Capstone Implementation: Maxim Speech Enhancement and Source Separation for Wearable Devices

By

Stella Cao, Renjie Liu, Huaxuan Wang, Curtis Zhuang

Master of Science in Analytics

University of Chicago

Instructor: Utku Pamuksuz

Abstract

Microphones will often record clean speech and noise at the same time, but clean speech is the only matter of interest. Therefore, a systematic approach is needed to improve the speech quality. Speech enhancement aims to increase speech quality through machine learning algorithms. This technology plays an important role in hardware and software audio improvement. Various research have been conducted using a wide range of machine learning algorithms both in academia and industry. In this paper, we specifically focus on using convolutional neural networks to improve speech related to Maxim Integrated's hardwares.

List of keywords

speech enhancement, convolutional neural network, real-time applications, hardware optimization

Executive summary

The goal of this project is to separate noise, either stationary or non-stationary from the speaker's voice and improve speech quality. The expected output is a series of speech enhancement neural networks compatible with the MAX78000 device with high power efficiency and low latency. To be more specific, we have to design and implement a system with minimal latency (less than 10msec) and produce and enhanced SNR.

We target CNN as our baseline model and then develop models with better performance, as well as the lowest latency and battery power consumption. Since noise is an important factor, we will start from training the model with a single stationary noise, followed by multiple non-stationary noises.

One of the challenges in this project is how to separate the speaker's voice and the voice in the background, especially when there are multiple people speaking, it is challenging to separate the main speaker from all the others.

Since speech enhancement is still a developing technology, we have to walk through a lot of papers, from data processing to model selection, for example, converting raw speech signals to time-frequency representation and models and real-time convolutional neural networks etc.

1. Introduction

With the work from home trend continuing, meeting people virtually has become a part of our life. Undoubtedly, online meetings save people both money and time on commuting. The effects of noise in our daily life are showing up with more and more conferences or activities held online. The definition of noise in this project is different from the one always mentioned in noise pollution. The latter one usually refers to sounds which are at or above 85 decibels, and would put people at risk of hearing loss. While all the noises mentioned in this project refer to everyday sounds, including but not limited to horns honking, kids yelling, dogs barking or whining etc. In face to face communication, humans could focus on one stimulus while mute all the other interferences , which is known as the cocktail party effect. In order to improve the meeting quality and efficiency, let the audience hear and understand the speaker more clearly, how to cancel the background noise becomes quite important.

Speech enhancement aims to improve human speech quality by distinguishing the speaker's voice from background noise and cutting back all the other sounds. Speech enhancement could not only be used in online meetings, but also be used in hearing aids, speech recognition etc. Companies like Microsoft, Google, Nvidia, research groups like Imperial College London, MIT, IEEE etc. all started working on speech enhancement these days. Google even establishes a deep learning model which can output only the audio from the selected speaker in the video.

Our client, Maxim Integrated, a leading edge-computing technology company that manufactures and sells a wide portfolio of high-performance analog and mixed-signal products and technologies. They also apply speech enhancement to their wearable devices. How to cancel the noise from the background and transfer it to wearable devices is quite important. While current deep learning models like recurrent neural networks and convolutional neural networks introduce unwanted latency and undesirable battery power consumption in operation. Besides, due to the limitation of the hardware, some traditional deep learning models like attention or RNN with multiple layers may not fit in the low-bit and low-power accelerators (MAX78000). To avoid undesirable costs and improve user experience, conducting a light and portable model becomes the best solution.

Analysis/Research Goals:

The goal of this project is to design a machine learning algorithm that improves the audio quality of Maxim Integrated's hardwares.

The experiment design is as follows:

Input and Output: The model's input will be an audio file consisting of clear speech and noise, and the output will be as clear as possible audio.

Data: The model will be trained by using Microsoft Deep Noise Suppression Challenge data and Open Speech and Language Resources.

Evaluation: The data will come in as audio file format. In the beginning phase, the model will be evaluated based on human perception. Furthermore, the model will be evaluated by MSE, MOS, and PESQ.

Scope: The findings present in this paper have several scope limitations. The machine learning algorithm is best suitable for devices similar to Maxim Integrated since there are computing power limitations. In addition, because of the type of speech we use to train the algorithm, the same level of performance would not be expected for other audio categories.

2. Background and Literature Review

2.1 Overview of Speech Enhancement

In the real-life, people often hear noisy speech when they talk via Zoom or the cell phone. It is hard to understand the speech when the speech-to-noise ratio (SNR) is low. Noisy speech contains clean speech and noise. Noise can be grouped into stationary noise and non-stationary noise. Street noise, train noise are examples of non-stationary noise, while examples of stationary noise include white noise. Then the relation between speech and noise with respect to time t can be written as:

$$y(t) = x(t) + n(t)$$

where $y(t)$ is noisy speech, $x(t)$ is clean speech signal with convolutional noise or room impulse response (RIR), and $n(t)$ is additive noise. The RIR is a transfer function between the original sound source and the clean speech signal. In the time domain, Speech Enhancement (SE) task is to directly recover the target speech $x(t)$ from the noisy speech $y(t)$. On the other hand, there is a time-frequency domain. In order to transfer the speech signal from time domain to TF domain, researchers apply short-term Fourier transform (STFT) on the speech signal as the input of the Speech Enhancement Model and Inverse-STFT on the output of the model to get the enhanced speech.

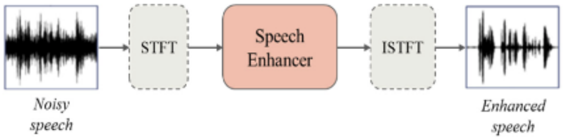


Figure 1. Speech Enhancement System Overview.

Speech Enhancement aims to improve the intelligibility and the quality of the speech signal. In recent years, SE plays an important role in the fields of robust automatic speech recognition (ASR), teleconferencing, and hearing aids design.

2.2 Speech Process

In the traditional Speech Enhancement works [Loizou], most works make strong statistical assumptions on characteristics of the speech signals, and aim to estimate the underlying target speech using mathematical algorithms. These works include Spectral-Subtractive Algorithms, Wiener Filtering, and Subspace Algorithms, etc. They often require less data volume, have less latency, and perform well under specific assumptions, even if the algorithm is in a new environment. Since they heavily depend on the assumptions of the speech signals, they have difficulty handling the speech signals that have non-stationary noise. In the more recent methods, researchers tend to use deep-learning modeling and depart from knowledge-based modeling. Most of these works treat the Speech Enhancement task as a supervised problem. Deep learning techniques require less statistical assumptions of the speech signal, however, their performances rely on lots of data, and design of the architecture of the Deep-learning model. In addition, deep-learning models may not perform well if they are in a new environment. Deep-learning Speech Enhancement Systems can be grouped into two categories: Time Domain, and Time-Frequency Domain. In each Domain, there are single-channel based SE and multichannel based SE. Single-channel SE addresses the problem of recovering clean speech from a noisy speech that is captured from one microphone, while multichannel SE's noisy speech is from multiple microphones. In this section, we will only investigate the single-channel based SE based on the business problem. And single-channel SE in low SNR condition is considered as the one of the most challenging problems in recent studies.

2.1.1 Time-Frequency Domain

In many Speech Enhancement works, researchers get their input features and targets by converting raw speech signals to time-frequency representation. These methods are operated on a 2-dimensional time-frequency spectrogram. They treat the spectral magnitude as a training target. And combine the phase with enhanced magnitude to generate enhanced speech signals by applying ISTTF.

The Microsoft Team proposed a flexible and scalable Convolutional Recurrent Networks (CRN) for speech enhancement (Tang and Chuanxin, 2020, 3816). While state-of-the-art models can achieve outstanding performance in SE, the main challenge is to get a compact model that has the same performance as a state-of-the-art model and is applicable in real-time. The Microsoft Team built a small size recurrent and convolutional-recurrent networks architecture that have reasonable inference time by skip-connections and parallel RNN grouping.

2.1.2 Time Domain

In recent studies, researchers found that TF-Domain has two main problems. First, STFT transforms the speech signal into magnitude and phase. The phase is ignored during the denoise process. But the phase information is important in speech quality. Second, using STFT will slow the speed of denoising. Time domain methods avoid these two problems by directly predicting target speech from the noisy speech.

Choi and his team proposed a Tiny Recurrent U-Net that its size only has 363 Kilobytes (KB) and it matches the performance of current state-of-art models (Choi and Hyeong-Seok, 2021 5789). The model takes the Per-channel energy normalization (PCEN) as its inputs. For its architecture, the model has the encoder-decoder format. The encoder is composed of 1D Convolutional Neural Networks blocks (1D-CNNs) and a frequency-axis Gated Recurrent Unit block (FGRU). The decoder is composed of a Time-axis GRU and 1D Transposed CNN blocks. They also applied skip-connection and the Phase-aware β-sigmoid mask in the model.

2.3 Model Architecture

Regardless of input, as discussed in the review (Asri et.al 2021, 20), there are two main categories of methods in speech enhancement training: mapping based and masking based. Mapping-based method takes SE as a regression question where the model maps noisy speech into clean speech. On the other hand, a masking-based approach treats SE as a classification question where a mask is estimated and applied for filtering noisy speech to clean output.

As for the distinction between those two methods, many models can be used for both methods with DNN, and GAN being one of the representations. For the rest of this part, we will go over some of the most prevalent models that have been adopted in the SE field: DNN, RNN-LSTM, DAE, GAN, CNN, and FCN.

2.3.1 DNN

DNN models are first introduced to replace conventional MMSE based methods as it is deployed as a nonlinear regression model with more complex modeling ability (Xu et al. 2015, 7). Compared with traditional SNR models used in AMS that cannot suppress sharp spectral peaks, the DNN model proposed by Xu was trained on more than 100 noise types and able to handle unseen noise environments. The authors normalized the input features to zero mean and unit variance before feeding it into the MMSE-based object function. And to overcome the over-smoothing issue caused by residual errors, the authors performed an equalization between global variance of the estimates and the clean speech.

Zhao (Zhao et al. 2018, 5075) proposed an enhanced model by incorporating speech perception model into loss function, which addresses the problem of MSE not reflecting speech quality and intelligibility. To create the mask, authors employ a log magnitude spectrum of the input noise to estimate IRM. And for their loss function, they compute a modified STOI value with 24-frame enhance magnitude spectrum Y and 24-frame clean magnitude spectrum X where $||\quad||_F$ denotes the Frobenius norm, and λ denotes a tunable hyperparameter:

$$L(m) \;=(1 \;-\; f(X_m^{24};\, Y_m^{24}))^2 \;+\; \lambda * \left\|X_m^{24} - Y_m^{24}\right\|_F /24$$

This proposed method improves STOI, PESQ and SDR performance compared with other DNN models, and the improvements in speech intelligibility does not sacrifice the speech quality.

2.3.2 RNN-LSTM

RNN-LSTM models, as compared to many of the neural network models, were introduced as a solution back in the days trying to solve ASR problems. But it has been proven to be more successful when compared with DNN models at various SNR levels (Asri et al. 2021, 21). A recent paper by Gao et.al (Gao et al. 2018, 5055) proposed a progressive learning framework with LSTM layer. The progressive learning framework allows the model to be trained on multiple SNR levels compared with direct mapping and by incorporating the LSTM architecture, the network will be able to hold long term information. The whole progressive learning structure links input and multiple target layers together allowing the progressive learning to happen and the multiple targets are averaged for improving the overall model performance. Just like most other RNN based models, this model is very computationally expensive and resource consuming.

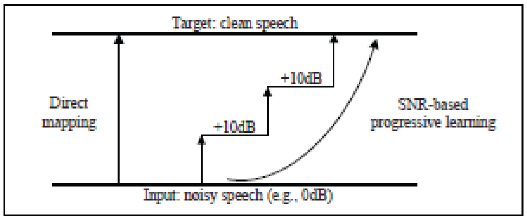
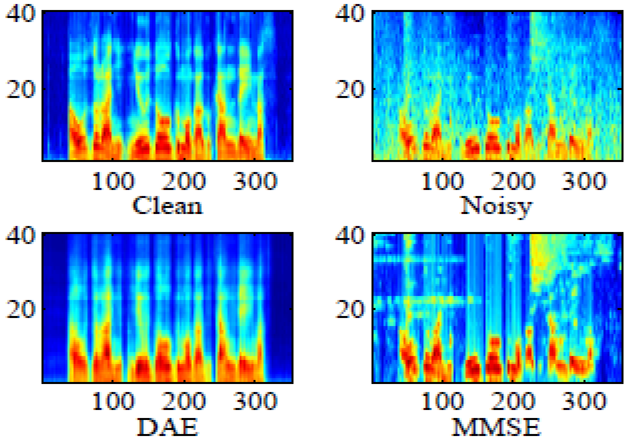


Fig. 1. Progressive learning for speech enhancement [26].

2.3.3 DAE

Deep autoencoders have already been used for extracting noisy features in image processing and the ASR realm when combined with recurrent structure. But when applied to noise cancellation problems, pre-training autoencoders with noisy-clean speech pairs and then stacking those AEs together into a DAE which serves a filter for noise cancelation. (Lu et al. 2013, 436). DAE is a very suitable model for SE given its ability to learn statistical differences between speech and noise automatically. For this model, large amounts of training data is preferred to reduce overfitting. Number of hidden layers within will increase the accuracy of the output but at the same time lead to overfitting issues. Depth of the DAE is expected to improve performance given a large enough training dataset. When compared with traditional approaches, DAE method leads to less distortion and clearer output, which is also shown by graph 2 .

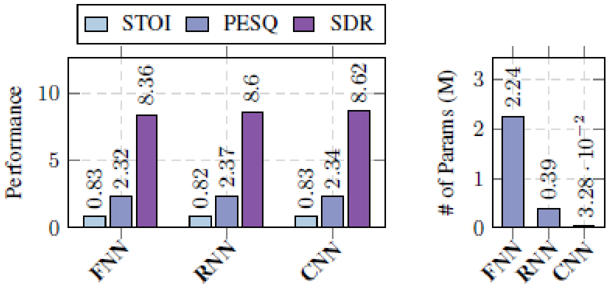


2.3.4 CNN

Compared with FNN and RNN models, CNN models are more effective when it comes to extracting speech features due to its lesser number of parameters. This important feature allows CNN models for SE to be implemented in embedded devices. Bhat el al. are able to implement low latency CNN-based SE models to be run realtime and offline on smartphones (Bhat et.al 2019, 78421).

Park and Lee proposed a modified CNN model, R-CED (Redundant Convolutional Encoder-Decoder) network, which is a fully convolutional network for speech enhancement (Park & Lee 2016). The R-CED network consists of no pooling layer, and instead it directly encodes the features into higher dimension with filters; as for the decoder layer, it applies filters systematically followed by a convolution layer as the last layer. By applying skip connections, R-CED results in SDR = 8.62 compared with 8.6 for RNN networks with corresponding network size. And more significantly, CNN models are 12 times smaller than RNN.

CNN models are also capable of processing raw waveforms directly with a network structure, as we have discussed in the Time Domain part. One such model is called WaveNet, this avoids the loss by conducting Fourier transformation and also reduces computational complexity significantly (Oord et al. 2016).



2.3.5 GAN

The generative adversarial network can map features of noisy speech into clean speech where then the discrimination network comes in as a binary classifier and decides whether the samples come from the clean speech of the output enhanced speech. This has received increased attention given GAN's generative ability. Speech enhancement GAN systems typically utilize one generator for the enhancement process and a study by Phan et al. proposed two multi-stage GAN models, ISEGAN (Iterated SEGAN) and DSEGAN (Deep SEGAN), via multiple enhancement mappings (Phan et al. 2020, 1701). ISEGAN has a common mapping for all its stages while DSEGAN has independent mapping for each stage and is updated accordingly. Results show that DSEGAN performs optimally compared with both SEGAN and ISEGAN across all speech quality metrics. Also DSEGAN is able to increase in performance with more signals passing to generators while ISEGAN shows a downward trend.

2.3.6 FCN

Fully convolutional network (FCN) is similar to CNNs except that it does not have a fully connected layer, which allows FCNs to adopt wave-form both as input and output (Fu et al. 2017, 011). FCNs can further reduce the number of parameters in the neural network compared with CNN which will be more adaptable in mobile devices. But equally importantly is FCN's ability to model both high and low frequency components of raw waveforms simultaneously, allowing the accurate characterization of high frequency components which are often lost due to fully connected layers.

2.4 Industry Research and Practices

2.4.1 The Microsoft DNS Challenge

Microsoft's Deep Noise Suppression (DNS) challenge is created to advance development and research in the field of noise suppression to achieve superior perceptual speech quality. Challenge participants will use Microsoft provided dataset to train and test their machine learning algorithms. A subjective evaluation framework is also used as a tool to evaluate participant's performance. In this challenge, Microsoft provided over 20 hours of clean speech with singing and provided more information about the characteristics of the noise based on stationarity. Microsoft also provided over 100000 synthetic and real room impulse responses (RIRs) curated from other data sets.

2.4.2 Microsoft's Baseline Model

Microsoft used the SE method from xxx as a baseline model, which is based on Recurrent Neural Network (RNN). This method uses log power spectra as input to predict the enhancement gain per frame using a learning machine based on Gated Recurrent Units (GRU) and fully connected layers. NSNet is computationally efficient. It only takes 0.16ms to enhance a 20ms frame on an Intel quad core i5 machine using the ONNX run time v1.12 .

2.4.3 Participant's Model

The TU Braunschweig and Goodix Technology team used a fully convolutional recurrent network (FCRN) to enter the Interspeech 2020 Deep Noise Suppression (DNS) Challenge. They trained the FCRN with a multi-target loss accounting for differences in quality perception of noisy or reverberated speech by controlling the weight on desired dereverberation and denoising.

The team used a two-step training approach, first performing a pretraining with part of the WSJ0 speech and fine tuning the pretrained models with a subset of the DNS training data. This approach limits the amount of training time by testing multiple hyperparameter settings only for fine tuning on DDNS.

The TU Braunschweig and Goodix Technology team's method outperforms all reference methods of the preliminary test and ranks third for the realtime and second for the non-real time track amongst all submissions to the challenge.

2.4.4 Conclusion

The results of the INTERSPEECH DNS Challenge show we still have a long way to go in achieving superior speech quality in challenging noisy conditions.

2.5 Applications

2.5.1 Hearing Aid

N Shankar proposed a real-time dual-channel SE by voice activity detector (VAD) assisted minimum variance distortionless response (MVDR) beamformer for hearing aid applications using smartphones (Shankar and Nikhil, 2020, 952).

2.5.2 Automatic Speech Recognition (ASR)

A Pandey exploited a SE for improving a recurrent neural network transducer based ASR system. They used dense convolutional recurrent networks for complex spectral mapping based SE. They found it is helpful for ASR systems in two ways: a data augmentation technique, and a preprocessing frontend (Pandey and Ashutosh, 2021, 223).

2.6 Data

Data used to train the model mainly from Microsoft DNS Challenge, includes clean speech datasets and noise datasets.

- The clean speech dataset is derived from Librivox, a public audio book dataset which includes recordings from a group of worldwide volunteers. One of the limitations of this dataset is that some of the recordings have poor quality due to the background noise and reverberation.
- The noise dataset is mainly from Audioset and Freesound, Audioset includes sound clips from YouTube videos. Room Impulse Responses will not be taken into our consideration.

Data processing includes transfering the data by utilizing Short-time Fourier transform (STFT), which is used to determine time-localized frequency for the signal. And combine the noise set and clean speech dataset together with various speech noise ratios (SNR).

Measurements include mean opinion score (MOS), signal distortion (SIG), and background noise intrusiveness (BAK), all of them have a scale from 1 to 5, higher the better. And also segmental signal-to-noise ratio (segSNR), distance measures, source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI) etc.

Acknowledgement

Special thanks to Dr. Batuhan Gundogdu, Dr. Utku Pamuksuz, and our Teaching Assistant Chris Olen for their advice and assistance in our project. Also our sincere thanks to Maxim Integrated who provided us this research opportunity.

Reference

Bhat, Gautam S., Nikhil Shankar, Chandan KA Reddy, and Issa MS Panahi. "A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone." *IEEE Access* 7 (2019): 78421-78433.

Choi, Hyeong-Seok, Sungjin Park, Jie Hwan Lee, Hoon Heo, Dongsuk Jeon, and Kyogu Lee. "Real-time Denoising and Dereverberation with Tiny Recurrent U-Net." *arXiv preprint arXiv:2102.03207* (2021).

Fu, Szu-Wei, Yu Tsao, Xugang Lu, and Hisashi Kawai. "Raw waveform-based speech enhancement by fully convolutional networks." In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 006-012. IEEE, 2017.

Gao, Tian, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "Densely connected progressive learning for lstm-based speech enhancement." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5054-5058. IEEE, 2018.

Lu, Xugang, Yu Tsao, Shigeki Matsuda, and Chiori Hori. "Speech enhancement based on deep denoising autoencoder." In *Interspeech*, vol. 2013, pp. 436-440. 2013.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).

Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." *arXiv preprint arXiv:1609.07132* (2016)

Pandey, Ashutosh, and DeLiang Wang. "Dense CNN with self-attention for time-domain speech enhancement." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 1270-1279.

Pandey, Ashutosh, Chunxi Liu, Yun Wang, and Yatharth Saraf. "Dual application of speech enhancement for automatic speech recognition." In 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 223-228. IEEE, 2021.

Phan, Huy, Ian V. McLoughlin, Lam Pham, Oliver Y. Chén, Philipp Koch, Maarten De Vos, and Alfred Mertins. "Improving GANs for speech enhancement." *IEEE Signal Processing Letters* 27 (2020): 1700-1704.

Shankar, Nikhil, Gautam Shreedhar Bhat, and Issa MS Panahi. "Real-time dual-channel speech enhancement by VAD assisted MVDR beamformer for hearing aid applications using smartphone." In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 952-955. IEEE, 2020.

Tang, Chuanxin, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng. "Joint Time-Frequency and Time Domain Learning for Speech Enhancement." In *IJCAI*, pp. 3816-3822. 2020.

Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "A regression approach to speech enhancement based on deep neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, no. 1 (2014): 7-19..

Yuliani, Asri Rizki, M. Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. "Speech Enhancement Using Deep Learning Methods: A Review." *Jurnal Elektronika dan Telekomunikasi* 21, no. 1 (2021): 19-26.

Zhao, Yan, Buye Xu, Ritwik Giri, and Tao Zhang. "Perceptually guided speech enhancement using deep neural networks." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074-5078. IEEE, 2018.