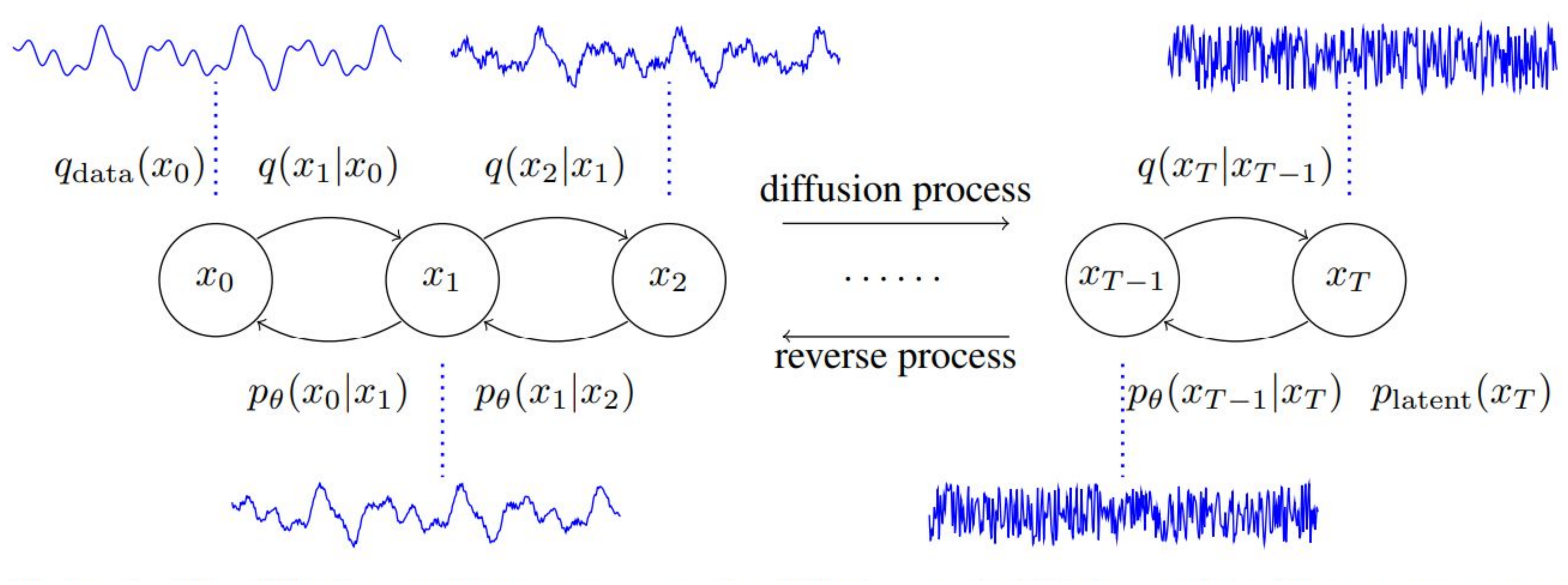
Diffusion Models for Speech Synthesis (with WS Audiology)

Thor Bjørn Olmedo Gabe, Giovanni Gomes Guerreiro, Agata Makarewicz, Mathias Høxbro Juel Vendt, Jacek Wiśniewski

Model

Diffusion Probabilistic Model



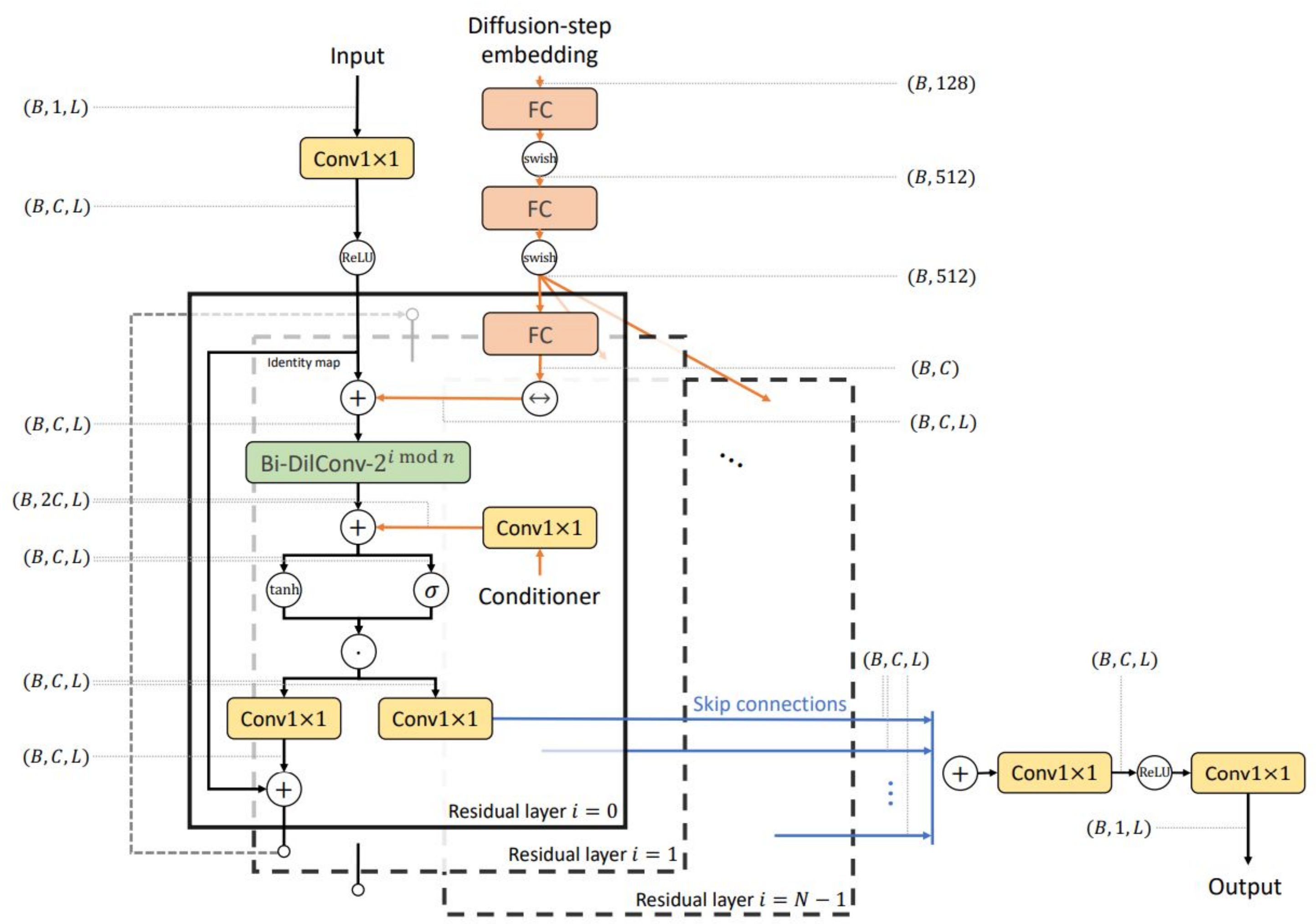
Diffusion process

$$q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

How noise is sampled?

$$\mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

DiffWave architecture



References:

[1] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," 2020.
[2] Calvin Luo, "Understanding diffusion models: A unified perspective," 2022.
[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," 2020
[4] diffwave, https://github.com/lmnt-com/diffwave

ELBO

ELBO-VDM = ELBO-MHVAE with 3 restrictions:[2]

1. The latent dimension is equal to the data dimension.
2. The structure of the latent encoder is a Gaussian distribution centered around the output of the previous timestep.
3. The Gaussian parameters change over time, so that the final latent is a standard Gaussian.

$$\underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\text{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

ELBO-DiffWave = Parameterized ELBO-VDM

Fixed schedule $\{\beta_t\}_{t=1}^{T} \rightarrow$ fixed variance: $\sigma_\theta(x_t, t) = \tilde{\beta}_t^{\frac{1}{2}}$, and let $\epsilon \sim \mathcal{N}(0, I)$

Gives a sum of KL divergences between tractable Gaussian distributions, which have a closed-form expression:

$$-\,\text{ELBO} = c + \sum_{t=1}^{T} \kappa_t \mathbb{E}_{x_0,\epsilon} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ t)\|_2^2$$

Minimizing the unweighted ELBO leads to higher generation quality[3], and thus the training objective of DiffWave becomes[1]:

$$\min_\theta L_{\text{unweighted}}(\theta) = \mathbb{E}_{x_0,\epsilon,t} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ t)\|_2^2$$

where t is uniformly taken from 1,...,T

Methodology

Implementation

Public implementation of DiffWave algorithm by Sharvil Nanavati (*diffwave* Python package)[4]

Datasets

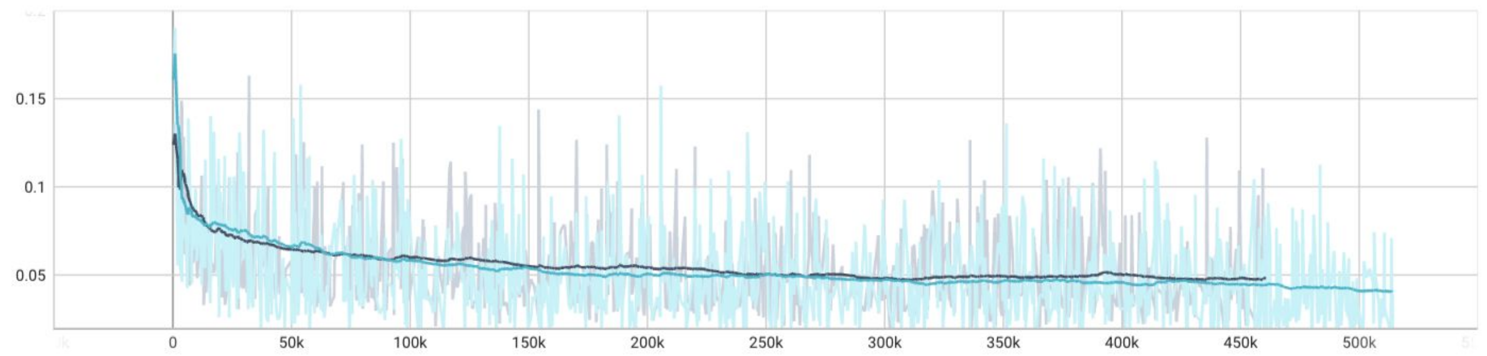
1. English - LJ Speech dataset (public domain set of 13,100 short audio clips; single speaker reading passages from books)
2. Portuguese - Common Voice dataset (public domain set of over 100 hours of audio; multiple speakers - volunteer collaborators around the world)

Models

1. Pre-trained model provided by the authors of the implementation, trained on LJ Speech dataset (22kHz sampling rate)
2. Model trained from scratch on the subset of LJ Speech dataset (~22% of the original dataset; 22kHz sampling rate)
3. Model trained from scratch on 4.8k Portuguese audio clips (32kHz sampling rate)

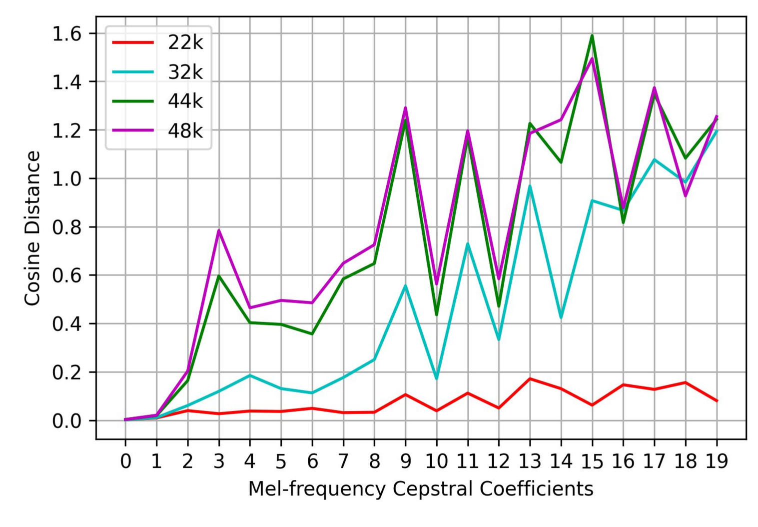
Learning curves of the models trained from scratch

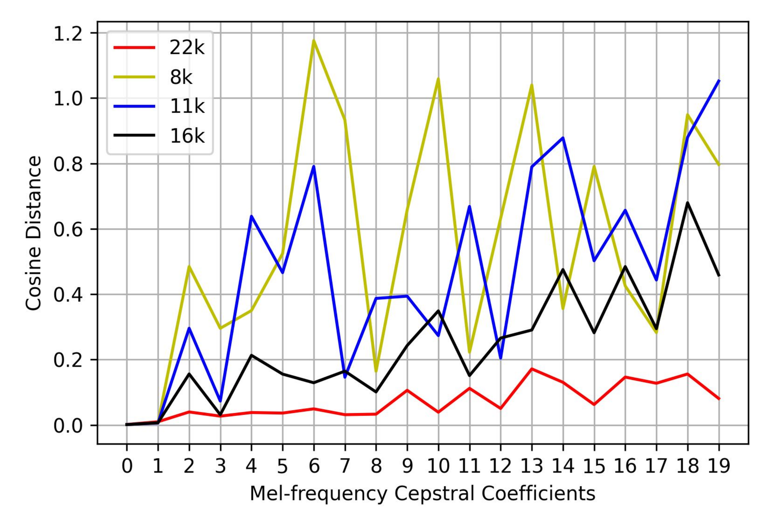
Number of steps vs loss



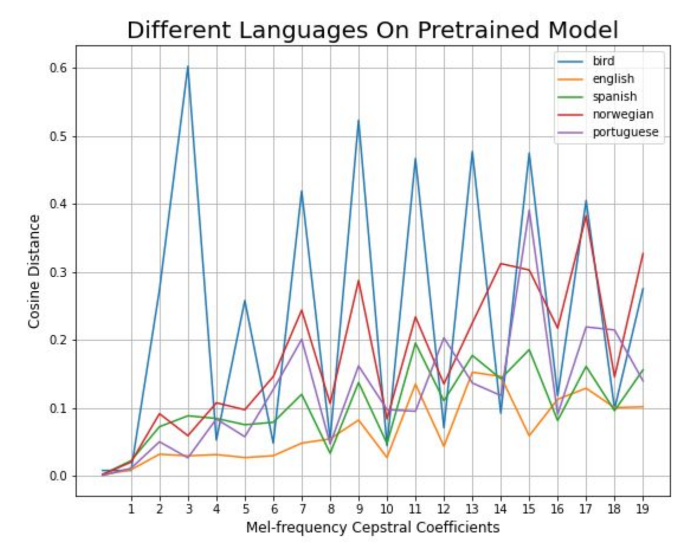
Results

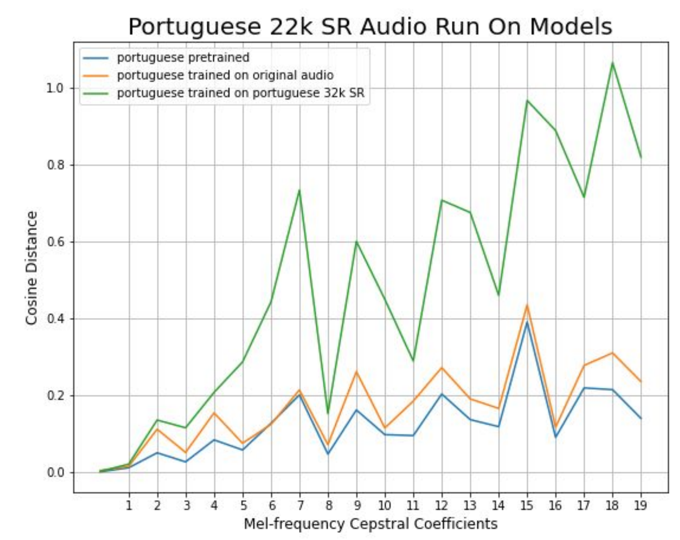
Sample-rate robustness

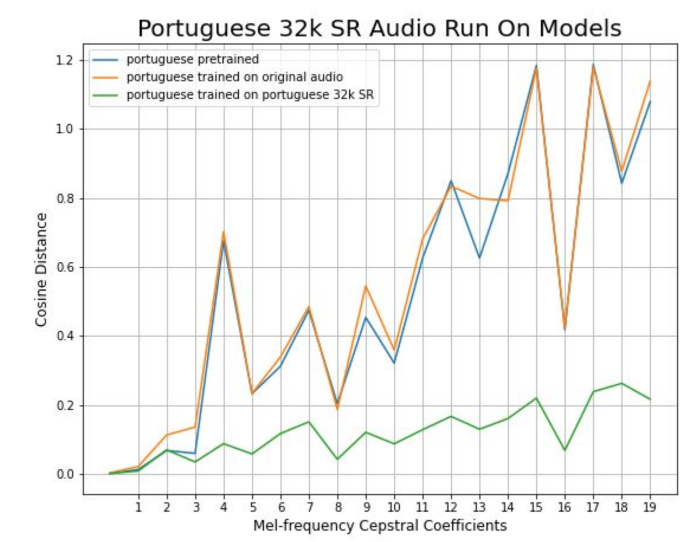




Different languages







Distorted audio

Amplitude	Sigma	SNR pure	SNR noise	SNR output
50	5	41.29	19.92	1.04
50	10	41.29	14.16	1.34
50	15	41.29	11.09	1.35
50	20	41.29	9.11	1.37
100	5	47.29	25.98	2.01
100	10	47.29	20.02	0.86
100	15	47.29	0.1	2.05
100	20	47.29	0.17	1.11
200	5	53.13	31.92	1.21
200	10	53.13	25.92	1.15
200	15	53.13	0.02	1.57
200	20	53.13	19.95	1.58