2.160 IDENTIFICATION, ESTIMATION, AND LEARNING

LECTURE NOTES NO. 3

3. Random Variables and Random Processes

So far we have dealt with "deterministic systems" for developing parameter estimation algorithms. This terminology, however, is misleading, since all systems are to some extent uncertain and noisy. If we can quantify uncertainties and utilize the quantified uncertainties for estimation, we would be able to do better or more rigorously. Within the framework of deterministic systems we did not use such information about uncertainties; we simply left it as prediction error.

An alternative approach is to **quantify uncertainty** for better prediction, estimation, and control. This requires modeling an uncertain process as a dynamical process perturbed by noise and characterizing the noise properties. The figure below illustrates a dynamical system where the plant is disturbed with process noise and the observed output is corrupted with measurement noise.

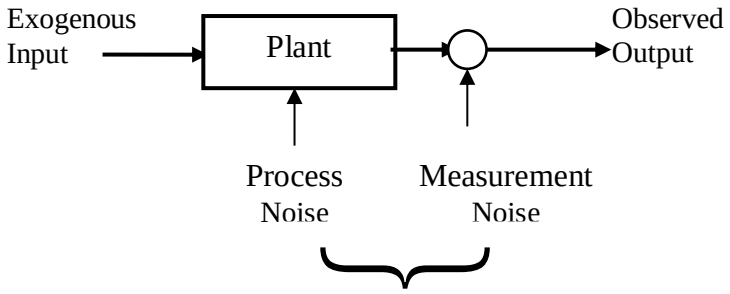


Figure 3-1 Dynamic process perturbed by noise

Objective of modeling uncertainties:

- Use stochastic properties of the process for better estimating parameters and state of the process, and
- Better understand, analyze, and evaluate estimation mechanisms.

Theoretical development of estimation, identification, and learning is heavily involved in quantification of uncertainties. In this chapter we begin with a quick review of probability, random variables, and random processes.

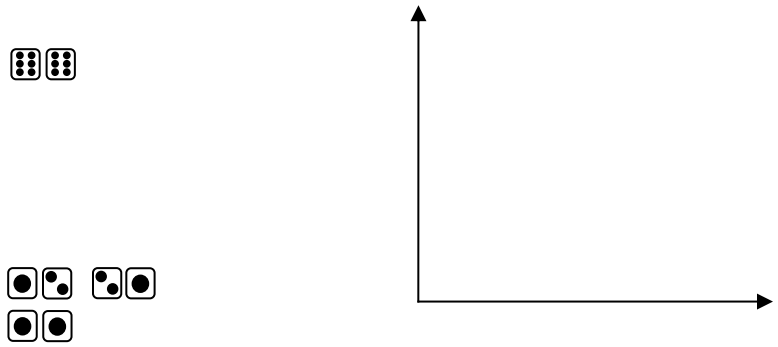
3.1 Random Variables

You may have already learned probability and stochastic processes in some subjects. The following are fundamentals that will be used regularly for the rest of this course. Check if you feel comfortable with each of the following definitions and terminology. (Check the box of each item below.) If not, consult standard textbooks on the subject. See the references at the end of this chapter.

The most rigorous way of introducing probability, random variables, and random processes is to begin with set theory, σ-field, and Lebesgue measure. We leave these topics to references and

focus on the only basic math that we will directly use in the following sections. Later on we will discuss more on random processes as needed in the identification sections.

1) Random Variable



X: Random variable is a function that maps every point in the sample space to the real axis line.

2) Cumulative distribution function (CDF), $F_X(x)$, and probability density function (PDF) $f_X(x)$;

$$F_X(x) = \text{Prob}(X \le x) \qquad (1)$$

$$f_X(x) = \frac{d}{dx} F_X(x) \qquad (2)$$

In the statistics and probability literature, the convention is that capital *X* represents a random variable while lower-case *x* is used for an instantiation/realization of the random variable. The following are basic properties of CDF and PDF.

$$\lim_{x\to\infty} F_X(x) = 1 \ , \ \int_{-\infty}^{\infty} f_X(x)dx = 1$$

For discrete random variables, we use the following Probability Mass Function (PMF):

$$p_X(x_i) = \text{Prob.}(X = x_i) \qquad (3)$$

3) Joint probability densities

Let *X* and *Y* be two random variables

For discrete random variables:

$$p_{XY}(x_i, y_j) = \text{Prob.}(X = x_i, Y = y_j, \text{ simultaneously}) \qquad (4)$$

For continuous random variables: $f_{XY}(x,y)$ (5)

4) Statistically independent (or simply Independent) random variables

$f_{XY}(x,y)= f_X(x)\, f_Y(y)$ for all x and y (6)

A similar expression can be given for discrete random variables. In the following only continuous random variables will be shown.

5) Conditional probability density

First consider the cumulative probability distribution of X when another random variable Y takes $y < Y \le y + \Delta y$. This can be written as:

$$\text{Prob}\left(X \le x, \text{ given } y < Y \le y + \Delta y \right) = \frac{\int_{-\infty}^{x} \int_{y}^{y+\Delta y} f_{XY}(x,y)\,dx\,dy}{\int_{y}^{y+\Delta y} f_Y(y)\,dy}$$

Dividing both numerator and denominator by Δy and letting $\Delta y \to 0$ yield the conditional probability density:

$$f_{X|Y}(x\,|\,y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{7}$$

= (Joint probability density divided by Probability density of Y at y)

If X and Y are independent,

$$f_{X|Y}(x\,|\,y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_X(x)\, f_Y(y)}{f_Y(y)} = f_X(x) \tag{8}$$

Occurrence of $Y = y$ does not influence the occurrence of $X = x$.

6) ***Bayes Rule***

$$f_{X|Y}(x\,|\,y) = \frac{f_{Y|X}(x)\, f_X(x)}{f_Y(y)} \qquad \because f_{X|Y}(x\,|\,y)\, f_Y(y) = f_{XY}(x,y) = f_{Y|X}(y\,|\,x)\, f_X(x) \tag{9}$$

7) ***Marginal Probability***

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy, \; f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx \tag{10}$$

8) Expectation

Expected value of X $\Rightarrow E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$ (11)

mean, average

For discrete random variables,

$E[X] = \sum_i p_i x_i$ where $p_i = p_X(x_i)$

9) Variance

$$VarX = E[(X - E(X))^2] = E[X^2 - 2XE(X) + (E(X))^2] = E[X^2] - (E(X))^2 \tag{12}$$

10) Moment

k-th moment of *X*

$$E[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)dx \tag{14}$$

k = 1 mean, k = 2, k = 3 higher moments

11) Normal (Gaussian) Random Variable

$X \sim N(m_X, \sigma^2)$: This means that random variable *X* has a normal distribution with mean m_X and variance σ^2

The first and second moments completely characterize the distribution.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - m_X)^2]$$
$$\sigma^2 = \int_{-\infty}^{\infty} (x - m_X)^2 f_X(x)dx \tag{15}$$

12) ***Correlation***

The expectation of the **product** of two random variables, *X* and *Y*, is called "Correlation".

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)\,dxdy \tag{16}$$

$f_{XY}(x,y)$ — Joint probability

If *X* and *Y* are independent

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\,dxdy$$
$$= \int_{-\infty}^{\infty} x f_X(x)dx \cdot \int_{-\infty}^{\infty} y f_Y(y)dy = E[X]E[Y] \tag{17}$$

Note that, although the correlation is zero, the two random variables are not necessarily independent.

13) ***Orthogonality***

X and *Y* are said to be orthogonal, if the correlation is zero $E[XY] = 0$ (18)

14) ***Covariance***

Covariance of *X* and *Y* = cov(*X*,*Y*) = $E[(X-m_X)(Y-m_Y)]$ (19)

Notice the difference between Covariance and Correlation. Sometimes the terminology is abused in literature. For example, "Covariance" resetting of the RLS algorithm with a forgetting factor is not covariance.

15) Correlation coefficient

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}} = E\left[\left(\frac{X-m_X}{\sigma_X}\right)\cdot\left(\frac{Y-m_Y}{\sigma_Y}\right)\right] \tag{20}$$

3.2 Random Processes

A Random Process is a family (ensemble) of time functions having a probability measure.

Random Variable X ⟶ Random Process $X(t)$

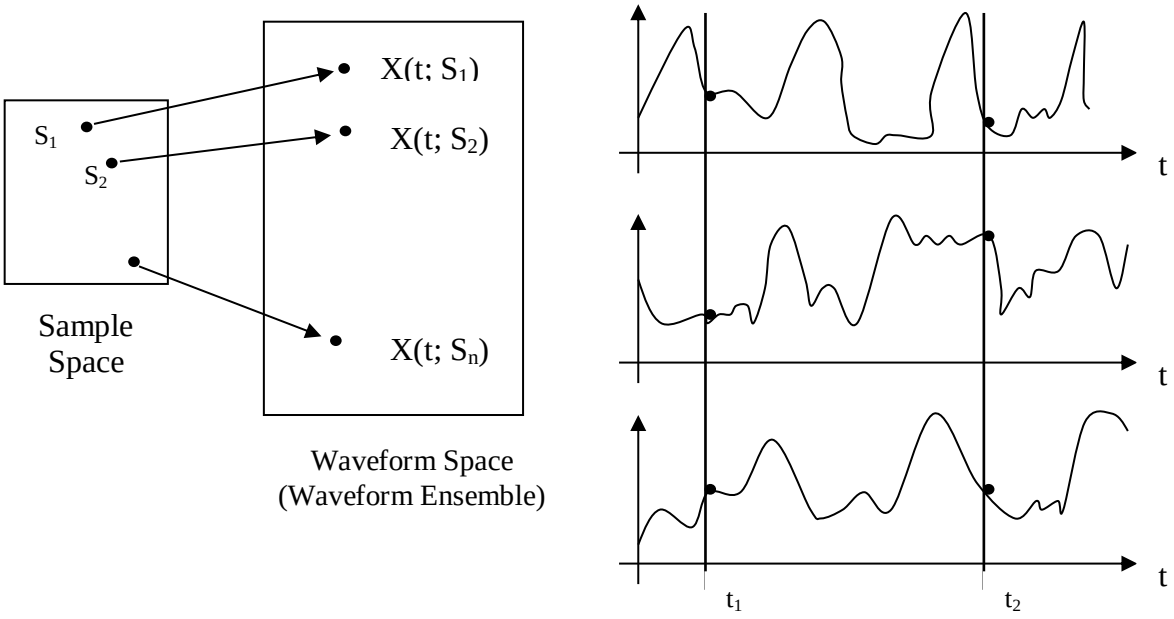


Figure 3-2 Ensemble of time profiles coming from a random process

Characterization of a random process

- First-order densities: $f_{X(t_1)}(x)$ abbreviated as $f_{X_1}(x_1)$;
- Second-order densities: joint probability density $f_{X1X2}(x_1,x_2)$

If the random process has some correlation between two random variables $X(t_1)$ and $X(t_2)$, it can be captured by the following autocorrelation: ("auto" means correlation within the same single random process)

$$R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)] = \iint_{-\infty}^{+\infty} x_1 x_2 \, f_{X_1X_2}(x_1,x_2)\,dx_1 dx_2 \tag{21}$$

If the auto-correlation function depends only on the time difference $\tau = t_1 - t_2$, it reduces to

$$R_{XX}(\tau) = E[X(t+\tau)X(t)] \qquad R_{XX}(\tau) = R_{XX}(-\tau) \text{ even function} \qquad (22)$$

Then the process is called "Wide Sense Stationary".

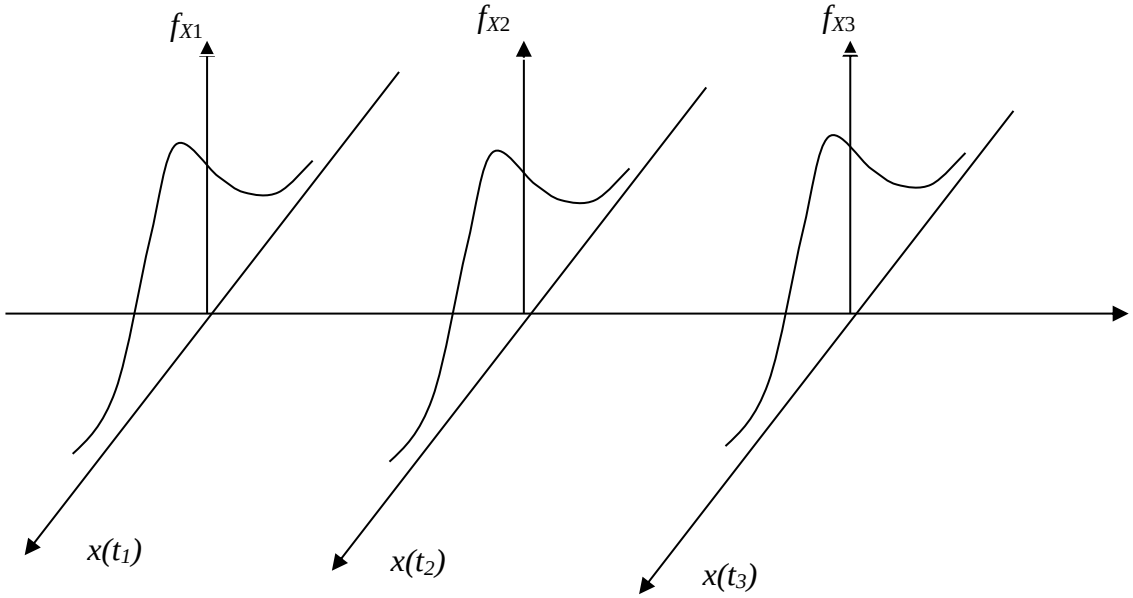


Figure 3-3 Probability densities at individual time slices

Auto-covariance:

$$\begin{aligned} C_{XX}(t_1,t_2) &= E[(X(t_1)-m_X(t_1))\cdot(X(t_2)-m_X(t_2))] \\ &= R_{XX}(t_1,t_2)-m_X(t_1)m_X(t_2) \end{aligned} \qquad (23)$$

- Higher-order densities

  Joint density $f_{X1X2....Xn}(x_1,x_2,...,x_n)$
  $E[X(t_1)X(t_2)\ldots X(t_n)]$ n-th order

  Total characterization: If joint densities of $X(t_1)$, $X(t_2)$, ….., $X(t_n)$ for all n are known, the random process is said to be totally (completely) characterized. … Unrealistic.

3.3 Multivariate Random Processes

In the previous section we characterized properties of a single random process. Now we extend it to multivariable random processes. Recall a family of time profiles, called an ensemble. Each of these waveforms is a realization of random process $X(t)$. Such a random

waveform can be seen in your oscilloscope, when you increase the gain of a ground signal. An ensemble of waveforms may be considered as a large collection of the same oscilloscope. If you have one hundred of the same type of oscilloscope, you have 100 waveforms of the ground noise. They come from the same ground signal, but the waveforms are all different.

In Section 3.2, stochastic properties of a single random process were described with the first and second order densities and auto-correlation functions. Now let us extend it to multiple random processes, say $X(t)$, $Y(t)$, and $Z(t)$, and characterize their properties. In the oscilloscope analogy, you can think about multi channels of signals, say channel 1 and channel 2, shown on a single oscilloscope display. See Figure 3-4. If there are a large number of the same oscilloscopes displaying both channels, an ensemble of multivariate random processes can be generated.

Let $f_{XYZ}(x_t,y_t,z_t)$ be the first-order density of a multivariate random process, $X(t)$, $Y(t)$, and $Z(t)$. The covariance is defined as

Covariance: Ensemble mean

$$C_{XYZ}(t) = E\left[\begin{pmatrix} X(t)-m_X(t) \\ Y(t)-m_Y(t) \\ Z(t)-m_Z(t) \end{pmatrix}\begin{pmatrix} X(t)-m_X(t) & Y(t)-m_Y(t) & Z(t)-m_Z(t) \end{pmatrix}\right] \quad (24)$$

If $m_X = m_Y = m_Z = 0$

$$C_{XYZ}(t) = \begin{bmatrix} E\left[X^2(t)\right] & E\left[X(t)Y(t)\right] & E\left[X(t)Z(t)\right] \\ E\left[X(t)Y(t)\right] & E\left[Y^2(t)\right] & E\left[Y(t)Z(t)\right] \\ E\left[X(t)Z(t)\right] & E\left[Y(t)Z(t)\right] & E\left[Z^2(t)\right] \end{bmatrix} \quad (25)$$

Second-order density:

Taking two time slices, t_1 and t_2, as shown in the above figure, the joint probability mass function is given by: $p_{X_1Y_1Z_1X_2Y_2Z_2}(x_{t_1}, y_{t_1}, z_{t_1}, x_{t_2}, y_{t_2}, z_{t_2})$

If $m_x=m_y=m_z=0$, the covariance is given by

$$C_{XYZ}(t_1,t_2) = \begin{bmatrix} E\left[X(t_1)X(t_2)\right] & E\left[X(t_1)Y(t_2)\right] & E\left[X(t_1)Z(t_2)\right] \\ E\left[Y(t_1)X(t_2)\right] & E\left[Y(t_1)Y(t_2)\right] & E\left[Y(t_1)Z(t_2)\right] \\ E\left[Z(t_1)X(t_2)\right] & E\left[Z(t_1)Y(t_2)\right] & E\left[Z(t_1)Z(t_2)\right] \end{bmatrix} \quad (26)$$

Note that the first order covariance in equation (25) can be viewed as a special case of the second order covariance in equation (26); $t_1 = t_2 = t$.

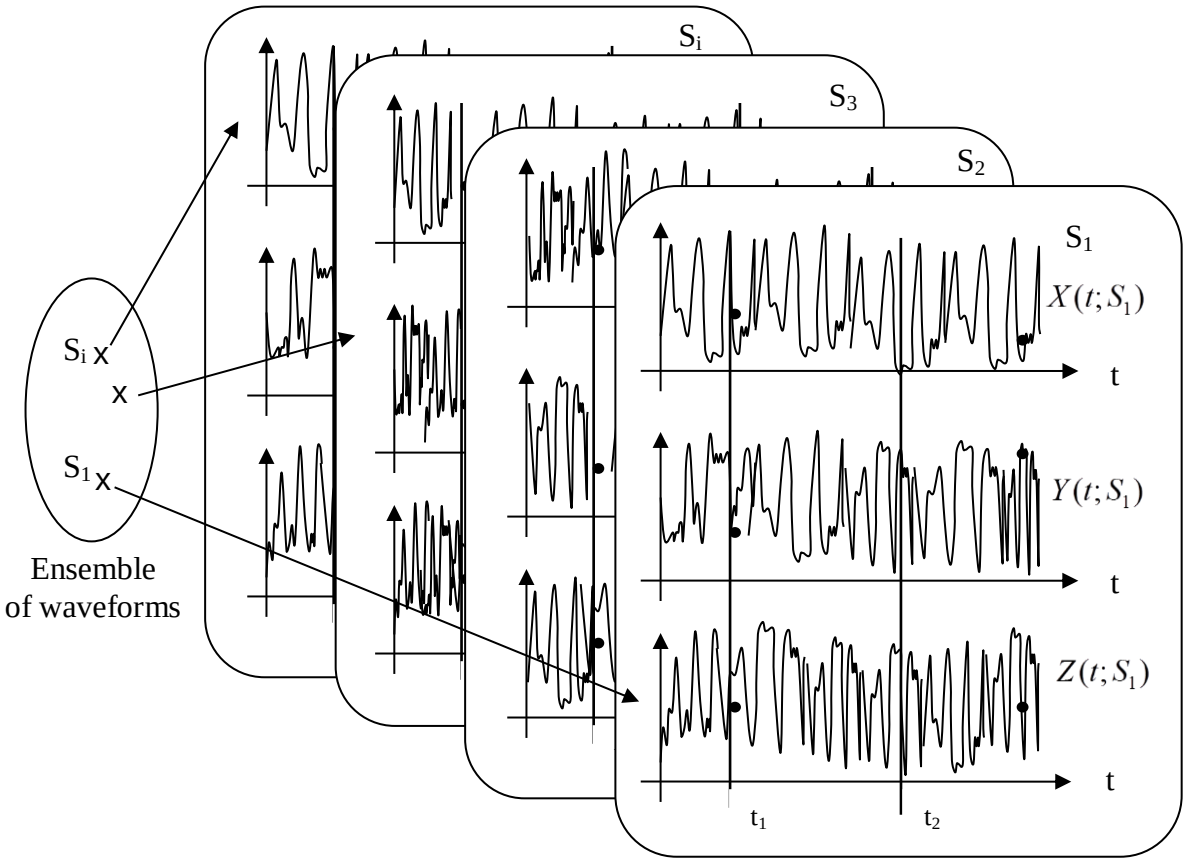


Figure 3-4 Multivariable random processes

Reference Textbook on Random processes

John A. Gubner, "Probability and Random Processes for Electrical and Computer Engineers", Cambridge University Press 2006, ISBN-13 978-0-521-86470-1

Lonnie Ludeman , "Random Processes - Filtering, Estimation, and Detection", Wiley 2003, ISBN 0-471-25975-6

Robert Brown and Patrick Hwang, "Introduction to Random Signals and Applied Kalman Filthering, Third Edition", Wiley 1997, ISBN 0-471-12839-2, TK5102.9.B75

3.4 Application: Adaptive Noise Cancellation

Let us consider a simple example of the above definitions and properties of random processes. Active noise cancellation is a technique for canceling unwanted noise by measuring the noise source as well as the signal source. Consider an airplane pilot talking to an air traffic controller. The pilot voice is quite disturbed by the engine noise and other acoustic disturbances. To cancel out the acoustic noise, another microphone is used for measuring the surrounding noise. See the figure below. A simple method of active noise cancellation is to generate an "anti-noise" signal by inverting the sign of the measured noise, and super impose it to the main microphone's signal, i.e. phase-cancellation. This cancellation technique, however, is limited, since we do not know exactly how the acoustic noise interferes with the pilot voice. To overcome this difficulty, "adaptive" noise cancellation, originated by David Widrow's research in the 60's, has been developed. In this system, the "interference dynamics" are identified in real time so that the noise may be best canceled out. The orthogonality of signals is the key to this adaptive noise cancellation.

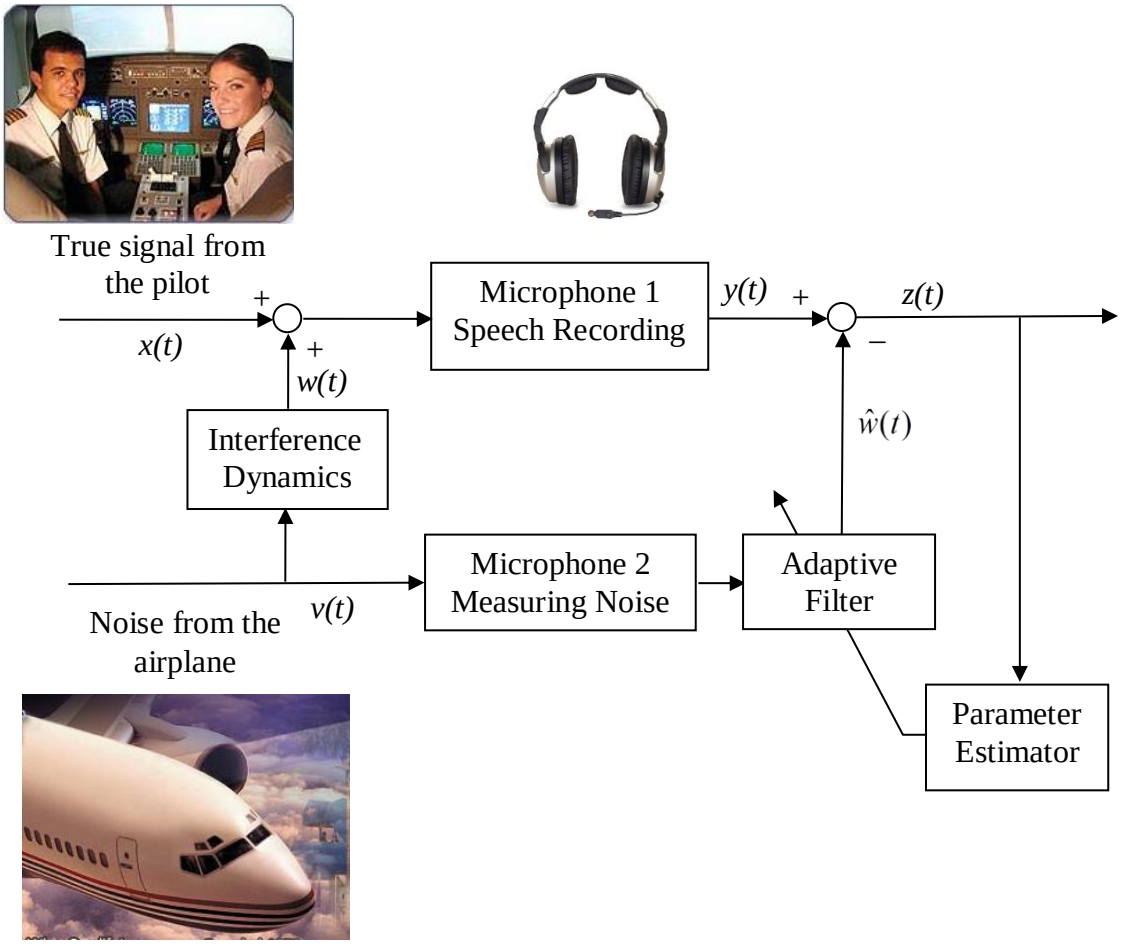


Figure 3-5 Block diagram of Adaptive Noise Cancellation Algorithm

Let $x(t)$ be the true pilot's voice and $v(t)$ be the airplane noise. The signal captured by the main microphone $y(t)$ is the mixture of the two:

$$y(t) = x(t) + w(t) \tag{27}$$

Let us assume that the true interference dynamics is given by the following form (FIR):

$$w(t) = b_1 v(t-1) + b_2 v(t-2) + \cdots + b_m v(t-m) \tag{28}$$

and we estimate the interference dynamics as:

$$\hat{w}(t;\hat{\theta}) = \hat{b}_1 v(t-1) + \hat{b}_2 v(t-2) + \cdots + \hat{b}_m v(t-m) = \varphi^T(t)\cdot\hat{\theta} \tag{29}$$

Based on the estimated interference model, we cancel the noise by subtraction:

$$z(t:\hat{\theta}) = y(t) - \hat{w}(t:\hat{\theta}) \tag{30}$$

Problem: Tune the FIR parameters $\hat{\theta}$ so that the recovered signal $z(t:\hat{\theta})$ is as close to the original true signal *x(t)* as possible. Assume that the true signal *x(t)* is uncorrelated with the noise *v(t)*.

$$E[x(t)v(t-\tau)] = 0, \quad \forall t, \tau \tag{31}$$

Solution: Consider the expectation of the squared output $z(t:\hat{\theta})$, i.e. the average power of signal $z(t:\hat{\theta})$,

$$\begin{aligned} E[z(t:\hat{\theta})^2] &= E[\{x(t) + w(t) - \varphi^T(t)\cdot\hat{\theta}\}^2] \\ &= E[x^2(t)] + 2E[x(t)\{w(t) - \varphi^T(t)\cdot\hat{\theta}\}] + E[\{w(t) - \varphi^T(t)\cdot\hat{\theta}\}^2] \end{aligned} \tag{32}$$

Our objective is to find the parameter vector $\hat{\theta}$ that minimizes the mean squared error $E[\{w(t) - \varphi^T(t)\cdot\hat{\theta}\}^2]$. Examining the second term:

$$\begin{aligned} &E[x(t)w(t)] = E[x(t)b_1 v(t-1)] + E[x(t)b_2 v(t-2)] + \cdots + E[x(t)b_m v(t-m)] = 0 \\ &E[x(t)\cdot\varphi^T(t)\hat{\theta}] = 0 \end{aligned} \tag{33}$$

These follow from the assumption (31).Therefore, minimizing the average power of $z(t:\hat{\theta})$ with respect to parameter vector $\hat{\theta}$ is equivalent to minimizing $E[\{w(t) - \varphi^T(t)\cdot\hat{\theta}\}^2]$,

$$\hat{\theta} = \arg\min_{\theta} E[\{w(t) - \varphi^T(t)\cdot\theta\}^2] = \arg\min_{\theta} E[z(t:\theta)^2] \tag{34}$$

since $E[x(t)^2]$ is not a function of the parameter vector and is not relevant to minimization of the squared error.

We can use the Recursive Least Squares algorithm with forgetting factor α ($0 < \alpha \leq 1$):

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{P_{t-1}\varphi(t)}{\alpha + \varphi^T(t)P_{t-1}\varphi(t)}\{y(t) - \varphi^T(t)\hat{\theta}(t-1)\}$$

$$P_t = \frac{1}{\alpha}\left(P_{t-1} - \frac{P_{t-1}\varphi(t)\varphi^T(t)P_{t-1}}{\alpha + \varphi^T(t)P_{t-1}\varphi(t)}\right)$$

Challenge Consider a live concert recording, as shown below. In addition to the main microphone for recording the performer's sound, another microphone is used for monitoring the audience noise. The difference from the above pilot microphone problem is that it is difficult to measure the audience noise separately. To some extent the second microphone does include the performer's sound. Therefore, the adaptive noise cancellation described above may not fully function. How can we alleviate the problem?

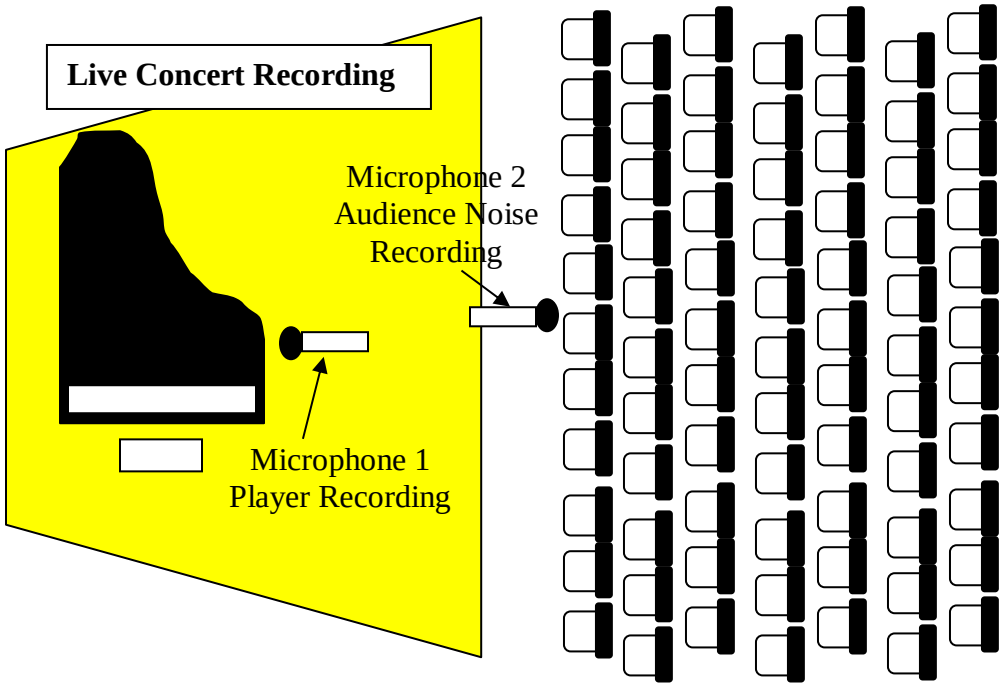


Figure 3-6 Live concert recording with double microphones

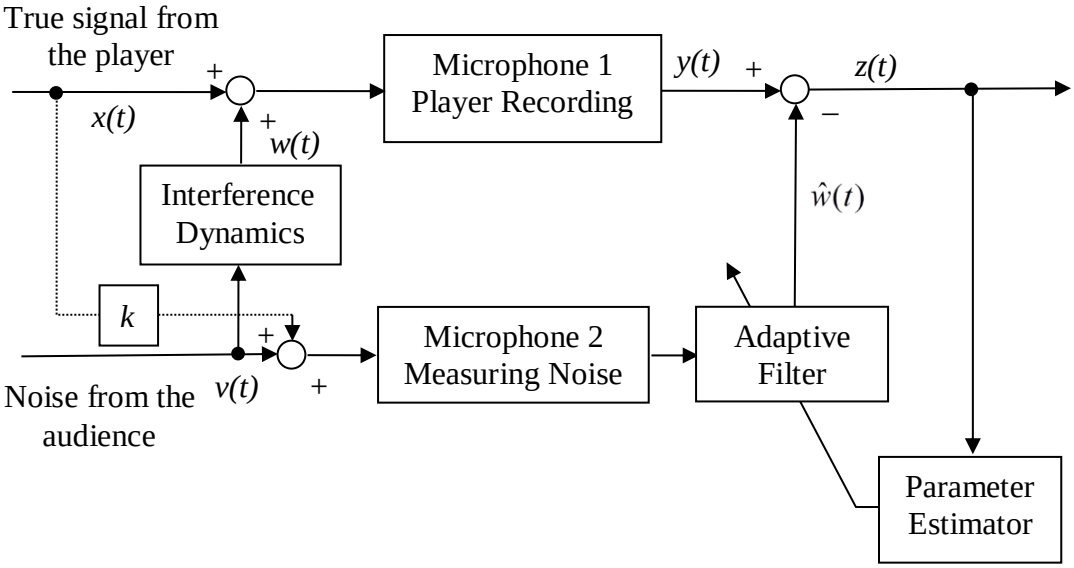


Figure 3-7 Block diagram of adaptive noise cancellation