2022 IDL Project Proposal

Microsoft Teams to Cloud Speech Enhancement

Brian Lim, Chanwoo Kim, Taeyoung Chang, Urvish Takker
Carnegie Mellon University

Abstract

With an increasing trend in remote work, real-time speech enhancement models are becoming important in conferencing platforms for improving the experience of online communication. However, the difficulty comes in the real-time aspect of these models, as it is hard to expect noise that arises in the future. Because noises can provide unsatisfactory quality of speech and eventually disrupt conversations, suppressing them is crucial for a stable meeting. Our proposed model, ¡¿, intended to promote collaborative research in real-time Speech Enhancement aimed to maximize the subjective (perceptual) quality of the enhanced speech (add how the model would process audio to extract clean speech). Our model aims to take a spectrogram of an audio that consists of both clean speech and artificial noise and To analyze the efficacy of our proposed model, we use metrics that tend to be better predictors of real-time performance over conventional objective metrics. For evaluation, we use a dataset generated using a large clean speech and noise corpus from the DNS challenge(Reddy et al. [2020]) for training the proposed noise suppression model and representative of real-world noisy data. (The paper would contain descriptions of the dataset, proposed model, and results of the test on the proposed model and metrics.)

1 Introduction

Speech enhancement has become one of the most important technologies in a remote working environment. Meetings done through video conferencing platforms such as Zoom or Microsoft Teams are more prevalent than ever, and extracting clean speech from a noisy environment in real time is becoming more important to avoid interruptions during a conference. However, solely extracting clean speech in real-time is a difficult task, especially with imperfect information on the level of noise in the future. Furthermore, suppressing noise can also lead to the suppression of clean speech as well, removing essential parts of a conversation. Losing essential signals from speech may further degrade its quality, and to prevent this, noise must be filtered out the moment it is detected in audio. Thus, we want to build a model which would perform better in real time.

Our goal is that when we get noisy speech data derived from synthesizing clean speech and noise, we want to build a model (where the basic structure is FullSubnet and some properties of Demucs are added in a clever way? Not determined yet...) that yields a speech close to the clean speech. In other words, we want to build a model that fulfills a noise suppression while perceptual quality and intelligibility of the output is at least competitive with the noise suppression output of Microsoft Teams.

Here, the input is a spectrogram transformed from the synthesized speech file. The spectrogram can be viewed as a 2D tensor of a shape with (T, F). For each time $t \in \{1, \cdots, T\}$, the wave form of short time segment near t is Fourier-transformed (STFT)

and the absolute value of this short term fourier transform at each frequency f is stored as F-dimensional vector. Our model is defined as mapping a spectrogram of noisy speech data to the training target cIRM which stands for complex Ideal Ratio Mask(Williamson et al. [2016]). Since the desired output of cIRM can be viewed as the complex ratio of clean speech spectrum to noisy speech spectrum, training via the loss of measuring distance of actual output and the desired output can be interpreted as knowing well where the noise exists in time-frequency domain so that we can suppress such noise.

2 Related Works

2.1 Demucs

The authors of this paper(Defossez et al. [2020]) utilize convolutional and recurrent neural networks to separate noise from the clean speech in waveform audio. By processing corrupted audio through convolutional encoder-decoder architecture with skip-connections they propose that clean speech can be extracted from latent representations of input audio. Furthermore, to account for sequential information in the audio a recurrent neural network is injected between encoders and decoders, which the decoder uses to reconstruct clean speech. To account for real-time noise suppression the authors accumulate standard deviations of audios only up to the current position for normalization, with paddings to account for a 3ms lookahead.

2.2 FullSubNet

This paper(Hao et al. [2020]) introduces a full-band and sub-band fusion model, FullSubNet, which enhances single-channel speech. Full-band model has its strength in finding global spectral patterns and cross-band dependencies. On the other hand, sub-band model has its strength in finding local spectral patterns and stationarity, which is an important factor that distinguishes noises from clean speech. Since both models have their own advantages, FullSubNet stacks both models to take both advantages. The whole speech spectrogram goes through a full-band model, and the output is concatenated with sub-band unit. Then the concatenated ones become frequency-wise inputs of the sub-band model. Hence, the sub-band model can find local patterns with complementary information captured from the full-band model. Besides, this model exceeds the high-ranked models in the DNS Challenge (INTERSPEECH 2020).

2.3 GeMAPS

This paper(Eyben et al. [2016]) suggests a minimalistic set of standard acoustic parameters such as Pitch, Jitter, Shimer, Loudness, Harmonics-to-Noise Ratio, and Harmonic difference. It is selected based on three criteria: the potential to discern physiological changes in voice production, the success in the past literature, and the theoretical importance. Through the experiments, the paper argues that, unlike the large brute-forced feature sets, the minimalistic parameter sets might reduce the danger of damaging generalization capabilities to unseen data. The authors proposed these parameters as the common baseline for evaluating future research. Using those metrics, we can measure the subjective qualities of the output from our model.

3 Dataset

We will use the dataset from the github repository of DNS Challenge 2020. There is an explanation for the dataset in Reddy et al. [2020].

For training data, we will get the noisy speech dataset by synthesizing clean speech and noise. The clean speech dataset comes from the public audiobooks dataset called Librivox. It has recordings over 10,000 public domain audiobooks by 11,350 speakers where majority of them are in English. Many of these recordings have excellent speech quality, but still some are of poor speech quality with speech distortion, background noise and reverberation. Thus, DNS Challenge 2020 has filtered the clean speech data based on the speech quality, which is measured by the Mean Opinnion Score (MOS), and chose only the upper quartile with respect to MOS as the clean speech dataset. The resulting dataset has 500 hours of speech from 2150 speakers and all the filtered clips are split into segments of 10 seconds.

The noise clips were selected from Audioset and Freesound. Audioset is a collection of about 2 million human-labeld 10 second sound clips. They are drawn from YouTube videos and belong to about 600 audio events. Since certain audio event classes such as music and speech are overrepresented in the Audioset, DNS Challenge 2020 has tried to balance the dataset by sampling so that each class audio event class has at least 500 clips. Also, DNS Challenge 2020 has removed the clips with any kind of speech activity since speech-like noise can bring out the suppression of speech while the trained model tries to suppress speech-like noise. The resulting noise dataset has about 150 audio classes and 60,000 clips. An additional 10,000 noise clips from Freesound are also augmented to the dataset.

The noisy speech dataset is created by adding clean speech and noise at various SNR levels. (Synthesizing the clean speech and noise : same method as the paper? Or new method?)

Although what we observe in the real world is not perfectly expressed by the synthetic dataset, we take advantage of the synthetic dataset since most of the speech enhancement (SE) models require a clean reference for utilizing objective metrics such as PESQ or STOI.

For the test set, we can use synthetic test dataset which adds the clean speech from the Graz University's clean speech dataset and noise clips from the Audioset and Freesound, which are not present in the training set. Since these synthetic clips come with ground truth references which are the clean speech data, we can evaluate the method using objective metrics such as PESQ and STOI. Whereas for the test set, or 'blind' test set, there is no ground truch references provided. We can use this set for the final evaluation using subjective metrics.

Here, for our problem, we can write $x = y + n$ where y is the clean speech, n is the noise, and the x is the synthesized noisy speech. Simply put, x is the input for our model and the y is the desired output of our model. If we denote $\hat{y}$ as the actual output of our model and $y^\star$ as the output from the Microsoft Teams platform, then we want for the divergence of $\hat{y}$ and y to be smaller than that of $y^\star$ and y. In this way, we can measure how well our noise suppression model works compared to Microsoft Teams.

4 Evaluation Metric

Evaluation Metric List	
Metric	Mathematical Expression
STOI	$$STOI = \frac{1}{JM}\sum_{j,m} d_{j,m}$$ • J : the number of one-third octave bands (frequency) • M : the total number of frames (time) • $d_{j,m}$: sample correlation between $\mathbf{x}_{j,m}$ and $\overline{\mathbf{y}}_{j,m}$ • $\mathbf{x}_{j,m}$: the short-time temporal envelope of the clean speech • $\overline{\mathbf{y}}_{j,m}$: the short-time temporal envelop of the (normalized and clipped) degraded speech
PESQ	$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind}$$ • D_{ind} : average disturbance • A_{ind} : asymmetric average disturbance • a_0, a_1, a_2 : a_0 is set as 4.5 and $a_1, a_2 < 0$ can be modified for the given task
fwSegSNR	$$fwSegSNR = \frac{10}{M}\sum_{m=1}^{M-1} \frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{\lvert X(j,m)\rvert^2}{(\lvert X(j,m)\rvert - \lvert\hat{X}(j,m)\rvert)^2}}{\sum_{j=1}^{K} W(j,m)}$$ • K : the number of bands • M : total number of frames • $\lvert X\rvert$: weighted (by a Gaussian shaped window) clean signal spectrum • W : weight on j-th frequency and m-th frame, which is exponential $\lvert X\rvert$

STOI(Taal et al. [2011]) stands for a short-time objective intelligence measure. Essentially it measures the average of every correlation coefficient between each of short-time segments of DFT(discrete Fourier transform)-based band for clean speech and the output of noise reduction from the noisy speech input.

Perceptual Evaluation of Speech Quality (PESQ)(Rix et al. [2001]) is a metric used for obtaining Mean Opinion Scores (MOS) of speech in an audio. This objective metric, ranging from -0.5 to 4.5, measures the quality of an audio based on a sum of disturbances

in the audio, and while it is not intended for speech enhanced by noise suppression, we follow Braun et al. [2021], our baseline model, to compare our model's performance.

SDR or Speech to Distortion Ratio: Distortion is unwanted signal correlated with the input signal. It's absent until the signal appears, and is usually not significant until the signal is occupying most of the dynamic range of the system. This can be a very good metric to track as distortion is something we can reduce and it might lead to clean speech results.

fwSegSNR(Hu and Loizou [2008]), Frequency weighted Segmental Signal-to-Noise Ratio, is a generalized short time performance measure. It gives each frequency's SNR two kinds of weights. First, static frequency weighting that is derived from known psychoacoustic properties of hearing. Second, dynamic frequency weighting which is related to the speech production mechanism.

5 Baseline

Microsoft Teams is a business communication platform developed by Microsoft, of which usage share keeps increasing so that its number of users reaches 270million in 2022 (Source). It mainly provides lots of features that support live speech in meetings, calls, and so on. Hence, the need for speech enhancement is inevitable. So we aim to improve it more and adopt Microsoft Teams as the baseline of our model. Although papers from Microsoft including Braun et al. [2021] suggest Microsoft's research into an extension of the Demucs architecture, the exact model used in Microsoft Teams is not publicly available due to the confidentiality of their research. However, we can still set our model's baseline as the model used in Microsoft Teams by obtaining the suppressed audio by inputting the audio through their platform. To minimize the noise that could come from recording audio through a microphone, we inject the corrupted audio through a virtual microphone to simulate the inference stage of inputting audio files into Microsoft's model.

References

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. 2020. URL https://arxiv.org/abs/2005.13981.

Donald S. Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):483–492, 2016.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. 2020. URL https://arxiv.org/abs/2006.12847.

Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. 2020. URL https://arxiv.org/abs/2010.15508.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set

(gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. URL https://sail.usc.edu/publications/files/eyben-preprinttaffc-2015.pdf.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001.

Sebastian Braun, Hannes Gamper, Chandan K. A. Reddy, and Ivan Tashev. Towards efficient models for real-time deep noise suppression. 2021. URL https://arxiv.org/abs/2101.09249.

Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1): 229–238, 2008.