Just-Noticeable Speech Enhancement

Chitrank Gupta, Donald S. Williamson
baekgupta@cse.iitb.ac.in, williads@indiana.edu

Abstract

Many speech enhancing technologies claim to improve the quality of noisy speech. One part of my research involves finding out whether the improvement made in the quality of the speech perceivable to humans or not. This would also help in comparing the realistic performance of two speech enhancing algorithms- whether or not the difference in their improvement makes any difference to humans. Also, through deep learning we would be able to enhance signals just noticeably better or to the extend as desired.

Overall Objectives

- **Listening Study.**
  The primary objective of this research was to determine the minimum audible difference that causes a signal sound just better than some reference signal. Because of the unavailability of such data, a listening study had to be conducted that would help in obtaining responses from the people. They had to be as accurate as possible and also had to be collected on the data as generalised as possible.
- **GAN for Speech Enhancement.**
  Lastly, from the responses collected, a general adversarial network (GAN) had to be trained that would help in enhancing signals just noticeably better.

Concepts

Humans assess the speech quality on factors like preference, loudness and intelligibility, including others. Adding noise mostly affects the preference for, and the intelligibility of, the speech. This makes the comparison among the noisy signals, varying only in their SNR, the best way to assess the speech quality only for their preference. SNR is generally measured in decibels (dB) which is given by the formula- $\text{SNR}_{dB} = 10 \log \left(\frac{Power_{\text{signal}}}{Power_{\text{noise}}} \right)$.

Future works

The work may be extended to different languages and also targeting for a specific group of people- for example for people with hearing disability.

References

[1] Pascual, Santiago and Bonafonte, Antonio and Serrà, Joan, SEGAN: Speech Enhancement Generative Adversarial Network, 2017.

[2] Rothauser, E. H. "IEEE recommended practice for speech quality measurements." IEEE Trans. on Audio and Electroacoustics 17 (1969): 225-246.

Acknowledgements

A part of this research is funded by the PI's IU research accounts. The support is gratefully acknowledged. Thanks to all the people who voluntarily participated in the task of listening study.

Listening Study

Listening Study was the central part of my research. It involved the collection of the response of subjects on certain evaluations and the analysis of this data. For the listening study, an app on MATLAB® was developed, which acted as an interface to collect such responses from the subjects. The protocol used for the study was-

1. The subjects were provided eleven signals in addition to a reference signal.
2. They had to listen to the reference signal and one by one compare its quality with the eleven other in the order from "signal 1" to "signal 11".
3. The first time they hear a difference in the quality, they had to indicate it as their "JND signal" response. Similarly, the signal of the best quality among the latter eleven had to be provided as their response for "Best signal".

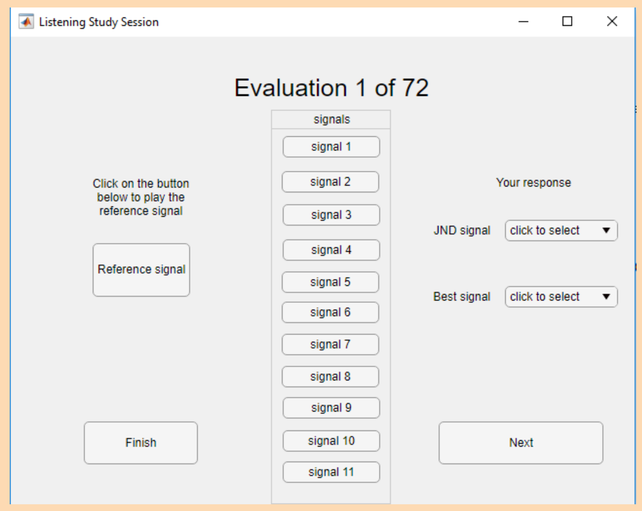

Figure 1: GUI for Listening study

Each of the latter eleven signals was 0.5 decibels (dB) better, in SNR than the previous one and the SNR of baseline signal varied from -6 dB to 15 dB. The evaluations not only differed in their noise type, with three different noise types- babble, airport and restaurant. Much thought needs to be given while designing an elegant experiment. Examples are described in figures.

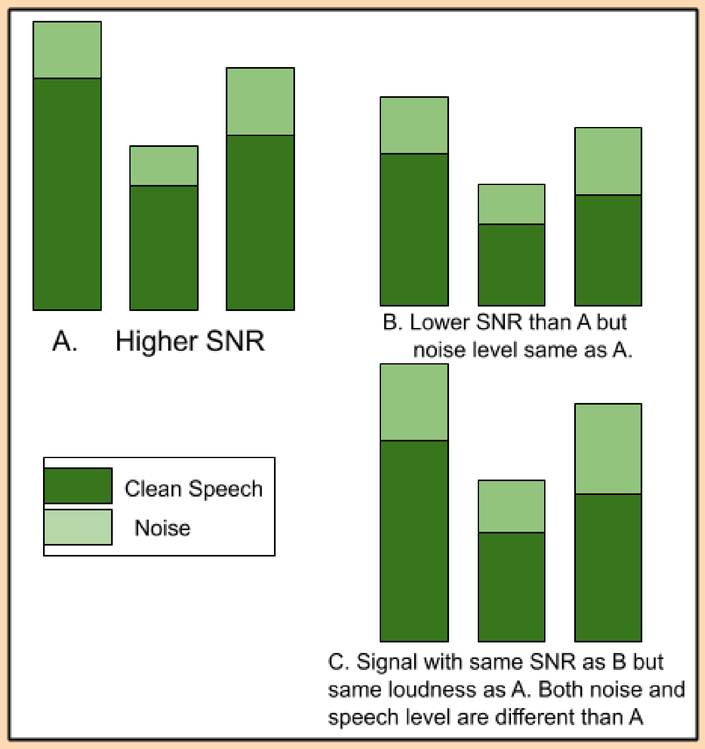

Figure 2: Level Roving: The comparison between the signals A and B is not the same as comparing the signals A and C.

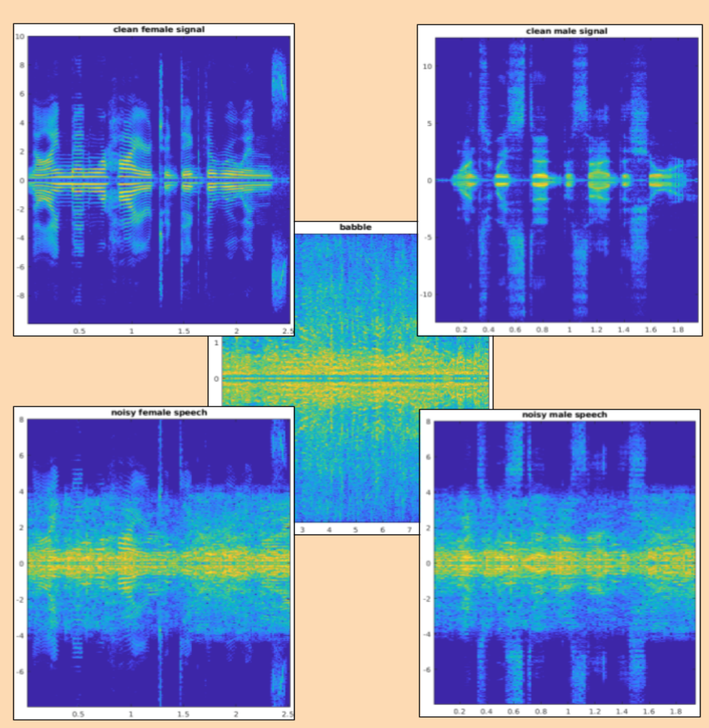

Figure 3: Bottom two signals of SNR -6 dB. Babble, with its louder lower frequency components, obstructs the male speech more than the female speech.

GAN for speech enhancement

A fully convolutional GAN was trained to enhance noisy speech just noticeably better. Most of the model was mimicked from the SEGAN. The generator was similar to an autoencoder while the discriminator's architecture was same as that of generator's encoder part.

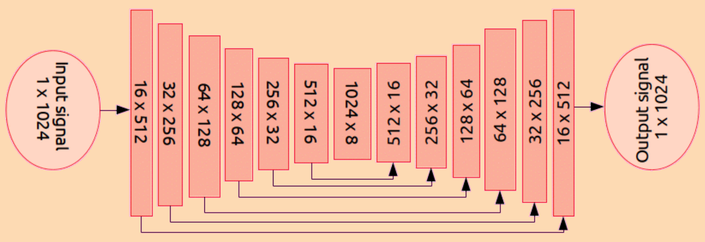

Figure 4: Autoencoder architecture of the Generator. The arrows between decoder and encoder blocks represent concatenation in feature maps. `F` × `L` in any block represents `Feature Maps` × `Length`

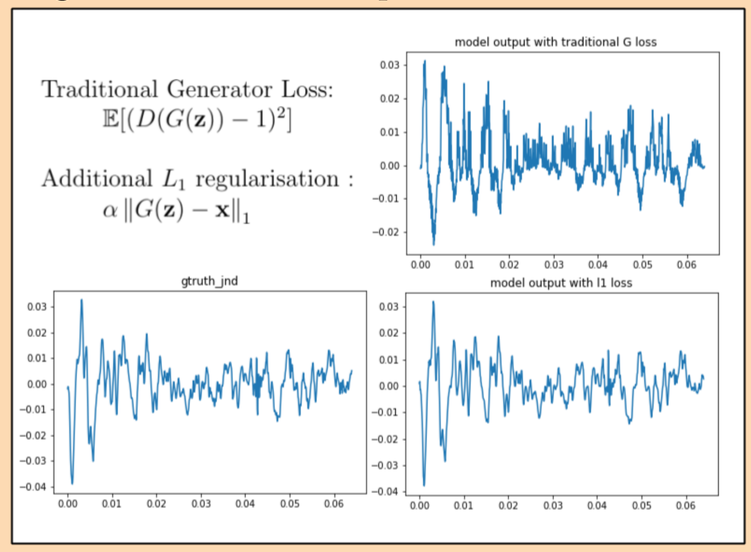

Figure 5: L_1 loss component increases performance of GAN. Both models trained for 10 epochs.

Conclusions

Currently, the formal study is yet to begin. Rough analysis from the some preliminary data is given alongside. All are measured in dB.

► Table 1a,1b: Preliminary data analysis

Base SNR	JND SNR
-6	Beyond 6
-3 to +3	3 to 4

Base SNR	JND SNR
+6 to +12	4 to 5
+15	2 to 3