Speech Understanding (CSL7770)
Assignment 3
Shubh Goyal (B21CS073)

Towards Sub-millisecond Latency Real-Time Speech Enhancement Models on Hearables

1 Summary

The paper is based on resource-constrained low-latency (sub-millisecond level) speech enhancement deployable on hearing devices. The authors have proposed an analysis and synthesis pipeline capable of providing algorithmic latency as low as 0.34 ms, a latency not achieved ever before. The proposed pipeline uses LSTM network to compute FIR taps and process samples one by one. They have evaluated the pipeline using low-power DSPs and have observed good results on both objective (same train and test dataset as well as generalization) and subjective evaluations.

2 Architecture

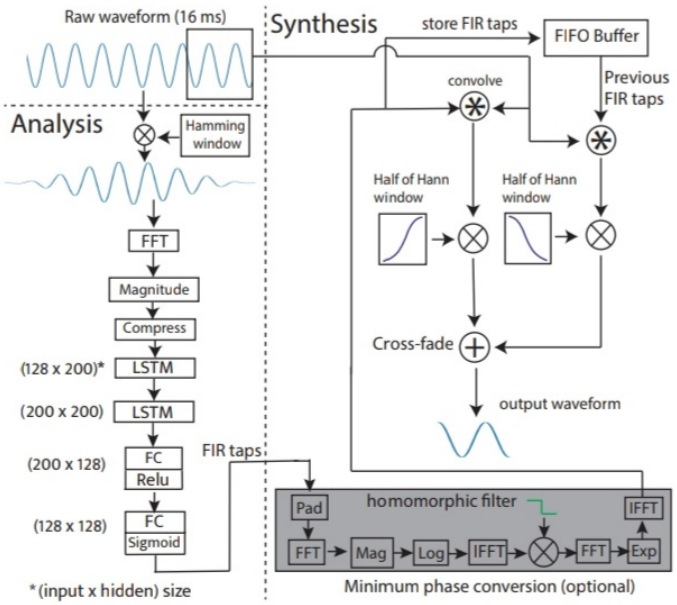

Fig. 1. Architecture and Pipeline - Fig 2 in paper - *Inference time causal Deep FIR signal processing diagram, divided into synthesis and analysis. A new FIR filter is estimated every hop.*

The architecture used in the pipeline is very simple. It contains two LSTM layers with 200 units each followed by two fully connected layers which finally predict a 128 dimensional FIR tap.

The inputs to the LSTM are features obtained from the FFT of a window of the audio sample.

The output from the LSTM is processed through a minimum phase conversion pipeline and then used to convolve the audio window followed by a hann window and a 50% cross fade.

3 Technical Strengths

Some technical strengths observed in the proposed pipeline are as follows:

- **Sub-millisecond latency:** The algorithmic latency reports ***(0.34 ms)*** is the lowest ever reported mono speech enhancement latency as claimed.

- **Compact LSTM model:** The model used for FIR tap prediction has just ***626k parameters***, and a size of ***0.58 MB*** after quantization. This can be thus used on low power low memory devices.

- **Possible real-time deployment opportunity:** The authors have tested their pipeline on ***i.MX RT600***, a low-power audio DSP. Decreased end-to-end latency of *3.35 ms.*

- **Novelty - FIR estimation:** The authors devised a new approach of *dynamically calculating FIR taps* using a LSTM Network.

4 Technical Weaknesses

- **Model recomputation:** For very short hops, recomputing the model is wasteful and is like an extra overhead.

- **Noise leakage:** Low-latency filtering as stated in the paper causes noise leakage during speech.

- **Limited generalization:** The pipeline is good in mono-channel speech enhancement. But results over generalized enhancement is still a bit low as compared to LSTW.

5 Minor Questions

- What is the performance change when under extremely noisy or reverberant conditions?
- What results were obtained using quantized models, not specifically mentioned?
- Was quantization aware training tried and how would it affect performance?

6 Review

6.1 Suggestions

The authors can consider the following to deal with limitations:

– Use of adaptive hop size or reuse of FIR can be considered to deal with the unnecessary recomputation for non-noisy audio segments.
– Authors can focus on generalization of the pipeline using some other training data or techniques. It will help the pipeline be feasible for real world use.

6.2 Rating and Justification

Rating - 3/5. The paper shows a great improvement in resource-constrained low latency speech enhancement. The pipeline is quite innovative and the thoughts involved behind are considerable. But it would have been better if the authors explored the generalized real world scenarios as well and also reported results on performance of quantized models.

The code can be found here: GitHub Repository