Exercise 5: Enhancement and Evaluation

1 Introduction

The objective of this exercise is to implement basic speech enhancement techniques and evaluate and visualize the quality of the enhancement. Briefly put, we implement four different filtering methods: Spectral-subtraction, Wiener-filter, linear filter and a VAD based filter. In all these filters, (1) a constant average magnitude noise model and (2) ideal noise estimate, which is the true noise you generate to create the noisy signal, are used. The enhanced signals are evaluated by computing the signal-to-noise ratios- global SNR and segmental SNR. To visualize the results, the segmental SNRs of all enhanced signals are plotted. Besides this, the spectrograms of the clean, noisy and the three enhanced results and plotted and visually inspected.

- In this exercise, you must implement and return the following functions:
  1. `ex5_main.py`: The assignment can be run from this file and each smaller function can be called using this script.
  2. `ex5_funcs.py`: Script comprises of all the function skeletons required to complete the assignment.
- Return your answers to MyCourses by 23:59 on Tuesday, October 19, 2021

Method (`ideal_noise`) stands for the method wherein the noise estimate is the ideal estimate, ie: the noise estimate is equal to the true noise. Additionally, method (`avg_noise_model`) is elaborated in the next section.

The following sections provide the necessary basic theory to implement the required the exercise tasks.

2 Noise model estimation

In this example, we use a constant average noise model, which means that as an estimate of noise energy, for each frequency, we use the average noise energy. Specifically, if $V(f, k)$ is the noise spectrum in frame k for frequency f, then the average energy is

$$|\hat{V}(f)|^2 = \frac{1}{N} \sum_{k=1}^{N} |V(f, k)|^2. \tag{1}$$

3 Filtering

To attenuate noise in the signal, we will multiply the signal with weighting factors, which are calculated for each frequency and each frame. The weighting factors are dependent on both the observed energy and the noise model.

3.1 Spectral subtraction

Conventional spectral subtraction is defined such that we subtract the estimated noise energy $|\hat{V}(f)|^2$ from the energy of the observation $|X(f,k)|^2$. We do not have an estimate of the phase, whereby we do not modify the signal phase. The energy of the estimated clean signal is then

$$|\hat{S}(f,k)|^2 = \begin{cases} |X(f,k)|^2 - |V(f)|^2, & \text{when } |X(f,k)|^2 > |V(f)|^2 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Above we used a threshold $|X(f,k)|^2 > |V(f)|^2$, because otherwise the estimated signal energy would be negative. Since negative energy is physically impossible, we put the estimate to zero always when that would occur.

Furthermore, we want to keep the phase of the original signal, whereby

$$\angle\hat{S}(f,k) = \angle X(f,k) = \frac{X(f,k)}{|X(f,k)|}. \tag{3}$$

The final estimate is then

$$\begin{aligned} \hat{S}(f,k) &= |\hat{S}(f,k)| \cdot \angle\hat{S}(f,k) \\ &= \begin{cases} X(f,k)\sqrt{\frac{|X(f,k)|^2 - |V(f)|^2}{|X(f,k)|^2}}, & \text{when } |X(f,k)|^2 > |V(f)|^2 \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

3.2 Wiener filter

Above we found that spectral subtraction can be written in the form $\hat{S}(f,k) = g(f,k)X(f,k)$, where $g(f,k)$ is a positive scaling coefficient. We can optimize $g(f,k)$ such that the output error energy is minimized, whereby we obtain

$$g(f,k) = \frac{|X(f,k)|^2 - |V(f)|^2}{|X(f,k)|^2}. \tag{5}$$

The Wiener estimate is then

$$\begin{aligned} \hat{S}(f,k) &= |\hat{S}(f,k)| \cdot \angle\hat{S}(f,k) \\ &= \begin{cases} X(f,k)\frac{|X(f,k)|^2 - |V(f)|^2}{|X(f,k)|^2}, & \text{when } |X(f,k)|^2 > |V(f)|^2 \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

3.3 Linear filter

Subtracting energies is a rather heuristic approach, whereby any other heuristic approach could be potentially just as good. One such heuristic approach is linear subtraction, where we subtract magnitudes $|V(f)|$ instead of energies $|V(f)|^2$. The linear estimate is thus

$$\begin{aligned}\hat{S}(f,k) &= |\hat{S}(f,k)| \cdot \angle \hat{S}(f,k) \\ &= \begin{cases} X(f,k)\frac{|X(f,k)|-|V(f)|}{|X(f,k)|}, & \text{when } |X(f,k)|^2 > |V(f)|^2 \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{7}$$

3.4 VAD-based filter

The objective is to demonstrate one of the many possible applications of a VAD. Here, we use the ideal VAD outputs, which you utilized in exercise 3 as the target outputs, to remove noise from frames. If the target output is 1, it implies that the frame is predominantly a speech frame, and then we use the Wiener filter to remove noise from such a frame. If the target output is 0, implying that the frame is mostly a silence frame, we remove all the energy in the frame. Mathematically, this is represented as:

$$\begin{aligned}\hat{S}(f,k) &= |\hat{S}(f,k)| \cdot \angle \hat{S}(f,k) \\ &= \begin{cases} X(f,k)\frac{|X(f,k)|^2-|V(f)|^2}{|X(f,k)|^2}, & \text{when } VAD_{output} == 1 \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{8}$$

4 Signal-to-noise ratio

To quantify the performance of enhancement, we can measure the signal-to-noise-ratio (SNR) between the clean signal and the estimated signal

$$\text{SNR} = \frac{\|S\|^2}{\|E\|^2} = \frac{\|S\|^2}{\|S-\hat{S}\|^2}. \tag{9}$$

Here S, $\hat{S}$ and E are the clean signal, estimated signal and the estimation error $E = S - \hat{S}$. A central choice in the application of SNR-measures is to decide whether to apply them for the whole signal (global SNR) or window-by-window (segmental SNR).

4.1 Global SNR

To determine the global SNR, we calculate the energy of the original signal s_n and the estimation error e_n, over the whole sound sample.

$$\text{SNR}_{\text{global}} = \frac{\sum_{k=1}^{N}|s_n|^2}{\sum_{k=1}^{N}|e_n|^2} = \frac{\sum_{k=1}^{N}|s_n|^2}{\sum_{k=1}^{N}|s_n-\hat{s}_n|^2}, \tag{10}$$

where s_n and $\hat{s}_n$ are the the original and estimated time-signals and N is the length of the whole sound sample.

4.2 Segmental SNR

Though the global SNR is simple to calculate, it does not take into account how humans perceive signal energy over time. Specifically, a high-energy section of the signal will dominate the whole SNR estimate, such that errors in low-energy areas are not taken into account properly. Still, for humans, SNR in low-energy areas can be just as important as the SNR in high-energy areas.

To make sure that we can measure the SNR such that it is independent of the energy in any one frame, we can first calculate the SNR in each frame and then take the average over all frames. Specifically, we first calculate the SNR for each frame k as

$$\mathrm{SNR}(k) = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |E(f,k)|^2} = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |S(f,k) - \hat{S}(f,k)|^2}, \tag{11}$$

where $S(f,k)$ and $\hat{S}(f,k)$ are the spectra of the clean and estimated signals and F is the number of frequency components.

In this exercises, you should thus plot $\mathrm{SNR}(k)$ over all k.

Usually, we would also calculate the mean of the frame-wise or segmental SNR as

$$\mathrm{SNR}_{\mathrm{segmental}} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{SNR}(k), \tag{12}$$

where K is the number of frames.

5 Steps in implementation

1. Generate a noisy signal, in which the noise is additive white Gaussian noise of power -35dB

2. Estimate the noise for the noisy signal, based on 1) ideal estimate 2) avg magnitude model, by completing `noiseEst`. Note that this function should return estimates of dimension same as the input noise matrix.

3. Enhance the noisy signal by implementing the filtering functions 1) spectral subtraction: `spectralSub`, 2) Wiener filter: `wiener`, 3) Linear filter: `linear`, 4) VAD based filter: `vadEnhance`

4. Compute the global SNR and the frame-wise segmental SNR of the enhanced signals by computing 1) `snrGlb` 2) `snrSeg`

5. Plot and visualize the results.