# Enhancing Acoustic Model for Children with Generative Adversarial Network

Zhaodong Wang, ricwang@Stanford.edu Video: https://youtu.be/w-1pygVS6uE

## Introduction and Problem

In automatic speech recognition, acoustic model is used to decode audio signals to phonemes and linguistic units with learning from transcriptions.

Acoustic model for children has been a challenge task in industry

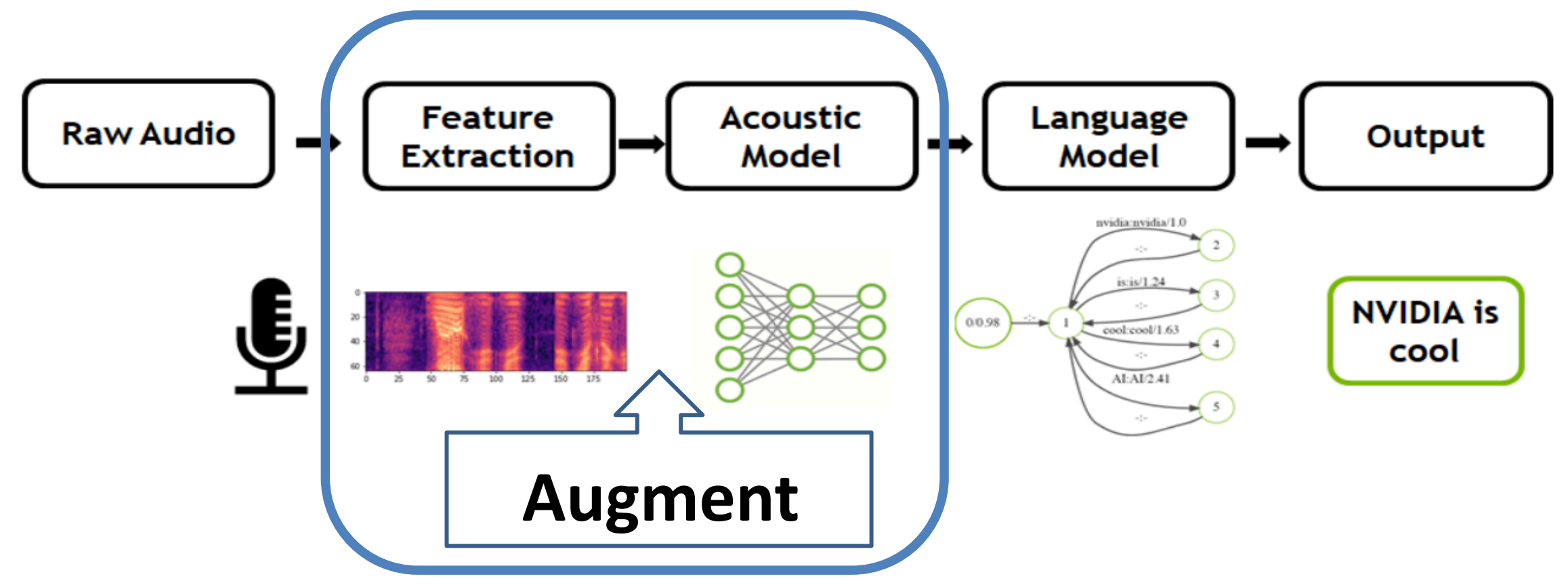
a. Data is very limited due to the high cost
b. Noisy conditions due to children's behavior

In this project we propose to **use generative adversarial network to denoise the audios together with light gated recurrent units to boost children's acoustic model**.

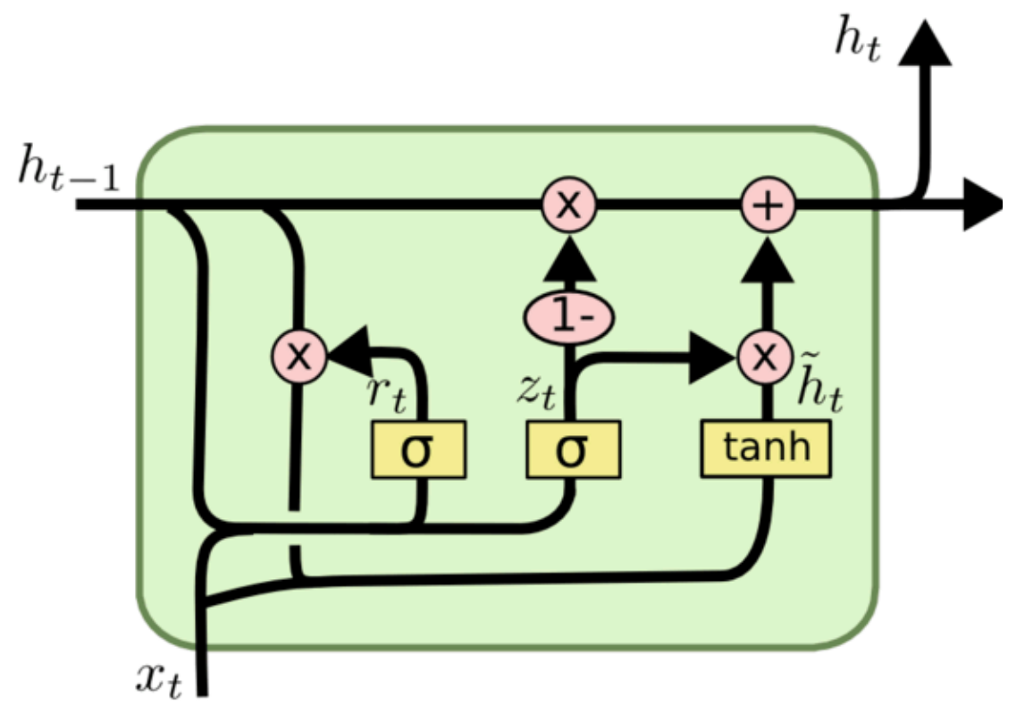
Evaluation of acoustic model: Word Error Rate

## Background

Data augmentation has been proved to be effective in acoustic model training such as speed perturbation and noising/denoising.

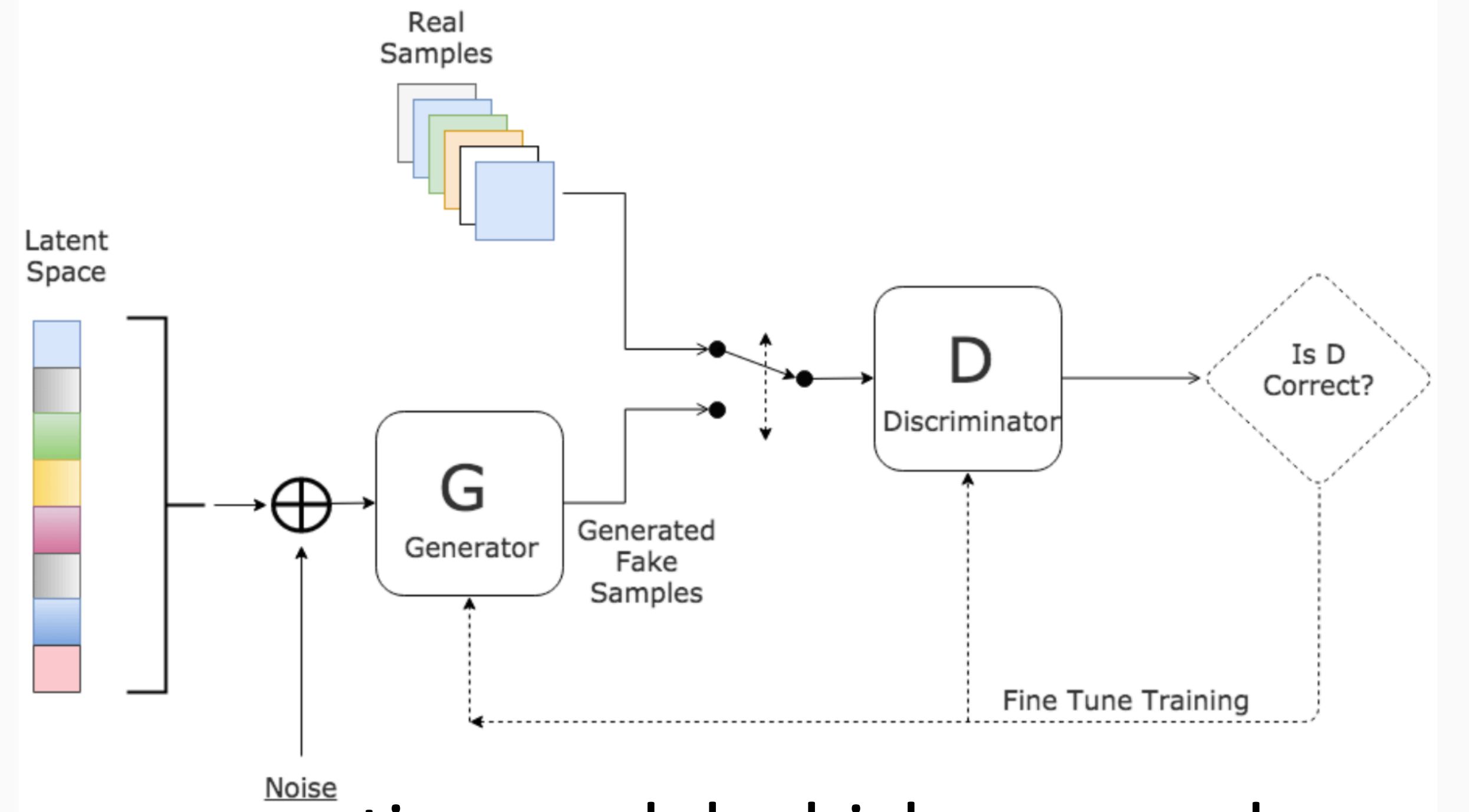


Recurrent neural network (RNN) with gated recurrent unit has (GRU) is powerful for acoustic model. Recent proposed Light-GRU achieved better learning efficacy.



$$z_t = \sigma(BN(W_z x_t) + U_z h_{t-1}),$$
$$\widetilde{h}_t = \mathrm{ReLU}(BN(W_h x_t) + U_h h_{t-1}),$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \widetilde{h}_t.$$

## Generative Adversarial Network



Generator: a generative model which can produce a sample

$$\min_G V(G,D) = E_{\mathbf{x}\sim p_{\mathrm{data}}}[\log D(\mathbf{x})] + E_{\mathbf{x}\sim p_G}[\log(1 - D(\mathbf{x}))]$$

Discriminator: a classification model trying to classify real data or data from Generator

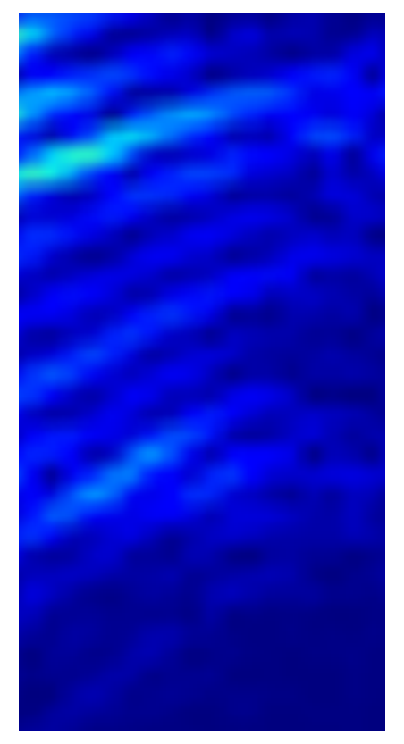
$$\max_D V(G,D) = E_{\mathbf{x}\sim p_{\mathrm{data}}}[\log D(\mathbf{x})] + E_{\mathbf{x}\sim p_G}[\log(1 - D(\mathbf{x}))]$$

Alternating optimization:

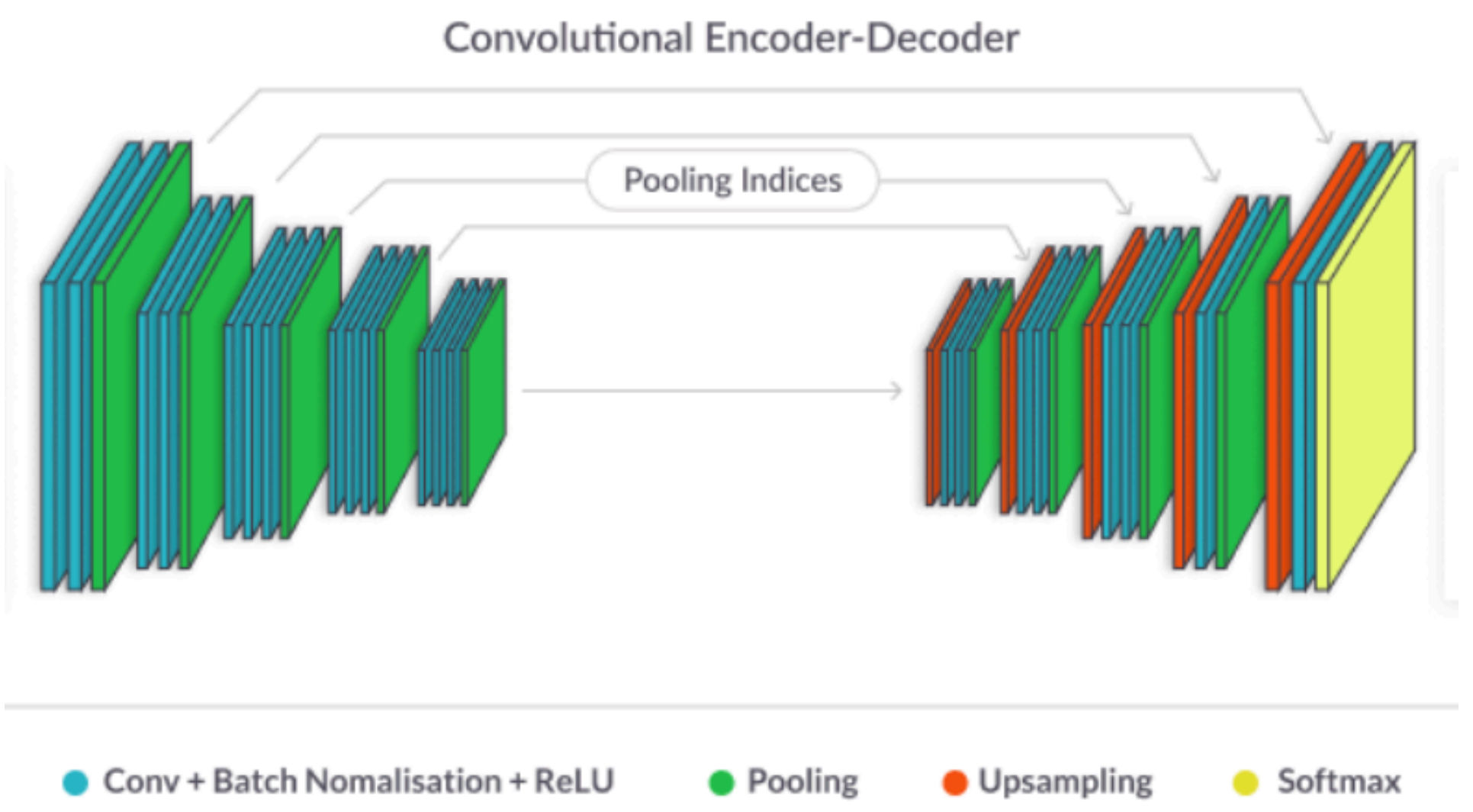
$$\min_\theta \max_\phi V(G_\theta, D_\phi) = E_{\mathbf{x}\sim p_{\mathrm{data}}}[\log D_\phi(\mathbf{x})] + E_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D_\phi(G_\theta(\mathbf{z})))]$$

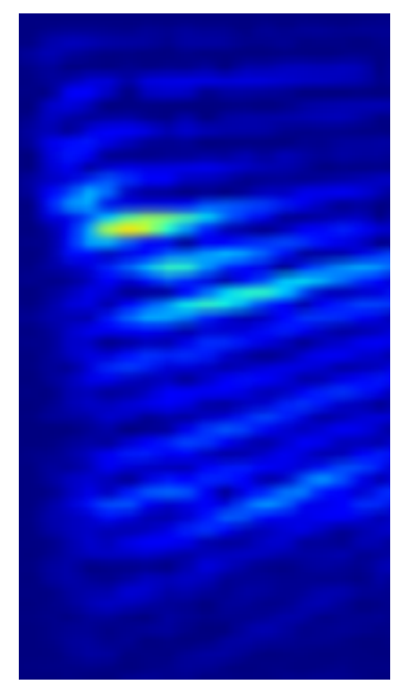
## Implementation of Denoising GAN

- Target: to train a generator to output clean signal when inputting a noisy signal, with the help of discriminator
- Discriminator: 1 convolutional layer , 1 fully connected layer, to classify the noisy sample as 1, the clean sample as 0
- Generator: a encoder-decoder with 5 convolutional layers, trying to encode-decode the noisy audio into clean one and get approval from discriminator



noisy


Convolutional Encoder-Decoder
Pooling Indices
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax



'clean'

•Evaluation of GAN output

Subjective evaluation from expert: obvious difference on the background noise, including line noise, microphone explosive noise, environmental echo, etc. No significant on multi-speech cases.

## Train with Light GRU network

Data set:

- 200-hour audios from in English, PCM, 8 Kbps, 8 bit, 5-30 seconds per utterance
- Processed 20 hours clean audio from the denoising GAN

Model structure:

- Features: Feature space Maximum Likelihood Linear Regression (FMLLR) and Cepstral mean and variance normalization (CMVN) used
- 5 Li-GRU layers, 550 nodes
- Dropout rate: 0.2, with batch normalization
- Optimization: RMSProp

## Results and Discussions

- Evaluation of Acoustic Model with Li-GRU

| | Total Words (ground truth) | Error of Insertion | Error of Deletion | Error of Substitution | Word Error Rate |
|---|---|---|---|---|---|
| 5-layer Li-GRU | 194917 | **11471** | **17406** | **41293** | **36.00%** |
| 5-layer Li-GRU with Denoising GAN | 194917 | **13587** | **12642** | **37395** | **33.64%** |

- Denoising GAN improves significantly on deletion and substitution: phoneme are clearer than in noisy audios
- Denoising GAN misrecognizes noises into short phonemes and words, 'the', 'hi', increases the error of insertions

Future Works:

- Can we use GAN to add noise into audios to boost acoustic model?

**Reference**

[1] Light Gated Recurrent Units for Speech Recognition, Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, Yoshua Bengio, https://arxiv.org/pdf/1803.10225.pdf

[2] SEGAN: Speech Enhancement Generative Adversarial Network, Santiago Pascual, Antonio Bonafonte1 , Joan Serra, https://arxiv.org/pdf/1703.09452.pdf