# 6 Signal Processing at the Acoustic Human-Machine Interface

## 6.1 Overview

For convenient human/machine interaction, acoustic front-ends are required which allow seamless and hands-free audio communication. The general objective of acoustic front-ends for distant-talking interfaces is to acquire, extract, and enhance the signals uttered by the desired speakers while attenuating interference.

This exercise provides an elementary overview of some of the algorithms used in acoustic front-ends applied for distant-talking speech interfaces. Such interfaces have been developed, e.g., for DICIT[4] ('Distant-talking Interfaces for Control of Interactive TV'), a project which was funded by the European Union. Such interfaces enable users to control a TV set via voice commands without the use of a close-talking microphone. The acoustic front end helps to maximize the speech recognition performance in noisy and reverberant scenarios, e.g., in typical living rooms.

The algorithms that we will consider in this exercise are

- Beamforming,
- Acoustic Echo Cancellation (AEC),
- and Acoustic Source Localization (ASL).

Note that in the scope of this exercise, we consider only simple methods which show the basic concepts.

In the directory `./SHARED_FILES/spsa/Exercise6/` exists a subdirectory `wav` including two five-channel speech files and one five-channel audio file (sampling rate $f_s = 48\text{kHz}$).

Copy the subdirectory `wav` into your home directory `spsaX`. The multichannel `wav` files `desired_speech.wav` and `interferer_speech.wav` contain the microphone signals of the desired speech only and the interference speech only, respectively. `bf_aec.wav` contains the microphone signals consisting of a mixture of the desired speech and loudspeaker audio.
To get started, the template file `SPSA_Ex6_template.m` might help you.

## 6.2 Acoustic Source Localization (ASL)

### 6.2.1 General

ASL aims at extracting the localization information of one or several sound sources from signals captured by a number of spatially distinct microphones (spatial diversity) without any prior knowledge about the observed acoustic scene. In many

[4] *http://dicit.fbk.eu/*

applications accurate localization of one or several sound sources provides the basic information needed by other processes, e.g., for steering a beamformer. In this exercise we will consider the conceptually simple **Cross-Correlation (CC) method**[5] which is based on Time Differences Of Arrival (TDOA). Each TDOA corresponds to a Direction of Arrival (DOA) $\vartheta$ as depicted in *Fig.* 11.

Firstly, one or several time delays between different pairs of microphones (i.e., the TDOAs) are estimated. Then, the TDOA estimates as well as the microphone array geometry are used to calculate the position of the sources.

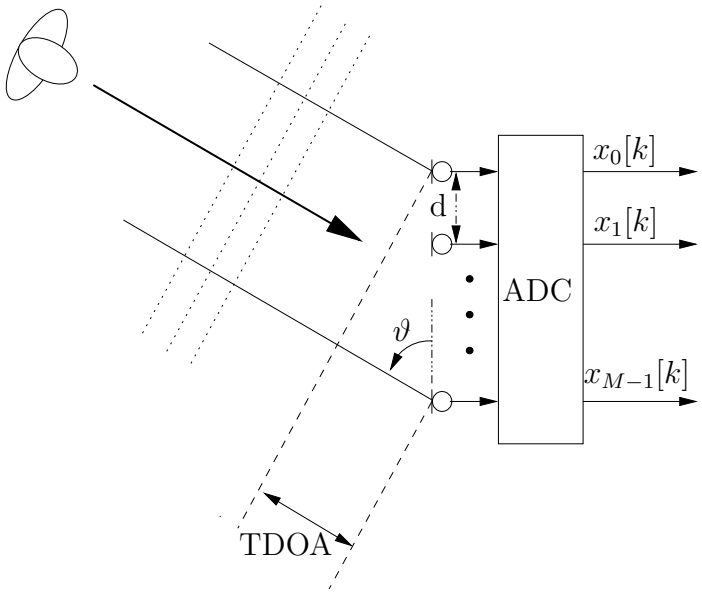


Fig. 11: TDOA given two selected microphones and a source in the far-field

The CC method relies on a single-source, single-path signal model:

$$x_m[k] \;=\; s[k - t_m] + n_m[k]. \tag{27}$$

where $x_m$ is the $m$-th microphone signal, $s$ is the source signal, and $n_m$ is the noise signal at $m$-th microphone. $t_m$ is the propagation delay with respect to the $m$-th microphone. The CC method estimates the TDOA of source signal $s$, for a pair of microphones $\{i,j\}$, by determining the value $\tau$ which maximizes the CC function defined by:

$$\hat{R}_{x_j x_i}[\tau] \;=\; \hat{E}\{x_j[k+\tau]x_i[k]\} \tag{28}$$

where $\hat{E}$ is an approximation of the expectation operator. Assuming that the source signal $s$ and the noise signals $n_m$, $m \in \{i, j\}$ are mutually uncorrelated (the noise

[5] A wide variety of algorithms exist, each addressing different acoustical scenarios depending on the nature of the source, the room reverberation or the amount of background noise.

processes are also assumed to be mutually uncorrelated), the CC function can be expressed as:

$$\hat{R}_{x_j x_i}[\tau] \quad = \quad \hat{R}_{ss}[\tau - \tau_{ij}], \tag{29}$$

which exhibits a maximum peak at $\tau = \tau_{ij}$. In the CC method, the TDOA estimate $\hat{\tau}_{ij}$ is obtained by:

$$\hat{\tau}_{ij} \quad = \quad \arg\max_{\tau} \hat{R}_{x_j x_i}[\tau]. \tag{30}$$

In a practical system, the CC has to be estimated based on a data segment of finite length. Adjacent peaks therefore appear in the CC estimate, in particular under noisy conditions. The detection becomes increasingly more difficult as the input signal decreases in bandwidth, necessitating higher Signal-to-Noise Ratio (SNR) values.

### 6.2.2 Experiments on ASL

- Given that the array consists of five microphones with a uniform spacing of $d = 0.04$m, compute the total length of the array. Use the outermost microphone signals to compute the TDOAs for both the desired speaker and the interferer using the CC method.
- What is the delay, in samples, between the outermost microphone signals for the desired source and the interferer?

  ▷ _____
- What are the DOAs of the desired speaker and the interferer?

  ▷ _____

## 6.3 Beamforming

### 6.3.1 General

Beamforming techniques are commonly used, where multiple microphones are jointly processed to form a beam, i.e., a region of increased selectivity, and steer it toward the desired source. The beamformer exploits the spatial distribution of desired sources and interferers in order to attenuate the latter.

The beamforming design which we will consider in this exercise is the classical **Delay-and-Sum Beamformer (DSB)**. The DSB design is based on the idea that the desired output contribution of each of the array microphones will be the same,

except that each one will be delayed by a different amount. Therefore, if the output of each of the sensors is delayed and weighted appropriately, the signal originating from a desired spatial region will be reinforced, while noise and interfering signals from other spatial regions will generally be attenuated. It is typically used for narrowband signals[6].

*Fig.* 12 depicts DSB with a linear, uniformly spaced array consisting of $M$ microphones. The source signal is captured by the microphones and digitized by analog-to-digital converters. The digitized signals are then weighted and delayed before being combined to produce the output.

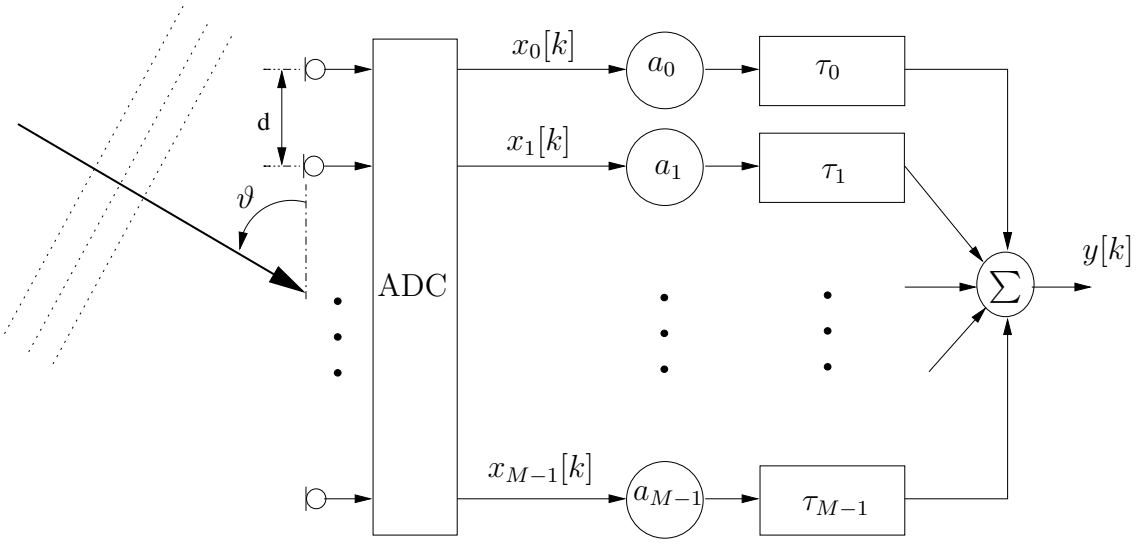


Fig. 12: Beamforming with a linear microphone array

The DSB response for a frequency $\omega$ and DOA $\vartheta$ is given by

$$B(\omega, \vartheta) = \sum_{m=0}^{M-1} a_m e^{-j\omega\tau_m(\vartheta)}, \tag{31}$$

where $\tau_m(\vartheta) = md\cos\vartheta/c$, $d$ is the distance between two adjacent array microphones, and $c = 342\frac{\mathrm{m}}{\mathrm{s}}$ is the speed of sound in air. The magnitude square of the beamformer response is known as the beampattern of the beamformer. The beampattern describes the beamformer's ability to capture acoustic energy as a function of the DOA of the plane wave. It is defined as

$$P(\omega, \vartheta) = 20\log_{10}(|B(\omega, \vartheta)|). \tag{32}$$

*Fig.* 13 depicts the beampattern obtained using a DSB with a 5-element linear uniformly-spaced array, $d = 0.04$m. It can be clearly seen that the beam width of the main beam of the DSB design varies with frequency.

[6]In the case of broadband signal acquisition, a Filter-and-Sum Beamformer (FSB) which aims at ensuring a frequency-independent beam is typically used.

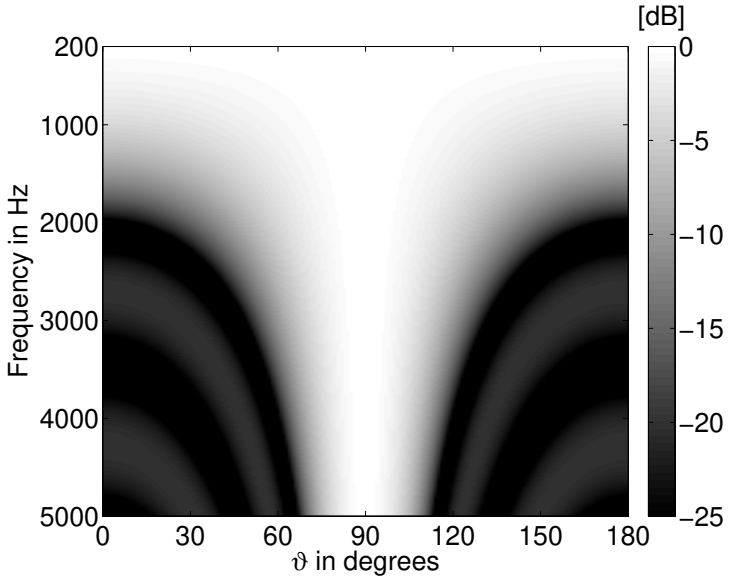


Fig. 13: DSB beampattern; $M = 5$, $d = 0.04$m, and $a_m = 1/5$

### 6.3.2 Experiments on Beamforming

- Using the array geometry ($d = 0.04$m) and the DOA of the desired speaker obtained in section 6.2.2, implement a DSB with uniform weighting, i.e., $a_m = 1/M$, which steers a beam to the user position (hint: you can use the command `fftfilt` for filtering).
- Listen to a microphone signal and the output of the beamformer for both the desired and interfering speaker. Which signal is attenuated at the beamformer output?

  ▷ _____

## 6.4 Acoustic Echo Cancellation (AEC)

### 6.4.1 General

Acoustic echoes occur due to the coupling between loudspeaker and microphones, i.e., due to the lack of acoustical barriers: apart from the speech uttered by the speaker $v[k]$ and a noise signal, the microphones in the receiving room also acquire the signal that is played back via the loudspeaker[7]. AEC aims at reducing the reverberated loudspeaker signal within a microphone signal $y[k]$.

In systems such as DICIT, the AEC is a crucial means to improve the recognition rate of an Automatic Speech Recognizer (ASR), providing the ASR with the echo-compensated signal $e[k]$ that should mainly contain the utterance of the desired speaker $v[k]$.

[7]For multichannel sound reproduction Multichannel-AEC is applied (see lecture notes)

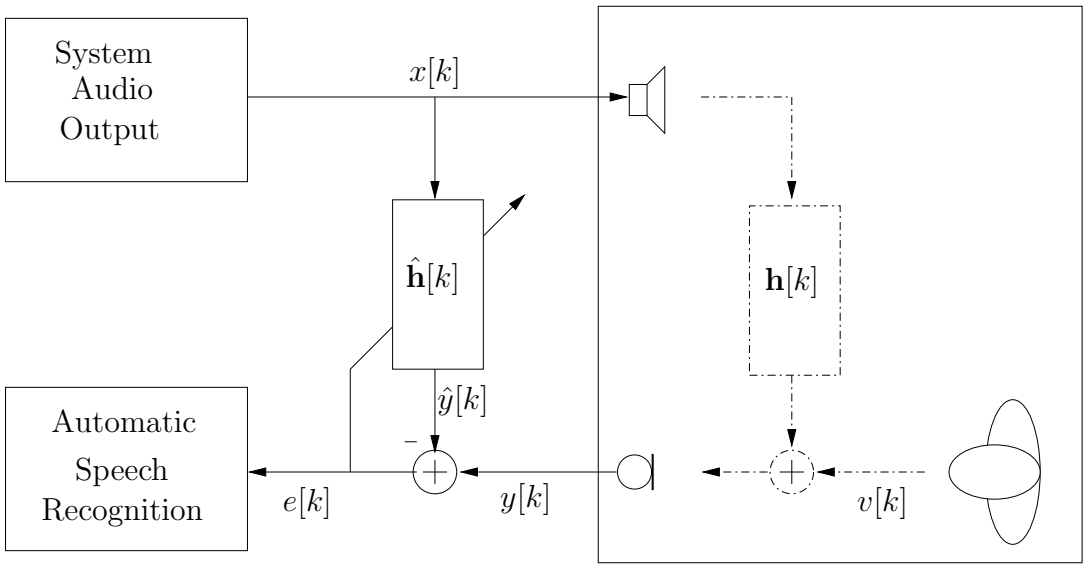


Fig. 14: AEC in Human-Machine-Interface System

The relation between the original loudspeaker signals and their contribution to the microphone signal $y[k]$ is established by the time-variant impulse responses of the Loudspeaker-Enclosure-Microphone (LEM) system $\mathbf{h}[k]$ – time-variance is due to continuous changes of the acoustic environment, e.g. caused by temperature changes, door openings or user movements. An Acoustic Echo Canceller (AEC) as depicted in *Fig.* 14 models the impulse response by means of an adaptive digital filter $\hat{\mathbf{h}}[k] = [\hat{h}_0[k], \hat{h}_1[k], \ldots, \hat{h}_{L-1}[k]]^T$. The echo replica $\hat{y}[k]$ computed via convolution of the AEC filter response with the known loudspeaker signal $x[k]$ is then subtracted from the microphone signal $y[k]$, leading to the desired echo reduction.

As to the design of $\hat{\mathbf{h}}[k]$, an adaptive filter is an adequate means to track the temporal variations of the LEM system. Several different algorithms have been developed for controlling the adaptive mechanism. In this exercise we consider the Normalized Least Mean Squares (NLMS) algorithm which is widely used because of its simplicity. The NLMS algorithm is practically used in the form:

$$\hat{\mathbf{h}}[k+1] = \hat{\mathbf{h}}[k] + \mu \frac{\mathbf{x}[k] e^*[k]}{\mathbf{x}^H[k]\mathbf{x}[k] + \delta}, \tag{33}$$

where $\mathbf{x}[k] = [x[k], x[k-1], \ldots, x[k-L+1]]^T$, $\delta$ is a regularization parameter for the case of small values of input signal power $\mathbf{x}^H[k]\mathbf{x}[k]$ and $\mu$[8] is the step size. The algorithm is based on the minimization of the mean squared error $\mathcal{E}\{e[k]^2\}$. This minimization in the time domain leads to the so-called Wiener-Hopf equations, describing the optimum filter coefficients $\hat{\mathbf{h}}_{\text{opt}}$. An iterative solution of these equations can be achieved by the gradient descent method and is approximated by the Least

[8]For stability reasons the following condition must be met: $0 < \mu < 2$

Mean Squares algorithm, which directly leads to the NLMS algorithm by introducing a normalized step size.

Please note that the filter update must only be performed for the case of single-talk, i.e., $v[k] = 0$. For the case of double-talk, i.e., $v[k] \neq 0$, the filter adaptation must be slowed down or halted in order to prevent a divergence of the AEC filters. For practical realizations a so-called double-talk detector therefore has to be employed.

### 6.4.2 Experiments on AEC

- Read the beamformer output signal `bf_aec.wav` and the original loudspeaker signal `Speaker.wav`.
- Implement an AEC based on the NLMS algorithm and use the beamformer output as input signal. The beamformer output contains the filtered loudspeaker signal (music) and near-end speech, $v \neq 0$, between $t1 = 20$s and $t2 = 25$s. To prevent filter divergence, halt the filter adaptation for $t1 < t < t2$ (assuming the availability of a perfect double-talk detector).
- Plot the error signal. Listen to a microphone signal and the output of the AEC. What is the difference?

  ▷ _____