The reconstruction of speech/voice with the use of noise cancelling algorithms and machine learning.

2023 Part IV Project Group: 58

Authors: Edward Chan & Timothy Cabrera

Project Introduction

Speech enhancement is the application of algorithms to deconvolve a speech signal from a noise-corrupted signal. This is most commonly done through statistical estimation or neural networks that provide filter coefficients or apply non-analytic filtering methods.

However, for both cases, the estimation of intelligible speech from signals with very high noise levels remains a challenge without using a microphone array to utilise spatiotemporal information of the signal.

Many standard communication devices need access to these multi-channel microphone configurations. Therefore, this research project aims to contribute to the field of single-channel speech enhancement by investigating the potential of a proposed hybrid deep neural network and model-based approach for single-channel speech enhancement.

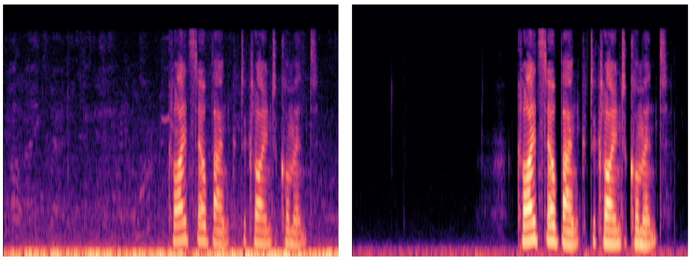
Project Goals

Our key project requirements are:

- A machine learning approach that can estimate noise and speech based on an observed mixture.
- A classifier that can distinguish between different signal-to-noise levels.
- The resulting enhanced signal from mixtures with SNR levels <0dB should be intelligible.

Background Theory

The signal-to-noise ratio (SNR) is a ratio that measures the presence of a signal relative to the noise in the signal. Figure 1 shows a spectrum of a mixture considered low SNR, while Figure 2 shows the opposite.



Figures 1 and 2: Spectrograms representing a speech audio signal's low SNR and high SNR.

To enhance the signal, it is common practice to use a collection of pre-defined noise/speech spectra candidates (codebook), where each candidate is scaled such that the sum of all candidates will result in the desired spectrum of noise and speech, as shown in Figure 3. However, this only works for some types of mixtures.

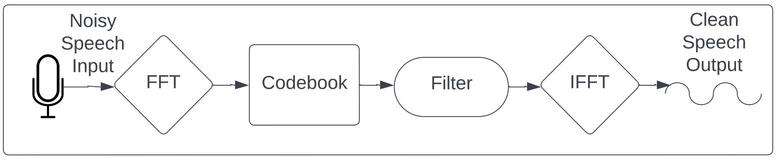


Figure 3: Conventional Speech Enhancement Algorithm

Results

A confusion matrix of the testing data is used to distinguish between the different SNR levels from the CNN classifier model, as seen in Figure 5. The accuracy is 86.17%.

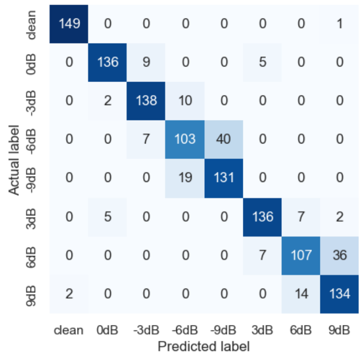


Figure 5: CNN classification model accuracy when testing using a confusion matrix.

To show the performance of the proposed design, a graph is shown in Figure 6, where the estimated spectrum is compared against the real spectrum. The SNR of the mixture is -9dB.

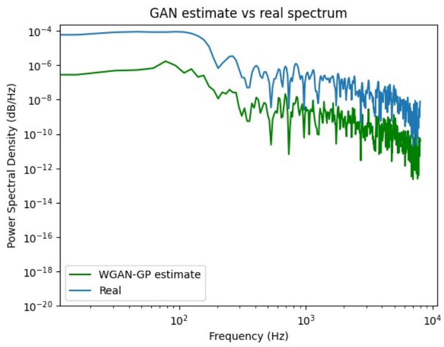


Figure 6: GAN spectrum vs. real noise spectrum.

Discussion and Conclusion

The proposed design was able to accurately estimate the spectrum of noise at a range of different SNRs. In contrast, it could not reliably estimate the speech spectra at different SNRs. As such, a custom filter was designed such that only the noise spectra are needed for spectral attenuation. However, results show that it still results in distortion, and unintelligible speech frequencies are also attenuated.

Future considerations:

- Increase the training data for the GAN.
- Add more classifications, such as no speech with more background noises.
- Further, develop the GAN so that speech spectra can also be estimated reliably to be compatible with a Wiener filter.

Generative Adversarial Network Design & Implementation

Our system implementation enhances speech using a generative adversarial network, as shown in Figure 4.

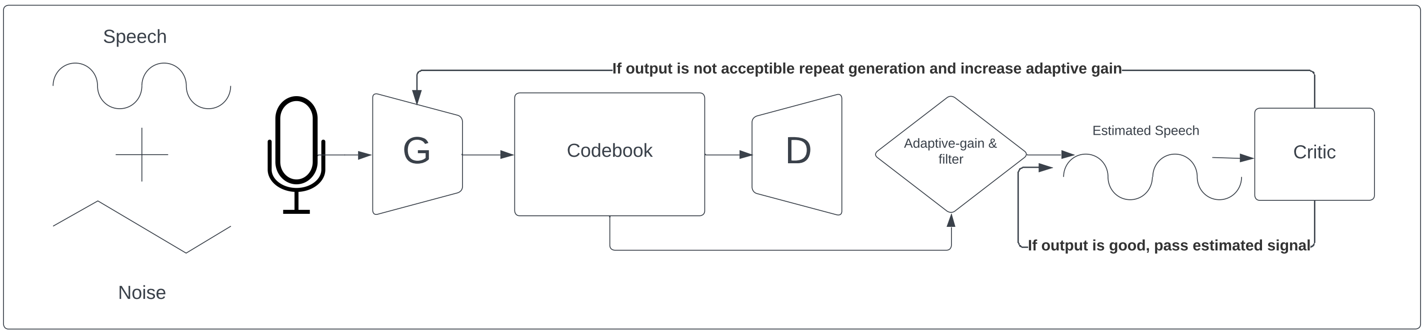


Figure 4: Flow chart of the Preposed Design

- Generative adversarial networks (GAN) are neural networks used to generate realistic images through small input vectors.
- A GAN has been designed from scratch to take in the mixture's power spectral density and generate a codebook based on the observed mixture.
- The codebook is scaled and used through an adaptive gain filter to attenuate the noise frequencies.

Supervised by: Catherine Watson and Yusuke Hioka