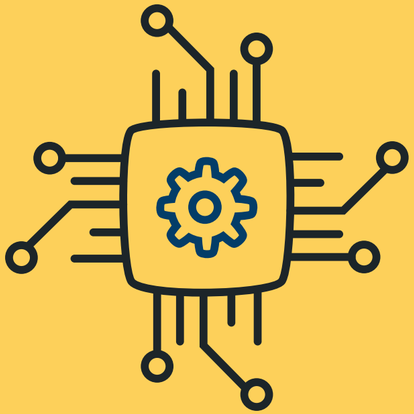
MAXIM INTEGRATED

SPEECH ENHANCEMENT

Stella Cao, Renjie Liu, Huaxuan Wang, Curtis Zhuang

Clean speech is better



Project Roadmap

Background → Problem Statement → Knowledge Expansion → Candidate Model → Data and Methodology → Findings and Discussion → Next Steps

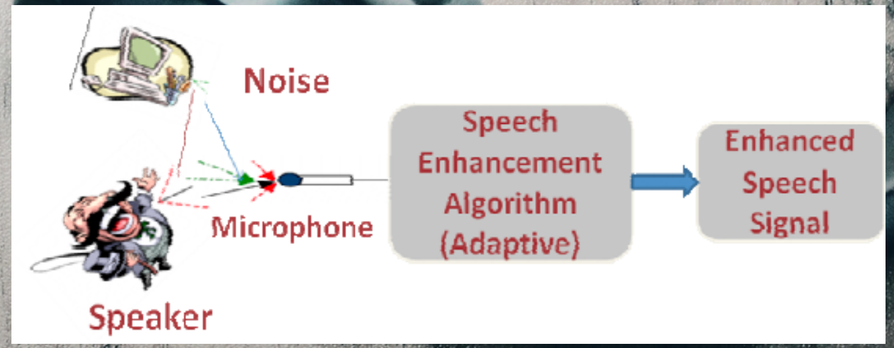
Working With An Analog Device Company

MAXIM INTEGRATED

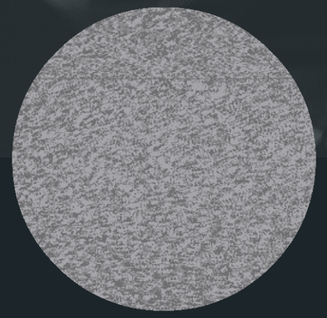
- Technology manufacture
- MAX78000 accelerator

SPEECH ENHANCEMENT

- Improve mixed speech
- Zoom and Microsoft Teams



Speech Enhancement Is A Harder Problem



Noise Cancellation Headphones

- Reduce ambient sounds using active noise control

Hearing Aid Requirements

- Enhance speech
- Aesthetic concern
- Power-efficient

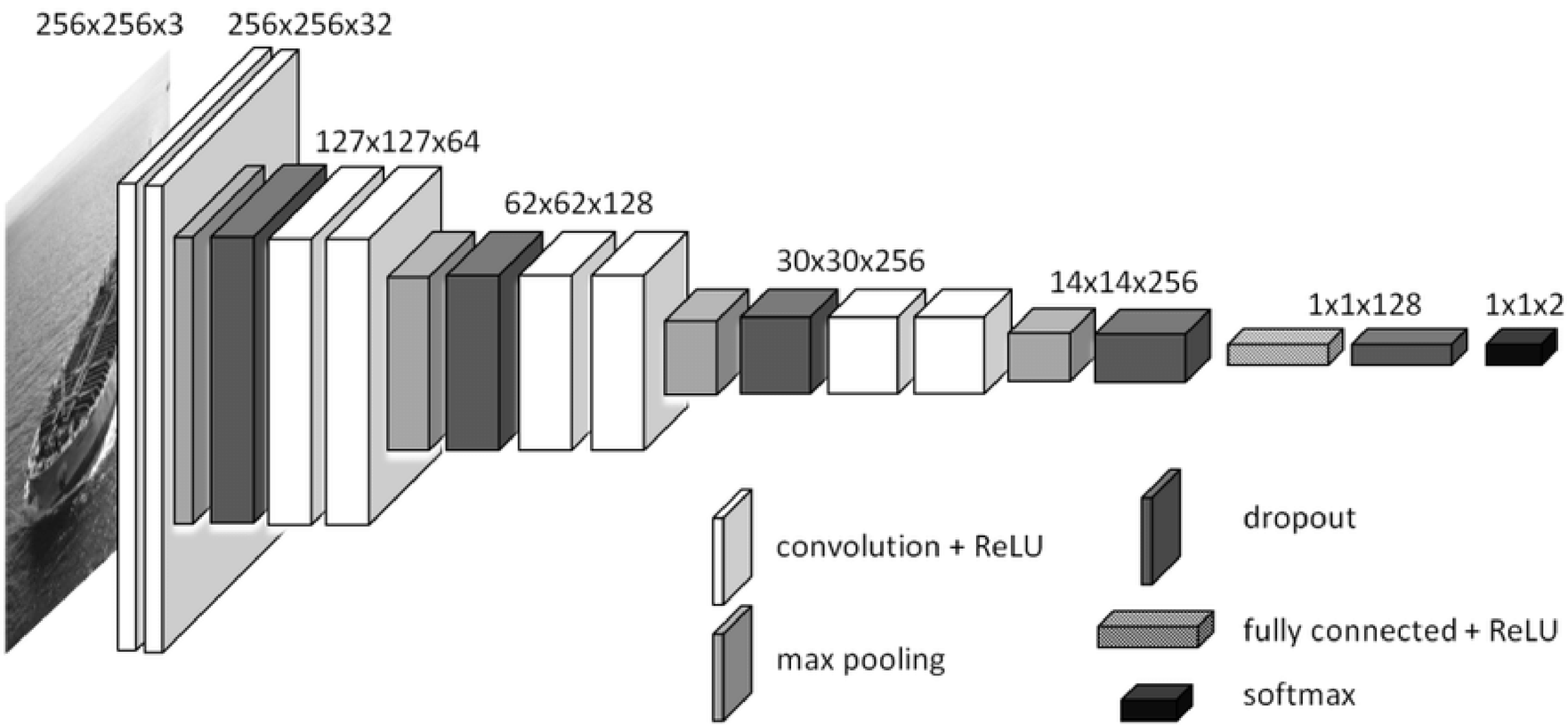
Huge Revenue Potential For Maxim

We developeded a speech enhancement algorithm that enables Maxim Integrated to participate in a **$300 Million** total addressable market.

About 28.8 million U.S. adults could benefit from using hearing aids

Max78000 priced at around $10

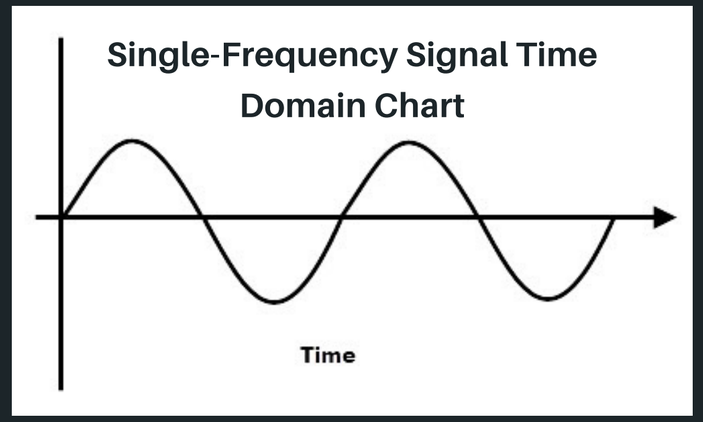
$300 Million

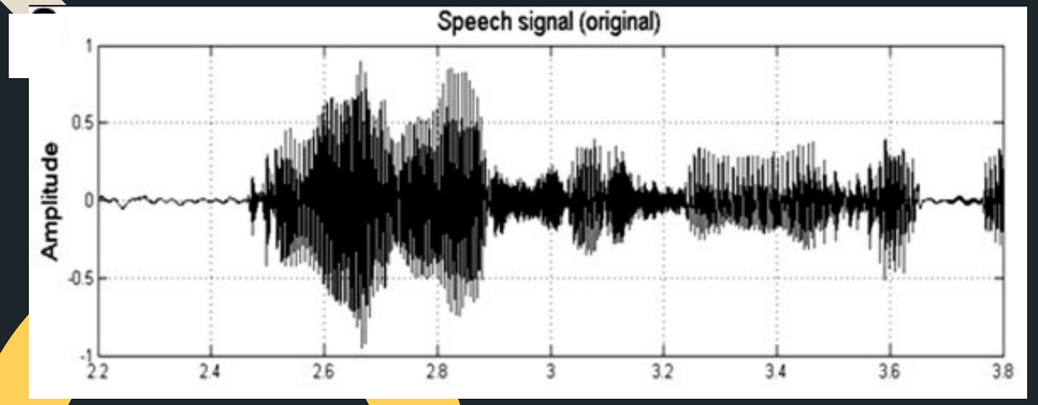


PROBLEM STATEMENT

Develope and implement machine learning methods, compatible with Maxim Integrated's MAX78000, to enhance speech for hearing aid devices with latency less than 10ms*.

Signal Is Basic of Speech





Signal

Signal is a variation of air pressure changed over time

Recording Signals

To digitally record signals, air samples are taken over a period of time. The sampling rate commonly is 44.1k samples per second, aka 44.1k Hz

Single-Frequency Signal

In real-life, the signal consists of different single-frequency signals

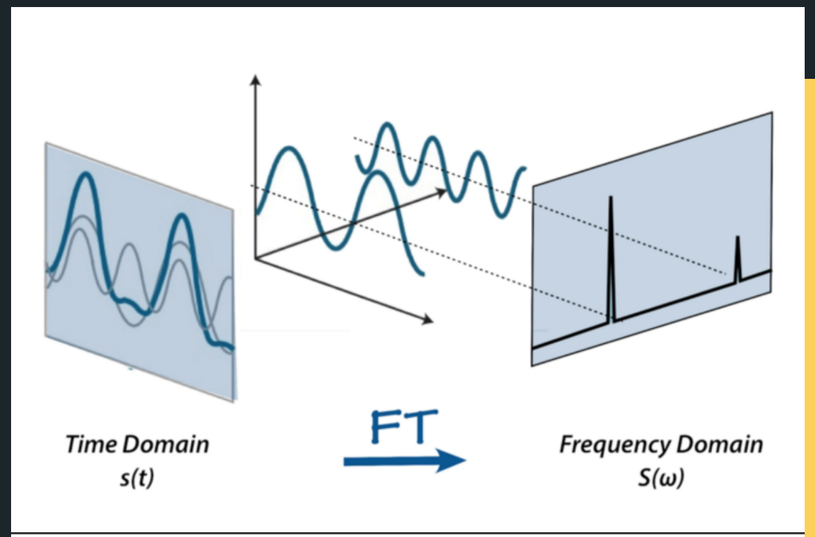
What Is Fourier Transform

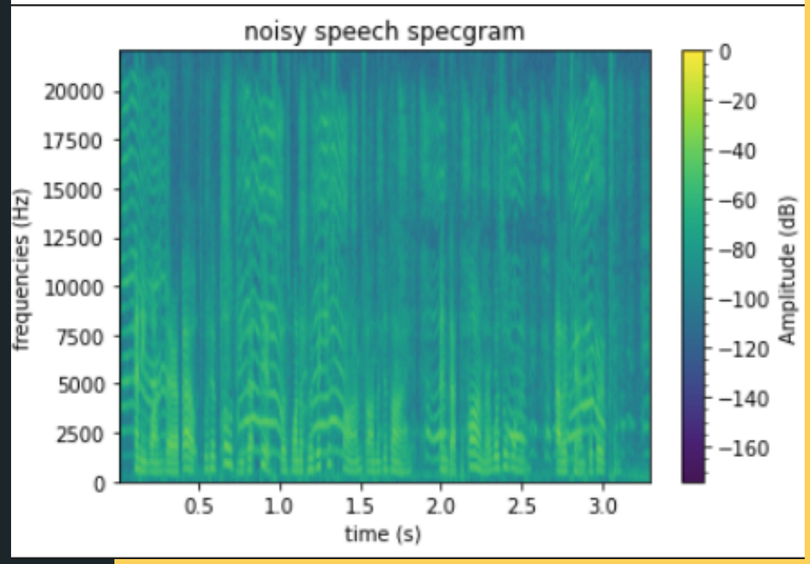
Fourier Transform

Apply the Fourier transform (FT) to decompose a signal into individual frequencies and its amplitude. But our speech's frequency content varies over time

Short-Term Fourier Transform

Short-term Fourier transform (STFT) performs Fast FT on windowed segments of the signal. The result can be visualized by Spectrogram





Current Strategies for Speech Enhancement

Mapping-based (End to End)

(Asri et.al 2021)

- Map a non-linear function from the noisy speech into the enhanced clean speech
- Less sensitive to SNR variations
- Reconstruct the target output from the observed input

Mask-based

(Asri et.al 2021)

- Estimate a mask to filter to the input spectrogram
- Different types of mask
  - Ideal binary mask (IBM)
  - Ideal ratio mask (IRM)

Convolution Models Suit better than GRU and GAN for its efficiency

R-CED - CNN Model

(Park & Lee 2016)

- CNN models have lesser number of parameters and can be used in embedded devices
- CNN models has encoder and decoder structure which is suitable for generation tasks
- Can perform either mapping-based or mask-based tasks

U-Net

(Stroller, Ewert & Dixon 2018)

- U-Net strucutre is a symmetric convolutional network investigating end-to-end source separation
- The architecture's encoder captures context and the symmetric expanding path enables precise localization with its skip-connection which retains information both from the bottlenet and the convolutioned counterpart

We mixed 6 noise types with clean speech

Speech

OpenSLR LibriSpeech ASR corpus data which contains corpus of read English speech divided into train (100 hours) and validation (5 hours).

Noise (from Youtube)

Stationary:

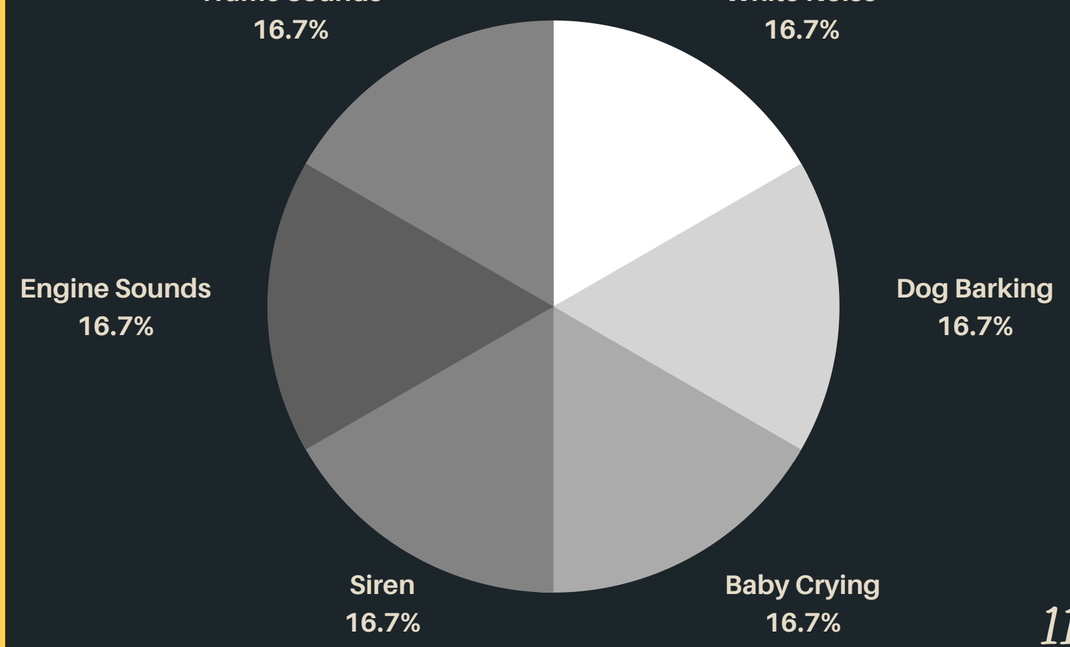
White

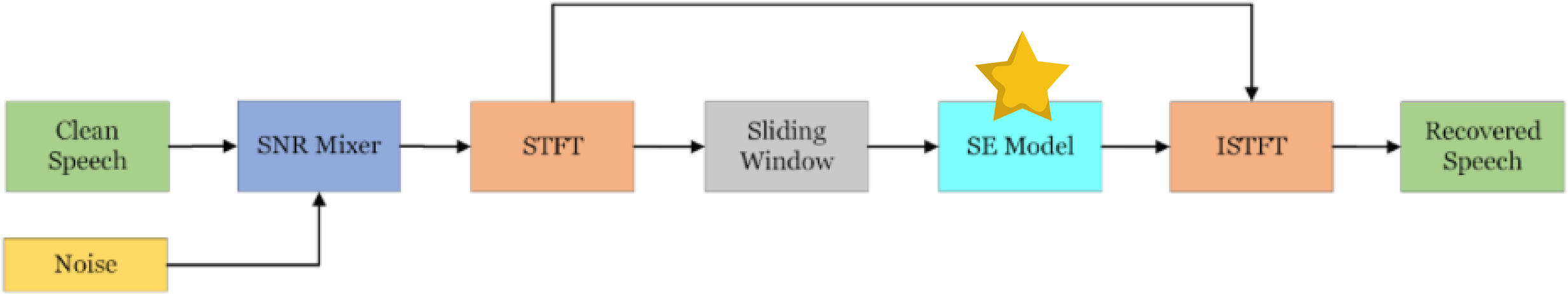
Quasi-Stationary:

Siren

Non-Stationary:

Traffic, Baby crying, Dog barking, Diesel truck

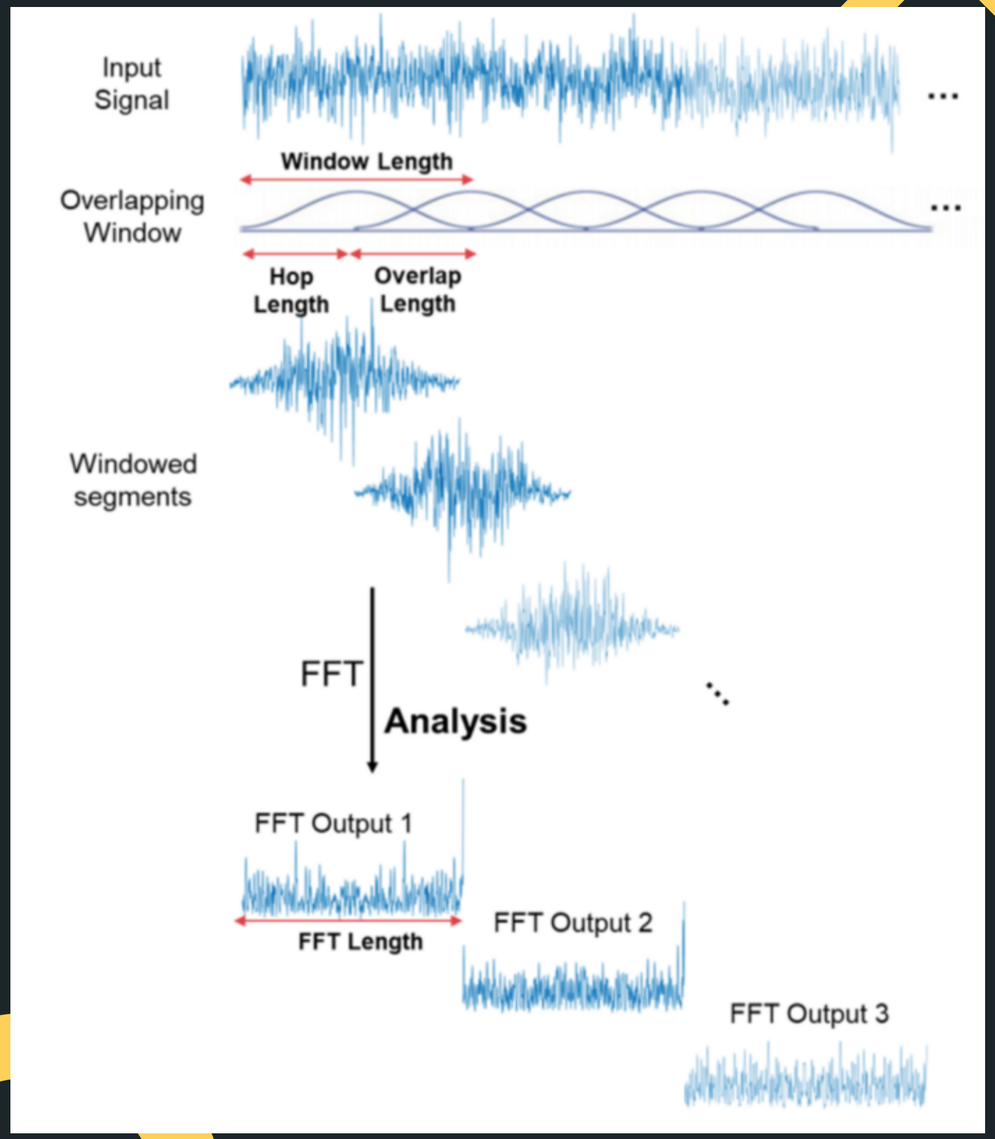




End-to-End Speech Enhancement

PyTorch Data Pipeline Modules

1. Read clean speech and noise
2. Generate noisy speech
3. Apply STFT
4. Perform Sliding Window
5. Fit Speech Enhancement Model
6. Recover speech using ISTFT



Feature Engineering
- Fourier Transform

Parameter of STFT:

- Window Length: 512 (32 ms)
- Number of FFT: 512
- Hop Length: 128 (8 ms)

Transform speech to magnitude (real-valued) and phase (complex-valued)

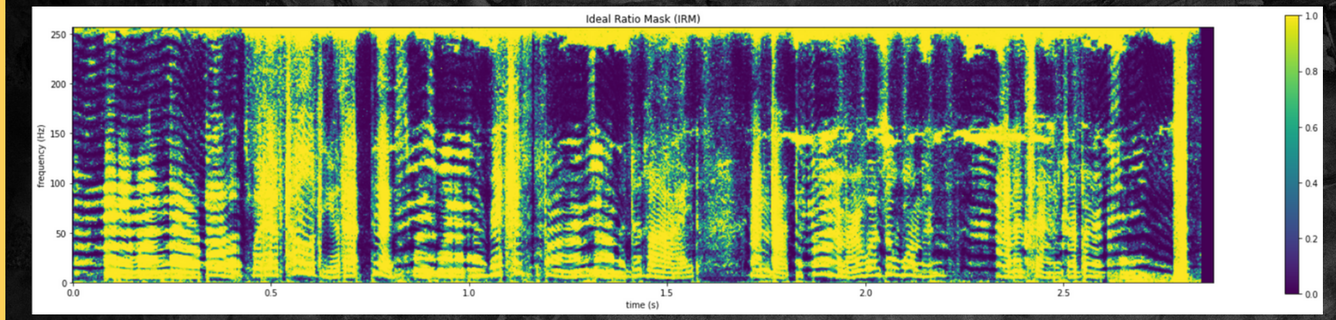
Feature Engineering
- Sliding Window & Masking

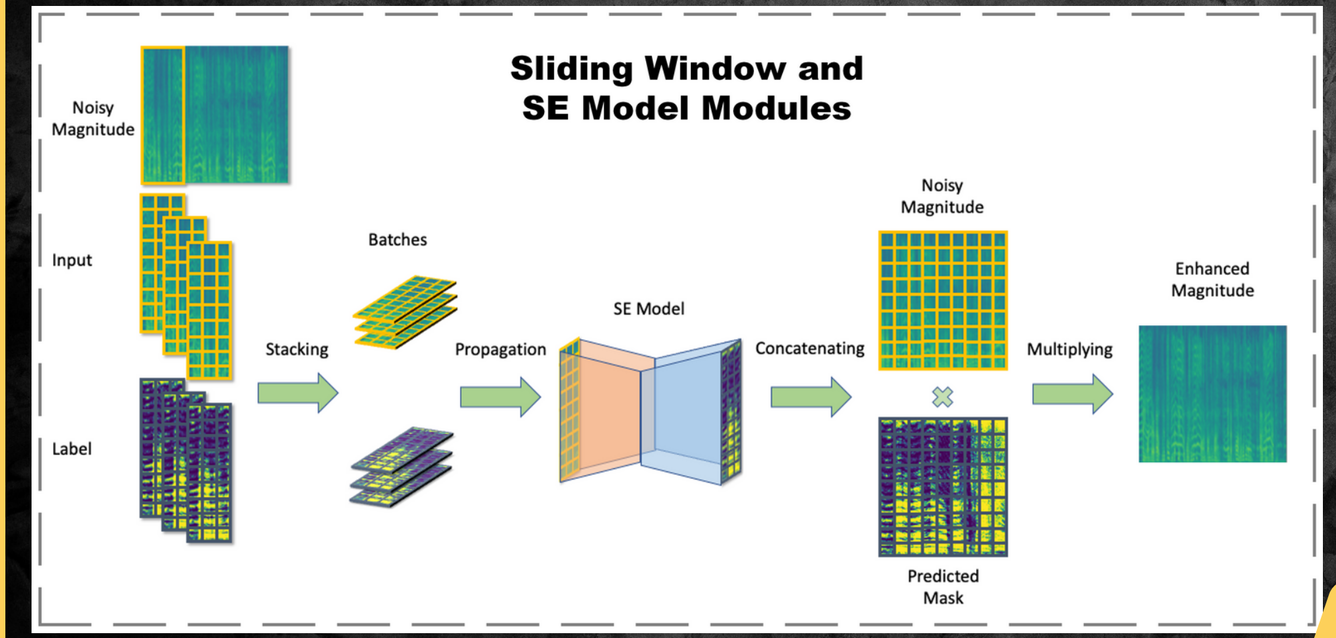
Sliding Window:

- Fix time dimension
- Create temporal context as model inputs

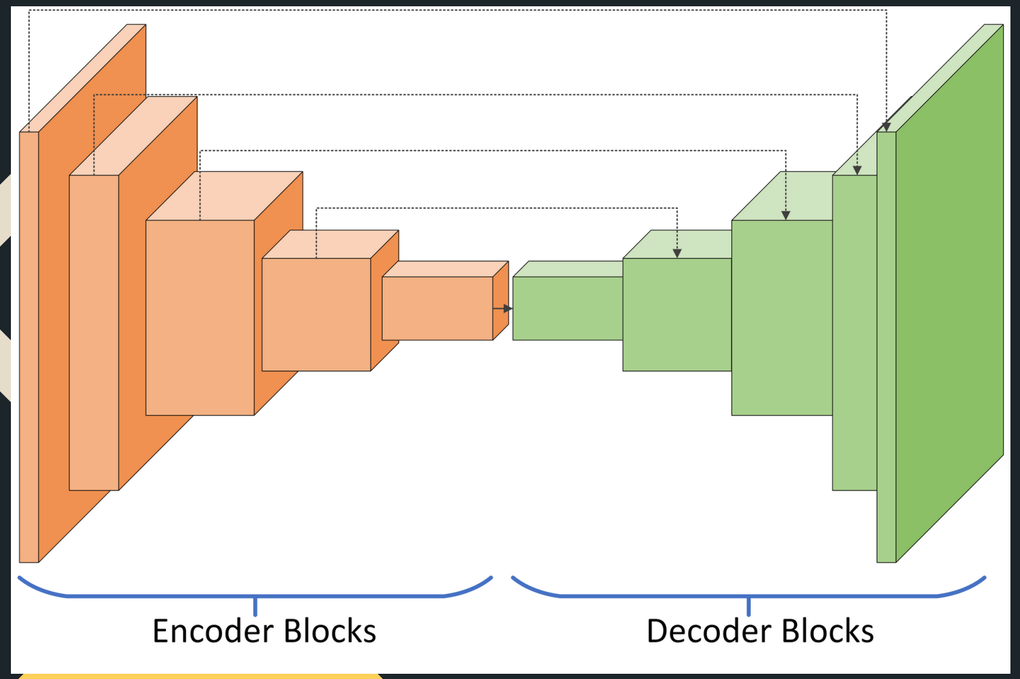
Ideal Ratio Mask (IRM):

- Label of the model
- Smoothed version of IBM





What the Basic Model Looks like



Proposed Model

- Encoder-Decoder based CNN

Objective

- Remove specific noise types
- Lightweight

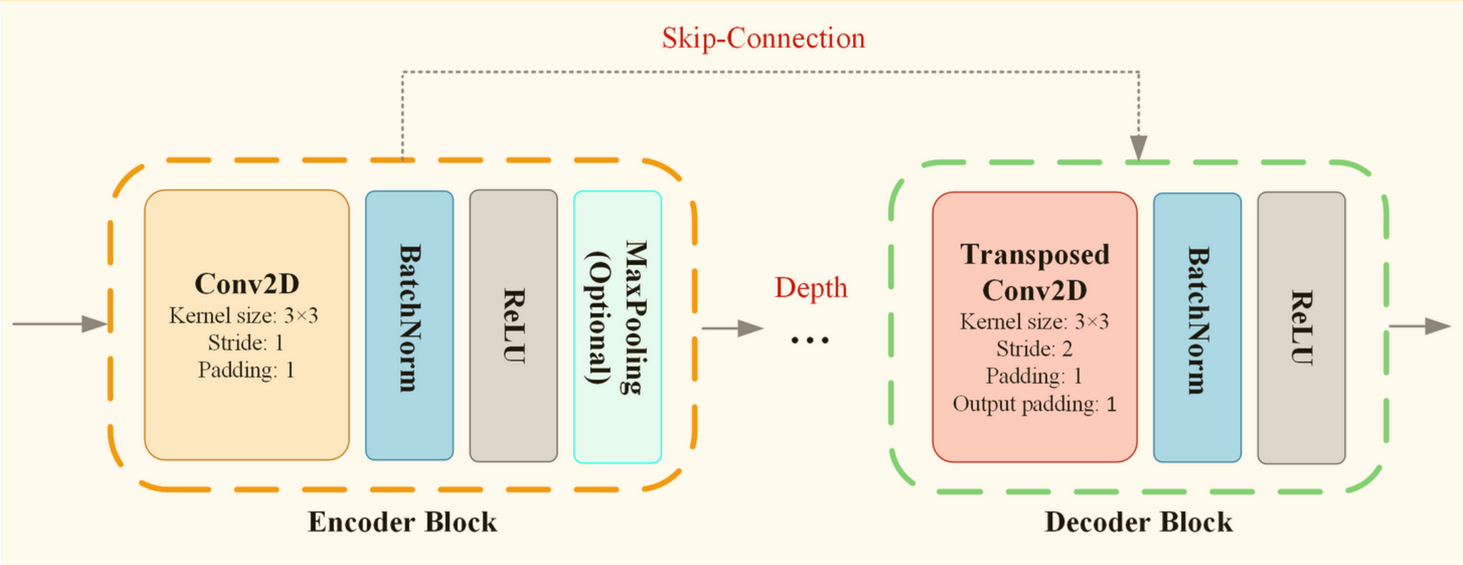
Architecture

- Encoder Block:
  - Conv2d, BatchNorm, Max-Pooling, and ReLU
- Decoder Block:
  - TransposedConv2d, BatchNorm, and ReLU

PyTorch

Final Model Details

Efficient UNet-128 parameters based on MAX78000 hardware



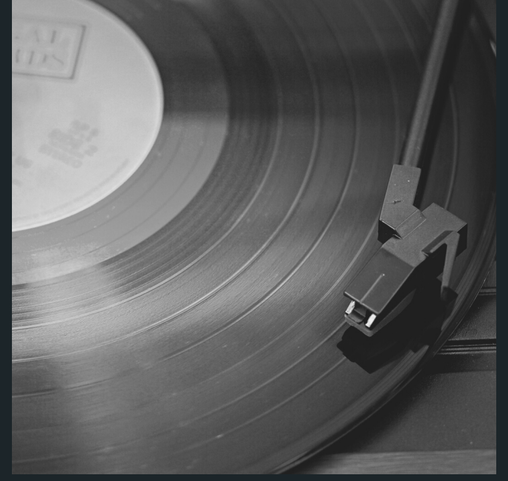
Hardware limitation

- Kernel Size: 3x3
- Conv2d:
  - Stride: [1, 1]
  - Dilation: 1

Efficient UNet-128 parameter

- 5 encoder blocks
- 5 decoder blocks
- Max channel size: 256
- Kernel size: 3x3

How to Evaluate Speech Properly



Noise Reduction

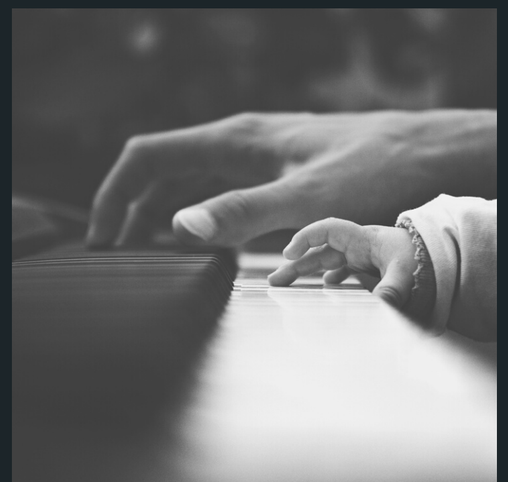
Segmental Signal-to-Noise Ratio (segSNR)

- Signal / Noise
- Higher segSNR, lower noises

Speech Quality

Perceptual Evaluation of Speech Quality (PESQ)

- From -0.5 (bad) to 4.5 (good)

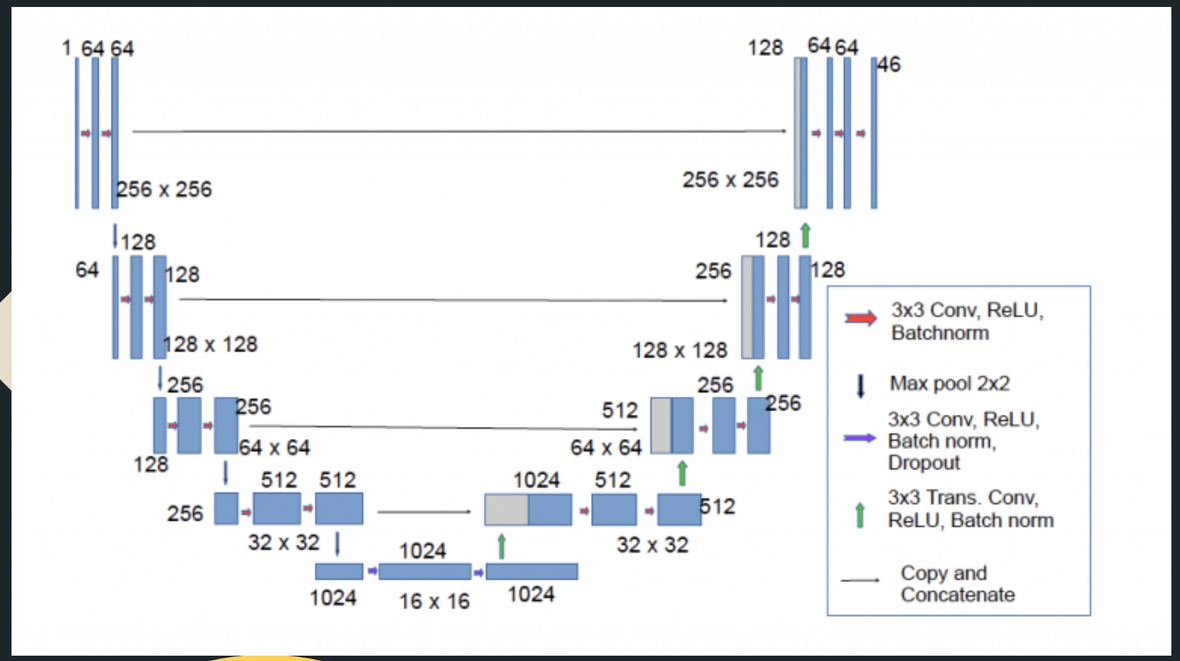


Speech Intelligibility

Short-Time Objective Intelligibility (STOI)

- From 0% (bad) to 100% (good)

U-Net Stands Out

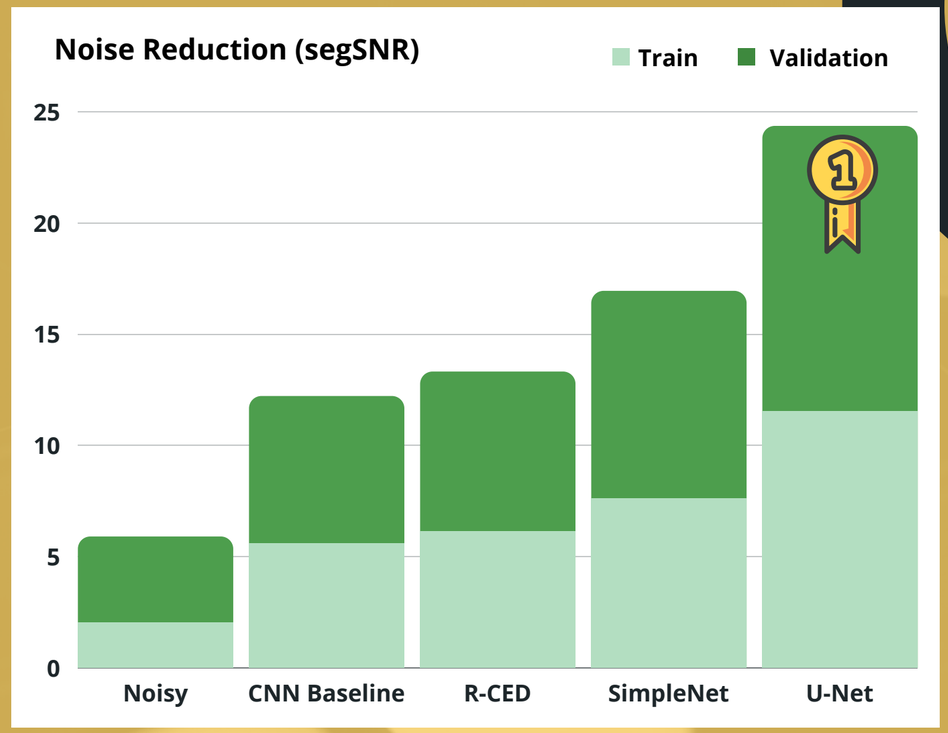


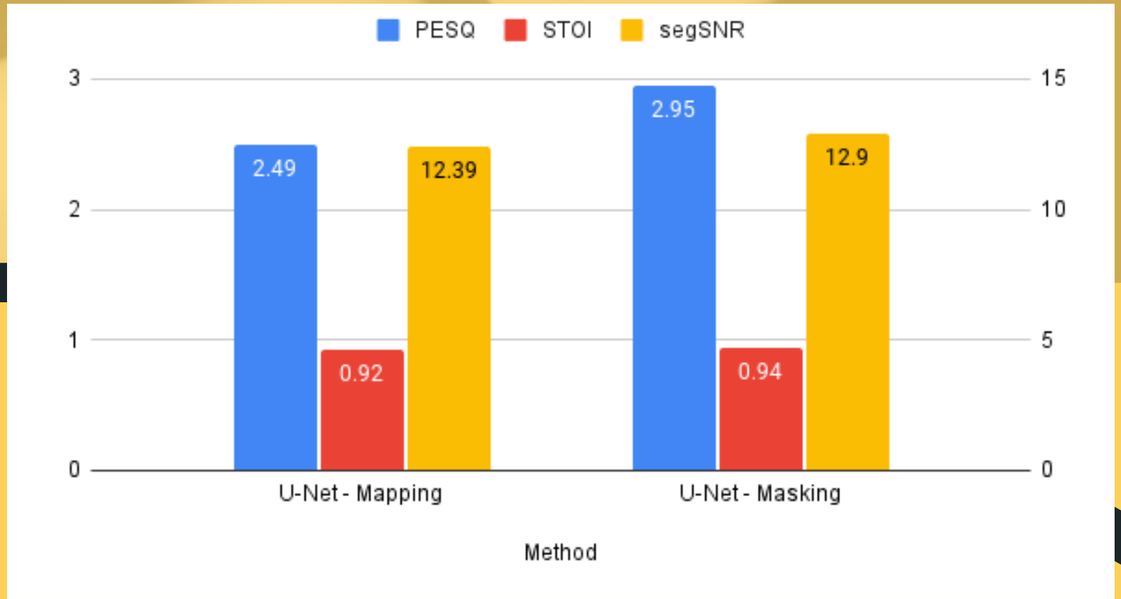
Hypothesis:

U - Net will be small enough to be embedded on Max78000 and achieve required speech enhancement performance.

- Flexible Size
- High performance architecture

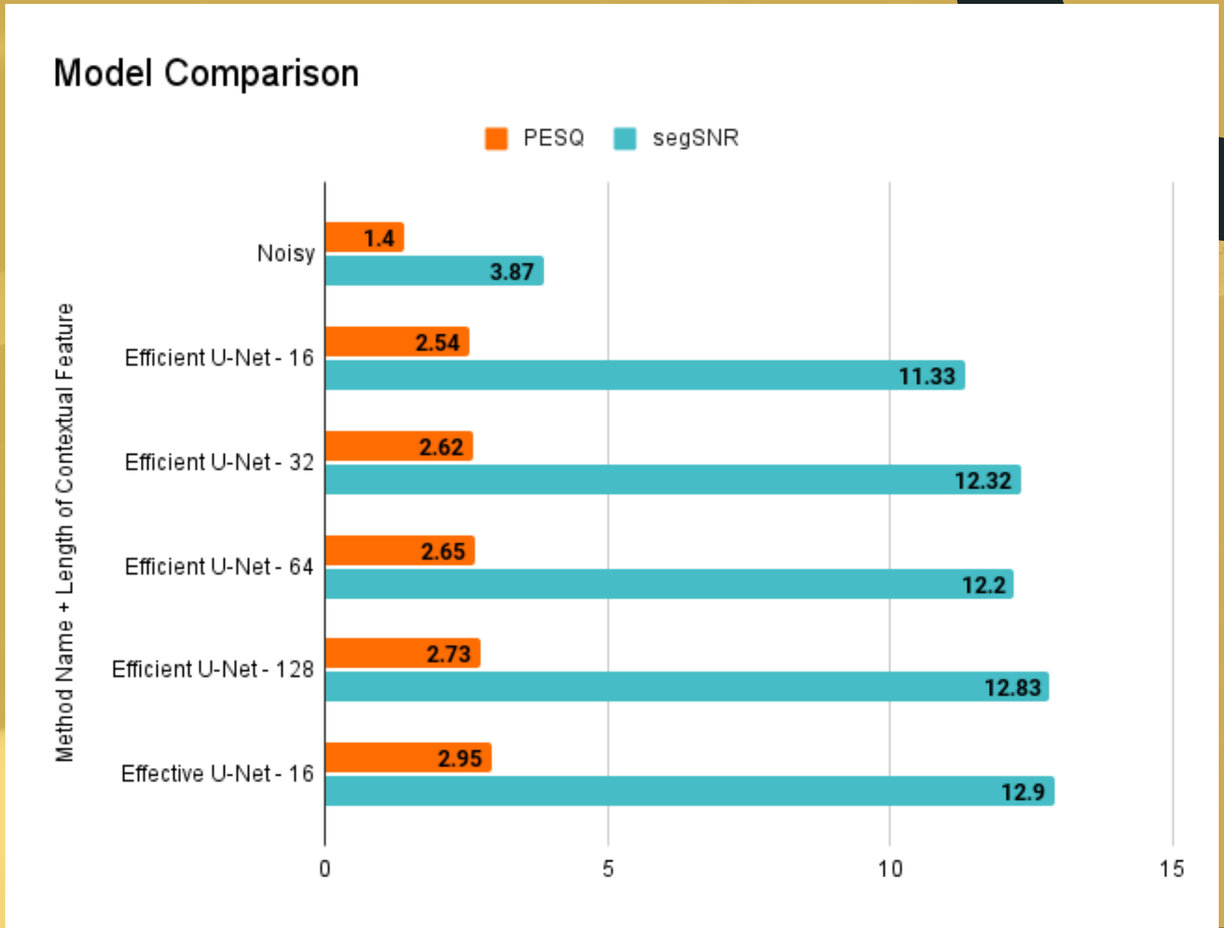
- **U-Net**
  - Significantly improves all 3 metrics
  - Benefit from flexible of encoder-decoder architecture and skip-connection
- **Mapping Method**
  - Unclear speech
- **Masking Method**
  - Better performance





U-Net With Masking Performs The Best

- **Effective UNet-16**
  - The best performance
  - Too large to deploy on MAX78000
- **Efficient UNet**
  - More contextual features, better performance
- **Efficient UNet -128**
  - Achieve similar performance with Effective UNet-16
  - Small enough to embed on MAX78000
  - Latency less than 3.3 ms



Efficient U - Net -128 Will Be Embedded

Significant Improvement In Different Noise Types

PERFORMANCE ON DIFFERENT NOISE TYPES

Small UNet-128

White Noise

SSNR increased 8.89 dB
PESQ increased 1.01
STOI increased 6%

Baby Crying

SSNR increased 10.61 dB
PESQ increased 1.80
STOI increased 14%

Dog Barking

SSNR increased 0.32 dB
PESQ increased 0.33
STOI increased 6%

Siren

SSNR increased 16.46 dB
PESQ increased 1.99
STOI increased 9%

Traffic Sounds

SSNR increased 9.41 dB
PESQ increased 0.92
STOI increased 11%

Engine Sounds

SSNR increased 10.71 dB
PESQ increased 1.25
STOI increased 8%

Model Structure & Size Have Huge Impact on Performance

Metric Evaluation

The model captures stationary noises well, while for non-stationary noises, the performance varies, especiallly dog bark.

Metrics may not capture all human perceivable components.

U-Net Structure

U-Net structure allows the upsampling process to retrieve data from not only the bottleneck but retain speech resolution from the downsampling counterpart in encode-decoder.

Mapping vs Masking

Masking method has better speech intelligibility and quality as it is less prone to error caused by latency which is preferred in real-time.

Model Size

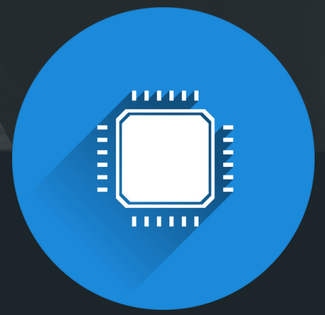
Larger kernels, with more channels capture more features and the use of pooling layers in small models affect the upsampling process when retaining information.

Large models have small temporal context.

Our Contribution to MAX78000

High Performance

- Reduce unwanted ambient sounds using active noise control
  - Average improvement by 8 times (segSNR increased by 8.96 dB)
  - Latency less than 3.3 ms



Compatibility

- Adjust model to MAX78000
- Model deployment on MAX78000 is currently in progress
- High marketing potential

Current model opens space for futher growth

Model: self learning

Hardware

Latency

More noise types

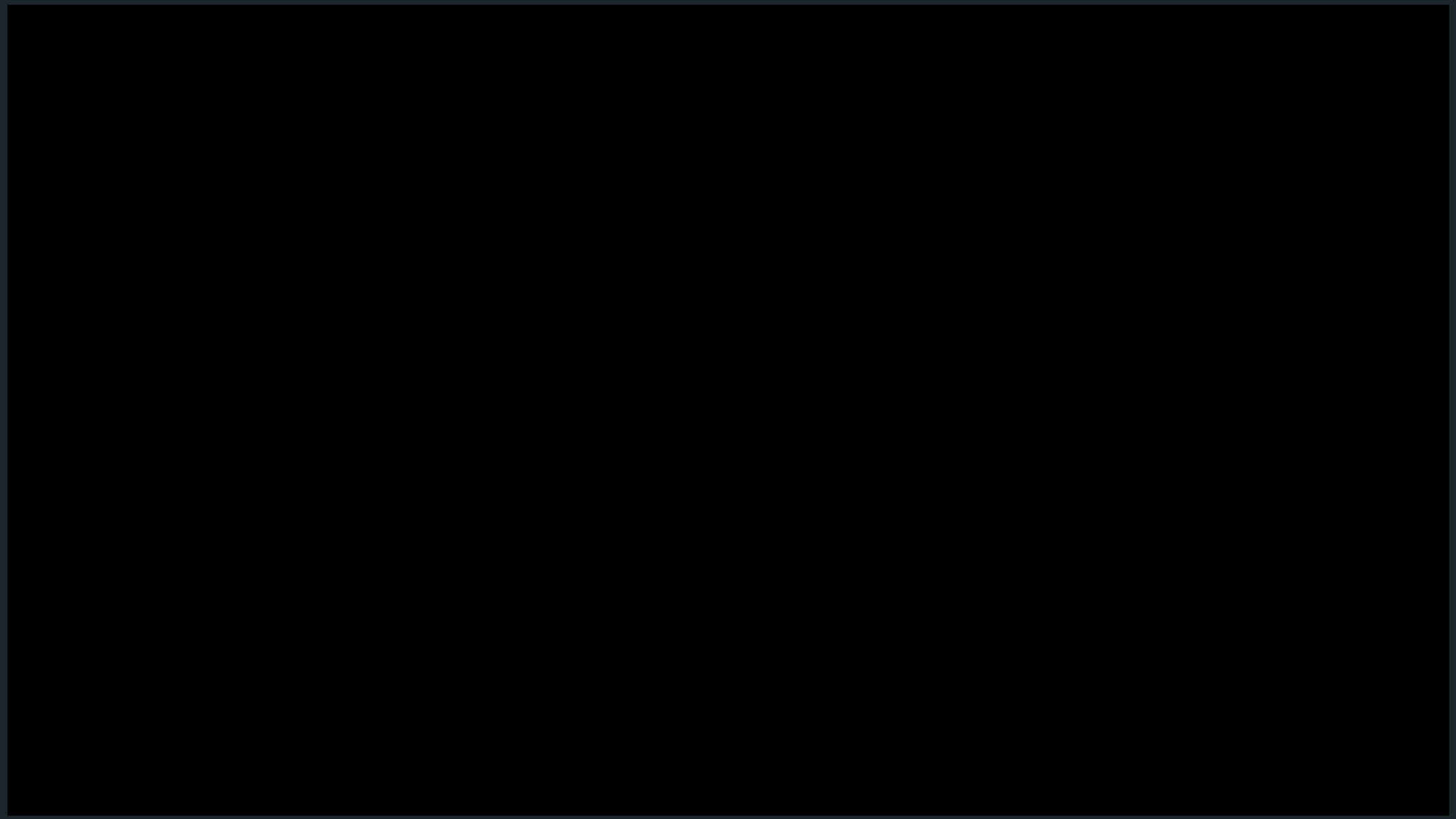
Languages

Mixed noises

Next Steps on Speech Enhancement for MAX78000

- Domain Adverserial Training
- Optimize with MAX78000
- Add more noise types and mix inputs
- Add more languages

Flask Demo



Questions, comments, or suggestions?

Reach out to us!

STELLA CAO

zijcao@uchicago.com

RENJIE LIU

renjiel@uchicago.edu

CURTIS ZHUANG

shengyan@uchicago.edu

HUAXUAN WANG

huaxuan@uchicago.edu

THANKS FOR WATCHING

Special thanks to Dr. Batuhan Gundogdu, Dr. Utku Pamuksuz,
Chris Olen and Scott Shepard for their advice and assistance in our project

Appendix

Q & A

- *Generate Noisy Speech(P30)*
- *ISTFT Model(P31)*
- *Ideal Ratio Mask(P32)*
- *Why Don't Use Phase(P33)*
- *Hearing Aid Future Market(P34)*
- *Masking and Mapping (P35)*

Generate Noisy Speech

- A custom PyTorch dataset module
- Total 30 conditions are considered:
  - 6 noise types
  - 5 different SNR each (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB
- SNR Mixer function generate noisy speech given specific SNR value

SNR Mixer function:

$$alpha = \frac{RMS_{clean}}{RMS_{Noise} \times 10^{\frac{SNR}{20}} + eps}$$

$$NoisySpeech = CleanSpeech + alpha \times Noise$$

ISTFT Module - Recover Back to Speech Audio

- A custom module that wraps up PyTorch's istft function
- The predicted mask are multiplied with noisy magnitudes to get enhanced magnitude
- Applied transformation back (log transform, log powered transform, and quantization) on magnitude
- The istft function that has same parameter setups with stft function takes both magnitude and phase to generate enhanced speech audio

```python
def forward(self, dt):

    if self.target == 'mask':

        dt = self.mask_recover(dt)
    else:
        if 'pred_mask' in dt:
            dt['pred_y'] = dt['pred_mask']

        # if dt['pred_y'].shape[2] == 256:
        #     dt['pred_y'] = F.pad(dt['pred_y'], (0, 1, 0, 0))

    if self.transform_type == 'logmag':

        for key in ['mixed_mag', 'clean_mag', 'pred_y']:

            dt[key] = torch.expm1(dt[key])
            dt[key] = torch.clamp(dt[key], min= 0)

    elif self.transform_type == 'lps':

        for key in ['mixed_mag', 'clean_mag', 'pred_y']:

            dt[key] = torch.sqrt(10 ** dt[key])

    elif self.transform_type == 'normal':

        for key in ['mixed_mag', 'clean_mag', 'pred_y']:

            dt[key] = torch.clamp(dt[key], min= 0)
            dt[key] = normalize_quantized_spectrum(dt[key])

    if self.cnn == '1d':

        dt = self.cnn1d_recover(dt)

    elif self.cnn == '2d':

        dt = self.cnn2d_recover(dt)

    for key in ['mixed_y', 'true_y', 'pred_y']:

        dt[key] = torch.istft(dt[key].cpu().detach().T,
                              n_fft = self.n_fft,
                              hop_length=self.hop_length,
                              win_length=self.win_length,
                              window=self.window)

    return dt
```

Ideal Ratio Mask

- **Ideal Binary Mask (IBM)**
  - Make the speech enhancement to binary classification task
- **Ideal Ratio Mask (IRM)**
  - smoothed version of IBM
  - Predicting the ratio between clean speech and noisy speech
  - Widely used in SE community
- **Other Masking strategies**
  - Take account of phase
    - Complex Ideal ratio mask (cIRM)
    - Phase sensitive mask (PSM)

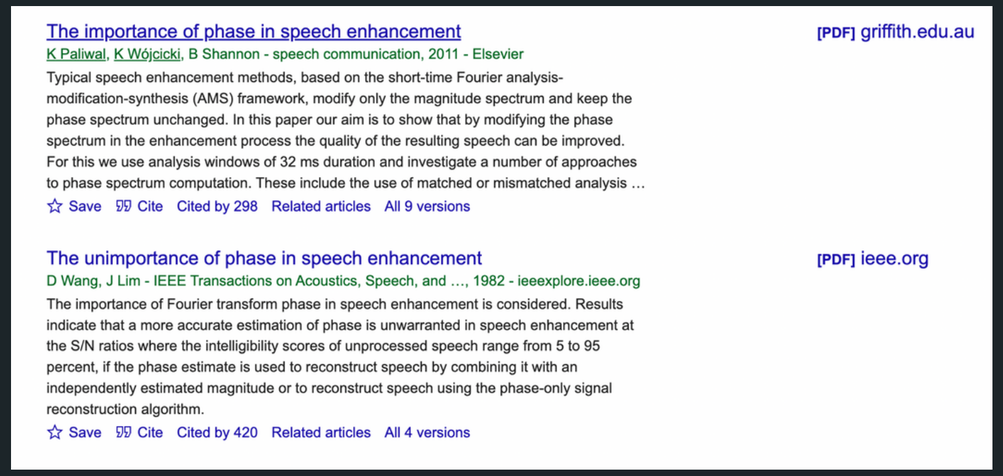
$$\mathrm{IBM}(t,f) = \begin{cases} 1, & if\ |S(t,f)|^2 - |N(t,f)|^2 > \theta \\ 0, & otherwise \end{cases}$$

$$\mathrm{IRM}(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2} \right)^{\beta}$$

$$\mathrm{M} = \frac{Y_r S_r + Y_i S_i}{{Y_r}^2 + {Y_i}^2} + i\,\frac{Y_r S_i - Y_r}{{Y_r}^2 + {Y_i}^2}$$

Why Don't Use Phase

- Phase may or may not contain useful information for Speech Enhancement task
  - The unimportance of phase in speech enhancement (Wang, D., & Lim, J. , 1982)
  - The importance of phase in speech enhancement (Paliwal, K., Wójcicki, K., & Shannon, B. , 2011)
- Involved complex-valued computation
  - Not suitable for MAX78000 device



The importance of phase in speech enhancement [PDF] griffith.edu.au

K Paliwal, K Wójcicki, B Shannon - speech communication, 2011 - Elsevier

Typical speech enhancement methods, based on the short-time Fourier analysis-modification-synthesis (AMS) framework, modify only the magnitude spectrum and keep the phase spectrum unchanged. In this paper our aim is to show that by modifying the phase spectrum in the enhancement process the quality of the resulting speech can be improved. For this we use analysis windows of 32 ms duration and investigate a number of approaches to phase spectrum computation. These include the use of matched or mismatched analysis …

Save Cite Cited by 298 Related articles All 9 versions

The unimportance of phase in speech enhancement [PDF] ieee.org

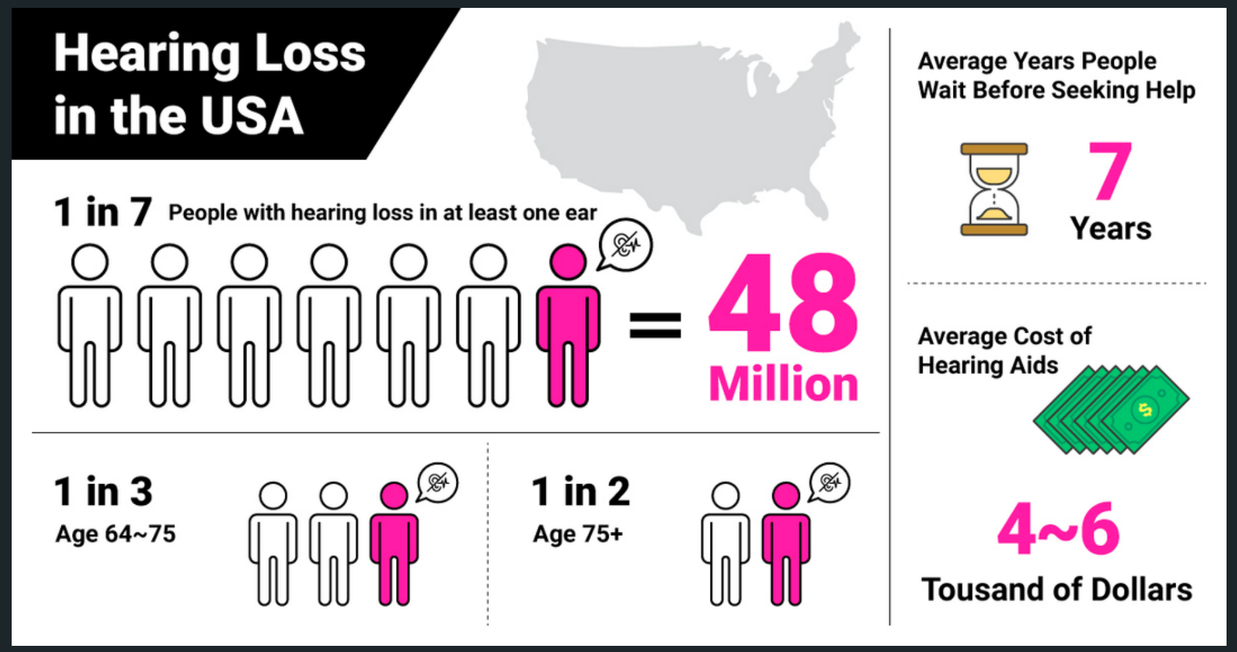
D Wang, J Lim - IEEE Transactions on Acoustics, Speech, and …, 1982 - ieeexplore.ieee.org

The importance of Fourier transform phase in speech enhancement is considered. Results indicate that a more accurate estimation of phase is unwarranted in speech enhancement at the S/N ratios where the intelligibility scores of unprocessed speech range from 5 to 95 percent, if the phase estimate is used to reconstruct speech by combining it with an independently estimated magnitude or to reconstruct speech using the phase-only signal reconstruction algorithm.

Save Cite Cited by 420 Related articles All 4 versions

Hearing Aid Future Market

The global “hearing aids market” size is expected to reach $11 billion by 2028, exhibiting a CAGR of 7.4 % during the forecast period.



Masking vs Mapping

CNN and CDAE are less affected by training target and show masking and mapping approaches are approximately giving the same performance.

'The performance of masking targets was shown to be much better at high SNR for all architectures, leading to a higher variance for all the evaluation metrics than that of the mapping targets, which makes mapping targets less affected by SNR changes.' (Nossier, 2020)