Improving Speech recognition with Neural enhancement of features

Shivam Kumar
Mentor- Abhinav Garg

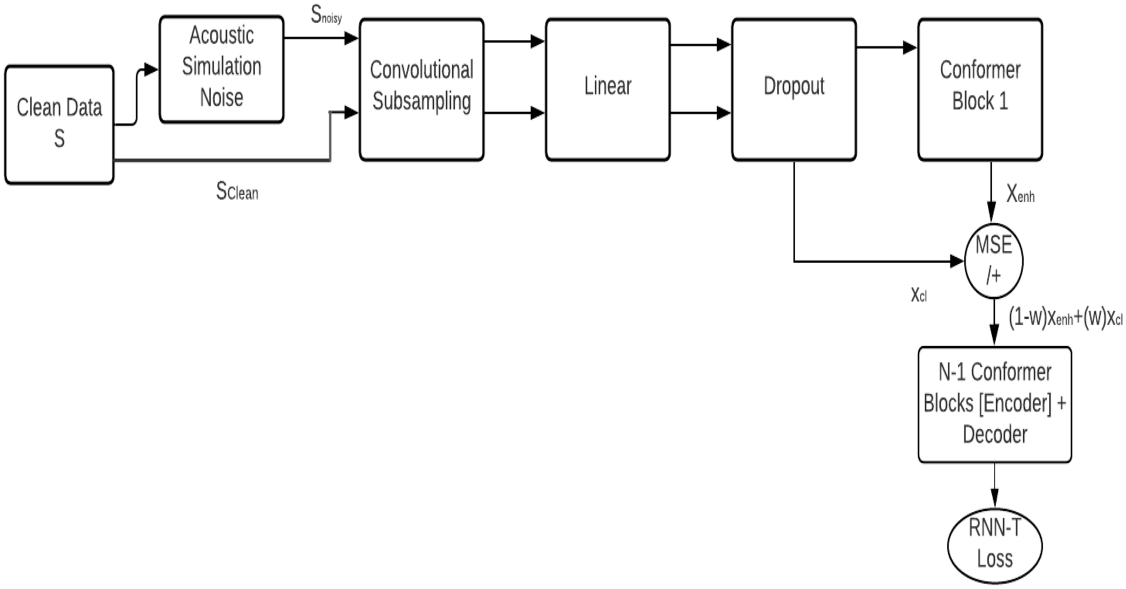
July 20, 2021

1 Motivation

Earlier all the models were trained on clean data sets that is without any noise, recorded from some close microphones without any disturbances in lab settings so when tested on real world datasets they did not perform well and if we tried to improve by training on noisy data sets the performance on clean data sets deteriorated so the idea is to train the model jointly with a linear combination of clean and noisy datasets so that we do preserve the performance on clean data and also improve our performance on noisy or far field data.

2 Model

2.1 Flow



2.2 Implementation

Output of the network fed to encoder/decoder is given by:

$$(1-w)X_{enh} + wX_{cl} \tag{1}$$

- Gradual application of enhanced features (GAEF):We always want to train any model first on easier examples easiest one will be clean signal. We slowly start to increase the weightage of the

Xenh and correspondingly decreasing the weight of Xcl. In 8 full epochs w goes from 1 to zero linearly as

$$w = \begin{cases} 1 - \frac{Epoch-1}{7} & \text{if } Epoch \leq 8 \\ 0 & \text{else} \end{cases}$$

The loss function is given by RNNT loss+ λ*MSE Loss

- Gradual Reduction of Enhancement Loss(GREL):Mean squared Error loss which is computed between Xenh Xcl also termed as enhancement loss. Now again as the name Gradual reduction of enhancement loss suggests the associated factor lambda goes from 1 to zero linearly in 8 epochs

$$\lambda = \begin{cases} 1 - \frac{Epoch-1}{7} & \text{if } Epoch \leq 8 \\ 0 & \text{else} \end{cases}$$

2.3 Results

- Trained the model with a batch size of 4 for 10 epochs
- Transducer loss reduced from 4098 to 2094 in 10 epochs which was constantly going down which was a good sign
- Now this setup may be utilised in practical settings which generally has inputs from a far field source for a better experience

3 Refrences

- Chanwoo et al., "Streaming End-to-end Speech Recognition With Jointly Trained Neural Feature Enhancement", 2021
- Anmol et al., "Conformer: Convolution-augmented Transformer for Speech Recognition", 2020