TU Delft
Faculty of Electrical Engineering, Mathematics, and Computer Science
Circuits and Systems Group

ET 4386 Estimation and Detection

ASSIGNMENT

Multi-Microphone Speech Enhancement

1 Context

The use of mobile speech processing applications (e.g., mobile phones, hearing aids, speech recognition systems) has significantly increased. As a side effect, such applications are often used in acoustically noisy environments. To be able to succesfully use these systems, most of these applications apply noise reduction before the actual speech signal is used by the speech recognition system or before it is played back to the user. When multiple microphones are present, so-called multi-microphone noise reduction can be applied. This is a technique where the fact is exploited that the microphones are at different physical locations. As a consequence, the target signal is recorded at each microphone with a slightly different delay. Compensating for this delay and combining the different microphone signals then results in constructive addition of the target signal. Signals coming from different directions than the target signal will then be added desctructively and thus reduced. This exercise consists of two parts: (a) derive and implement a suitable estimator for the target signal; and (b) study the performance of the estimator.

In a group of 2 students, make a short report (4-5 pages; pdf file) containing the required Matlab scripts, plots, and answers. Also, prepare a short presentation to explain your results.

System model

In Fig. 1, we show a high level block diagram of a multi-microphone noise reduction system. The microphones record the noisy signal, which can be modeled as

$$y_m(n) = s_m(n) + w_m(n),$$

where $s_m(n)$ is the target speech signal at sample-time index n and microphone $m \in \{1, ..., M\}$. Similarly, the disturbance is modeled by $w_m(n)$. We assume that $w_m(n)$ is a random processes. Due to the efficiency of the discrete Fourier transform (e.g. using the FFT), speech enhancement algorithms are often implemented in the frequency domain. Since speech signals are non-stationary, prior to a transformation to the frequency domain, the signals is devided in short overlapping time frames (typically with 50 % overlap and 20 ms long) and windowed (e.g. using a Hann window). In Fig. 1 this is given by the vector $\mathbf{y_m}(l)$ for microphone m and time-frame l. After the frequency transform, we obtain the DFT coefficients $Y_m(l, k)$, where $k \in \{1, ..., K\}$ indicates the frequency-bin index. The signal model in the frequency domain is then given by

$$Y_m(l, k) = S_m(l, k) + W_m(l, k).$$

The propagation of the target source from its source location so the different microphones is given by the acoustic transfer function (ATF) $A_m(l, k)$. Using the ATF, we can write the above signal model as

$$Y_m(l, k) = A_m(l, k)S(l, k) + W_m(l, k),$$

where $S(l, k)$ is now the target source at the source location. Stacking the DFT coefficients for the different microphones in a vector we obtain

$$\mathbf{Y}(l, k) = \mathbf{A}(l, k)S(l, k) + \mathbf{W}(l, k),$$

with $\mathbf{Y}(l, k) = [Y_1(l, k), ..., Y_M(l, k)]^T$, $\mathbf{A}(l, k) = [A_1(l, k), ..., A_M(l, k)]^T$ and $\mathbf{W}(l, k) = [W_1(l, k), ..., W_M(l, k)]^T$. The clean target can be estimated by applying a complex weighting $\mathbf{w} \in \mathbb{C}^M$ to the noisy DFT coefficients at the M microphones, that is, $\hat{S}(l, k) = \mathbf{w}^H(l, k)\mathbf{Y}(l, k)$.

The goal now is to estimate the target speech signal in Matlab. The dataset provided to you corresponds to 4 recorded noisy microphone signals at a sampling frequency of $f_s = 16$ kHz. For simplicity, assume that the ATF of the target source is given by $A_m(l, k) = 1 \; \forall \; k, l, m$. This basically means that the target source at each microphone has the same delay and damping (i.e., the target source is in the far-field perpendicular with respect to the array). The noise in the frequency domain can be assumed to be zero-mean Gaussian distributed, that is, $W_m(l, k) \sim \mathcal{CN}(0, \sigma^2_{W_m(l,k)})$. The speech signal is assumed to be deterministic, but unknown. Furthermore, it can be assumed that the noise is stationary across time. The structure of the matlab variable Data is such that the time domain signals are given along the first dimension and the microphone index is along the second dimension.

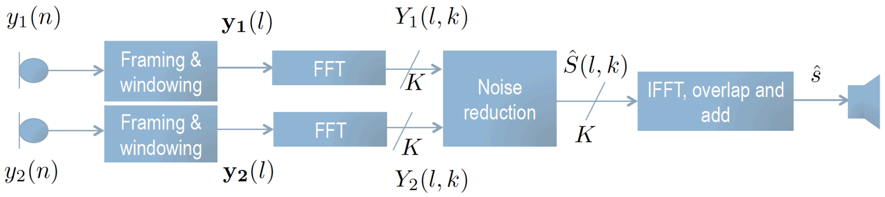

Figure 1: Overview of a Dual-microphone speech enhancement system.

2 Assignment

Assume that the noise is independently (but not identically) distributed across the microphones. To determine the noise variance in the frequency domain ($\sigma^2_{W_m(l,k)}$), you can use the time frames during the first second, as these are noise-only.

1. Derive and implement an estimator for the target signal $s[n]$ of your choice. Motivate the reason for selecting the estimator that you have implemented.

2. Determine emperically (using the data, averaged across frequency and time indices) the variance of the estimator. Do this by calculating $\text{var}_{emp} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}|\hat{S}(l,k) - S(l,k)|^2$. Plot var_{emp} as a function of the number of microphones that are involved in the estimator. First use only the first microhone. Then the first 2, then the first 3. Etc. Untill you finally have used all microphone signals.

3. Determine the Cramer Rao bound per frequency band. Plot the Cramer Rao bound averaged across frequency as a function of the number of microphones. Does you estimator reach the Cramer Rao bound?

4. Assume that the speech signal is not anymore deterministic but also a random process. Explain how you could further improve the estimator in that case.

3 Consultant

dr. ir. Richard C. Hendriks
Email: ET4386-EWI@tudelft.nl
Room: HB 17.080
Office hours: By prior appointment only.