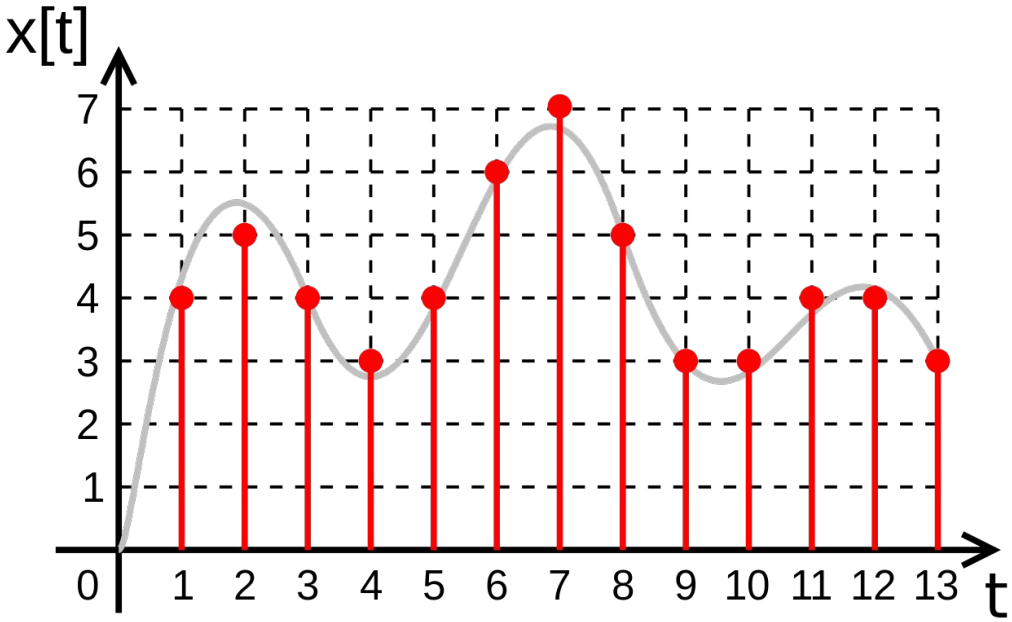
- Text To Speech
- Automatic Speech Recognition
- Acoustic Scene Recognition
  - Source Separation
  - Source Localization
- Audio Enhancement
  - Denoising
  - Bandwidth extension
  - Echo cancellation
- ...

2020 Samsung Research. All rights reserved

Introduction to Signal Processing

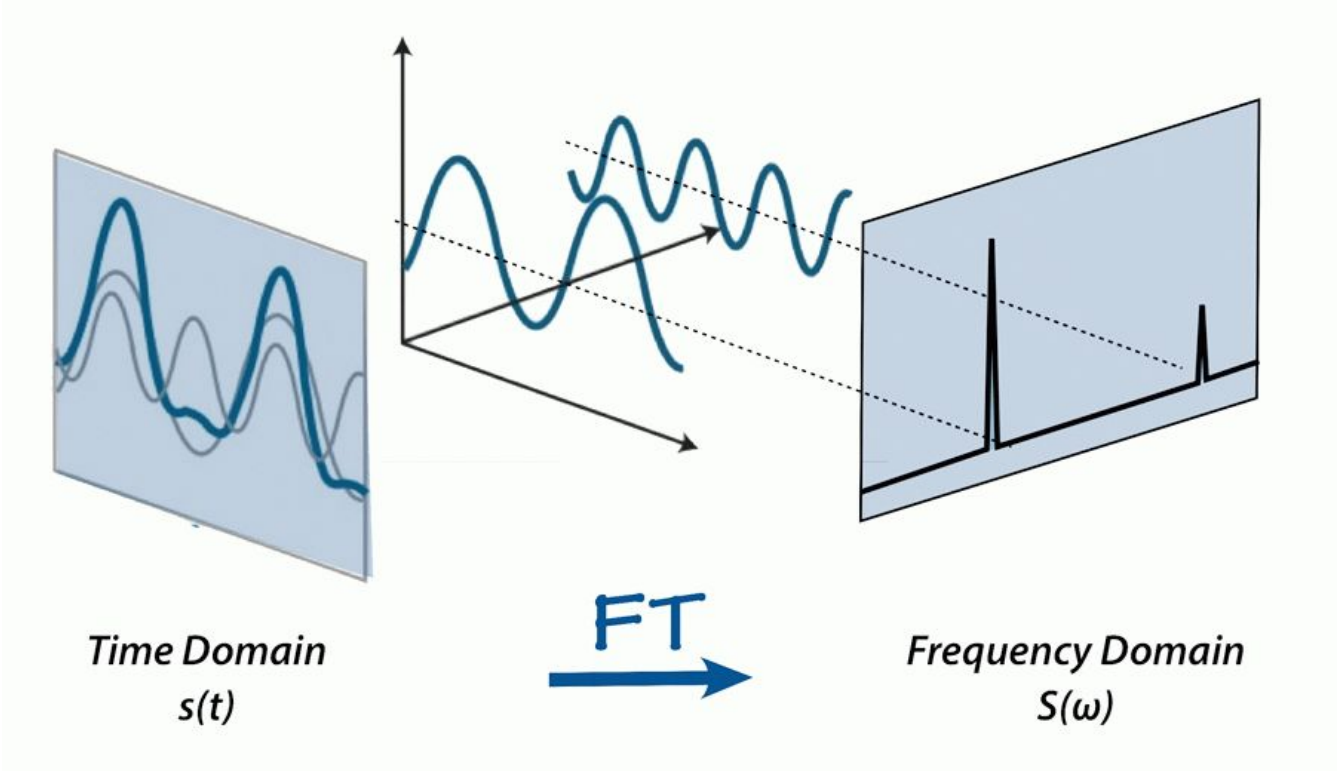
2020 Samsung Research. All rights reserved

Digital representation of signal



2020 Samsung Research. All rights reserved

Fourier Transform



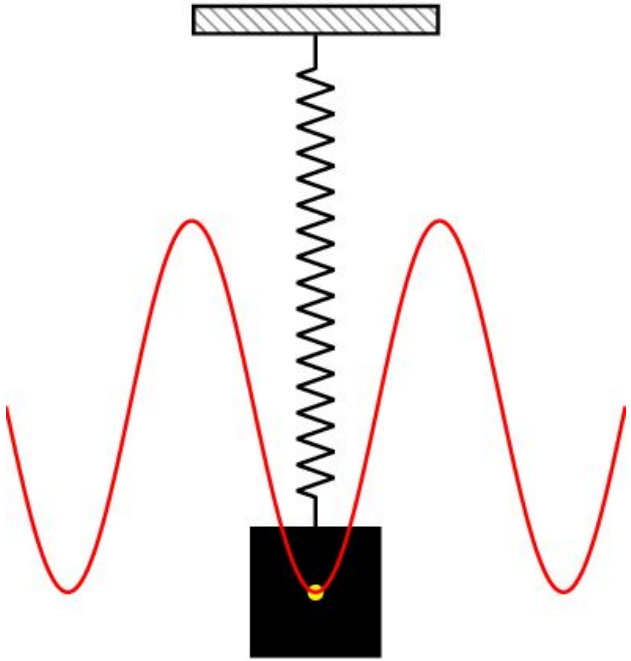
2020 Samsung Research. All rights reserved

Fourier Transform: physics intuition

$$\frac{\partial^2 x(t)}{\partial t^2} = -\frac{k}{m} x(t)$$

$$x(t) = A\cos(\phi + \omega t)$$

amplitude phase (cycle) frequency



Fourier Transform

$$x(t) = \sum_k A_k \cos(\phi_k + \omega_k t)$$

2020 Samsung Research. All rights reserved

Fourier Transform: mathematics strikes back

$$x(t) = \sum_k A_k \cos(\phi_k + \omega_k t)$$

$$e^{i\omega t} = \cos(\omega t) + i \sin(\omega t)$$

$F(\omega_k)$ Fourier coefficient

$$x(t) = \sum_k \overbrace{A_k e^{i\phi_k}} e^{i\omega_k t}$$

2020 Samsung Research. All rights reserved

Inverse Fourier Transform

$$x(t) = \sum_k F(\omega_k) e^{i\omega_k t}$$

$$F(\omega_k) = \int x(t) e^{-i\omega_k t} dt$$

2020 Samsung Research. All rights reserved

Fourier Transform

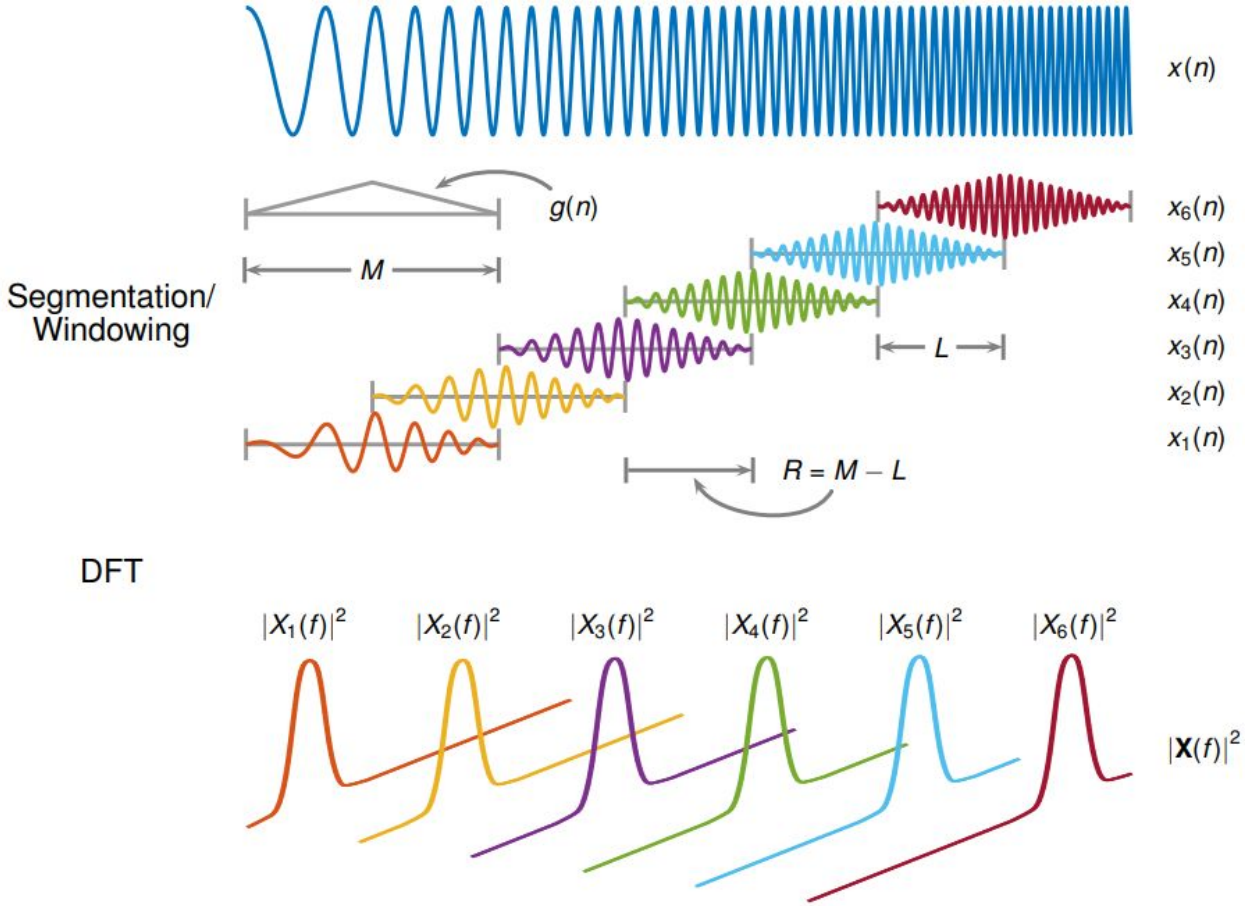
2020 Samsung Research. All rights reserved

Fourier Transform: Continuous vs. Discrete

Time Duration		
Finite	Infinite	
Discrete FT (DFT)	Discrete Time FT (DTFT)	discr.
$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\omega_k n}$	$X(\omega) = \sum_{n=-\infty}^{+\infty} x(n) e^{-j\omega n}$	time
$k = 0, 1, \ldots, N-1$	$\omega \in (-\pi, +\pi)$	n
Fourier Series (FS)	Fourier Transform (FT)	cont.
$X(k) = \int_0^P x(t) e^{-j\omega_k t} dt$	$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt$	time
$k = -\infty, \ldots, +\infty$	$\omega \in (-\infty, +\infty)$	t
discrete freq. k	continuous freq. ω	

2020 Samsung Research. All rights reserved

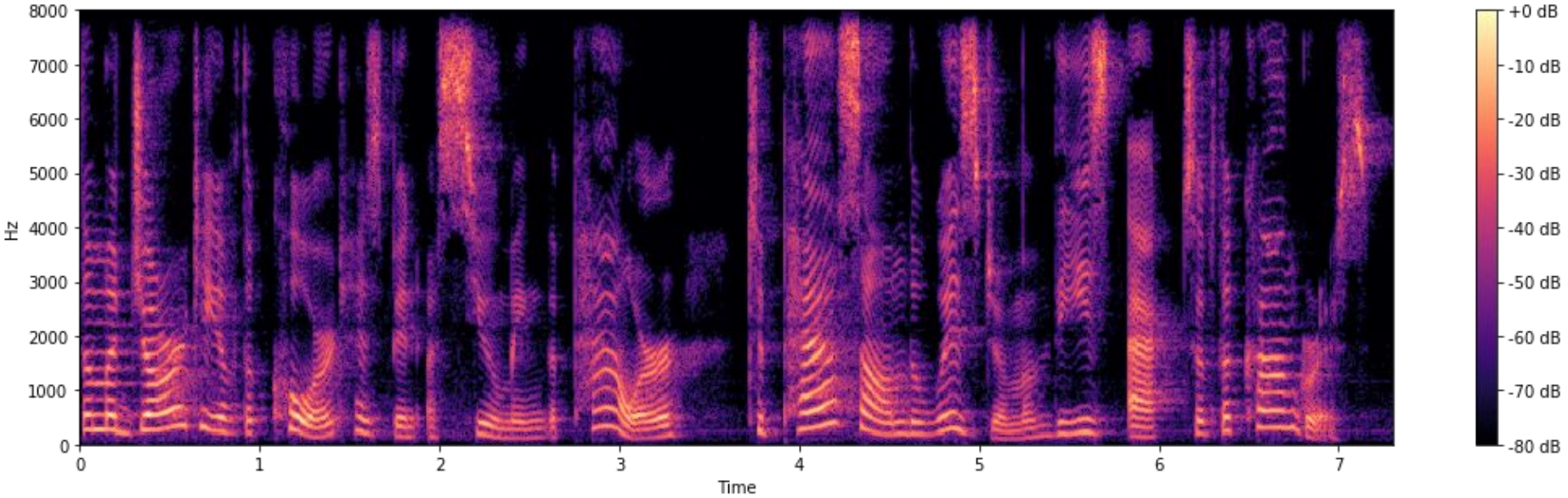
Short-Time Fourier Transform



2020 Samsung Research. All rights reserved

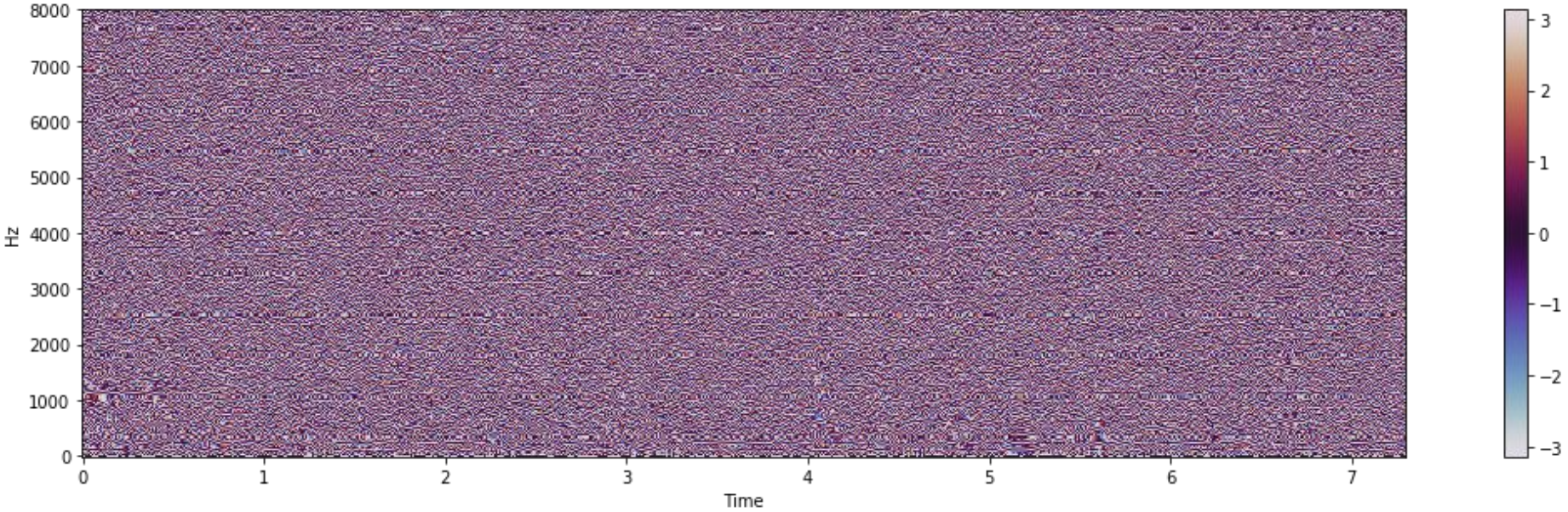
Spectrogram

Amplitude spectrogram



2020 Samsung Research. All rights reserved

Phase spectrogram



2020 Samsung Research. All rights reserved

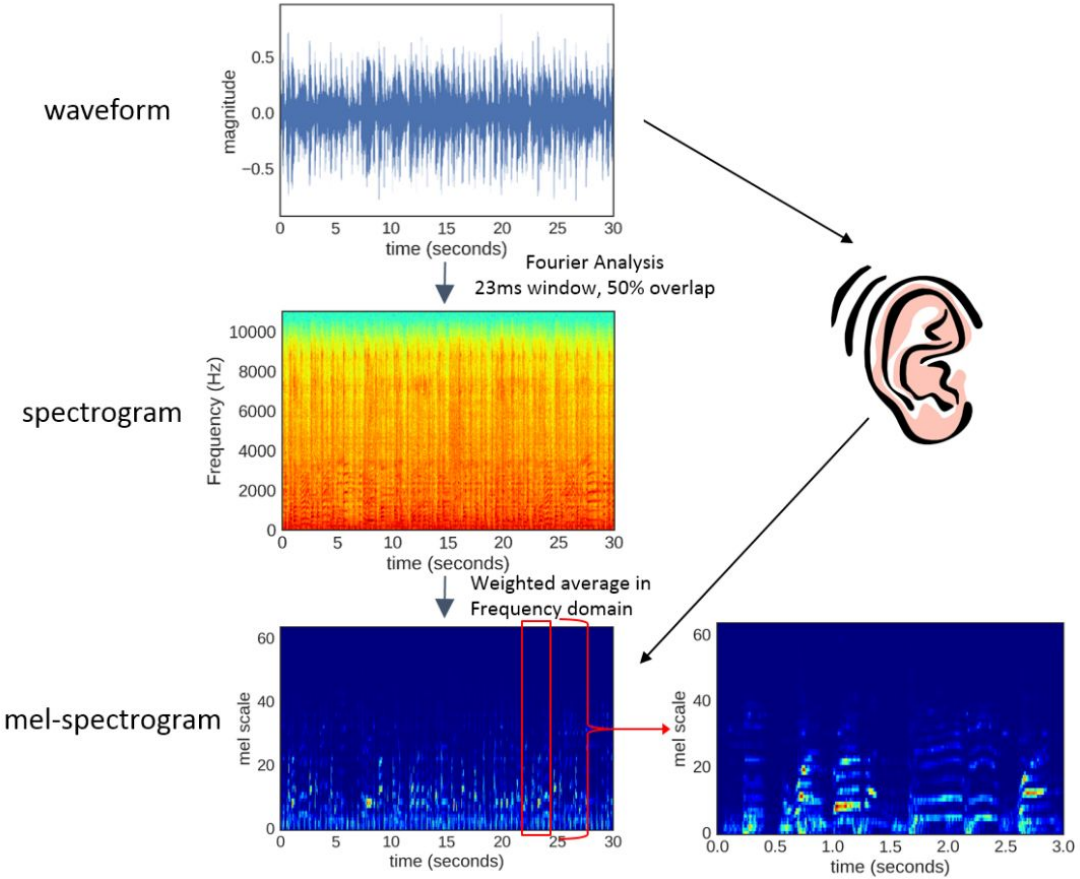
Inverse Short-Time Fourier Transform

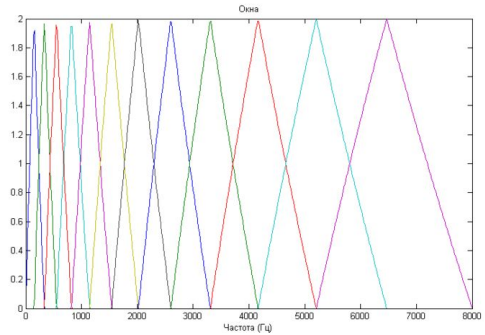
Griffin-Lim

```
def griffin_lim(F, max_iters=32):
    phi = Uniform(0, 2 * pi, size=F.shape)
    for _ in range(max_iters):
        y = ISTFT(F * exp(1j * phi))
        phi = angle(STFT(y))
    return y
```

2020 Samsung Research. All rights reserved

Mel-spectrogram



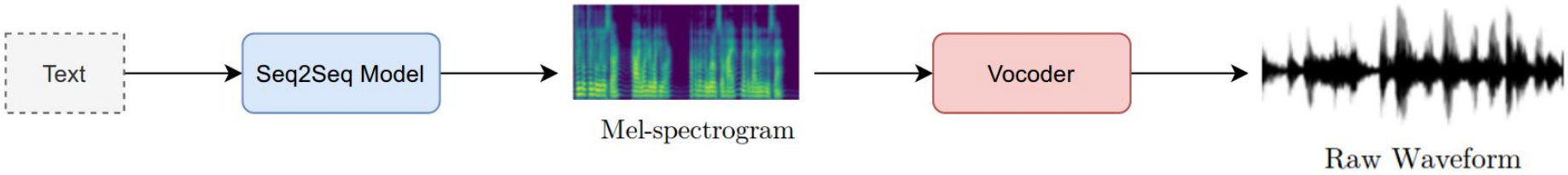


2020 Samsung Research. All rights reserved

Audio Processing Problems

2020 Samsung Research. All rights reserved

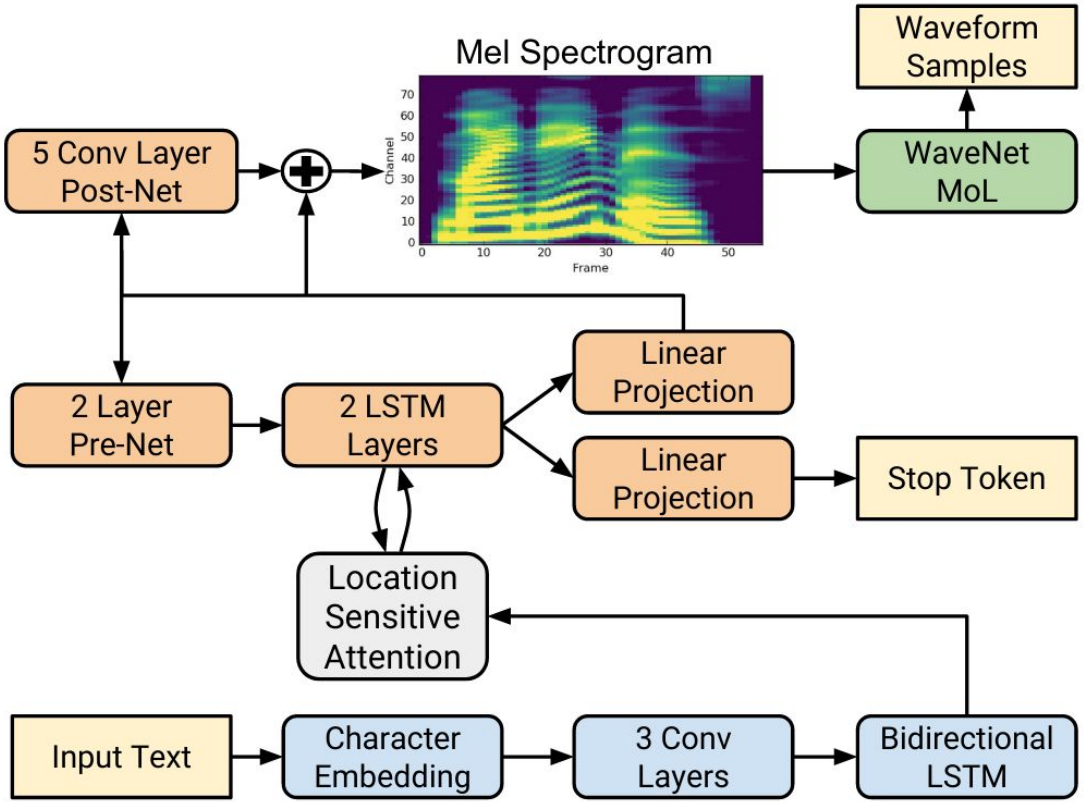
Text-to-speech (TTS)



Mel-spectrogram inversion

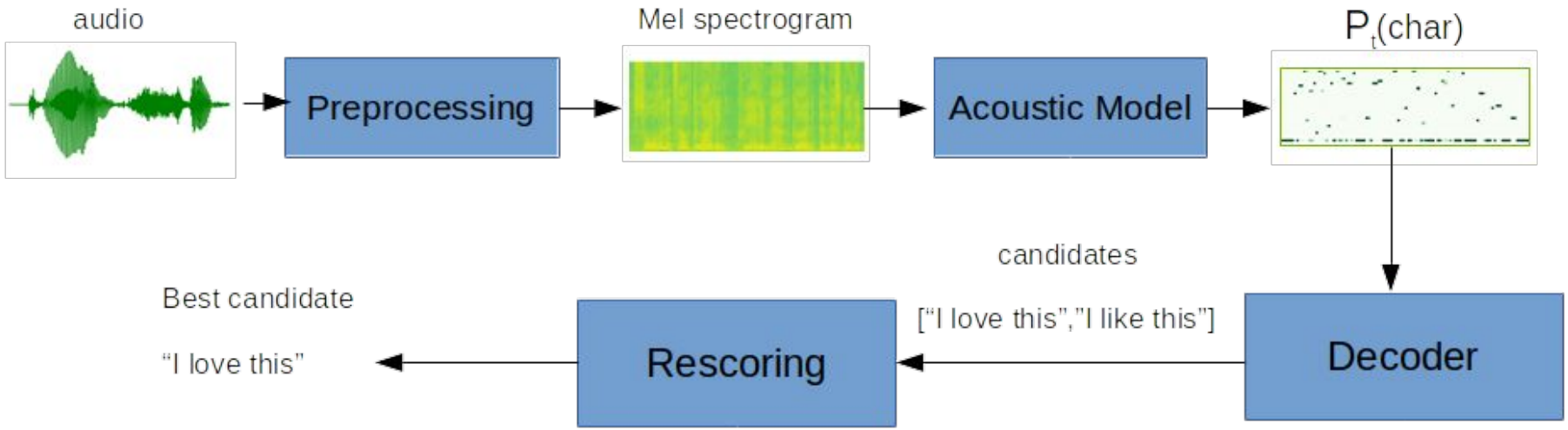
2020 Samsung Research. All rights reserved

TTS Example - Tacotron 2



2020 Samsung Research. All rights reserved

Automatic Speech Recognition (ASR)



2020 Samsung Research. All rights reserved

ASR Example - Deep Speech 2

audio

Preprocessing

- Data augmentation
  - additive Gaussian noise
  - time stretch (resampling)
- Windowing 20-25 ms, stride 10 ms
- FFT, log
- Normalization

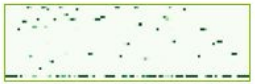
spectrogram

DNN

- 2-3 convolutional layers
  - ch=32, ks=[11, 41], s=[2, 2]
  - ch=32, ks=[11, 21], s=[1, 2]
  - ch=64, ks=[11, 21], s=[1, 2]
- 3-7 bi-/uni- directional GRUs/LSTMs
- 1 row conv layer (for unidirectional)
- 1-2 fully connected layers

- BN/dropout for regularization

$P_t(c)$



training

CTC Loss

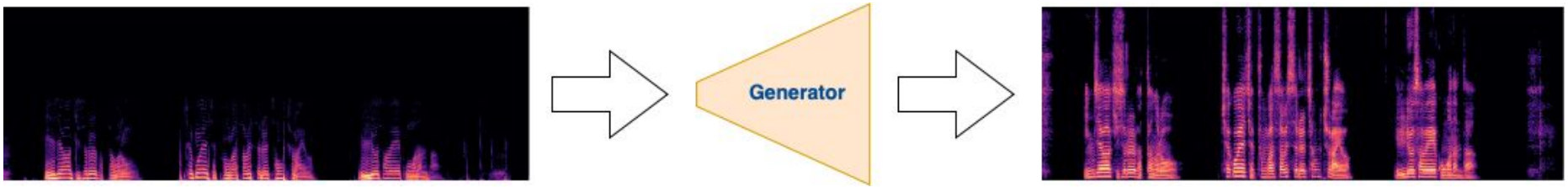
inference

Decoder

- greedy (argmax)
- beam search
- beam search with language model (width=2000-8000)

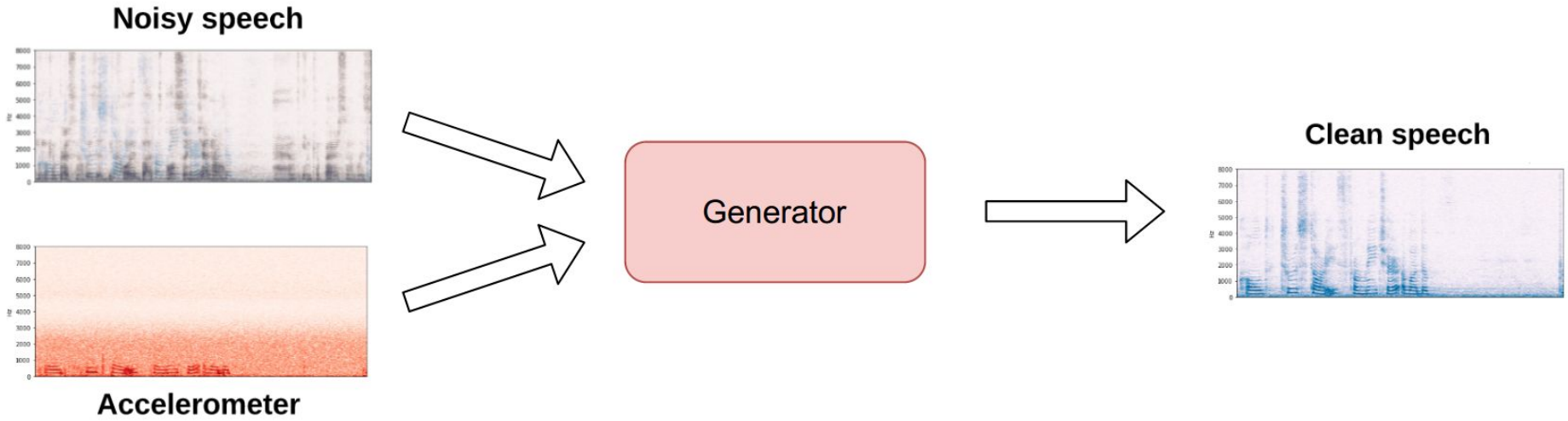
2020 Samsung Research. All rights reserved

Bandwidth Extension



2020 Samsung Research. All rights reserved

Multi-modal Speech Enhancement

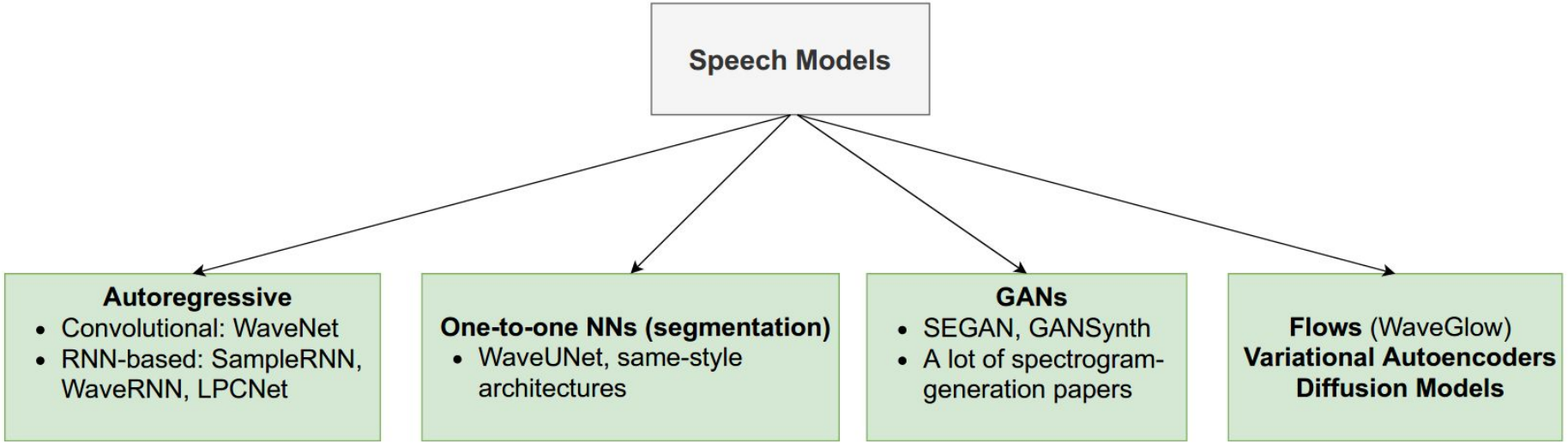


2020 Samsung Research. All rights reserved

Architectures

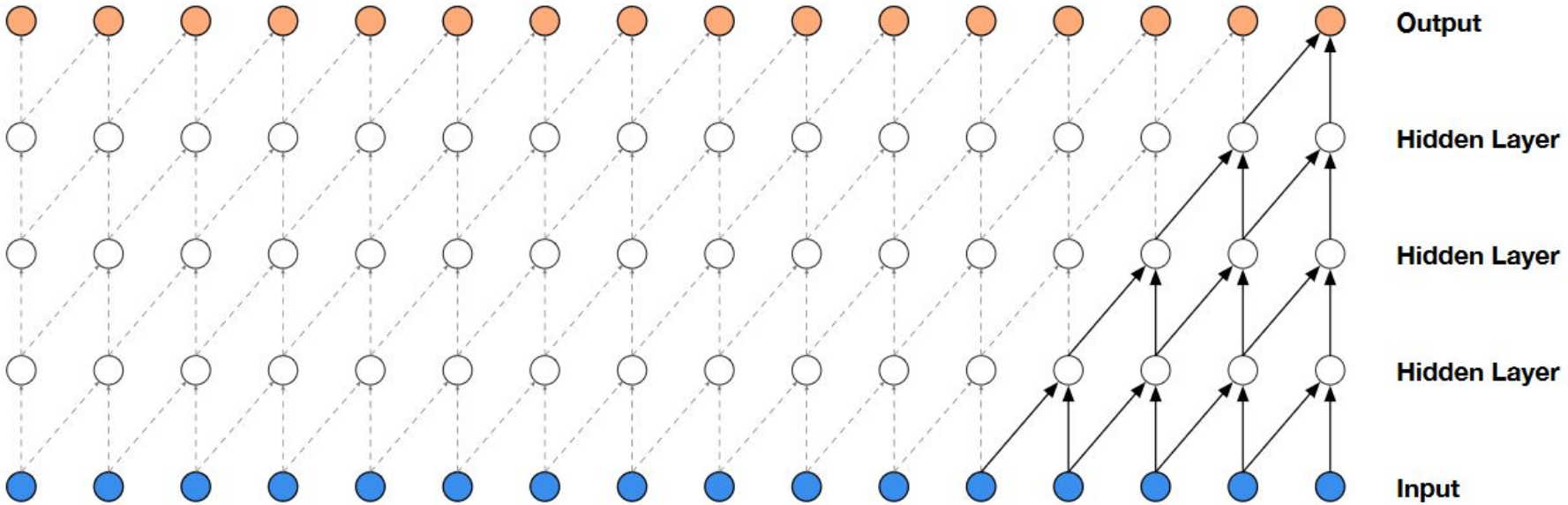
2020 Samsung Research. All rights reserved

Conditional Generative Models



2020 Samsung Research. All rights reserved

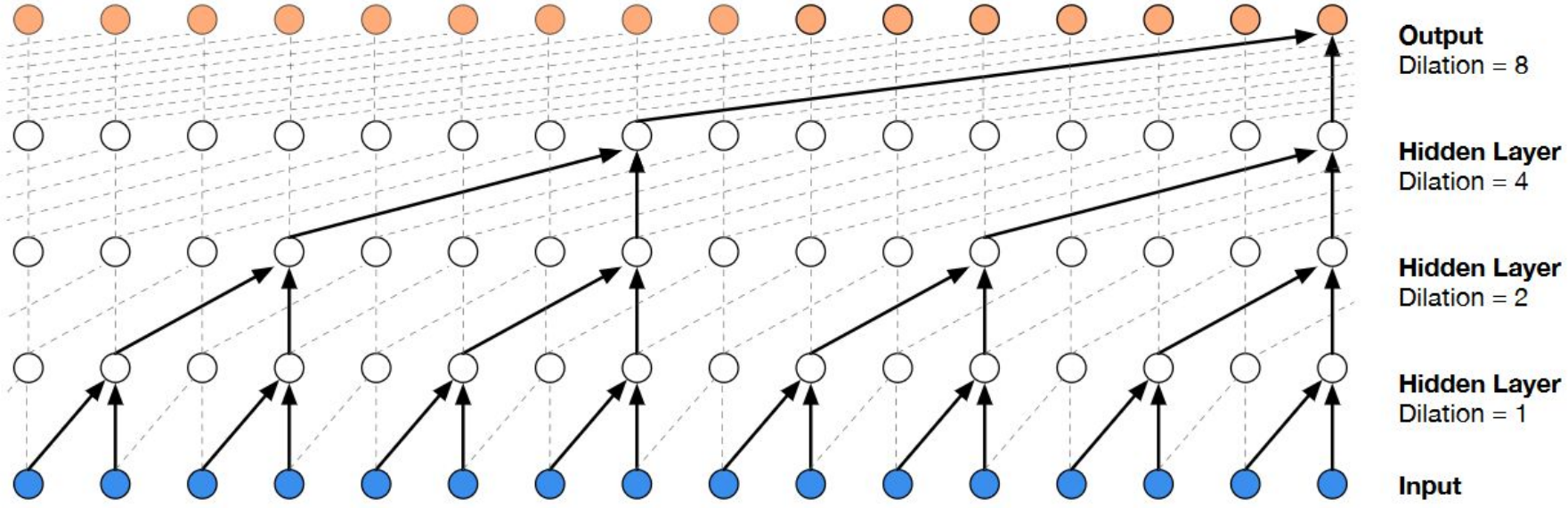
Causal Convolutions have small receptive field



2020 Samsung Research. All rights reserved

WaveNet

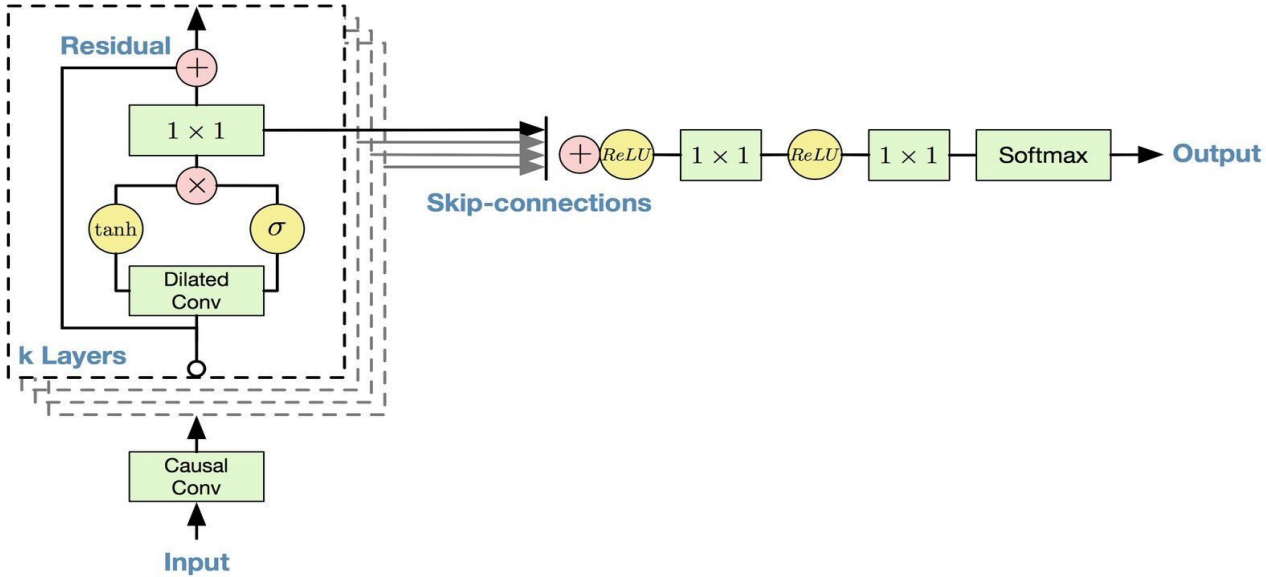
WaveNet uses causal *dilated* convolutions



Dilations grows only up to some limit: 1, 2, 4, …, 512, 1, 2, 4, …, 512, 1, 2, 4, …, 512

2020 Samsung Research. All rights reserved

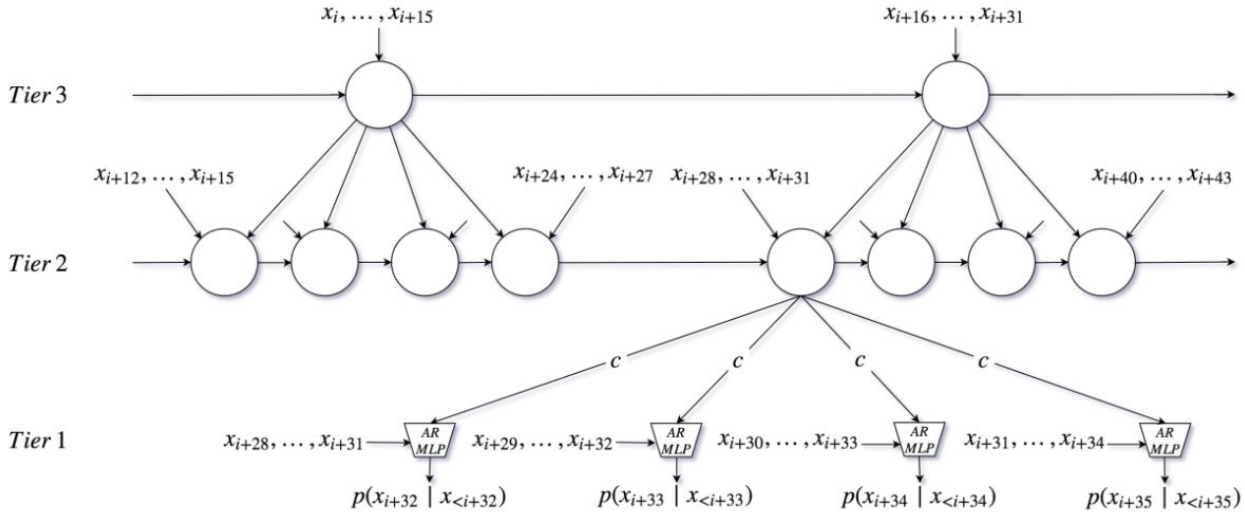
WaveNet: one block structure



2020 Samsung Research. All rights reserved

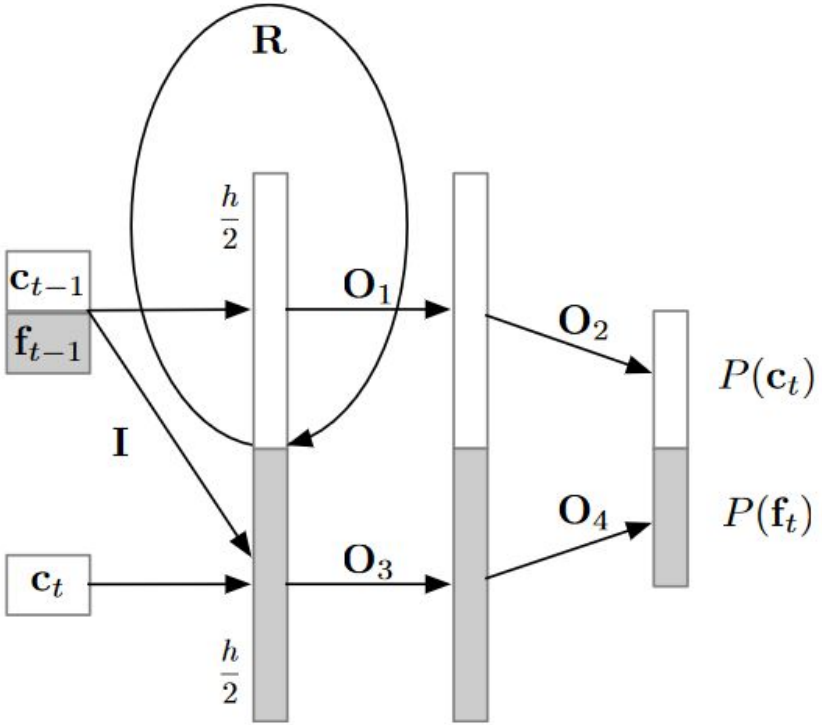
Multiscale RNNs

neither popular nor very effective



2020 Samsung Research. All rights reserved

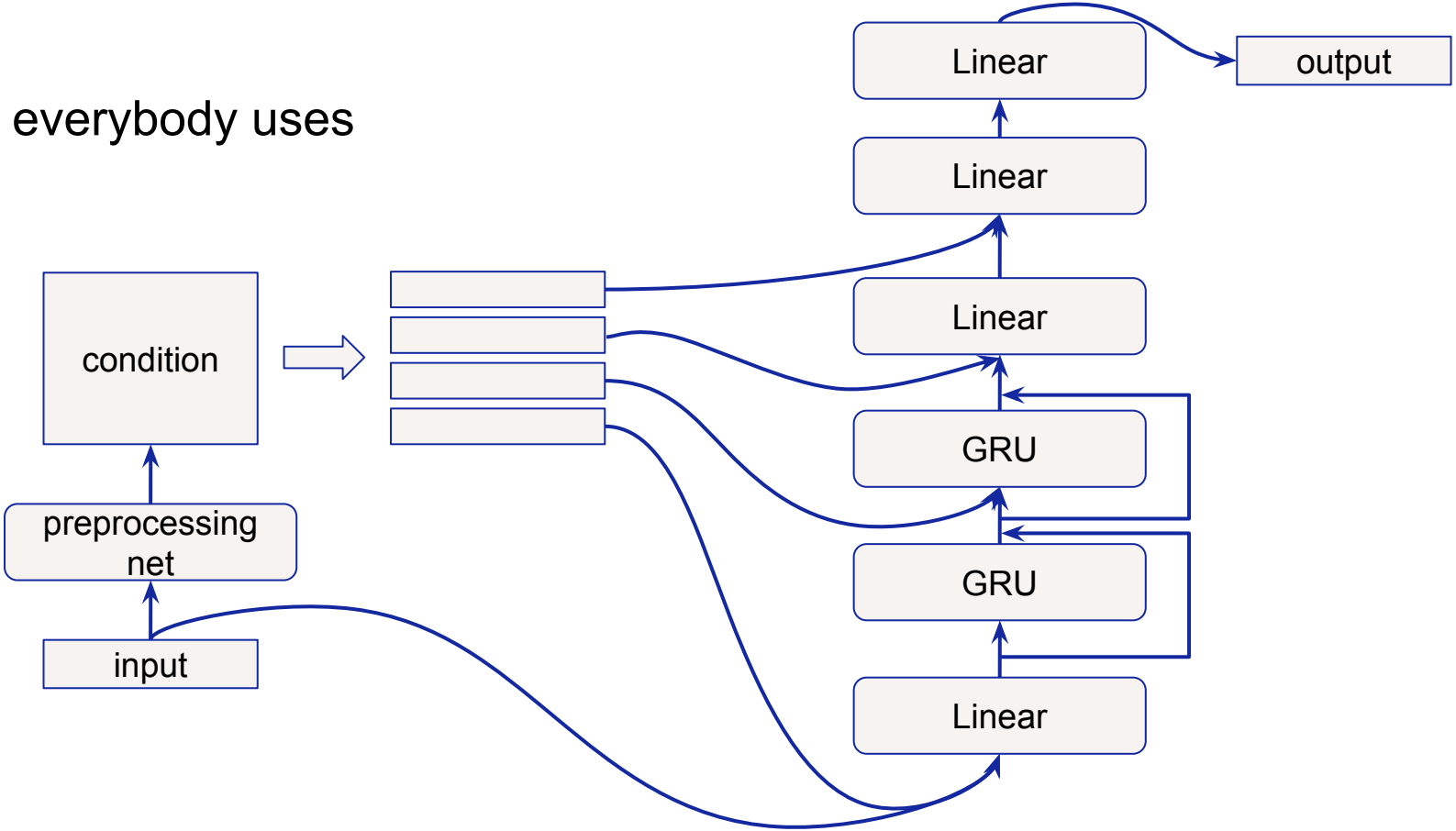
Original net: audio 8bit -> 16bit



2020 Samsung Research. All rights reserved

WaveRNN

What everybody uses



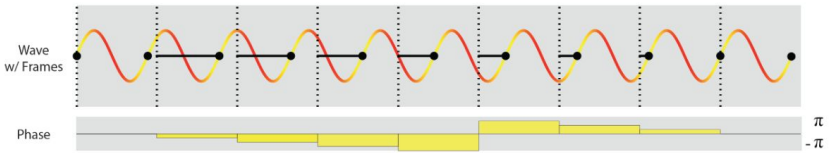
2020 Samsung Research. All rights reserved

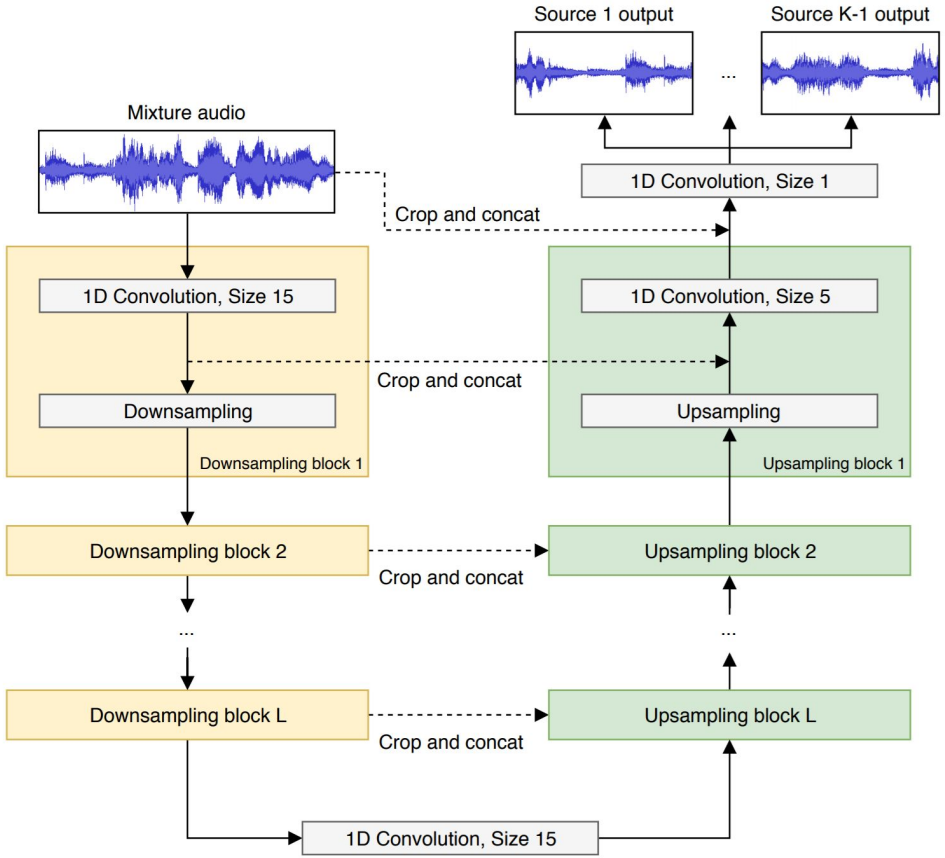
WaveUNet

Pros:

+ lightweight
+ operates on several scales
+ fully convolutional

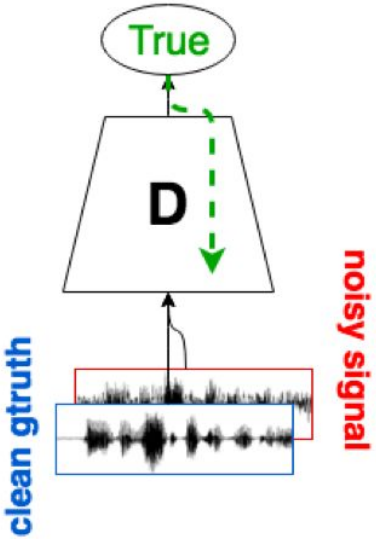
Cons:

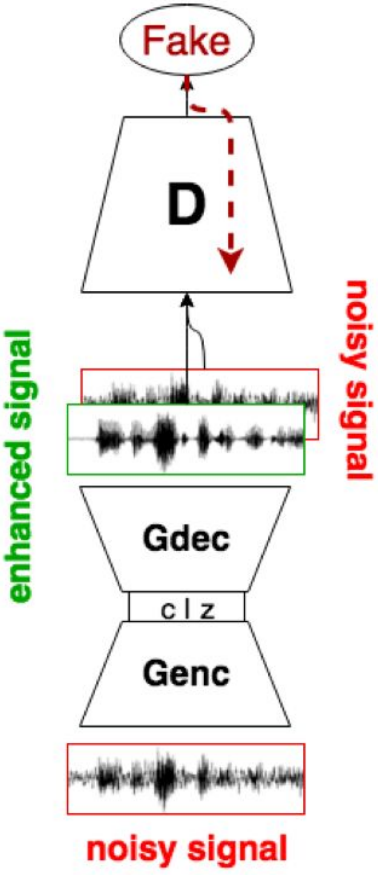


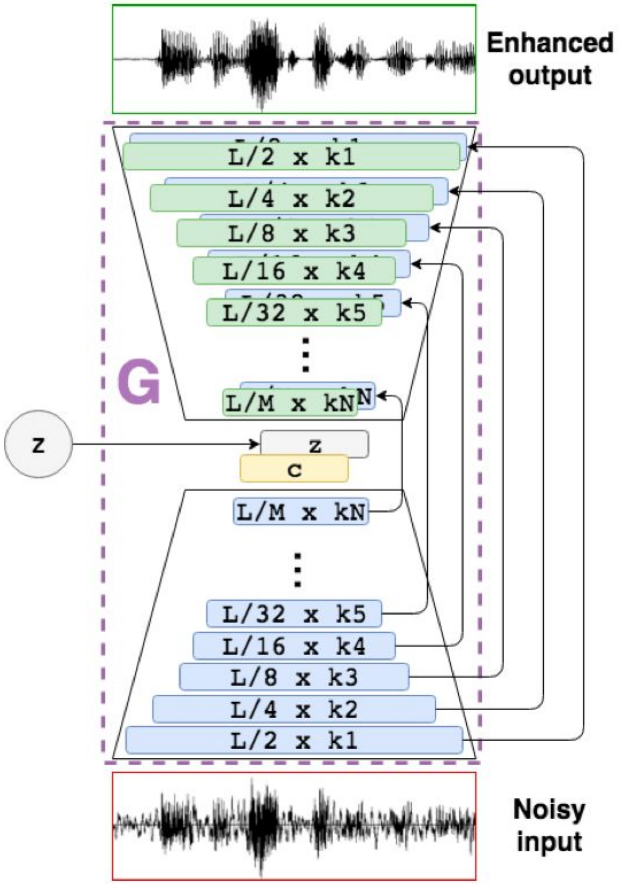


2020 Samsung Research. All rights reserved

GANs (*)





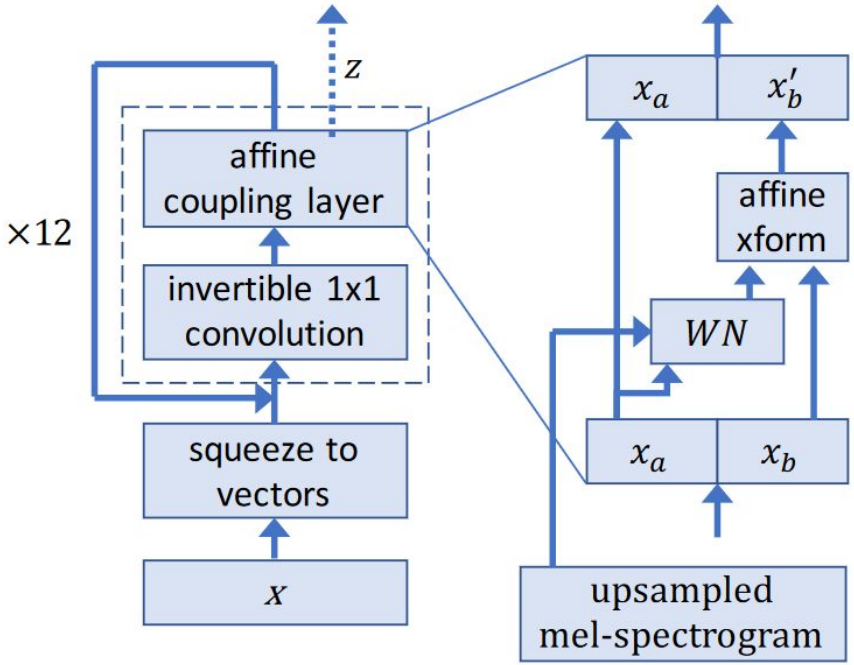


2020 Samsung Research. All rights reserved

WaveGlow

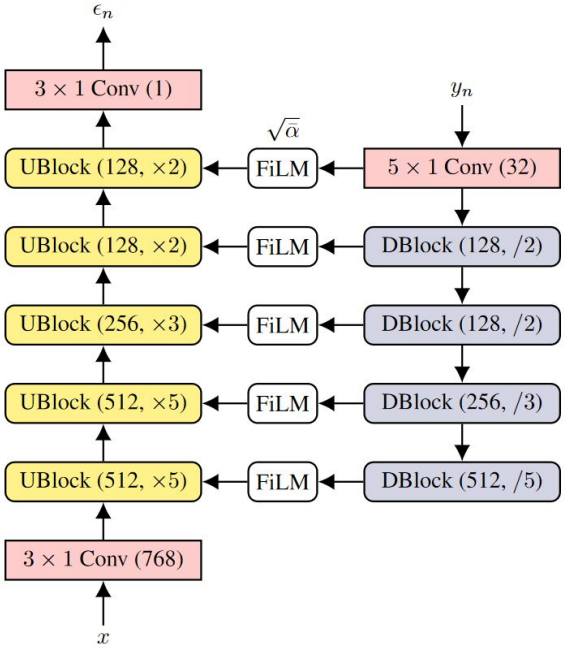
Pros:
+ highly parallel
+ fast on GPUs

Cons:
- slow on CPUs



2020 Samsung Research. All rights reserved

WaveGrad (*)

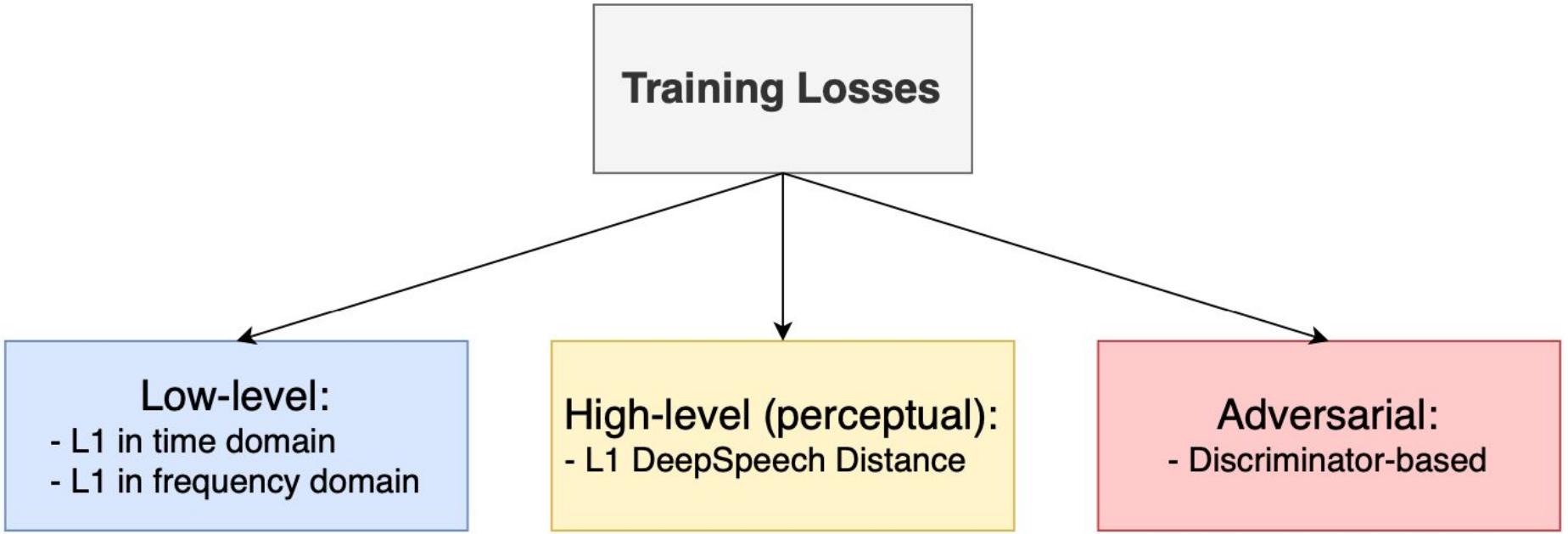


2020 Samsung Research. All rights reserved

Loss Functions

2020 Samsung Research. All rights reserved

Training Losses



2020 Samsung Research. All rights reserved

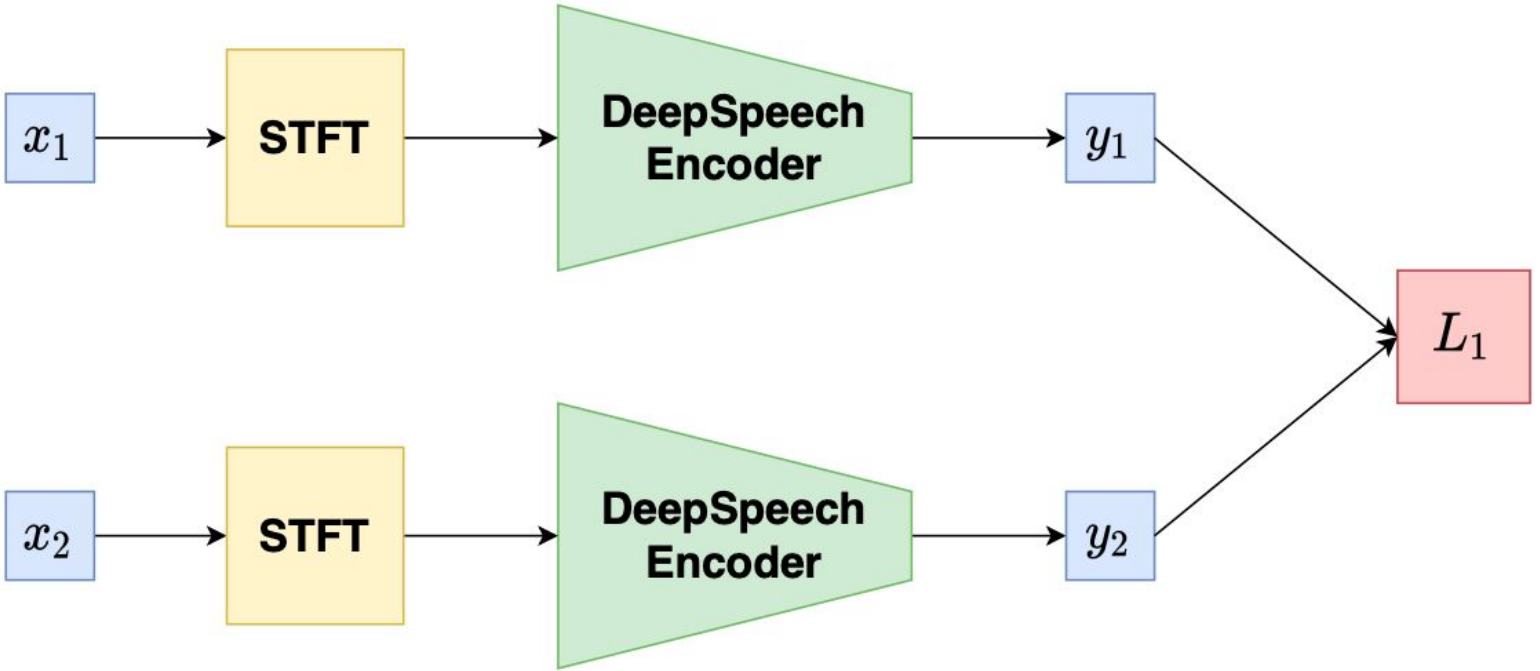
Low-Level Losses

$$L_1(y, \hat{y}) = \|y - \hat{y}\|_1$$

$$L_1^{stft}(y, \hat{y}) = \|\log(1 + |STFT(y)|) - \log(1 + |STFT(\hat{y})|)\|_1$$

2020 Samsung Research. All rights reserved

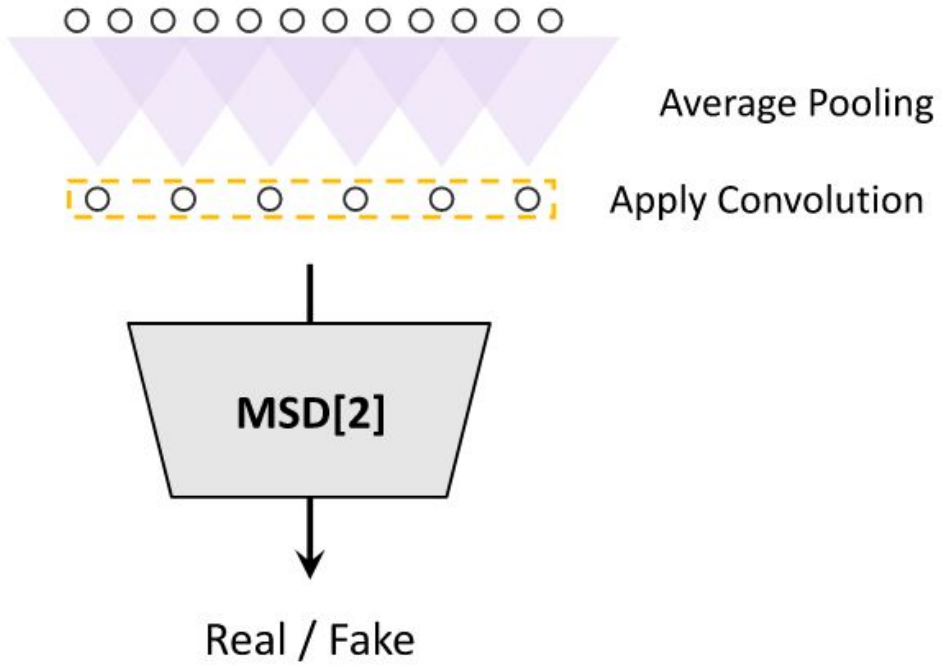
High-Level Loss



2020 Samsung Research. All rights reserved

Adversarial Loss

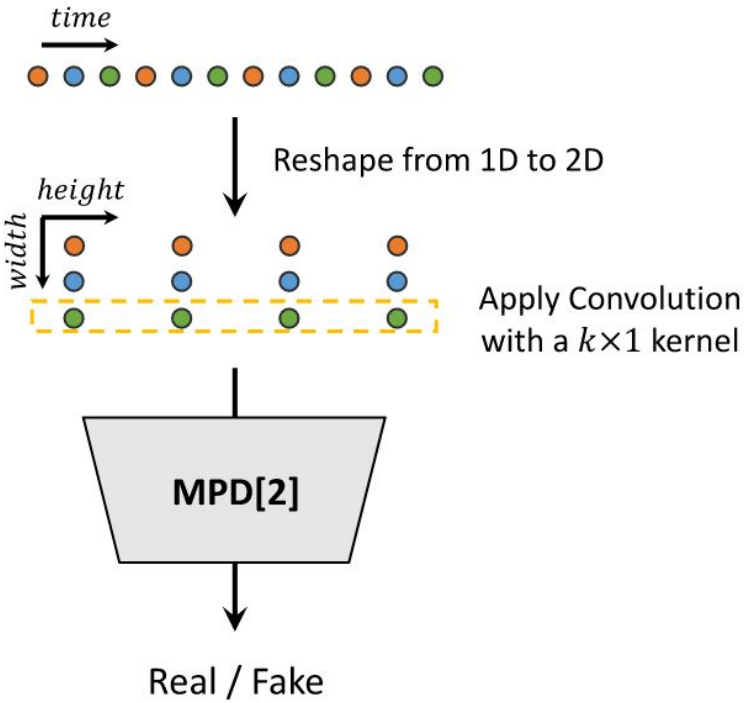
Multi-Scale Discriminators (MSD)



2020 Samsung Research. All rights reserved

Adversarial Loss

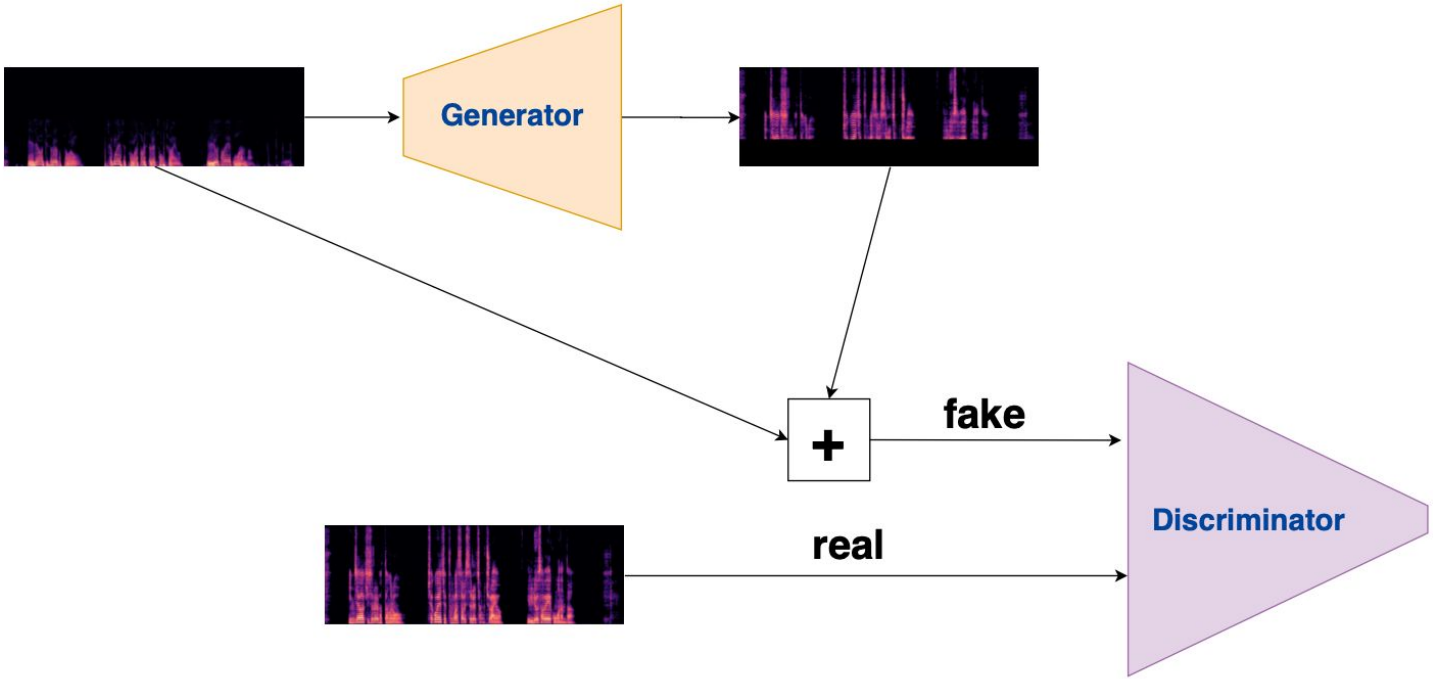
Multi-Period Discriminators (MPD)



2020 Samsung Research. All rights reserved

Adversarial Loss (Add more)

Spectrogram-based Discriminators



2020 Samsung Research. All rights reserved

Adversarial Objective Functions

Hinge Loss:

$$\min_{D_k} \mathbb{E}_x \Big[\min(0, 1 - D_k(x)) \Big] + \mathbb{E}_{s,z} \Big[\min(0, 1 + D_k(G(s, z))) \Big], \; \forall k = 1, 2, 3$$

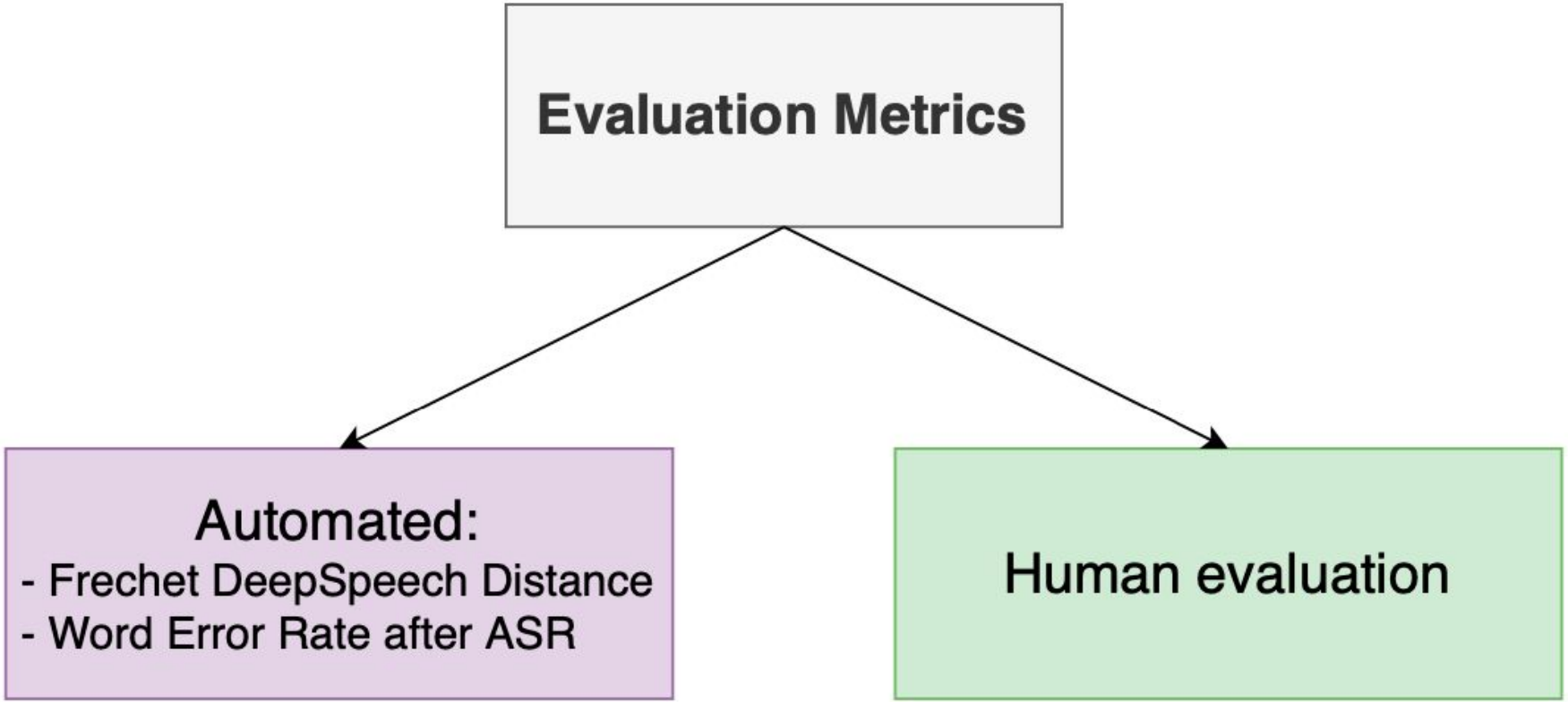
$$\min_{G} \mathbb{E}_{s,z} \left[\sum_{k=1,2,3} -D_k(G(s, z)) \right]$$

Feature Matching Loss:

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

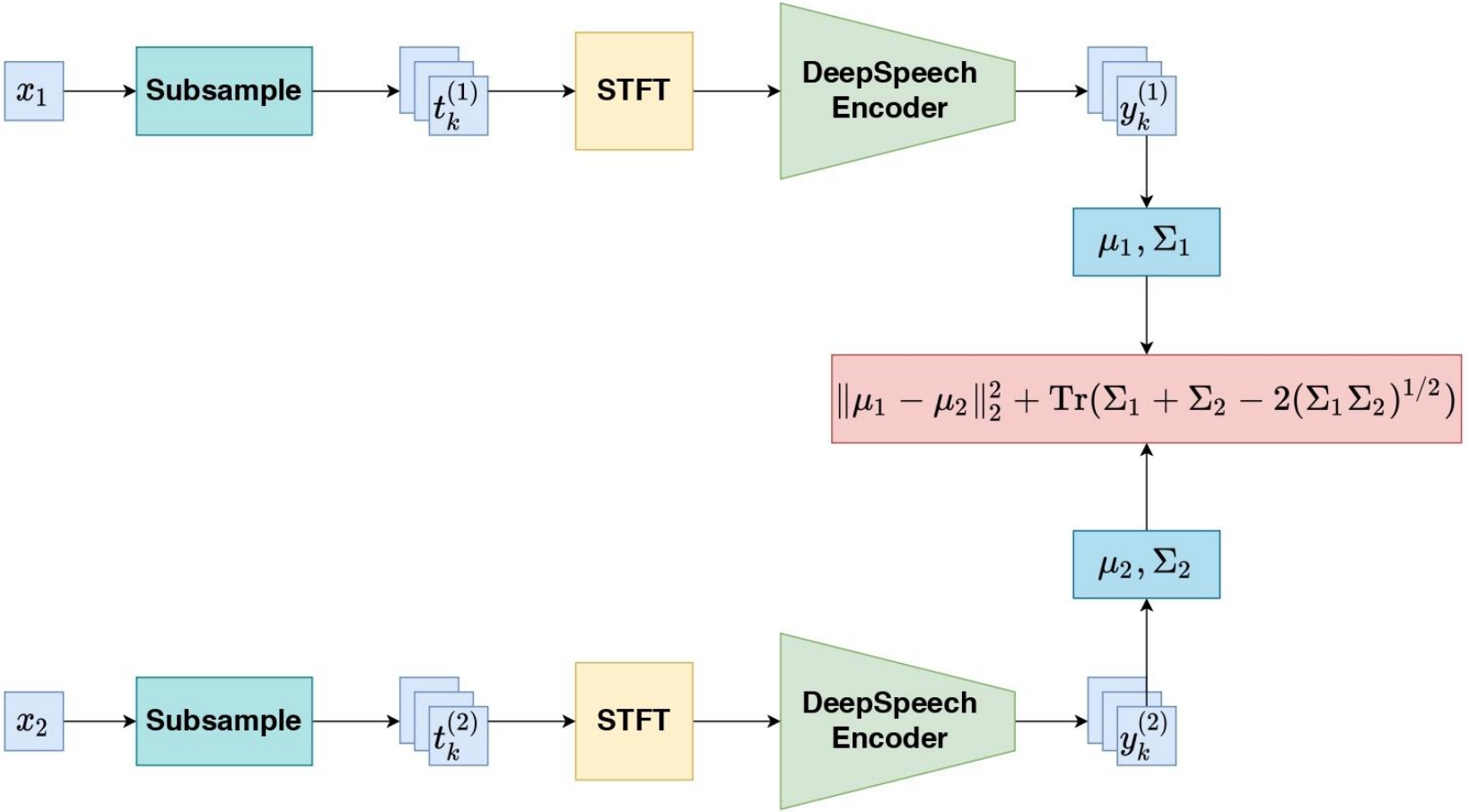
2020 Samsung Research. All rights reserved

Evaluation Metrics



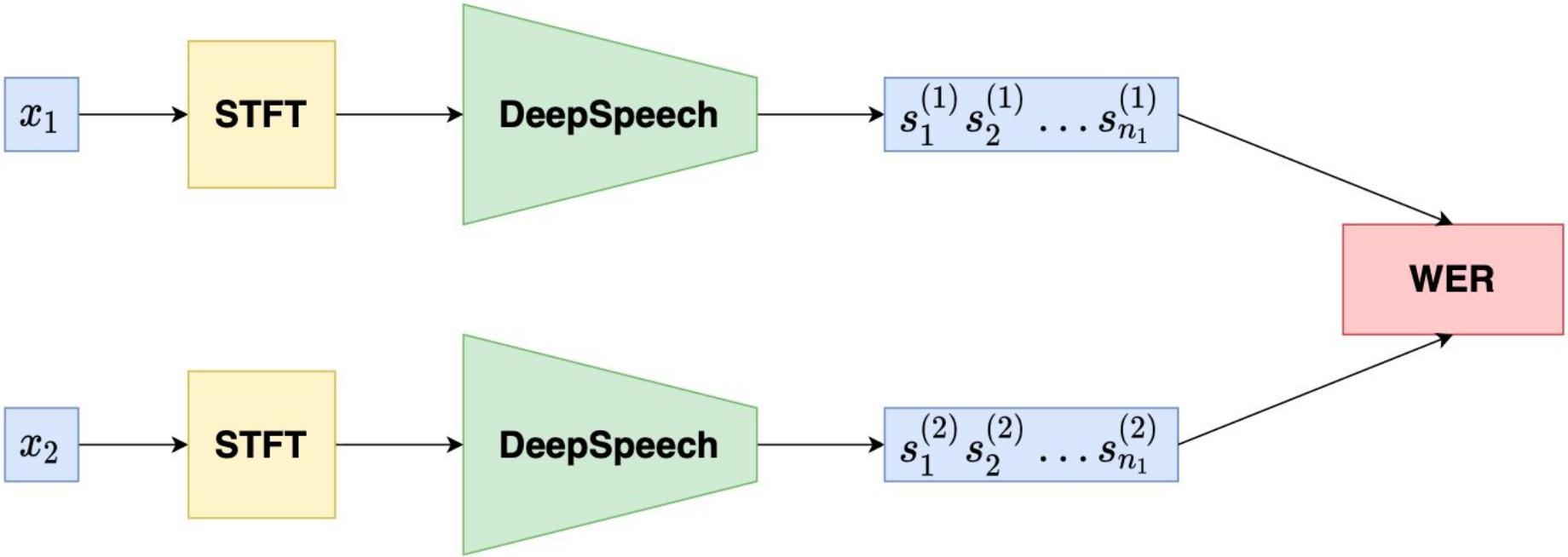
2020 Samsung Research. All rights reserved

Frechet DeepSpeech Distance



2020 Samsung Research. All rights reserved

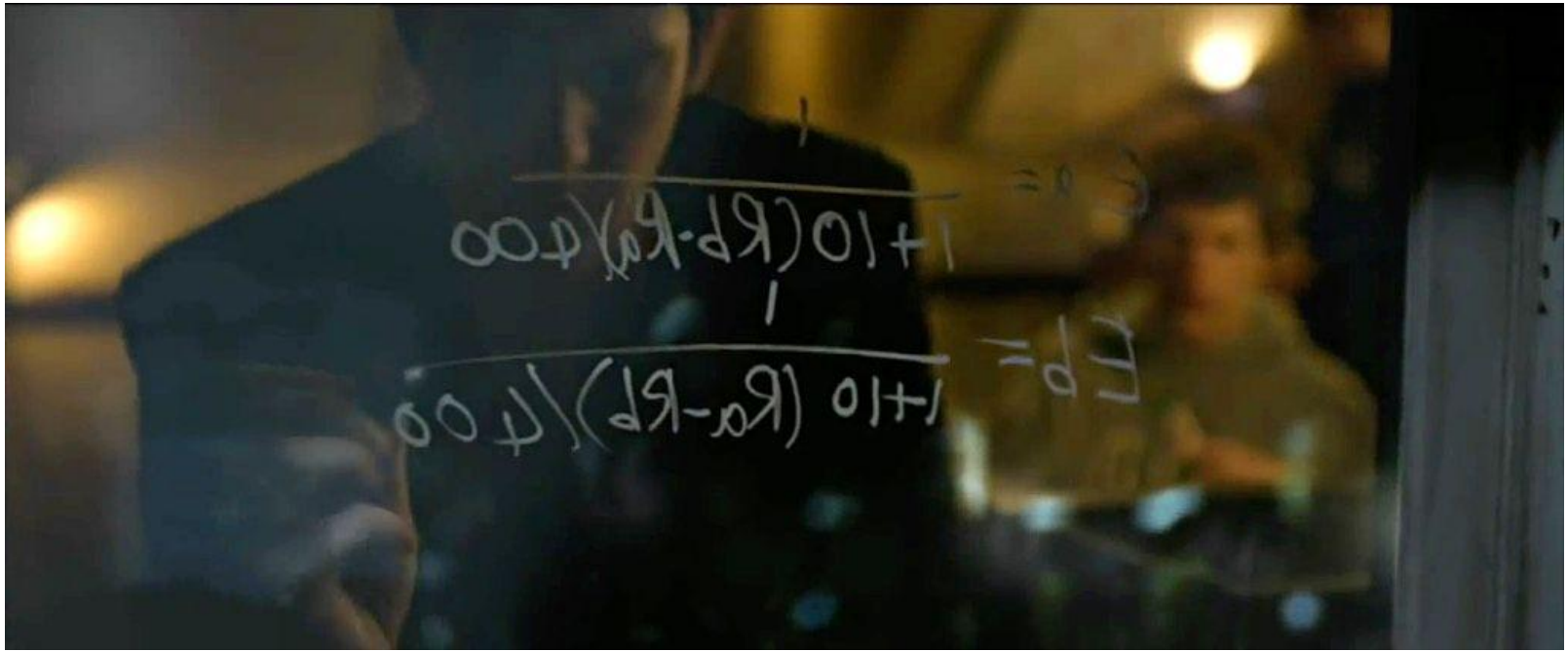
Word Error Rate after ASR



2020 Samsung Research. All rights reserved

Human Evaluation

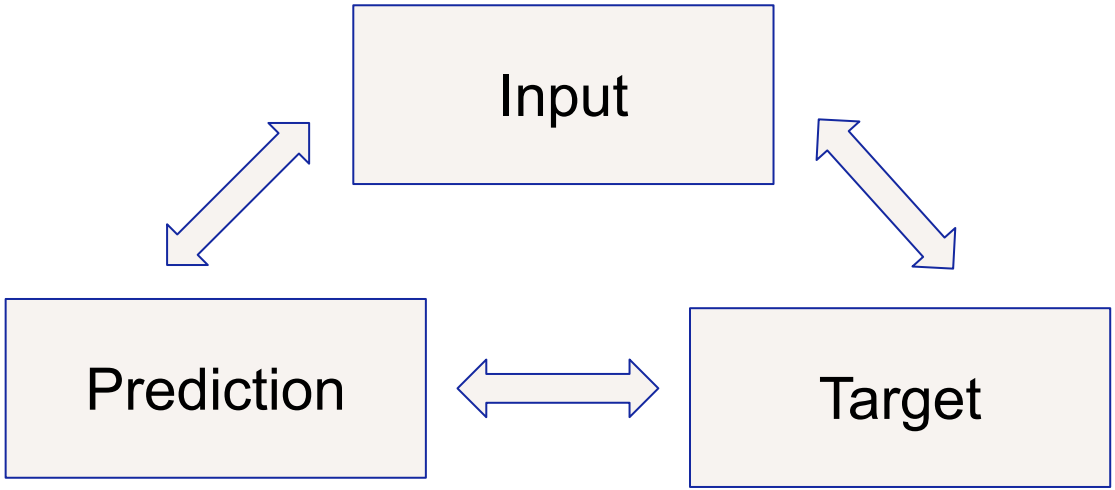
Idea: pairwise comparisons



2020 Samsung Research. All rights reserved

Human Evaluation

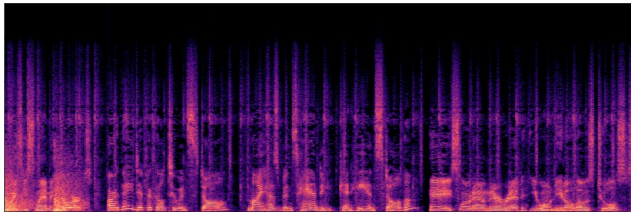
Answers: <, >

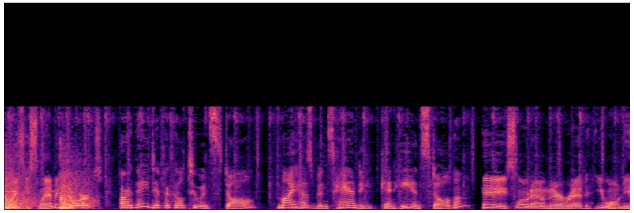


2020 Samsung Research. All rights reserved

Human Evaluation

Problem

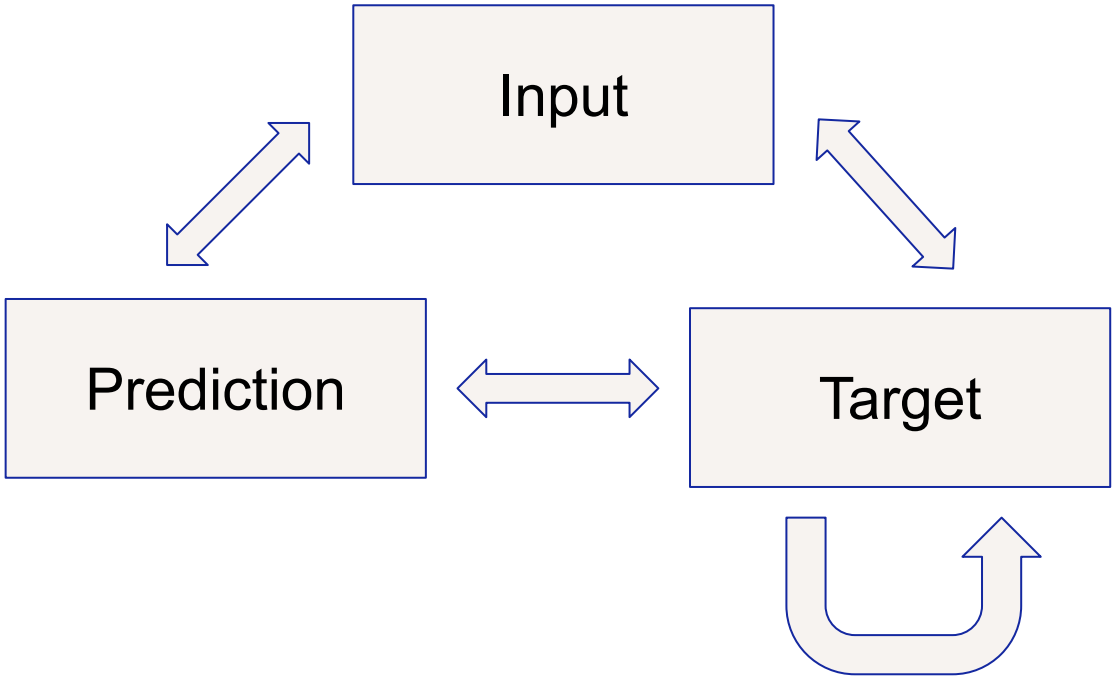




2020 Samsung Research. All rights reserved

Human Evaluation

Answers: <, >, **=, ≠**

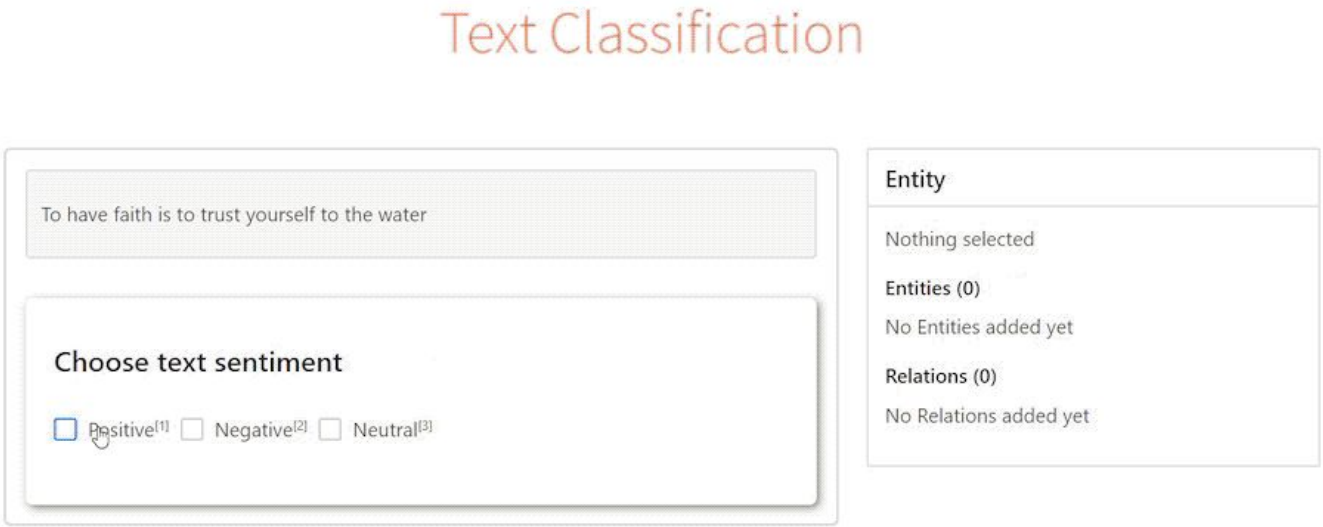


2020 Samsung Research. All rights reserved

Human Evaluation

Label Studio

https://labelstud.io/



2020 Samsung Research. All rights reserved

Datasets

2020 Samsung Research. All rights reserved

Speech Datasets

LJ Speech
1 speaker,
~ 24 hours

VCTK
110 speakers
~ 44 hours

Mozilla Common VoiceSet

- Allows commercial usage
- Two months of train data (~64Gb)
- 17 train and 10 validation languages
- Highly imbalanced (languages, genders)

2020 Samsung Research. All rights reserved

Speech Datasets

LJ Speech
1 speaker,
~ 24 hours

VCTK
110 speakers
~ 44 hours

Mozilla Common VoiceSet

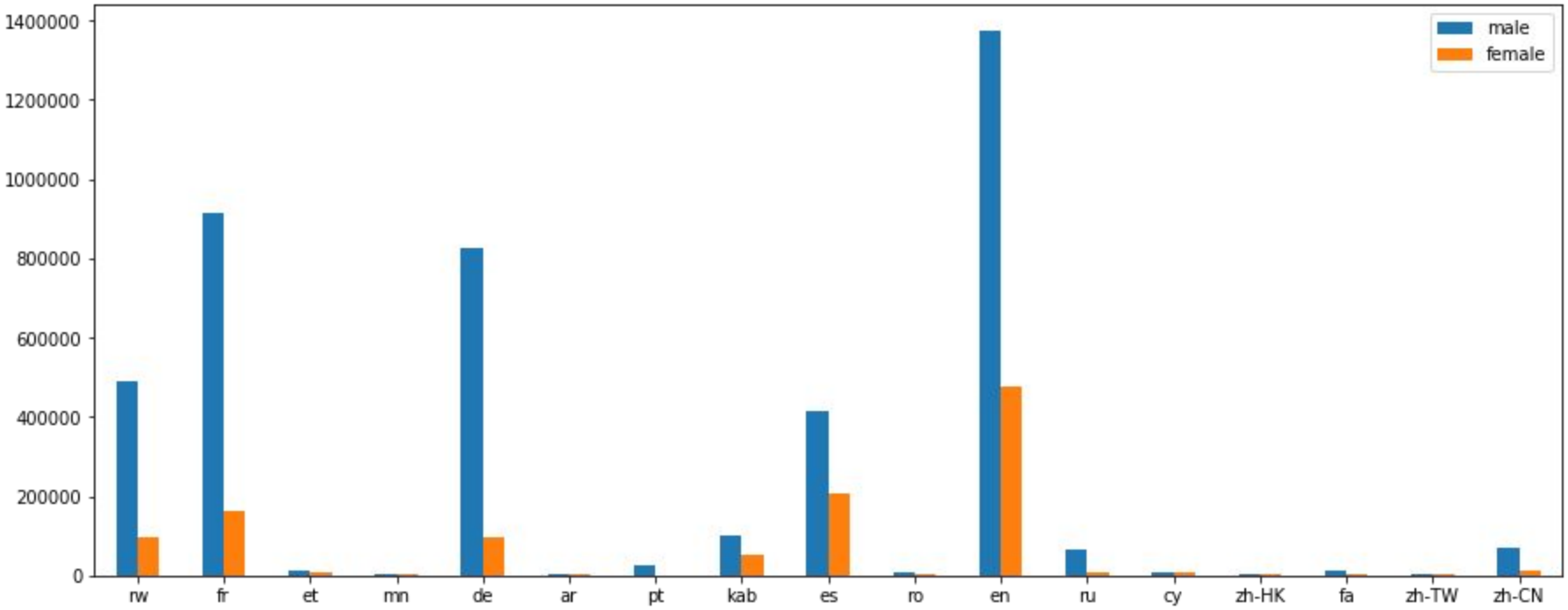
- Allows commercial usage
- Two months of train data (~64Gb)
- 17 train and 10 validation languages
- Highly imbalanced (languages, genders)
- Credits to Jules

Jules Churkina consulted us while we balanced languages

2020 Samsung Research. All rights reserved

Speech Datasets

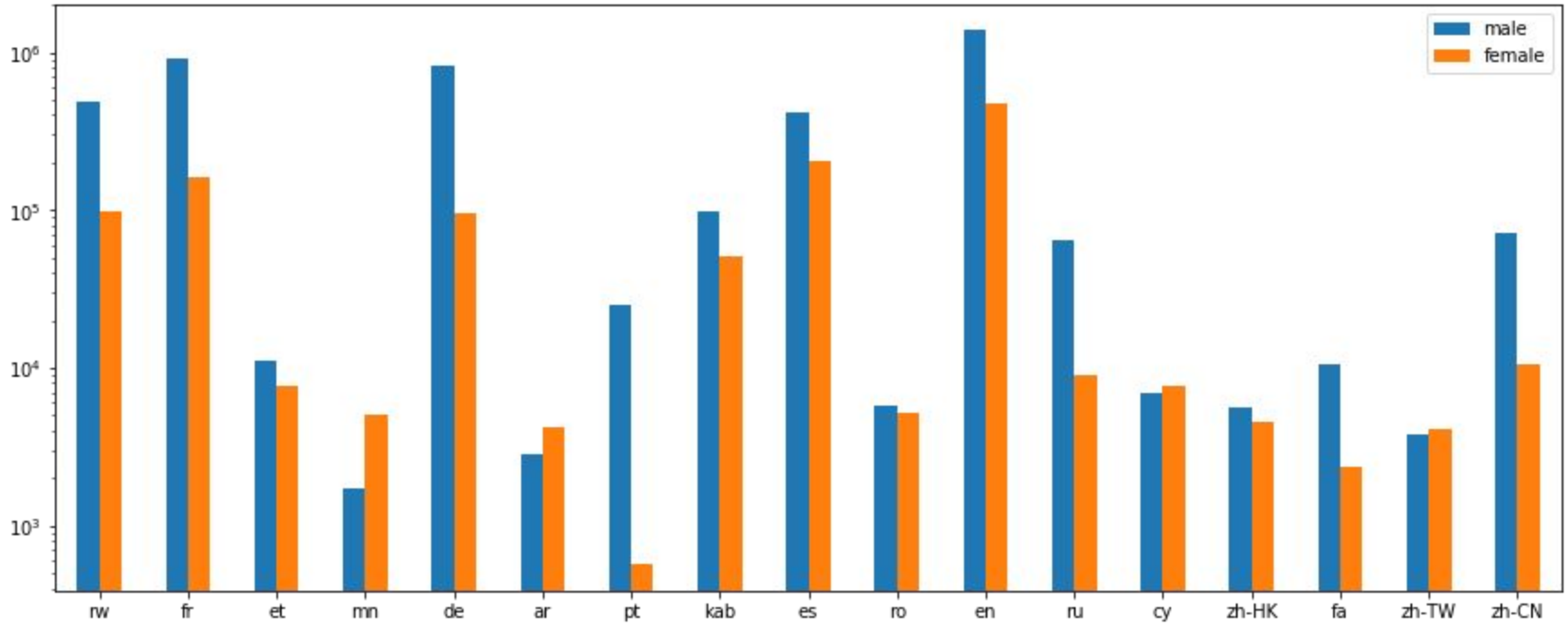
VoiseSet Imbalance: Train Set



2020 Samsung Research. All rights reserved

Speech Datasets

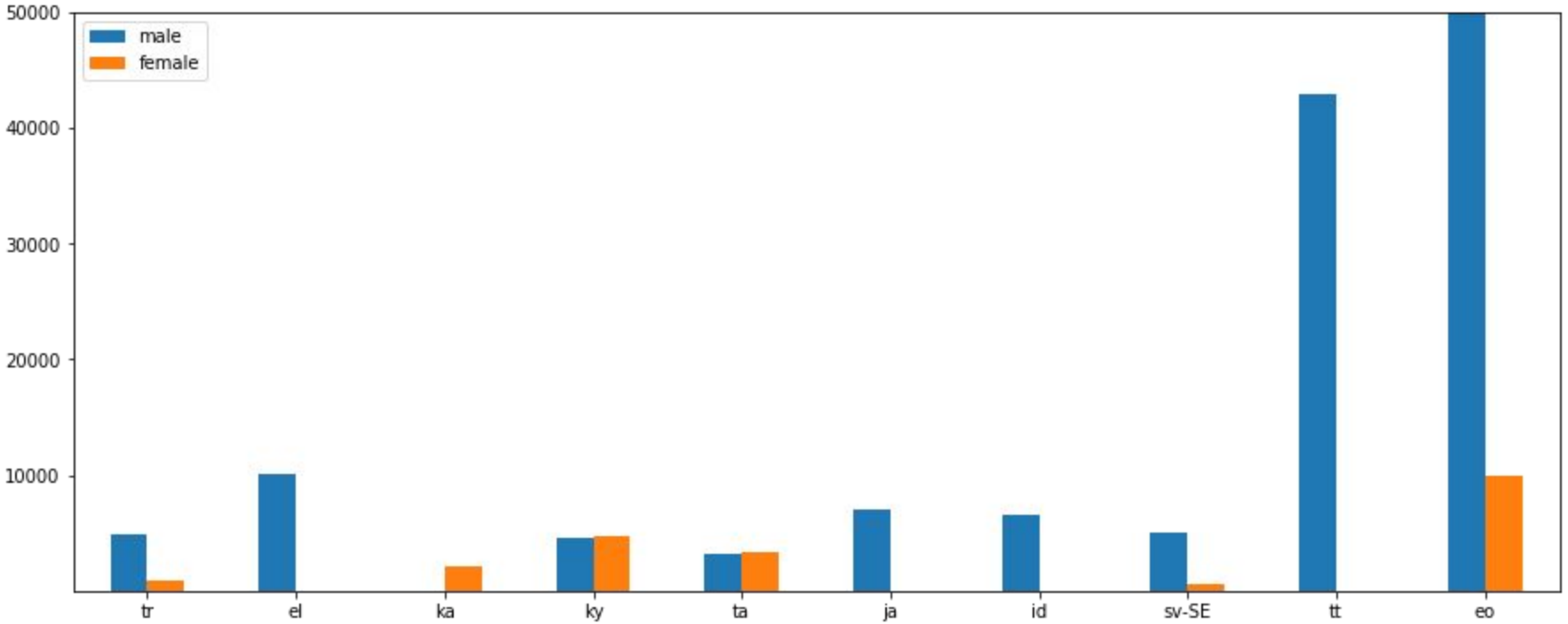
VoiseSet Imbalance: Train Set



2020 Samsung Research. All rights reserved

Speech Datasets

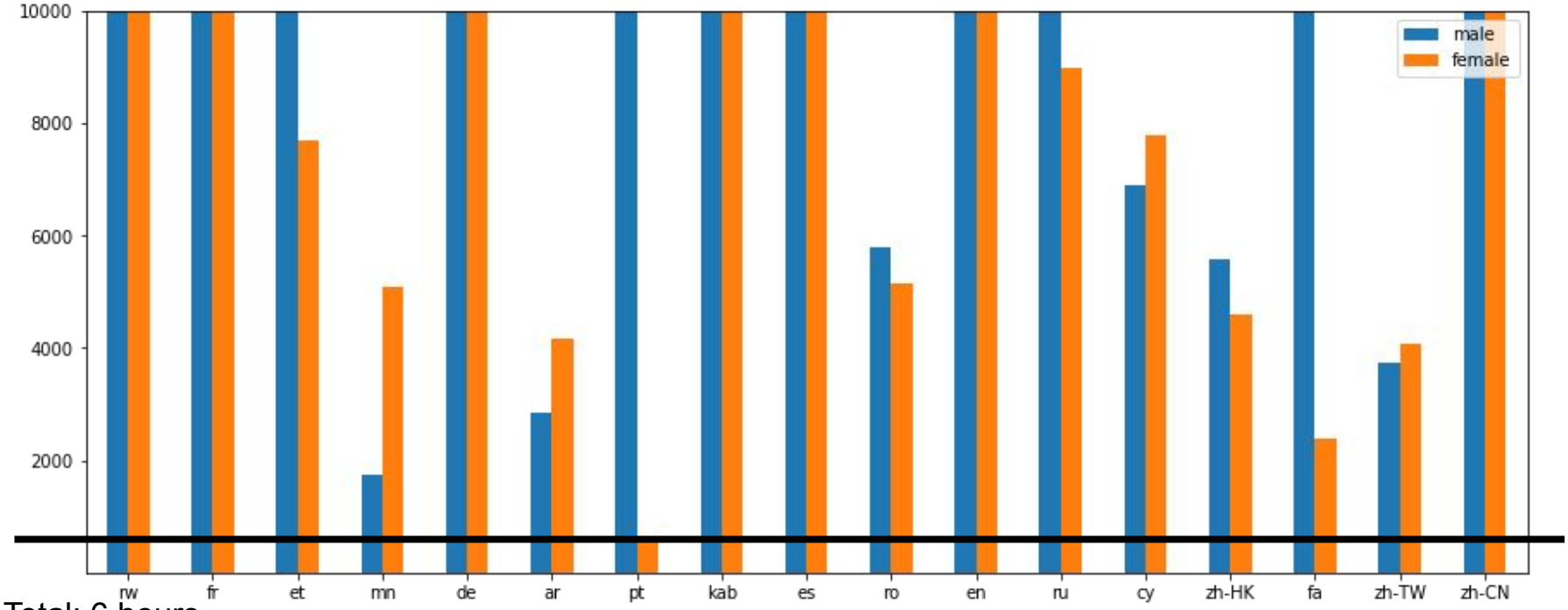
VoiseSet Imbalance: Validation Set



2020 Samsung Research. All rights reserved

Speech Datasets

VoiseSet Imbalance: Train Set



Total: 6 hours

2020 Samsung Research. All rights reserved

Universal Sound Datasets

AudioSet
~ 20K audios
Length of each audio is 10s
Labelling: audio event tags
No raw audio
Downloading is illegal

FreeSound

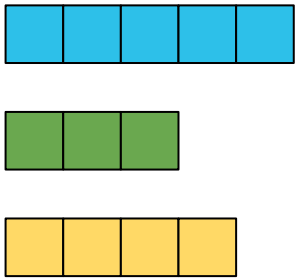
Each audio is either CC0, CC-BY, Sampling+ or CC-BY-NC
Tags, description and other labels are available
~13Gb of uncompressed filtered wavs
~ 1 day of train set

2020 Samsung Research. All rights reserved

Big Data for Slow FileSystem

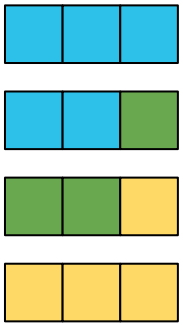
Data Pool

Audios splitted to frames



Batches

All frames of audio are used



- Fast
- Highly correlated batches

One frame per audio is used

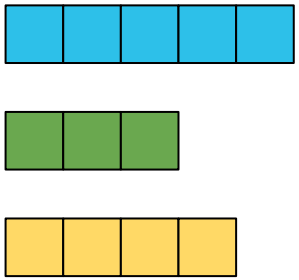
- Slow (most read frames are dropped)
- Uncorrelated batches

2020 Samsung Research. All rights reserved

Big Data for Slow FileSystem

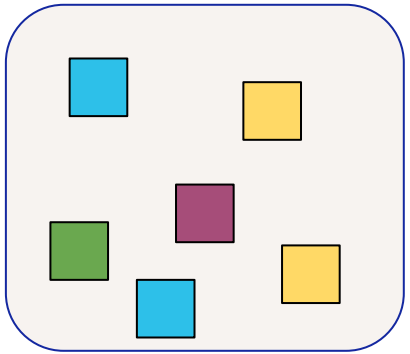
Data Pool

Audios splitted to frames



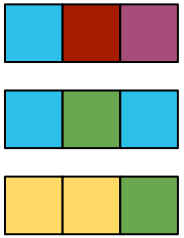
read audios until required pool size is reached

Pool



Invariant:
there are at least K frames in Pool before sampling

sample batches randomly from the pool



- Fast
- Weakly correlated batches

2020 Samsung Research. All rights reserved

Big Data for Slow FileSystem

Databases for Large Datasets

File System

HDF5 (Hierarchical Data Format)

LMDB (Lightning Memory-Mapped Database)

WebDataset (https://github.com/tmbdev/webdataset/)

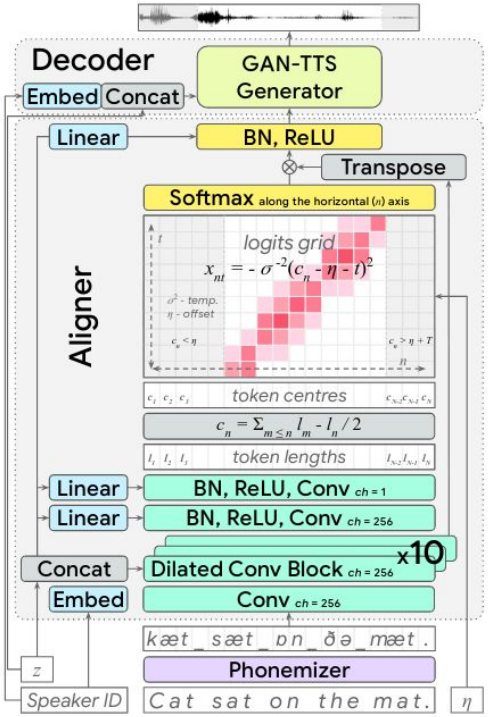
2020 Samsung Research. All rights reserved

SOTAs

2020 Samsung Research. All rights reserved

TTS Problem

End-to-End Adversarial Text-to-Speech (ICLR 2021 oral)



- Data - Multispeaker (MS) or Single Speaker (SS)
- Inputs - characters (Ch) or phonemes (Ph)
- Random Window Discriminators (RWD)
- Mel-spectrogram Discriminator (MSD)

2020 Samsung Research. All rights reserved

End-to-End Adversarial Text-to-Speech

Model	Data	Inputs	RWD	MSD	$\mathcal{L}_{\text{length}}$	$\mathcal{L}_{\text{pred}}$	Align	MOS
Natural Speech				-				4.55 ± 0.075
GAN-TTS (Bińkowski et al., 2020)				-				4.213 ± 0.046
WaveNet (van den Oord et al., 2016)				-				4.41 ± 0.069
Par. WaveNet (van den Oord et al., 2018)				-				4.41 ± 0.078
Tacotron 2 (Shen et al., 2018)				-				4.526 ± 0.066
No $\mathcal{L}_{\text{length}}$	MS	Ph	✓	✓	×	$\mathcal{L}''_{\text{pred}}$	MI	[does not train]
No $\mathcal{L}_{\text{pred}}$	MS	Ph	✓	✓	✓	×	MI	[does not train]
No Discriminators	MS	Ph	×	×	✓	$\mathcal{L}''_{\text{pred}}$	MI	1.407 ± 0.040
No RWDs	MS	Ph	×	✓	✓	$\mathcal{L}''_{\text{pred}}$	MI	2.526 ± 0.060
No Phonemes	MS	Ch	✓	✓	✓	$\mathcal{L}''_{\text{pred}}$	MI	3.423 ± 0.073
No MelSpecD	MS	Ph	✓	×	✓	$\mathcal{L}''_{\text{pred}}$	MI	3.525 ± 0.057
No Mon. Int.	MS	Ph	✓	✓	✓	$\mathcal{L}''_{\text{pred}}$	Attn	3.551 ± 0.073
No DTW	MS	Ph	✓	✓	✓	$\mathcal{L}_{\text{pred}}$	MI	3.559 ± 0.065
Single Speaker	SS	Ph	✓	✓	✓	$\mathcal{L}''_{\text{pred}}$	MI	3.829 ± 0.055
EATS (Ours)	MS	Ph	✓	✓	✓	$\mathcal{L}''_{\text{pred}}$	MI	4.083 ± 0.049

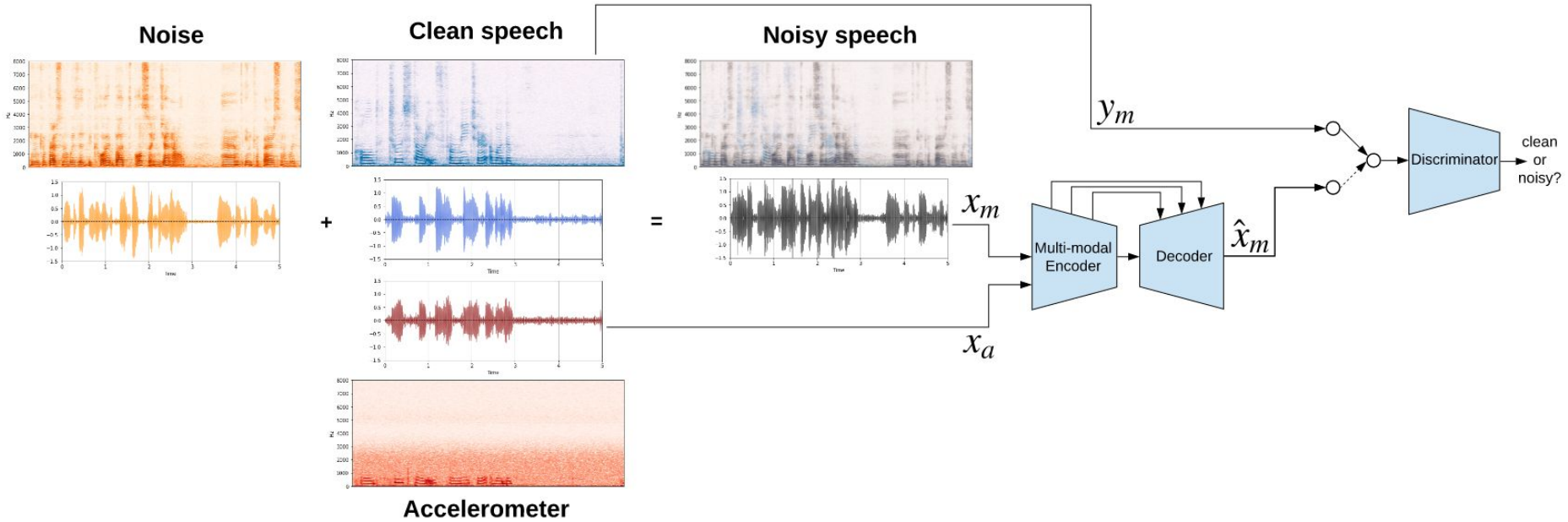
Input text: *In this work, we take on the challenging task of learning to synthesise speech from normalised text or phonemes in an end-to-end manner, resulting in models which operate directly on character or phoneme input sequences and produce raw speech audio outputs.*

Main model: Single speaker: No RWDs: No MelSpecD: No Discriminators:

2020 Samsung Research. All rights reserved

Multi-modal Speech Enhancement

SEANet: A Multi-modal Speech Enhancement Network



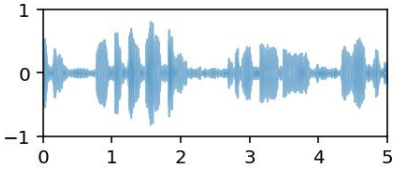
2020 Samsung Research. All rights reserved

SEANet Results

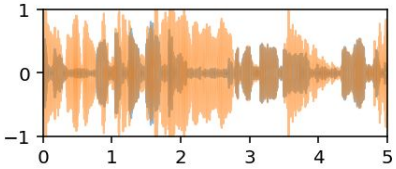
scenario	split	SEANet audio + accel	SEANet audio only
Mixed noise	1	9.9 ± 0.2	8.4 ± 0.2
	2	8.0 ± 0.2	7.9 ± 0.1
	3	8.3 ± 0.1	7.2 ± 0.2
	4	8.8 ± 0.1	8.1 ± 0.1
	5	9.9 ± 0.1	8.4 ± 0.1
	avg.	8.9	8.0
Mixed speech	1	10.1 ± 0.1	-0.9 ± 0.1
	2	8.6 ± 0.1	-0.9 ± 0.1
	3	9.2 ± 0.1	-0.7 ± 0.0
	4	9.0 ± 0.2	-1.0 ± 0.1
	5	11.1 ± 0.2	-0.9 ± 0.1
	avg.	9.6	-0.9

Mixed speech example:

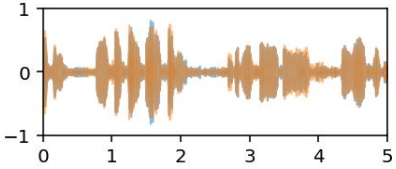
Clean sample:



Noisy sample:



Denoised sample:



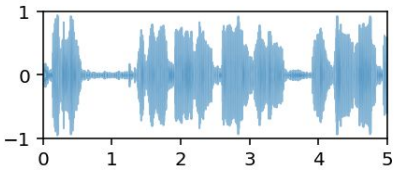
2020 Samsung Research. All rights reserved

SEANet Results

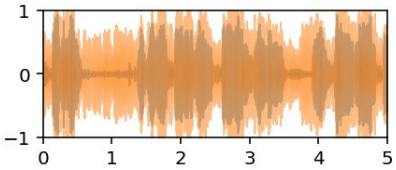
scenario	split	SEANet audio + accel	SEANet audio only
Mixed noise	1	9.9 ± 0.2	8.4 ± 0.2
	2	8.0 ± 0.2	7.9 ± 0.1
	3	8.3 ± 0.1	7.2 ± 0.2
	4	8.8 ± 0.1	8.1 ± 0.1
	5	9.9 ± 0.1	8.4 ± 0.1
	avg.	8.9	8.0
Mixed speech	1	10.1 ± 0.1	-0.9 ± 0.1
	2	8.6 ± 0.1	-0.9 ± 0.1
	3	9.2 ± 0.1	-0.7 ± 0.0
	4	9.0 ± 0.2	-1.0 ± 0.1
	5	11.1 ± 0.2	-0.9 ± 0.1
	avg.	9.6	-0.9

Mixed noise example:

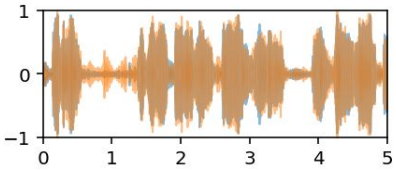
Clean sample:



Noisy sample:



Denoised sample:

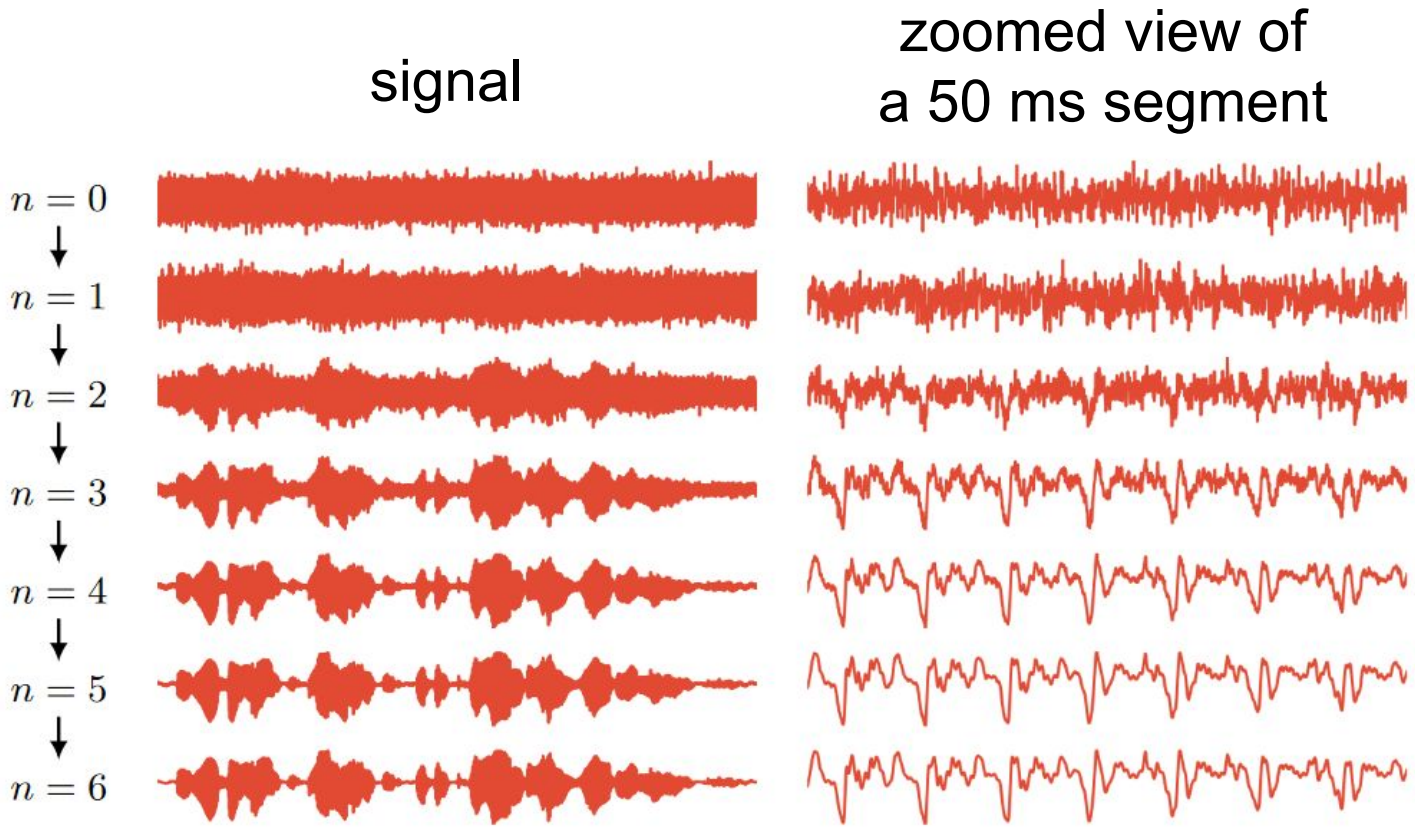


2020 Samsung Research. All rights reserved

Diffusion Models

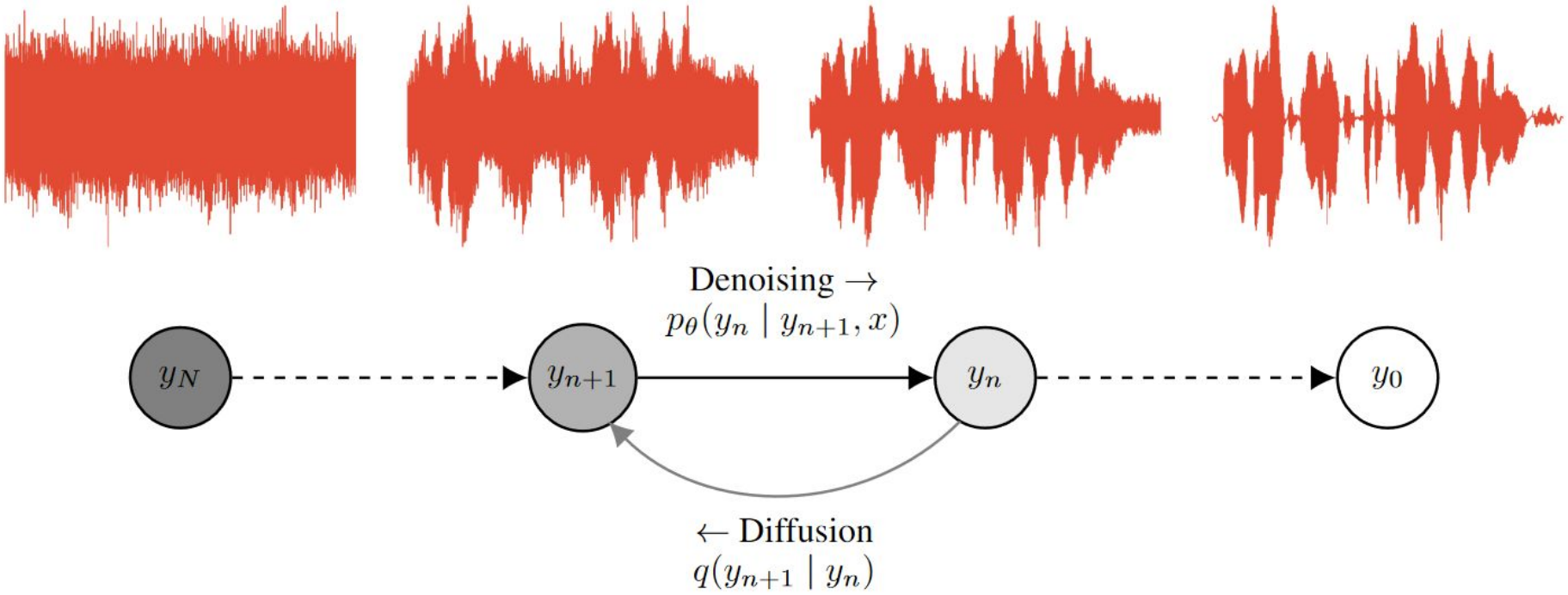
2020 Samsung Research. All rights reserved

Diffusion Models



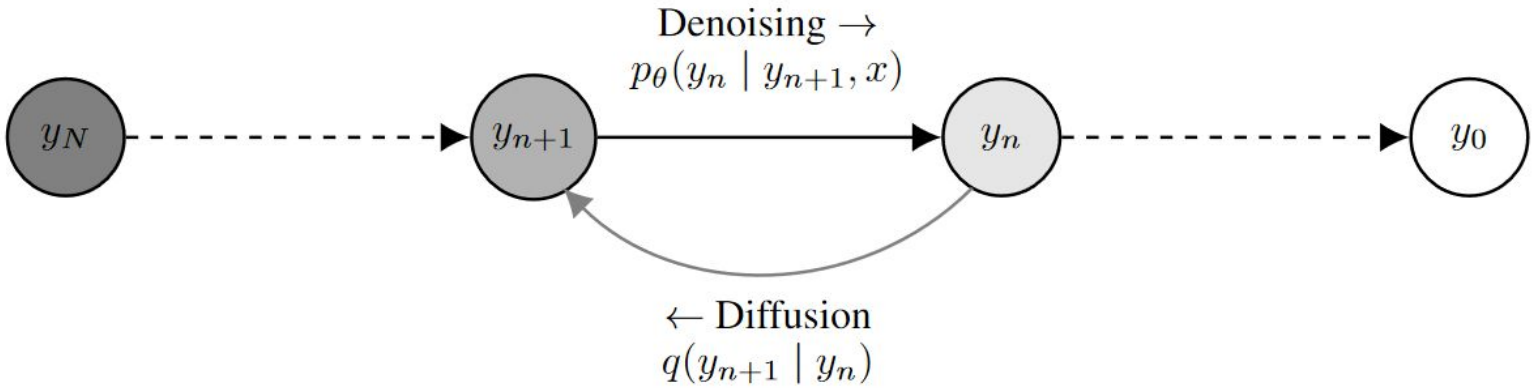
2020 Samsung Research. All rights reserved

Diffusion Models



2020 Samsung Research. All rights reserved

Diffusion Models



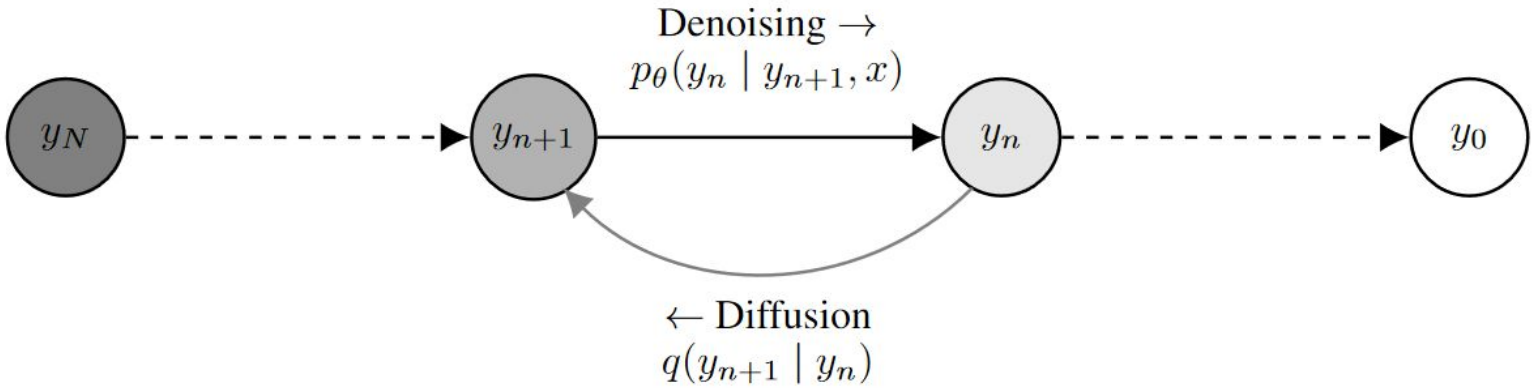
$$q(y_{n+1}|y_n) = \mathcal{N}(y_{n+1}|\sqrt{1-\beta_n}y_n, \beta_n I)$$

$$p(y_n|y_{n+1}) = \mathcal{N}(y_n|\mu(y_{n+1}, n), \sigma(y_{n+1}, n)^2 I)$$

$$p(y_N) = \mathcal{N}(y_N|0, I)$$

2020 Samsung Research. All rights reserved

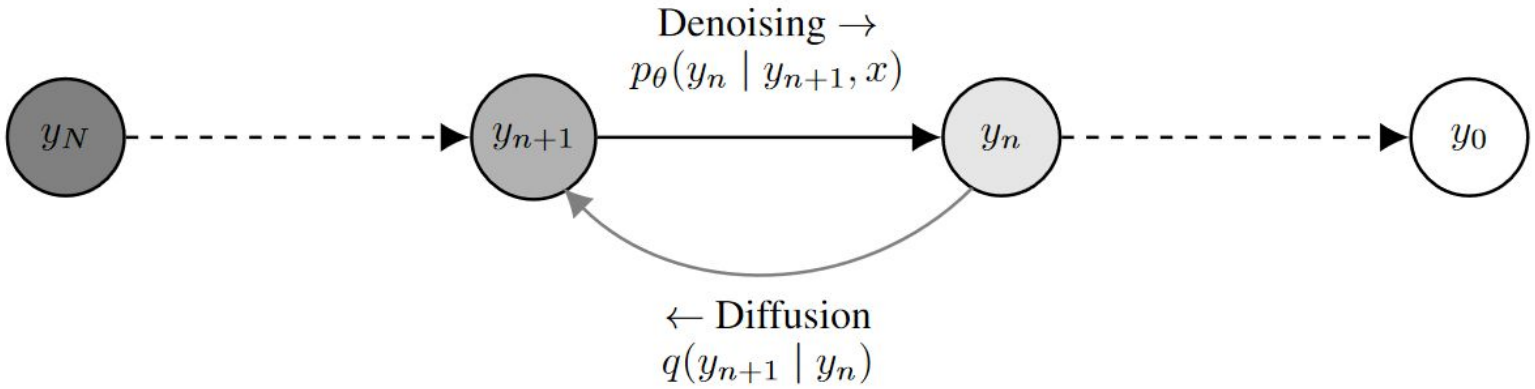
Diffusion Models



$$p_\theta(y_0) \geq \mathbb{E}_{y_1,\ldots,y_N \sim q(y_1,\ldots,y_N|y_0)} \log \frac{p_\theta(y_0, y_1, \ldots, y_N)}{q(y_1, \ldots, y_N|y_0)}$$

2020 Samsung Research. All rights reserved

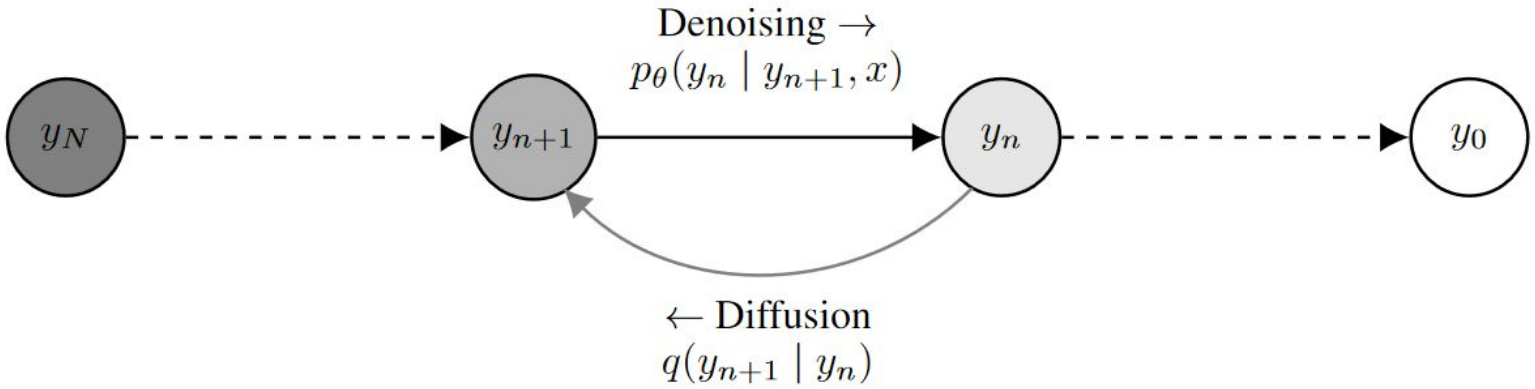
Diffusion Models



$$\mathbb{E}_{y_1,\dots,y_N \sim q(y_1,\dots,y_N|y_0)} \log \frac{p(y_N)p_\theta(y_{N-1}|y_N)\cdots p_\theta(y_0|y_1)}{q(y_1|y_0)\cdots q(y_N|y_{N-1})}$$

2020 Samsung Research. All rights reserved

Diffusion Models



$$\mathbb{E}_{y_1,\dots,y_N \sim q(y_1,\dots,y_N|y_0)} \left(\sum_{n=0}^{N-1} \log \frac{p_\theta(y_n|y_{n+1})}{q(y_{n+1}|y_n)} + p(y_N) \right)$$

2020 Samsung Research. All rights reserved

References (ICLR 2021)

- WaveGrad (Google Brain)
- DiffWave (Baidu Research, NVIDIA)
- For images: Score-based Generative Modeling Through Stochastic Differential Equations (Stanford University, Google Brain)

2020 Samsung Research. All rights reserved

Thank You

2020 Samsung Research. All rights reserved