ET4147: Signal Processing for Communications

Homework 3

Erik Hagenaars
4272404

Teun de Smalen
4321014

TODO LIST

I. Introduction

The use of microphones and audio devices are becoming more relevant every year. Speech audio is used to communicate humans and recently often with devices as well. Hearing loss and noisy spaces are an increasing problem which makes it harder to communicate. It is therefor important to come up with techniques to enhance the desired speech signal from a noisy environment. This comes with many challenges including correlating noises, dynamic noise energies and interference noise. This report discusses the work done on the mini-project for the course "IN4182 - Digital Audio and Speech Processing" in which a single microphone speech enhancement system is designed.

The first objective is to create an architecture of the system, which is discussed in Section II. From this, five subsystems are designed and implemented in the Sections III to VII. After designing the system, its performance will be tested and concluded in Section IX.

As a bonus, a multi microphone system is designed with the implementation of beamformers in Section VIII.

II. System

A. Signal model

Before the architecture of the system can be designed, the signal model and assumptions are made. For the signal model, Additive White Gaussian Noise (AWGN) is expected. If Y is defined as the signal with AWGN, S the desired source signal and N the noise, the model can be expressed as shown in Eq. 1 and Eq. 2. In which the time domain and frequency domain expressions are shown.

$$Y_t[n] = S_t[n] + N_t[n] \quad \text{(time domain)} \tag{1}$$
$$Y_k[l] = S_k[l] + N_k[l] \quad \text{(frequency domain)} \tag{2}$$

To simplify the model more, some assumptions are made. The first assumption is that the source signal and noise are uncorrelated (Eq. 3). This allows for the autocorrelation of the received signal to be simplified. Since the source and noise are uncorrelated, the autocorrelation of the received signal can be expressed by the addition of the autocorrelation of the signal and the noise (Eq. 4). The second assumption is that the speech signal is wide-sense stationary (WSS) in small frames 5). Since a speech signal can be seen as a periodic signal or noise, this assumption holds in theory. In practise, speech is not stationary but the performance of the enhancement system is sufficient.

$$R_{S_t N_t}(n, m) = 0 \quad \text{(uncorrelated)} \tag{3}$$
$$R_{Y_t Y_t}(n, m) = R_{S_t S_t}(n, m) - R_{N_t N_t}(n, m) \tag{4}$$
$$R_{Y_t Y_t}(n, m) = R_{Y_t Y_t}(m - n) \quad \text{(wide-sense stationary)} \tag{5}$$

An important property of audio is the power spectrum density (PSD). The PSD is defined as the Fourier Transform (FT) of the autocorrelation (Eq. 6). And with the assumption of uncorrelated noie and source, this can be expressed as the PSD of the signal and noise added as in Eq. 7. Since the frames are of finite length, an estimation of the PSD needs to be made. The estimation of Eq. 8 is called the periodogram of the signal. To enhance this estimation, Bartlett's method can be used shown in Eq. 9.

$$P_{YY,k} = \lim_{L\to\infty} \sum_{m=-L/2}^{L/2} R_{Y_t Y_t}(m) e^{-j2\pi \frac{km}{K}} \tag{6}$$

$$= P_{SS,k} + P_{NN,k} \tag{7}$$

$$\hat{P}^P_{YY,k}(l) = \frac{1}{L} |Y_k(l)|^2 \tag{8}$$

$$\hat{P}^B_{YY,k}(l) = \frac{1}{M} \sum_{m=l-M+1}^{l} \hat{P}^P_{YY,k}(m) \tag{9}$$

B. System overview

From the signal model, a few components of the system can be derived. Firstly, to hold the WSS assumption, the signal sequence needs to be split in multiple segments. This process will be expressed as Framing. Since the PSD is an important property of the signal, it need to be converted to the frequency domain. A problem is that the FT can cause problems in the sidelobes of the frames due to Gibb's phenomenom. To avoid this, the framing function needs to window this signal frame to suppress the edges of the frame. The frames will become overlapping. At the end of the system, the time domain representation of the signal will be recovered using the inverse FT and the frames will be merged using the overlapp add method where a deframing window is used. To estimate the source, some other properties need to be estimated from the received signal. These important properties include the estimated noise PSD, the estimated source PSD (by estimating the SNR) and if the source is present. With these properties, a gain filter can be created for the received signal which should suppress the noise and interference. The resulting system design is shown in Fig. 1.

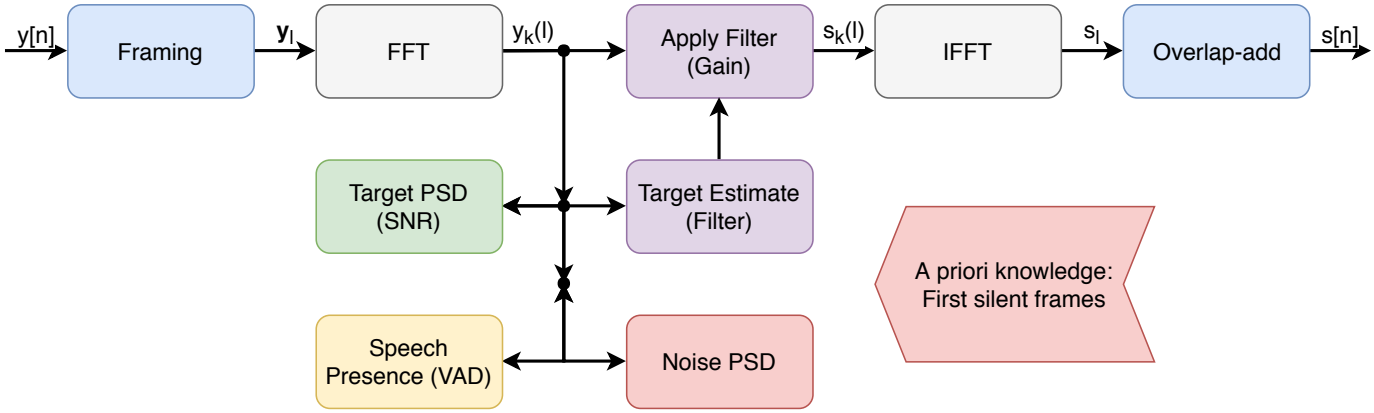

Fig. 1: Overview of the system.

This system is focused on a single microphone setup. With a multiple microphone setup, the speech enhancement can be improved upon. Direction estimation and spatial filtering with beamformers can be implemented to filter the unwanted interference and noise from certain directions. This subsystem can be placed before the single microphone speech enhancement system.

The system's functions will be divided into six Sections. The framing and overlap add function will be implemented in Section III. The noise PSD estimation and the source PSD estimation are discussed in Section IV and Section V respectively. The voice activity detection will be discussed in Section VI. With all this information, the target estimation where the filters are designed is discussed in Section VII. And lastly, the spatial filtering will be discussed in Section VIII.

III. Framing & Overlap Add

The first step, as described in Figure 1, is the framing of the audio file. This is done according to Equation 10. Where l is the frame index (the l-th frame), n is sample number, R is the hoplength. $w[n]$ is the window used to smoothen the signal in such a way that the signal does not become discontinuous and cause wrong sidelobes wen applying the FT.

$$y_l[n] = y[n + lR]w[n], \quad n = 0, \ldots, N-1 \tag{10}$$

The last step of the system is the Overlap Add-block. The windowing is removed after which the samples are added back together to one file.

$$y[n] = \sum_{l=1}^{k} y_l[n]/w[n] \tag{11}$$

There are various windows that are suited for framing an audio signal. The standard Hamming and Hann window and the square-root Hann window used in a paper by Hendriks[1] was evaluated. When overlap adding all the frames without applying any changes, the signal should be recovered without any trouble. The overlap added windows should then be equal to an ones vector. Using different values for the overlap percentage, the corresponding overlap add vector is generated and compared to an all-ones vector. An important variable for the framing function is the overlap percentage. The results in Fig. 2 show the negative mse of these two vectors. The peak value is around 70% vor the square-root Hann.

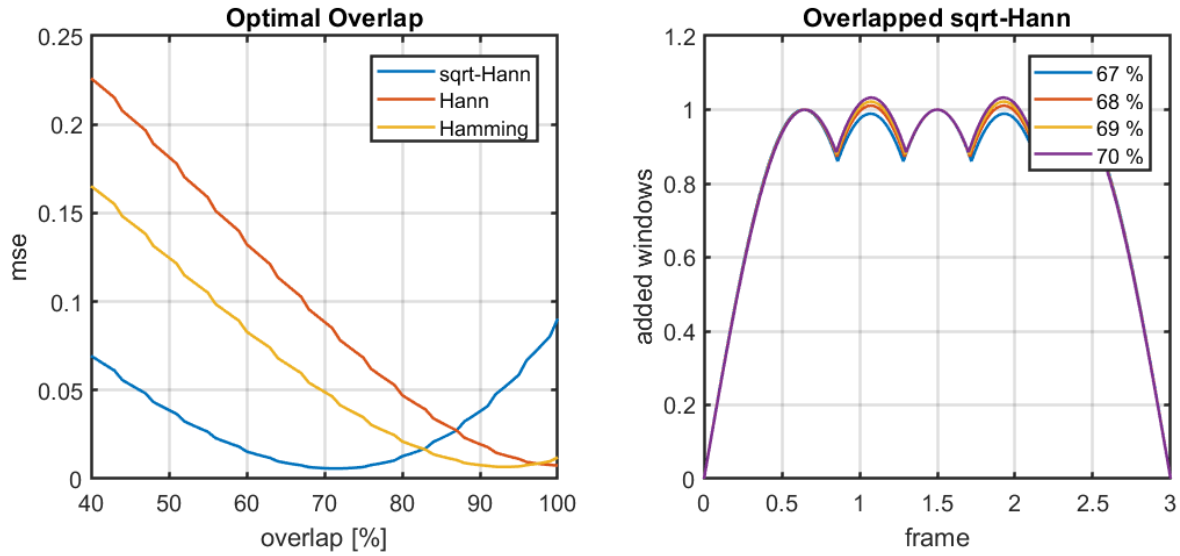

Fig. 2: Optimal overlap percentage for the square-root Hann window.

IV. Noise Estimation

For the noise estimation, it is considered that speech and noise are independent and uncorrelated. From this Equations 12 and 13 can be derived.

$$H_0 : Y_K(l) = N_k(l) \quad \text{(speech absence)} \tag{12}$$
$$H_1 : Y_K(l) = S_k(l) + N_k(l) \quad \text{(speech presence)} \tag{13}$$

$$f(x) = \begin{cases} \alpha\hat{\sigma}^2_{N,k}(l-1) + (1-\alpha)\left|y_k(l)\right|^2 & \text{when } H_0(l) \\ \hat{\sigma}^2_{N,k}(l-1) & \text{when } H_1(l) \end{cases} \tag{14}$$

$$\mathbf{Q} = \{P_{YY,k}(l-M+1) \quad \dots \quad P_{YY,k}(l)\} \tag{15}$$
$$\hat{\sigma}^2_{N,k}(l) = Q_{min} \tag{16}$$

$$\widehat{\sigma^2_N}(l) = \alpha\widehat{\sigma^2_N}(l-1) + (1-\alpha)E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|\right] \tag{17}$$
$$E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|\right] = P\left(H_{0,k}(l)|y_k(l)\right) E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|, H_{0,k}\right] + P\left(H_{1,k}(l)|y_k(l)\right) E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|, H_{1,k}\right] \tag{18}$$

$$P\left(H_{0,k}(l)|y_k(l)\right) = 1 - P\left(H_{1,k}(l)|y_k(l)\right) \tag{19}$$
$$E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|, H_{0,k}\right] = \left|y_k(l)^2\right| \tag{20}$$
$$E\left[\left|N_k(l)\right|^2 \left|y_k(l)\right|, H_{1,k}\right] = \widehat{\sigma^2_N}(l-1) \tag{21}$$

$$P\left(H_{1,k}(l)|y_k(l)\right) = \frac{P\left(H_{1,k}(l)\right)\right) p_{Y|H_1}}{P\left(H_{1,k}(l)\right)\right) p_{Y|H_1} + P\left(H_{0,k}(l)\right)\right) p_{Y|H_0}} \tag{22}$$

$$p_{Y|H_0} = \frac{1}{\widehat{\sigma_N^2}\pi} \exp\left(-\frac{|y^2|}{\widehat{\sigma_N^2}}\right) \tag{23}$$

$$p_{Y|H_0} = \frac{1}{\widehat{\sigma_N^2}(1+\xi_{H_1})\pi} \exp\left(-\frac{|y^2|}{\widehat{\sigma_N^2}(1+\xi_{H_1})}\right) \tag{24}$$

V. SNR Estimation

$$\xi = \frac{\sigma_{S,k}(l)^2}{\sigma_{S,k}(l)^2} = \frac{P_{SS,k}}{P_{NN,k}} = \frac{E\left\{|S_k(l)|^2\right\}}{E\left\{|N_k(l)|^2\right\}} \tag{25}$$

$$\xi_k(l) = \frac{E\left\{|Y_k(l)|^2\right\}}{E\left\{|N_k(l)|^2\right\}} - 1 \tag{26}$$

$$= \frac{\hat{P}^B_{YY,k}(l)}{\frac{1}{L}E\left\{|N_k(l)|^2\right\}} \tag{27}$$

$$\xi_k(l) = \alpha\frac{E\left\{|S_k(l)|^2\right\}}{E\left\{|N_k(l)|^2\right\}} + (1-\alpha)\left(\frac{E\left\{|Y_k(l)|^2\right\}}{E\left\{|N_k(l)|^2\right\}} - 1\right) \tag{28}$$

$$|S_k(l)|^2 = \left|\hat{S}_k(l-1)\right|^2 \tag{29}$$

$$\frac{E\left\{|Y_k(l)|^2\right\}}{E\left\{|N_k(l)|^2\right\}} - 1 = \max\left[\left(\frac{|Y_k(l)|^2}{E\left\{|N_k(l)|^2\right\}} - 1, 0\right)\right] \tag{30}$$

VI. Voice Activity Detection

For the noise PSD estimation in Section IV and the SNR estimation in Section V, the pressence of speech is needed to increase the performance of the estimators. The Voice Activity Detector will evaluate the frame and give a speechflag wether speech is present or not.

This can be done by looking at the average energy of the frame. In the case of speech present and a SNR bigger than 0, this should be higher than in the case of noise only. To get a linear scaling with respect to the SNR, the log-energy is calculated. We then can express the energy difference as stated in Eq. 33 and Eq. 34. This likelihood function is shown in Eq. 31 and Eq. 32.

$$T(l) = \frac{1}{L}\sum_{k=1}^{k=L} log(\Lambda_k(l)) \gtrless_{H_0}^{H_1} \lambda \tag{31}$$

$$\Lambda_k(l) = P_{YY,k} \tag{32}$$

$$T(l)_{H_0} \approx \sigma_N^2 \tag{33}$$

$$T(l)_{H_1} \approx \sigma_N^2 + SNR \tag{34}$$

The threshold for the likelihood function could be a constant which is higher than the noise PSD. When noise is dynamic however, this can not be a constant. A dynamic value can be chosen based on the previous noise estimation and SNR estimation. A value between σ_N^2 and $\sigma_N^2 + SNR$ can be chosen.

The results of the VAD can be seen in Fig. 3. These results were made using the Noise and SNR estimates from the previous subsystems.

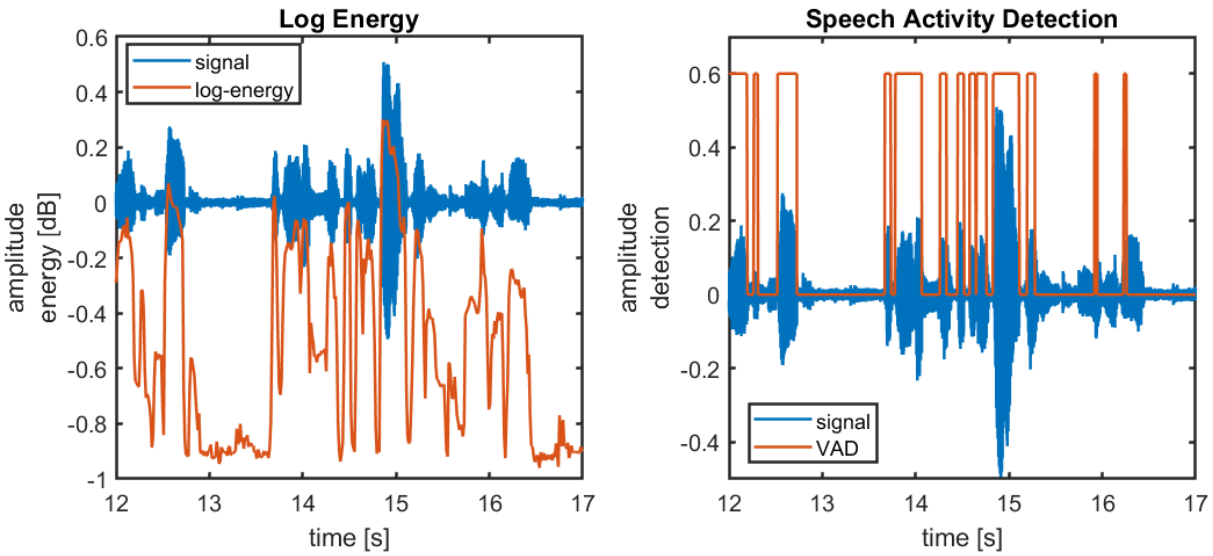

Fig. 3: Log-energy and VAD performance.

VII. Target Estimation

$$P_{SS,k}(l) = P_{YY,k}(l) - P_{NN,k}(l) \tag{35}$$

$$\widehat{|S_k(l)|}^2 = |Y_k(l)|^2 - |N_k(l)|^2 \tag{36}$$

$$\widehat{|S_k(l)|}^2 = \left(\max \left\{ 1 - \frac{E\left[|N_k(l)|^2\right]}{|y_k(l)|^2}, \epsilon \right\} \right)^{\frac{1}{2}} |y_k(l)| \tag{37}$$

$$S_k(l)^2 = \widehat{|S_k(l)|}^2 e^{j\angle y_k(l)} \tag{38}$$

$$\hat{S_k} = H_k \dot{Y_k} \tag{39}$$

$$H_k = \frac{P_{SY,k}}{P_{PYY,k}} \tag{40}$$

$$= \frac{SNR_k}{SNR_k + 1} \tag{41}$$

VIII. Multi-microphone beamforming

When using multiple microphones, the spatial properties of the incomming signals can be exploited. Since there is a (small) but relevant distance difference between the microphones, the time of arrival of the same signal differs. This time delay can be used to estimate the direction of the source and to filter other (interfering) directions.

Before the system is designed and the signals are discussed, some assumptions are made. First of all, the speech and noise are assumed to be WSS. Secondly a far field is assumed where the angles of arrival at every microphone is identical. These two assumptions can be used to simplify the incoming signals at each microphones. Since the signal is WSS, a time delay can be interpreted as a pahse shift in frequency.

Because there is a phase shift, spatial aliasing becomes important. This is where distance between microphones become too big where the periodic signal shifts a full time interval. To avoid this, the distance between microphones should be smaller than half of the wavelength. When assuming a maximum frequency of 8000KHz, the maximum distance between microphones is 2 milimiters.

Since the signal only differs in phase and a linear microphone setup, the signal model for each microphone can be defined as in Eq.

$$\mathbf{Y}_k(l) = \begin{bmatrix} S_k(l) & S_k(l)e^{-j2\pi \sin\theta} & \dots & S_k(l)e^{-j2\pi(M-1)\sin\theta} \end{bmatrix} + \mathbf{N}_k(l) \tag{42}$$

IX. Conclusion

References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement:A Survey of the State of the Art*. Morgan & Claypool, 2013. [Online]. Available: https://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6813348

APPENDIX A
MATLAB CODE